OXFORD

Briefings in Bioinformatics, 2024, **25(5)**, bbae382
https://doi.org/10.1093/bib/bbae382
Review

A single-cell multimodal view on gene regulatory network inference from transcriptomics and chromatin accessibility data

Jens Uwe Loers [1,2,3] and Vanessa Vermeirssen [1,2,3,*]

[1]Lab for Computational Biology, Integromics and Gene Regulation (CBIGR), Cancer Research Institute Ghent (CRIG), Corneel Heymanslaan 10, 9000 Ghent, Belgium
[2]Department of Biomedical Molecular Biology, Ghent University, Zwijnaarde-Technologiepark 71, 9052 Ghent, Belgium
[3]Department of Biomolecular Medicine, Ghent University, Corneel Heymanslaan 10, 9000 Ghent, Belgium

[*]Corresponding author. E-mail: vanessa.vermeirssen@ugent.be

Abstract

Eukaryotic gene regulation is a combinatorial, dynamic, and quantitative process that plays a vital role in development and disease and can be modeled at a systems level in gene regulatory networks (GRNs). The wealth of multi-omics data measured on the same samples and even on the same cells has lifted the field of GRN inference to the next stage. Combinations of (single-cell) transcriptomics and chromatin accessibility allow the prediction of fine-grained regulatory programs that go beyond mere correlation of transcription factor and target gene expression, with enhancer GRNs (eGRNs) modeling molecular interactions between transcription factors, regulatory elements, and target genes. In this review, we highlight the key components for successful (e)GRN inference from (sc)RNA-seq and (sc)ATAC-seq data exemplified by state-of-the-art methods as well as open challenges and future developments. Moreover, we address preprocessing strategies, metacell generation and computational omics pairing, transcription factor binding site detection, and linear and three-dimensional approaches to identify chromatin interactions as well as dynamic and causal eGRN inference. We believe that the integration of transcriptomics together with epigenomics data at a single-cell level is the new standard for mechanistic network inference, and that it can be further advanced with integrating additional omics layers and spatiotemporal data, as well as with shifting the focus towards more quantitative and causal modeling strategies.

Keywords: single-cell; multi-omics; gene regulatory networks; ATAC-seq; RNA-seq

Introduction

In eukaryotes, spatiotemporal gene regulation is orchestrated by specific combinations of transcription factors (TFs), which bind to regulatory elements (REs). A RE can be further subclassified as a promoter when it is in immediate vicinity to the TSS, as an enhancer when it is located distal to the TSS and has an activating effect, or as a silencer when it is located distal to the TSS and has a repressive effect [1]. TF binding is further controlled by nucleosomes occluding RE access, leading to open/accessible chromatin or closed/nonaccessible chromatin regions. Different chromatin accessibility remodeling mechanisms include multiple TFs together outcompeting nucleosomes; pioneer TFs binding to DNA within regions of closed chromatin; active recruitment of chromatin remodelers; and post-translational modifications and/or DNA methylation [2, 3]. TFs, in turn, recruit cofactors to REs, which include the mediator complex, chromatin remodelers, and histone modifiers, and which activate or repress RNA polymerase II at core promoters of target genes (TGs). Unraveling the complex, dynamic, and quantitative interplay between TFs, cofactors, REs, and TGs at the cellular level is crucial for the molecular understanding of development, cellular differentiation, and disease [4]. High-throughput technologies such as (sc)RNA-seq (single-cell RNA sequencing) and (sc)ATAC-seq (single-cell assay for transposase-accessible chromatin using sequencing) profile transcriptomics and chromatin accessibility, respectively, and enable to characterize transcriptional regulatory interactions [5]. However, to decode the complex mechanisms of gene regulation, computational methods are essential to integrate these multi-omics data and predict regulatory relationships thereof.

Biological networks that model regulatory relationships between TFs and TGs are called gene regulatory networks (GRNs). A GRN is formulated as a directed graph with a set of nodes representing TGs and TFs and a set of edges representing directed relationships **TF → TG** between them. More and more additional data modalities are included to capture a multimodal view on gene regulation, more accurately representing the intricate regulatory relationships across omics layers [6]. Especially TF binding to TF binding sites (TFBSs), chromatin accessibility of REs and promoters, and three-dimensional chromatin contacts provide information on the precise genomic location of TF binding. The physical interrelationship between RE, TF, and TG combined with expression levels of TFs and TGs allow to draft

Received: April 5, 2024. **Revised:** June 27, 2024. **Accepted:** July 23, 2024

enhancer GRNs (eGRNs). These can be formulated as a directed graph with nodes representing TFs, genes, and REs, and two distinct sets of edges leading to a directed triple relationship **TF → RE → TG**. (e)GRNs offer a comprehensive view on gene regulation circuitry and are valuable as a mechanistic model. However, network inference, especially at the individual interaction level, is prone to false positives and negatives, and therefore, predicted regulatory interactions need to be further experimentally validated. In addition to GRNs and eGRNs, specific GRN structures are defined (Fig. 1). A regulon is the set of all TGs regulated by a specific TF [7], while a cistrome refers to the genome-wide set of REs that is targeted by a specific TF [8]. An eRegulon combines both concepts by including all REs targeted by a TF and the TGs targeted by those REs [9]. Networks can be classified as bulk, single-sample, or single-cell networks. Bulk network inference depends on many different samples with comparable phenotypes to obtain sufficient information for statistical learning and generates in this way average, population-level networks. Single-sample network inference applies specific statistical measures on bulk networks to predict networks representative of a single sample or patient, which is highly desired in precision medicine [10]. Single-cell transcriptome data, on the other hand, with thousands of cells profiled, inherently contain the variability required to infer statistical dependencies between genes, hence allowing for patient-specific (e)GRNs and even cell type specific (e)GRNs [7]. Another critical aspect of (e)GRN inference is to go beyond statistical associations and identify direct causal relationships between TFs and TGs. Here, the term causality is used to describe which TF (causal factor) regulates which TG (target of regulation). This is done for example via Bayesian methods that infer statistical causality [11, 12], incorporation of biological constraints, and/or sophisticated experimental set-ups (e.g. TF binding information, chromatin accessibility, time-series data, single-cell pseudotime, genetic perturbation screens), hence leveraging event-related logic to reveal the regulatory direction [13–15].

In this review, we discuss state-of-the-art network inference methods that integrate transcriptomics and chromatin accessibility data for (e)GRN inference at the single-cell level. We first introduce technologies, data, and preprocessing steps necessary for (e)GRN inference. Next, we discuss different correlation, regression, probabilistic, and deep learning models of network inference that combine (sc)RNA-seq and (sc)ATAC-seq data, and on that base, identify shared and unique components for successful network inference. We further point out opportunities, pitfalls, and challenges and discuss some aspects that still lack consensus in the community concerning best practices. We end the review with an outlook of directions in which the (e)GRN inference field might evolve in the near future.

High-throughput technologies for transcriptomics and chromatin accessibility

Methodologies used to chart single-cell omics divide into plate-based and microfluidics droplet-based systems. Plate-based systems such as SMART-seq separate cells into wells on plates and obtain full-length transcript coverage per cell [16]. Droplet-based strategies such as Chromium single-cell gene expression from 10x Genomics enclose single cells in uniquely barcoded GEMs (Gel Bead-In Emulsions) utilizing microfluidic devices and generate larger numbers of sequenced cells, but with less transcripts per cell and coverage only from the 5′ or 3′ end of the transcript [17]. However, newer plate-based and droplet-based methods show significantly increased performance and throughput by reducing hands-on time in the lab [5, 18], increasing throughput via multiplexing [19, 20], or split-pool combinatorial barcoding [21]. For chromatin accessibility, the core principle of (sc)ATAC-sequencing for processing cells is the same as in (sc)RNA-seq, but targets DNA and involves the hyperactive transposase Tn5 to fragment open chromatin regions of cell nuclei and simultaneously tag the DNA with adaptors for sequencing [22, 23]. Methods are plate-based, such as sciATAC-seq [24, 25], or droplet-based, such as Chromium single-cell ATAC from 10x Genomics. Recent single-cell multi-omics approaches can jointly profile both modalities by separating the transcriptome and accessible chromatin libraries from the same cell [5, 18]. Examples are plate-based methods such as Paired-seq [26] and SHARE-seq [27] and droplet-based methods like SNARE-seq [28] and the 10x Genomics Multiome [5, 18].

Preprocessing of transcriptome and chromatin accessibility data

Preprocessing of single omics data starts from demultiplexing of raw sequence files and assigns reads to features (transcripts, genomic regions) and cells using cellular barcodes. The cardinal preprocessing steps include trimming and filtering of minimally expressed reads, filtering of poor-quality cells, Polymerase Chain Reaction (PCR) duplicate removal [e.g. based on unique molecular identifiers (UMIs)], read alignment, peak calling for scATAC-seq, and cell-by-feature count matrix with transcripts [29, 30] or genomic regions as features [28, 31, 32]. This count matrix is further normalized to control for nonmeaningful sources of variability (e.g. sequencing depth or batch effects). The next steps are feature selection and dimensionality reduction. For a more detailed view on processing steps, we refer to Supplementary text 1.

Single-cell count matrices are high-dimensional, sparse structures (e.g. typically containing >20 000 TFs/TGs or >100 000 REs), while the number of cells for individual samples are often much lower than the number of features. This concept is known as the curse of dimensionality, and linear methods such as principal component analysis, matrix factorization, and latent semantic indexing, and nonlinear methods such as autoencoders, aim to consolidate information from the high-dimensional space into fewer dimensions while preserving as much biological information as possible from the original data [33]. Although proper dimensionality reduction allows for effective noise removal and facilitates downstream analysis, it also alters the feature space [33]. Therefore, (e)GRN inference itself is mostly done in the original feature space, either directly on all features or on a set of selected features, based on highly variable genes (HVGs) or deviance for scRNA-seq [34] and differential or coaccessible regions (e.g. pyCisTopic [9]) for scATAC-seq. While some (e)GRN inference models are able to directly use filtered raw counts (e.g. SCENIC+ [9]), others need further normalization steps, e.g. log-normalized counts per million [35] or data transformations [36]. To address sparsity, imputation methods can estimate and replace missing count values; however, this is more often done on scATAC-seq than scRNA-seq data and might be detrimental [35].

GRN inference from single omics data

Originally, bulk GRN inference relied exclusively on gene expression profiles for inferring statistical dependencies between genes, using e.g. mutual information (CLR [37], ARACNE [38]), random forest regression (GENIE3, [39]), ensemble methods (Inferelator [40]), and module network inference (LemonTree, [41],

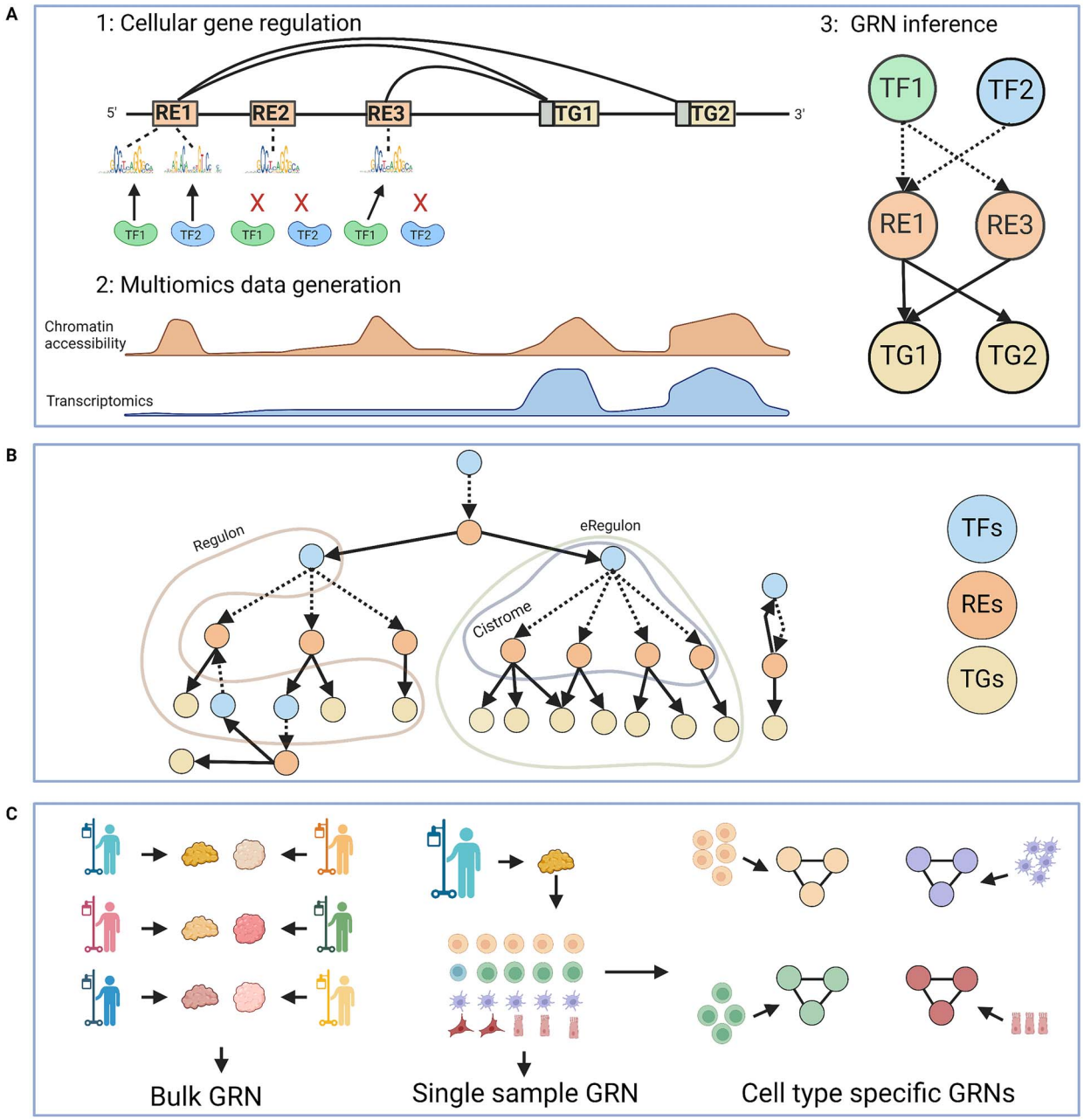

Figure 1. Combinatorial gene regulation between TFs, REs, and TGs can be modeled in eGRNs. (A) (1) In a simplified model, a TF binds to its TFBS within an active RE or promoter region, which is, or comes, in close contact with the TG promoter, and subsequently regulates TG transcriptional expression. (2) Chromatin accessibility and gene expression data provide complementary readouts of gene regulation. (3) These data can be used to reverse engineer eGRNs. (B) An example eGRN consisting of TF, RE, and TG nodes, connected with directed TF-RE and RE-TG edges. The set of all regulatory interactions derived from a tissue or cell type forms a context-specific eGRN. Important components of eGRNs are regulons (set of one TF and all its TGs), cistromes (set of genome-wide REs targeted by a specific TF), and eRegulons (one TF, its REs, and subsequently connected TGs). (C) Bulk GRN inference relies on many samples to be successful. Single-sample GRNs output one network per sample and are either inferred from bulk with single-sample network inference methods, or from single-cell data using all cells as observations. Cell type–specific network inference allows to resolve specific networks per cell type. Figure created with BioRender.com.

MERLIN [42]). Pioneering single-cell network inference methods mostly integrated or adapted methods that were designed for bulk network inference [7, 43]. One popular example is SCENIC [7], which combines regression trees–based approaches like GENIE3 or GRNboost and TFBS overrepresentation to generate cell state–specific GRNs. Later methods explored different inference techniques including deep learning (DeepDRIM, [44]), integration of time series (SINCERITIES, [45]), partial information decomposition (PIDC, [46]), ordinary differential equations (SCODE, [47]), or probabilistic matrix factorization (PMF-GRN, [48]). Some single-cell GRN inference methods were also optimized by incorporation of (prior-based) transcription factor activity (TFA), e.g. implemented as regulon pruning step [7], or TFA replacing TF expression as predictor [42, 49, 50]. One popular tool is decoupleR [49] implementing many different TFA inference algorithms, including VIPER [51], AUCell [7], or a consensus of several tools.

Many attempts have been undertaken to validate and benchmark GRN inference methods both at bulk [39, 52, 53] and single-cell level [43, 54–59]. While methods such as GENIE3, LemonTree, and SCENIC were shown to outperform other methods in public

benchmarks, GRN inference only achieved low performance, e.g. reflected in low area under the precision recall curve (AUPR) values across benchmarks. This is not only influenced by the incompleteness of ground truth data but also because of limited observations and incomplete modeling of regulatory mechanisms. Moreover, benchmarks revealed that methods relying purely on gene expression even struggle with outperforming random predictors [54, 60].

Besides transcriptomics data, (e)GRN inference can also be performed on chromatin accessibility data alone when harnessing TFBS within accessible REs. In the simplest case, a network is constructed by overlapping TFBS positions and open chromatin peaks and connecting them to genes within a specific genomic distance [61]. Detection of TFBS sites can be based on simple scanning or further be refined via overrepresentation analysis where occurrences are compared against a background model in which accessible regions are shuffled across the noncoding genome [7, 9, 62]. Alternatively, enhancer accessibility is correlated with core promotor or gene accessibility, compared against correlation scores of randomly selected regions to filter out nonsignificant interactions [63, 64]. Cicero, for example, computes the covariance of peaks within genomic blocks to produce a distance-weighted coaccessibility score, consecutively used to find RE-TG interactions [65]. However, (e)GRNs inferred from scATAC-seq alone differ from those inferred from scRNA-seq, as there is no perfect correlation between both modalities [66, 67].

(e)GRN inference from (sc)RNAseq and (sc)ATAC-seq data

Recently, multi-omics methods have started to dominate the GRN inference field. The combination of gene expression and chromatin accessibility contributes both omics-specific and complementary information. Thus, well-modeled multimodal integration of transcriptomics and chromatin accessibility addresses shortcomings of single-omics GRN inference and have been shown to improve the reverse engineering of regulatory interactions [9, 58, 68]. A variety of (e)GRN network inference methods combine ATAC-seq and RNA-seq data at the bulk or single-cell level, and their number is constantly increasing (Table 1, Supplementary file 1, Figs 2–4). Current state-of-the-art methods take different input types such as paired gene expression and chromatin accessibility measured in the same cell (e.g. TRIPOD [69], scREG [70]), measured in independent cells of the same sample (e.g. GRANIE [71], ANANSE [72]) or in metacells that algorithmically couple gene expression and chromatin accessibility (e.g. SCENIC+, FigR [73]). Some methods integrate additional data modalities, such as protein–protein interactions (PPIs) (SCORPION [74], HuMMuS [75]), histone modifications (ANANSE) or three-dimensional chromatin contacts (GRANIE). Methods that were originally developed for application on bulk data (e.g. ANANSE, GRANIE, or SPIDER [61]) were found to be applicable to single-cell data, especially upon transforming cell type–specific omics into pseudobulk data. Methods can also produce different types of output networks, including fully modeled eGRNs, GRNs, or only TG-RE or RE-TG interactions, and mostly follow mathematical principles of regression, correlation, probabilistic models and deep learning. In the following sections, we will review the important concepts and elements of (e)GRN inference based on examples from state-of-the-art (e)GRN inference methods.

Pseudobulk and metacells to tackle sparsity in single-cell data

To deal with sparsity for single-cell ATAC data (only 1%–10% of peaks are called in each cell [76]), imputation or transformation such as latent semantic indexing (LSI) [77], latent Dirichlet allocation (LDA) [78], or spectral embedding is applied. While some degree of sparsity is expected caused by biological effects, it is also introduced by drop-out values. Furthermore, sparsity of both RNA and ATAC-seq data can be addressed with pseudobulking. In pseudobulking, either all or a subset of reads (e.g. from a cell type) are mapped together to a reference. This reduces technical noise, enables differential gene expression or differential chromatin accessibility analysis, and allows to confidently define enhancers and even to detect TF footprints upon sufficient sequencing depth. However, pseudobulk loses the single-cell resolution, which makes it difficult to detect rare cell types or to explore the dynamic behavior across cells. A middle way between single-cell resolution and pseudobulking is done by grouping only similar cells into so-called metacells. We define a metacell here as an aggregated group of cells representing a distinct cell state, whereby within metacell variation is assumed to originate only from technical rather than biological sources [79, 80]. Metacells are usually inferred using a measure of similarity that is computed within the lower-dimensional embedded space of cells. For example, SEACells can infer metacells from scATAC- and/or scRNA-seq data. It constructs first a *k*-nearest neighbor graph based on an embedding, generates an affinity matrix with adaptive Gaussian kernel transformation to capture nonlinear cell relationships, and employs an archetypal analysis on the kernel matrix to identify clusters representing distinct cell states, subsequently labeled as metacells [80]. Another example of a metacells generating tool is Cicero, which samples highly similar cells based on a *k*-nearest-neighbor graph for scATAC-seq [65, 77]. However, it is important to note that metacells trade observations in exchange for an increase in counts, which might introduce a nonbiological bias. For example, for a metacell that includes an exceptionally low number of cells, an averaging effect might only increase technical variation in gene expression.

Computational pairing of scRNA-seq and scATAC-seq using sequencing data

Given the idea that cells with comparable properties can be grouped and might reflect similar cell states, it is also possible to computationally pair similar cells from unpaired scATAC-seq and scRNA-seq modalities. Ideally, both modalities should follow a similar distribution and are measured by the same process, while in reality, scRNA-seq follows a negative binomial distribution and scATAC a quasi-binary distribution with most regions being supported by one or two reads only [81]. However, tools rely on explicit assumptions on the underlying biology, and predict the 'activity' of each gene as a proxy for gene expression in each cell of the scATAC-seq data based on the accessibility of the gene's surrounding chromatin, and subsequently map cells from the two modalities into the same feature space based on gene expression and chromatin accessibility. For example, SIMBA jointly embeds single cells and their features into a common latent space allowing to link cells from different modalities [82]. DIRECT-NET first creates metacells by aggregating similar cells learned from a *k*-nearest neighbor graph for scATAC-seq and scRNA-seq, respectively, and consequently generates a single

Table 1. Overview of (e)GRN inference methods.

Method	Meta-cells	Additional inputs	RE definition	Algorithmic principles	Time	Perturbation	Cell type GRN	Output	Language
ANANSE	No	H3K27ac, ChIP-seq	Literature (ReMap)	Interaction score	Snapshot	No	No	GRN	Python
CellOracle	No	Base GRN	Data driven	Regularized regression	Snapshot	Yes	Yes	GRN	Python
CNNC	No	None	Data driven	Convolutional neural network	Snapshot	No	Yes	GRN	Python
DeepMAPS	No	None	Data driven	Graph autoencoder	Snapshot	No	Yes	GRN	HTML, Python, R
Dictys	Yes	None	Data driven	Differential equations	Pseudotime	No	Yes	GRN	Python
DIRECT-NET	Yes	None	Data driven	Gradient boosting, topic modeling	Snapshot	No	Yes	eGRN	R
FigR	Yes	None	Data driven	Spearman correlation	Snapshot	No	No	eGRN	R
GLUE	No	k-different omics, Hi-C, GWAS	Data driven, literature (ENCODE)	Variational auto encoder	Snapshot	No	Yes	RE-TG	Python, R
GRaNIE	No	Optional Hi-C	Data driven	Pearson correlation	Snapshot	No	No	eGRN	R
HuMMuS	No	PPI, Hi-C, snmC-seq, ...	Literature and data driven	Random walk with restart	Snapshot	No	Yes	eGRN	Python, R, CSS, HTML
Inferelator 3.0	No	None	Data driven	BBSR, StARS-LASSO, AMuSR	Snapshot	No	No	GRN	Python
Intrinsic	No	None	Data driven	Logarithmic saturation regression	Snapshot	Yes	Yes	eGRN	R, MATLAB
IReNA	No	None	Data driven	Pearson correlation	Pseudotime	No	No	GRN	R
LINGER	No	None	Data driven	Multilayer Neural Network + SHAP	Snapshot	No	Yes	eGRN	Python
MAGICAL	No	Hi-C	Data driven	Bayesian framework	Snapshot	No	Yes	eGRN	R, MATLAB
Merlin-P	No	ChIP, Knockout	Literature or data driven	Probabilistic graphical models	Snapshot	No	No	GRN	C
MTLRank	No	ChIP	Data driven	Multi-task—multilayer neural network, SHAP	Snapshot	No	Yes	GRN	Python
Pando	Yes	None	Conservation & data driven	Regression	Experimental time points	No	No	GRN	R
PECA2	No	PPI	Literature (ENCODE)	Bayesian, logistic regression	Snapshot	No	No	GRN	MATLAB
RENIN	Yes	None	Data driven	Elastic net regression	Snapshot	Yes	Yes	eGRN	R
scAI	No	None	Data driven	Perturbation correlation	Snapshot	No	Yes	eGRN	R, MATLAB
SCENIC+	Yes	None	Data driven	Gradient boosting, topic modeling	Snapshot	Yes	Yes	eGRN	Python
SCENT	No	None	Data driven	Poisson regression	Snapshot	No	Yes	RE-TG	R
scKinetics	No	None	Unknown	Ordinary differential equations	Pseudotime	No	Yes	GRN, velocities	Python
scMEGA	Yes	None	Data driven	Pearson correlation	Pseudotime	No	No	eGRN	R
scMTNI	No	Linage tree	Data driven	Bayesian probabilistic graphical model	Snapshot, Pseudotime	No	Yes	GRN	C, C++, python, R, MATLAB
SCORPION	No	PPI	Data driven	Message passing	Snapshot	No	No	GRN	R
scREG	No	None	Data driven	Joint matrix decomposition	Snapshot	No	Yes	eGRN	R
scREMOTE	No	Hi-C-database	Hi-C-database	Linear regression	Snapshot	Yes	No	eGRN	R
STREAM	No	None	Data driven	Steiner Forest problem, IRIS-FGM, Cicero	Snapshot	No	No	eGRN	R
Symphony	No	None	Data driven	Bayesian	Snapshot	No	No	GRN	Python
TimeReg	No	None	Data driven	Bayesian, logistic regression	Experimental time points	No	No	eGRN	MATLAB
TRIPOD	Yes	None	Data driven	Correlation	Snapshot	No	Yes	eGRN	R

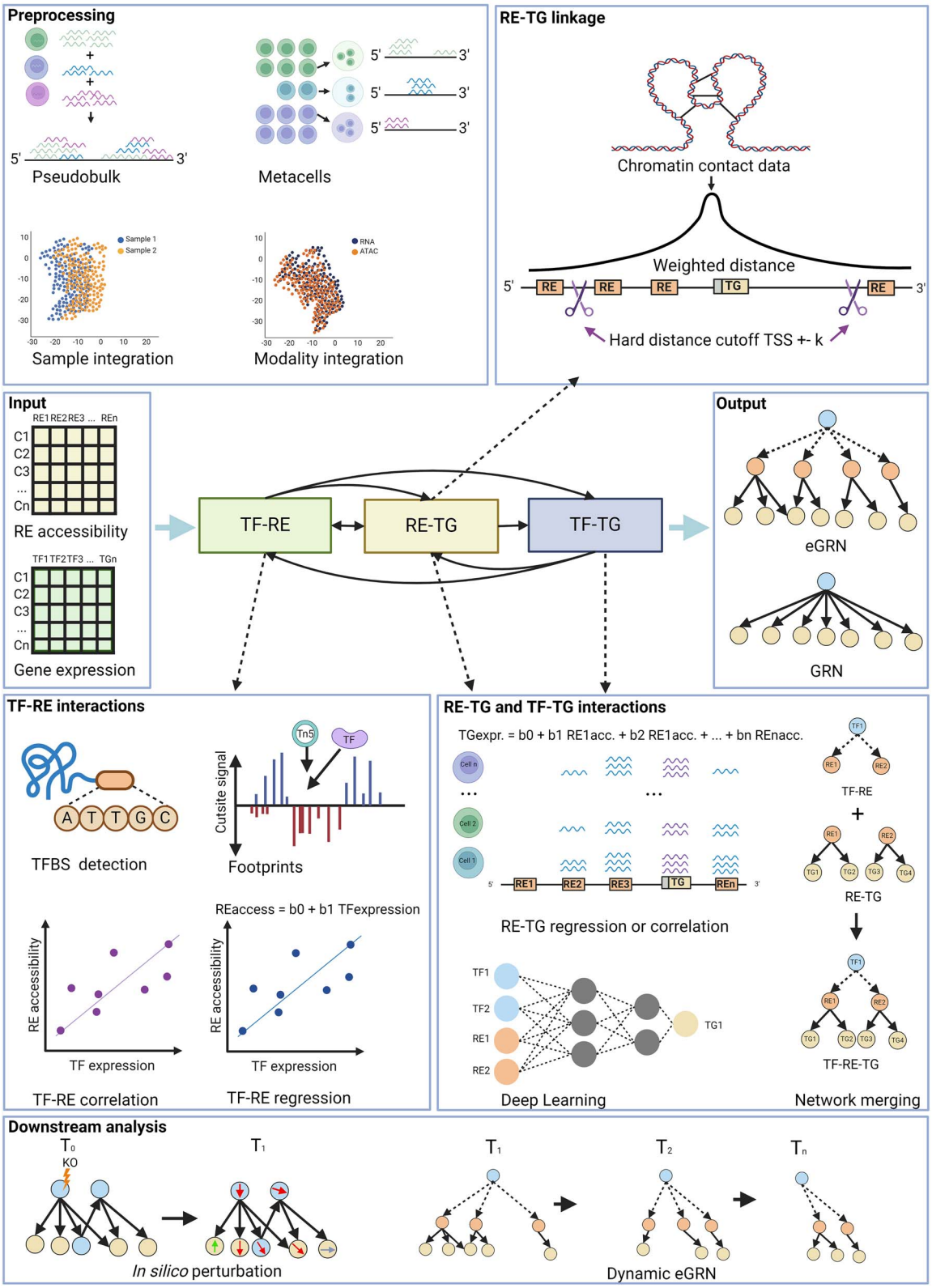

Figure 2. Schematic overview of (e)GRN inference from both transcriptomics and chromatin accessibility data. As input, (e)GRN inference methods take gene count and chromatin accessibility matrices. The core steps of (e)GRN inference include the detection of TF-RE, RE-TG, and TF-TG interactions, done in a stepwise or integrated manner. Preprocessing: Samples can be integrated together, e.g. to enable GRN inference at atlas level. Modalities can be integrated, e.g. by computationally coupling cells from a scRNA and scATAC-seq dataset. Pseudobulk increases detection strength of scATAC-seq peaks. RE-TG linkage: Potential TF-RE connections can be derived based on genomic distance or chromatin conformation data. TF-RE interactions: Requires the detection of TFBS within REs. Footprinting approaches aim to find evidence of TF binding events within REs. TF expression and chromatin accessibility can be further statistically investigated, mostly via correlation or regression. RE-TG and TF-TG interactions: Regression, correlation, probabilistic models and deep learning approaches are used to predict TG expression from TFs and/or REs. In case of stepwise GRN inference, TF-RE and RE-TG layers are merged via shared nodes. Output: While some methods model TF-RE-TG networks (eGRNs), others generate TF-TG interactions. Downstream analysis: "*In silico* perturbation" allows to investigate changes in TG expression propagated through the inferred GRN. Some methods also model (e)GRNs over experimental timepoints or pseudotime, and thus produce dynamic (e)GRNs. Figure created with BioRender.com.

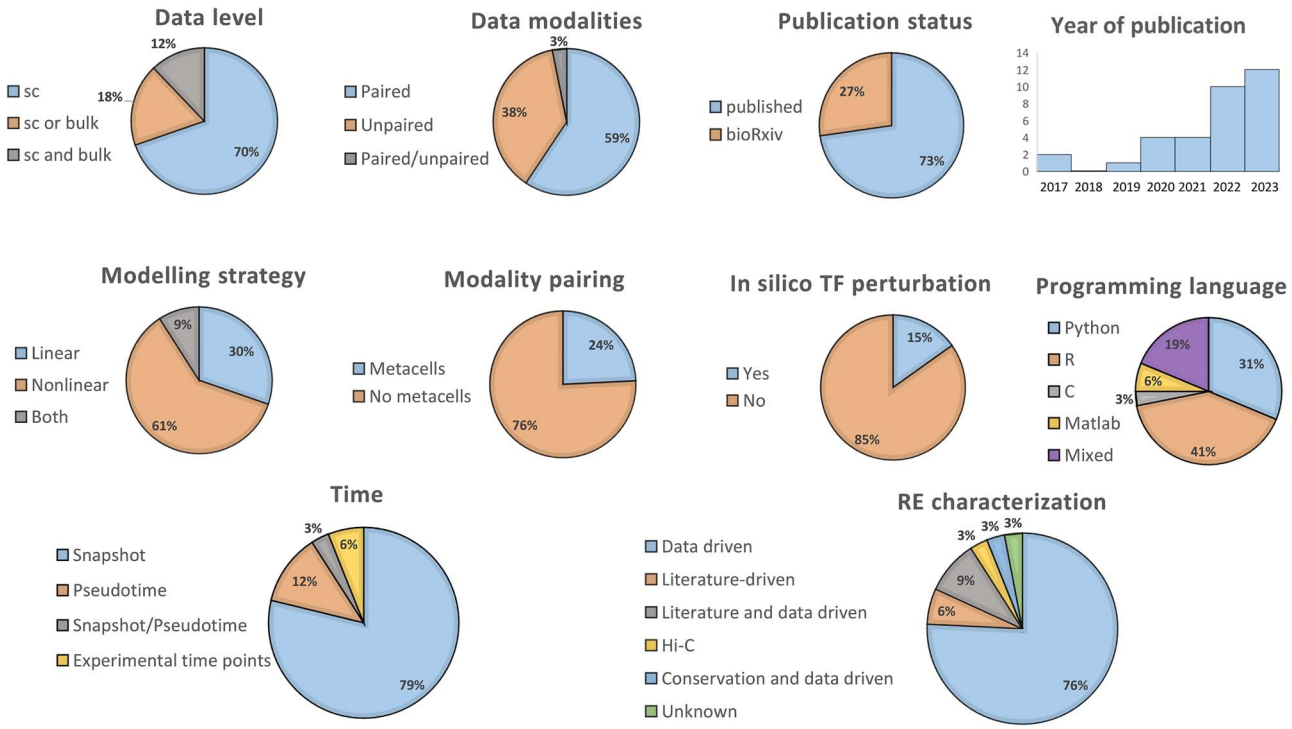

Figure 3. Distribution of (e)GRN inference method characteristics. All properties are based on our selection of inference methods listed in Table 1, which were developed between 2017 and 2023. Publication status threshold was 2023.

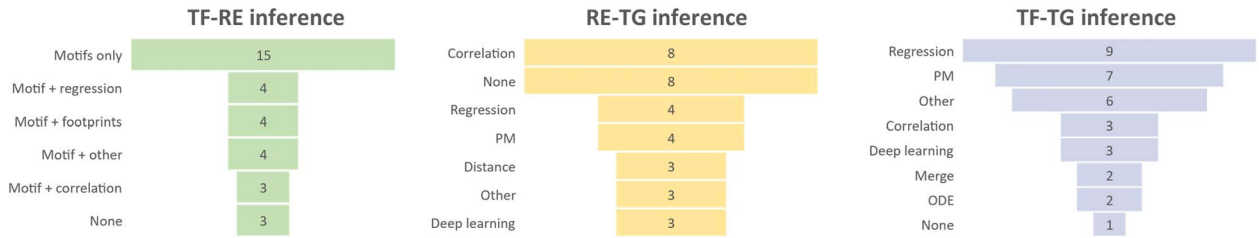

Figure 4. Regression and correlation approaches together make up the majority of strategies across all three inference steps. Approaches and strategies that were only used in one single method are summarized in the category "Other," while methods that do not model the corresponding step are put in the category "None." ODE = or dinary differential equation; PM = probabilistic model.

unified low-dimensional representation to couple similar cells between different modalities [83]. SCENIC+ samples a predefined number of cells from each data modality within the same cell type annotation label and averages the raw gene expression and imputed chromatin accessibility data across these cells to produce a multiome metacell [9]. FigR, on the other hand, applies the scOptMatch algorithm where cells from both assays are first placed in a shared coembedding using canonical correlation analysis and next sampled within a *k*-nearest-neighbors subgraph while accounting for cell number and cell type imbalances [73]. Pando uses a minimum-cost maximum-flow bi-partite matching in a canonical correlation space and summarizes matches into metacells [84]. GLUE won NeurIPS 2021 multimodal single-cell data integration challenge and models cell states as low-dimensional embeddings learned through modality-specific variational autoencoders [85]. It incorporates prior knowledge in the form of a guidance graph to explicitly model cross-modality regulatory interactions, e.g. between REs and TGs from scATAC and scRNA, respectively [85]. With increasingly available paired multi-omics data, mosaic integration, where only a subset of cells and/or features can serve as anchors, becomes feasible [86, 87]. Computationally matched cells can be used for (e)GRN inference and be processed similarly as if the features were directly measured in the same cells.

Characterization of REs

To infer eGRNs from a specific dataset at hand, REs need to be first identified by either literature-based or data-driven approaches. Literature-based approaches assign REs based on external sources of known enhancers, including ENCODE, VISTA, and Enhancer Atlas annotations, which are based, for example, on a consensus of ChIP-seq, ATAC-seq, MNase-seq, CAGE, ChIA-PET, and chromatin marks such as H3K27ac [88–91]. Also, additional biological prior information such as evolutionary conservation of REs can be considered [84]. On the other hand, data-driven approaches infer REs directly from ATAC-seq data. Reads are first pseudobulked either on all cells or preferentially per annotated cell type, and subsequent peak calling returns a set of regions serving as set of potential REs [9, 35, 76]. Overlapping peaks across cell types can also be merged and filtered based on peak scores to obtain a consensus [9]. Literature-based approaches offer a more reliable RE definition since they are independent from sequencing depth and noise. Data-driven approaches, however, add context specificity and sensitivity, even at the cell type level.

Literature-based as well as data-driven RE definitions can also be combined to define enhancers [72, 84].

Inference of TF-RE interactions

Linking a TF to one or more potential REs is based on binding evidence of the given TF to the DNA sequence of the RE. A direct way of measuring genome-wide TF binding is given by TF-specific binding assays based on immunoprecipitation such as *in vivo* ChIP-seq, CUT&RUN [92], CUT&TAG [93, 94], and their single-cell counterparts [95] or *in vitro* methods like protein binding microarrays and HT-SELEX [96]. While these assays measure context-specific TF binding, they can be expensive, require the availability of a TF-specific antibody, and detect mainly high-affinity binding. Hence, this experimental TF binding information is not available for all TFs and certainly not in all conditions [88]. Computational TFBS discovery approaches rely on DNA sequence information to search for overrepresented DNA patterns in genomic regions of interest. These patterns are often learned on TF binding experimental data, and prioritizing genomic regions with reduced crowdedness for general TF binding was shown to significantly improve this process [97]. Examples for *de novo* TFBS discovery and scanning algorithms include tools in the MEME Suite (e.g. FIMO [98]), RSAT [99], HOMER [100] (e.g. in PECA), the R package motifmatchr [101] (e.g. in SPIDER, IReNA, Cicero), and gimmemotifs [102] (e.g. in CellOracle). While computational approaches are cheap and already cover a wide range of TFs, they lack context specificity. Moreover, not all TFs have known TFBS, and some binding might be ambiguous for TF family members [102]. Besides, TF-TF or TF-cofactor interactions, DNA modifications and shape, and the orientation of a TFBS all influence TF binding [103, 104]. Restricting the search space to accessible REs and/or REs with overrepresented TFBS partially overcomes their lack of cell type or context specificity, as each cell type will now rely on their unique set of accessible and/or overrepresented TFBS (e.g. done by SCENIC+ and/or GLUE [9, 85, 105]).

It is worth noting that the way in which TFBS are represented can pose challenges due to varying model capacity as well as ease of use. Most TFBS detection methods rely on simple Position Weight Matrix (PWM) representing probabilities for nucleotide occurrences at specific positions, assuming statistical independence between distinct positions in the TFBS. Other representations exist that might capture more or complementary motif characteristics, including high-order PWMs, probabilistic graphical modes, support-vector machine models and deep neural networks [REF], although they are rarely seen in (e)GRN inference.

Open chromatin is no guarantee for TF binding [106]. Pioneer TFs also initiate binding in still closed chromatin regions and therefore cannot be detected only based on chromatin accessibility [107]. TFBS footprinting methods on ATAC-seq or pseudobulk scATAC-seq data offer a solution to obtain quantifiable evidence for those binding events. Briefly, a TF footprint is a pattern characterized by a lower read count at the TFBS compared to its surrounding bases as the transposase enzyme competes with the TF in the ATAC-seq assay. This information can be extracted by methods such as TOBIAS [108], PIQ [109], Wellington [110], or HINT-ATAC [111] and is also exploited in (e)GRN inference methods such as ArchR, IReNA, and Dictys [77, 112, 113]. Footprinting methods, however, need a deep sequencing coverage and therefore can only be applied at the bulk or pseudobulk level. Alternatively, TFBS detection can be restricted to regions with active chromatin marks such as H3K27ac and H3K4me1 [72]. Recently, deep learning approaches such as autoencoders and convolutional neural networks are on the rise that combine different assays, priors, and sequence information to improve condition-specific TFBS prediction [114] and *de novo* motif discovery [115]. For example, maxATAC makes use of cell line– and TF-specific ChIP-seq data from literature as ground truth to learn TFBS from ATAC-seq and DNA sequence [114].

Having known TFBS detected within open chromatin already suffices to characterize TF-RE interactions, as shown in the SPIDER seed network [61]. However, TF-RE interaction inference can be further improved by regressing or correlating the gene expression or TFA of the TF with RE chromatin accessibility followed by pruning based on the false discovery rate (FDR). GRANIE realizes this by comparing correlations of a TF within the (e)GRN to the correlation against a set of random TF-RE interactions not present in the (e)GRN [71]. FigR performs a peak set enrichment test for TF-RE matches based on a GC-content-matched permutated background peak set and correlates RE accessibility scores with gene expression levels of tested TFs [73]. scMEGA [116] correlates the binding activity for each TF estimated with chromVAR [105] on accessible chromatin regions with the TFs gene expression. However, care must be taken to not remove important true-positive interactions with too stringent filtering or if biological assumptions are not met, e.g. for pioneer factors, where chromatin remains accessible once it is opened, or in case of a temporal delay between the opening of chromatin and TF expression [117].

Linear and three-dimensional approaches to characterize RE-TG interactions

Besides their connection to TFs, REs also need to be linked to potential TGs. Chromatin three-dimensional architecture, e.g. obtained from ChIA-PET [118], Hi-C [119], or scSPRITE [120], gives experimental evidence of physical chromatin interactions between REs and TGs. However, these data are often not available for the same sample on which (sc)RNA- and (sc)ATAC-seq were performed and suffers even greater sparsity than scATAC-seq. Each Hi-C interaction requires two reads bound to different regions, resulting in a quadratic increase in sequencing depth to probe all pairings. While features such as Topologically Associated Domains (TADs) can be confidently detected, they possess a median size of ~1 Mb [121, 122]. This resolution is, however, insufficient to comprehensively call interactions between enhancers and promoters, since the majority of interactions occur within a range of 0.5 Mb around the TSS [63, 123, 124]. Methods like capture-C [125] or promoter capture-Hi-C [119] can increase resolution by only investigating interactions of specific promoter or enhancer bait sequences. Another problem is that chromatin contact, chromatin accessibility, and transcriptomics data cannot yet be measured together in the same cell. Methods that rely on Hi-C data such as scREMOTE [126] bypass this problem by using databases of measured Hi-C interactions as baseline, such as the 4D Nucleome Data Portal [127]. However, most methods instead constrain the potential number of RE-TG interactions by setting a distance cut-off on the linear genome starting from the TSS of a TG, either with a hard threshold (mostly ±100–250 kb from the TSS) or weighted by distance from the TSS (based on exponential- or partial decay, power-law function, or logarithmic functions (Fig. 5)) [70, 85]. The chosen distance cut-off has a strong impact on the resulting (e)GRN: if the threshold is set too small, relevant regulatory interactions might be missed; if the threshold becomes too large, regulatory interactions combinatorically explode with false-positive interactions. Generally, the ATAC-seq signal is enriched around the TSS, especially in the promoter region [128], and shows decreasing chromatin accessibility as well as a lower

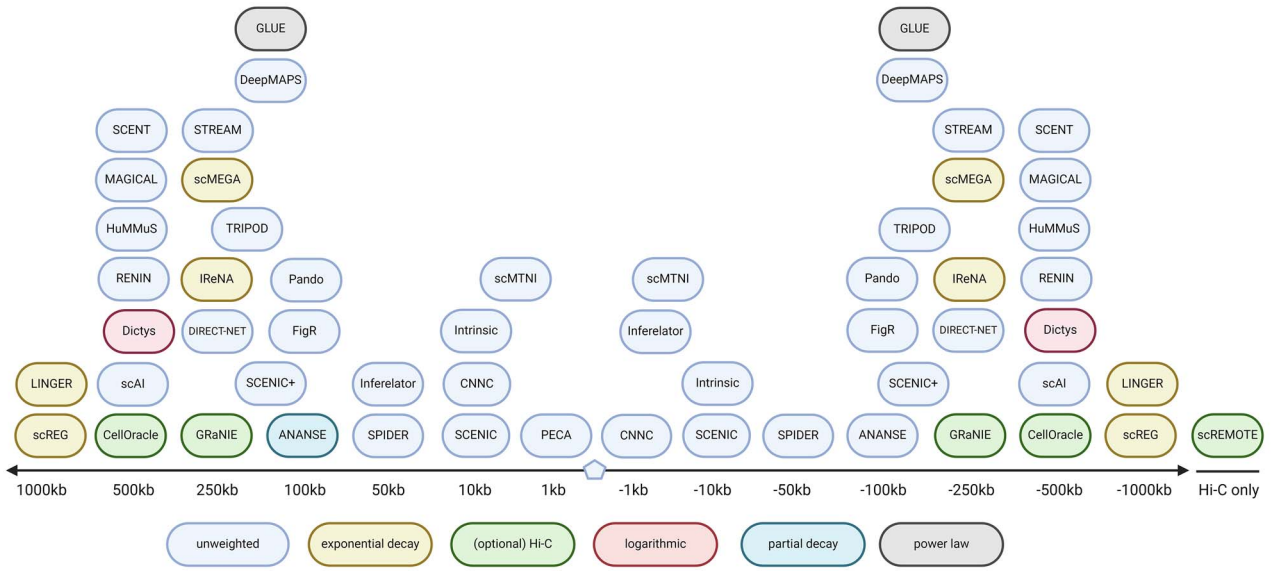

Figure 5. Different default distance thresholds for candidate RE-TG interactions. Methods that apply a hard threshold on the genomic distance around a TSS consider every interaction of all RE-TG pairs located within the defined window. The up- and downstream distance is identical for most but not all GRN inference tools. Other methods impose a distance-dependent weight function by altering the strength of RE-TG interactions dependent on their distance. These weights can follow distinct functions, including exponential decay, partial decay, logarithm, or a power law. Partial decay applies a weight function only for a defined genomic range, e.g. by only weighing interactions the region further away than ±5 kb from the TSS. Some methods are also capable of integrating chromatin contact information or entirely depend on them (scREMOTE). Figure created with BioRender.com.

number of chromatin contacts further away from the promoter [129, 130]. However, despite their overall lower frequency, long-range RE-TG interactions can be crucial for development, as the example of an 850 kb distant RE crucial for limb development in mice demonstrates [131, 132]. Distance-dependent weight functions cannot account for this and will most likely miss such interactions.

Inference of RE-TG interactions

While REs and their candidate TGs inferred based on a distance threshold or chromatin contacts already constrain the number of possible interactions, they need further refinement. Most methods solve this by applying correlation, regression, probabilistic models, or deep learning. These approaches differ in the way they model chromatin accessibility, either as a binary on/off signal or as a quantitative variable based on read counts. Correlation-based approaches mostly exploit the correlation between RE accessibility and TG expression, combined with multiple hypothesis testing. ArchR applies Pearson correlation on paired cells containing measures of gene expression and chromatin accessibility, respectively, and filters interactions based on correlation coefficient thresholds [77]. Methods such as GRANIE, Corces [63], FigR, scAI, and IReNA calculate correlation coefficients between candidate RE-TG pairs. To filter out low-confidence interactions, they calculate a background distribution based on the correlation of randomly sampled RE-TG pairs to estimate the FDR for filtering out low-confidence interactions [63, 71, 73, 112, 133]. Similarly, perturbation approaches set either gene expression or chromatin accessibility to zero and calculate the correlation with the other modality, while only keeping the RE-TG interactions with sufficient difference between perturbed and nonperturbed correlation coefficients, for example, scAI [133]. PECA performs a tissue crossing correlation by dividing bulk samples into groups of high and low openness and calculates a fold change of TG expression as well as promoter accessibility for the high openness group compared against a random group of the same size [134]. Some methods also perform additional pruning steps, such as removing interactions with negative correlation (GRANIE), prioritization TGs included in a high number of peak–gene interactions (FigR), or intersecting predictions with interactions from external sources such as ENCODE and ChIA-PET (PECA) [71, 73, 134].

Regression approaches employ predictive models to estimate a response variable, such as TG expression, TG promoter chromatin accessibility, or RE chromatin based on a regressor variable, such as TF expression or TF activity. If regression uses multiple regressors, the pairwise interaction strength between RE and TG is calculated either with importance scores (e.g. SCENIC+) or prediction coefficients (e.g. for multiple linear regression, CellOracle) that quantify how much a single predictor contributes to estimate a response variable. The interaction strength can then be used to rank and filter interactions. Methods such as DIRECT-NET and SCENIC+ apply gradient boosting to predict gene expression or promoter accessibility from enhancer accessibility [9, 83]. SnapATAC goes the other way around and uses TG expression to predict the binary state of potential REs with logistic regression [135, 136]. SCENT assumes that binarized RE accessibility follows a Poisson distribution and uses a Poisson-generalized linear model to predict gene expression, while at the same time modeling confounding factors such as mitochondrial counts, UMI counts, and batch effects [130]. The significance of coefficients for each RE is then assessed with a nonparametric bootstrapping procedure that allows for an effective type I error control and detection of causal variants in enhancers [130]. Overall, inference of RE-TG interactions is still mostly done using linear regression, and little effort is made to account for sparsity and nonlinearity.

Direct inference of TF-TG interactions

Another class of methods makes use of chromatin accessibility and transcriptomics data to predict TF-TG interactions without explicitly modeling REs in the resulting GRN. Prior based methods (e.g. MERLIN-P(-TFA), PANDA [137] or inTRINSiC [138]) construct matrices containing putative interactions based on chromatin

accessibility data and subsequently integrate them with transcriptomics data to infer a GRN. Recent benchmarking approaches demonstrated increased performance when integrating prior knowledge as opposed to expression data alone, especially for probabilistic models such as MERLIN-P [58]. Specifically developed for single-cell applications, CellOracle creates a base GRN of candidate regulatory interactions from scATAC-seq data by first linking TFs via TFBS detection within accessible regions and subsequently applying Cicero to couple accessible REs to TGs [35]. Interactions of this initial network are further pruned by predicting TG expression from TF expression with a regularized linear machine learning method. Also, deep learning approaches such as Convolutional Neural Network for Coexpression (CNNC) [36] transform gene expression data of gene pairs into image-like objects and infer relationships as well as causality between genes, while considering prior data such as chromatin accessibility. Finally, MTLRank utilizes a Multi-Task Learning Rank approach with scATAC-seq-derived TFA and scRNA-seq-derived gene expression based to predict TG velocity instead of gene expression [139]. The generated models are subsequently interrogated to rank TF-TG interactions with deep SHapley Additive exPlanations (SHAP) that measure the importance of TFs for regulation. GRN inference methods directly predicting TF-TG interactions incorporate RE information in the background. Since they process both data modalities independently, they are good choices when paired modalities are not available or if users want to incorporate prior information of any kind.

TF-RE-TG inference

To infer TF-RE-TG interactions, methods predominantly implement one out of two different strategies. The first strategy infers interactions in a transitive manner by connecting TF-RE and RE-TG interactions via the shared RE node. Examples of methods using this strategy are Direct-Net, GRANIE, and SCENIC+. Moreover, SCENIC+ additionally uses random forest regression or gradient boosting to calculate TF-TG importance scores and refines TF-RE-TG modules with a Gene Set Enrichment Analysis (GSEA) to filter out eRegulons with a small amount of TGs as well as RE-TG interactions with negative correlation [9].

The second strategy infers TF-RE-TG interactions in an integrated manner where all three components are simultaneously considered in the model. Considering three-way interactions enables the investigation of different modes of regulation and allows, for example, to distinguish cases where a TF acts as pioneer factor, opposed to a TF that binds to already open chromatin [69]. TRIPOD uses a unique non-parametric model-free Spearman's test for detecting conditional associations and three-way interactions between TF, RE, and TG [69]. For this, they use two tests that investigate relationships between RE and TG while keeping TF expression constant, as well as relationships between TF and TG while keeping RE accessibility constant [69]. Additionally, TRIPOD investigates whether the magnitude of change in the third variable depends on the association between the other pairwise variables [69]. PECA [134] aims to jointly characterize relationships between four elements, namely, TFs, TGs, REs and additionally expression of chromatin remodelers (CRs) by utilizing a probabilistic model conditioned on TG, RE and CR.

Pando uses TF-RE pairs as independent variables to predict TG expression with a generalized linear model. Here, transcriptomics and chromatin accessibility data are joint dependent variables, and fitted coefficients are further pruned using ANOVA with multiple hypothesis correction [84]. HuMMuS, on the other hand, first connects several omics layers such as TF-TF interactions, RE-TG interactions, TF-TG interactions and chromatin contact data, and subsequently applies a Random Walk with Restart to infer several outputs including TF-TG, TF-RE and RE-TG interactions as well as highly connected communities within the (e)GRN [75]. Deep learning approaches such as LINGER first learn weights of a neural network for different tissues from bulk data, and in a second run refine the learned weights with single-cell data [140]. Both TFs and REs are represented in the input layer and interconnect in the second network layer to predict expression of one gene in the output layer.

Dynamic (e)GRN inference

Many biological processes have developmental or time course aspects, including timing and initiation of developmental cues, gene expression dynamics, cellular fate decisions, cell-cycle dynamics, and homeostatic regulation to ensure a proper balance between cell proliferation and cell death, all tightly controlled by gene regulation. However, omics are profiled at a given instance in time, generating a "snapshot" of the underlying system. Even when measuring different time points of the same system, each measurement remains a discrete snapshot, and real-time resolutions are hard to achieve. While the majority of (e)GRN inference algorithms also model static networks based on snapshot data, alternative notions of time like "pseudotime" can be extracted from snapshot data, which position cells according to their current state of development. Early methods such as VeloCyto [141] and scVelo [142] use unspliced RNA as predictor for future spliced RNA to determine rate and direction of changes in gene expression over time, leading to pseudotime ordering. However, those early methods were shown to fail in some instances, and careful assessment of results is required [143]. CellRank tries to predict the future state of a cell by generating cell–cell transition matrices, which, in addition to pseudotime inference, allow the detection of initial and terminal cell states along trajectories [144].

The pseudotime ordering of cells is then considered by specific algorithms to improve (e)GRN inference or model temporal regulatory events. IReNA uses averaged gene expression of cells with similar pseudotime and demonstrates more accurate GRN inference in comparison to raw gene expression when compared against ChIP-seq and perturbation-based ground truth data. Also, deep learning approaches such as MTLRank demonstrated performance improvements predicting velocity instead of gene expression [139]. Dictys partitions cells along a trajectory into bins, generates bin-specific eGRNs, and combines those to one dynamic eGRN with Gaussian kernel smoothing [113]. scMTNI incorporates pseudotime information in form of a lineage tree and infuses this prior knowledge about cell state relationships into their model to subsequently infer lineage-specific GRNs [145]. Finally, scKinetics simultaneously aims to solve the GRN inference task and velocity prediction in an iterative manner using an expectation maximization algorithm [146].

Another approach towards more dynamic networks is to first build a predictive model from snapshot data and afterward perturb TF expression within that model. The resulting expression changes are then transmitted through the (e)GRN to predict future time points. Existing models usually implement regression frameworks, and examples include CellOracle, SCENIC+, Intrinsic, RENIN or scREMOTE [9, 35, 126, 138, 147]. After generating cells with perturbed gene expression vectors, they can be compared to original expression to investigate direction and magnitude of potential differentiation trajectories. However, predictive methods

are limited due to simplified assumptions (e.g. linearity) and are applied in a more exploratory context. Lastly, a few tools also try to incorporate multiple discrete time points into the (e)GRN inference or analysis process. For example, TimeReg first infers networks with PECA2 at each time point, and driver regulators for core regulatory modules are studied across time points [148]. Pando first integrated developmental time points into a unified embedding and associates specific (e)GRN modules with stages of organoid development and segregation events during brain regionalization, highlighting how valuable time series data can be to understand complex biological systems [84]. However, time-point integration so far remains limited to few tools, and there is still a great potential of improvement, in generating time series data, in modeling dynamic time in current data and integrating it into (e)GRNs.

Conclusions

The integration of gene expression and chromatin accessibility for (e)GRN inference has improved network quality and allowed to investigate more causal and mechanistic interactions. However, the high number of publications on archives indicates that the development of eGRN inference methods is still highly active. The process of network inference contains several critical sub-problems to solve, including the coupling of data modalities; the definition of a potential enhancer–gene interaction space; and the establishment of trustworthy TF-TG, RE-TG, and TF-RE-TG relationships. Methodologically, most methods pursue either a predictive strategy, such as correlation, regression, or probabilistic models to predict either GRNs or eGRNs, while upcoming methods also explore deep learning or try to consider temporal dynamics in the models [9, 69, 71, 139]. Interestingly, most methods already rely on paired modalities, while most (high quality) datasets still come from separate cells, which requires a good integration strategy.

For the identification of TF binding, TFBS detection and over-representation are still a common choice, but comes with the drawbacks that not all TFs have well-characterized nonambiguous TFBS, and the capacity of a PWM model is limited in fully characterizing binding sites (Supplementary File 1). However, the number of characterized TFBS significantly grew over the past years, as shown by the impressive collection of 1553 human TFs in SCENIC+ [9]. Deep learning approaches that combine chromatin accessibility with sequence encodings could increase condition-specific RE modeling, but still suffer from low TF coverage regarding training data [114]. We expect deep learning models to rise as predominant means to model REs, but this development has still to be accompanied by an increase in trustworthy, condition-specific training data (e.g. ChIP-seq).

Another controversial question raised when constructing (e)GRNs is the selection of potential regulatory regions for a TG. Most current methods apply a wide range of different thresholds and weights on the linear chromosome as workaround, but there is no consensus on which distance is best suited to balance the combinatorial explosion of RE-TG pairs against an increasing number of false positives. Moreover, mechanisms of enhancer-mediated regulation are not yet fully understood [2] and novel mechanistic discoveries are still frequently published [149]. Different thresholds also render (e)GRN inference methods less comparable across studies. We believe that current efforts in integrating (predicted) chromatin contacts might deliver the edge to resolve the choice of candidate regulatory interactions in the future [119, 150].

Sparsity of single-cell data poses a big problem, especially for genes with low gene expression or chromatin accessibility. When modeling RE-TG interactions explicitly, binary, or quantitative chromatin accessibility scores are prone to be missed due to data sparsity, thus limiting the number of interactions that can be recovered. Metacells and scaling up the number of observed cells, e.g. by creating a cell atlas that integrates different samples, studies, or omics assays [151, 152], are ways of overcoming this. While atlas-level data will prove an invaluable source to study development and disease, eGRN inference methods will also need to deal with atlas-specific features, e.g. the potential loss of biological variability when normalizing for technical biases as a result from the integration process. Also, the integration process of unpaired scRNA-seq and scATAC-seq data is challenging but crucial for GRN inference. Since many methods rely on correlations or regression between features of both omics, multimodal data integration needs to preserve biological dependencies across both feature vectors and account for batch effects. This process is additionally complicated by the potential nonlinear relationships, kinetic differences, and temporal shifts [153]. Moreover, overcorrection can occur especially under strong batch effects or when cell type proportions are mismatched between omics and lead to pairing cells that do not share the same cell type or state [154]. Combining algorithmic development, the fast-growing amount of data, and constant improvements in single-cell sequencing technologies will help tackle this problem and increase the robustness of (e)GRN inference.

Most methods still do not consider dynamic changes over pseudotime or real timepoints. Dynamic changes are already predicted by several tools; however, they usually infer pseudotime based on the same RNA-seq data as the network and thus introduce some bias and “double dipping” effects where data are used twice for two prediction tasks [155]. This could be overcome with multimodal data, where tasks are separated between orthogonal omics. A second approach to studying the dynamic behavior of rewiring in (static) networks is done by introducing perturbations. Perturbed cell states can also indicate a developmental direction of cells upon changes in TF expression, allowing to generate hypotheses for trajectories, drug response, or therapeutic escape.

Given an inferred (e)GRN, experimental validation remains a critical task. Comparison against literature curated and database regulatory interactions such as OmniPath or orthogonal data such as ChIP-seq or perturbation screens can be used to approximate ground truth but might have limited credibility since they often lack context specificity or are themselves derived from high-throughput technologies [88, 156, 157]. Additional measures such as TF protein abundance and post-translational modification, TF binding and cooperativity, regulatory activity of REs, and chromatin contacts and TF perturbation can give complementary views on mechanisms involved in gene regulation [153]. Overall, it is best practice to assess as many metrics as possible to gain a more accurate picture of how well a method performs. When interpreting different metrics, it is important to consider their specific meaning and scope, their limitations, and their origin (e.g. neutral benchmark versus benchmarking of novel methods by their authors) to not draw false conclusions and to avoid overoptimism [158].

Future perspectives

For the future of (e)GRN inference, many directions have been proposed, focusing on multimodal integration, scaling, sparsity,

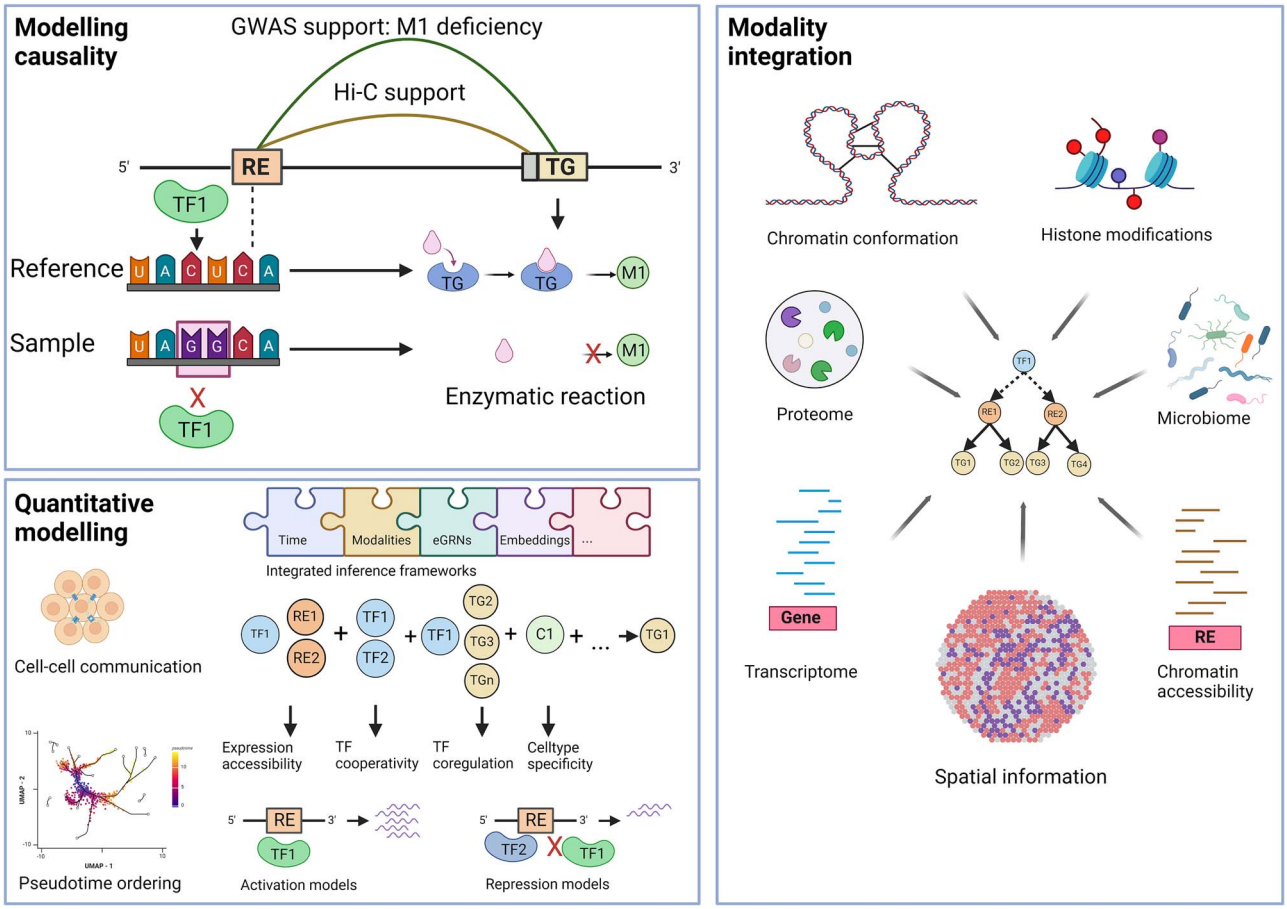

Figure 6. Future directions of (e)GRN inference. "Modeling causality," for example, establishing event chains from SNPs over gene regulation toward downstream biological processes such as misregulation of a disease-relevant metabolite (M1), in combination with publicly available GWAS data, not only add confidence to edges in an (e)GRN but also enable mechanistic interpretability. "Integration of multiple modalities" add complementary perspectives to (e)GRN inference. "Quantitative modeling" can integrate multiple regulatory influences into one integrated or multiple complementary models. Figure created with BioRender.com.

the 3D genome, TFBS predictions, causality and benchmarking [153, 159, 160]. We focus here on three main directions that we believe have great potential to further enhance our understanding of gene regulation, namely, quantitative modeling, causality, and multimodal integration (Fig. 6).

First, more quantitative modeling strategies can help to fine-tune (e)GRN inference. This could include modeling coregulation of multiple TGs, e.g. by incorporating knowledge of global TF binding patterns across the genome, or by directly modeling biophysical processes [161]. For example, right now repression and activation are often predicted with the same model and defined by the sign (positive or negative) of an interaction. However, both might have different underlying mechanisms, and using separate models might improve individual predictions. Another strategy could be to model (e)GRNs for a whole tissue without splitting (e)GRN inference tasks to one network per cell type, but instead model the cell type information as explicit parameter. Additionally, cells are often not isolated entities and communicate and interact with each other and influence gene regulation [162]. Considering additional models such as cell–cell communication, chromatin remodeling, or pseudotime inference directly in the inference step can further benefit the quality of these additional models as well as the (e)GRN inference [163]. Furthermore, direct integration of GRN inference with tasks such as dimensionality reduction, pseudotime inference, or timepoint incorporation is promising to preserve more biological signals compared to stepwise approaches, as recently demonstrated by scTIE [164]. Quantitative modeling can be a way of consolidating the current mechanistical knowledge into the models, which constrains them in a more biological meaningful way to get the most information out of the (still-often-limited) data and will allow to interpret the underlying biology more directly. This can be done by taking advantage of probabilistic as well as deep learning models or combinations of both. Finally, foundation models, such as those pioneered by scGPT, which are trained on vast amounts of data and fine-tuned to specific use cases, bear promise to effectively distill critical biological insights for which individual datasets lack statistical power [165].

Second, modeling of causality in (e)GRN inference needs to go beyond regression or correlation. While true, confounder-free causality will probably remain unachievable, the key is to collect as much evidence as possible and reconstruct the most likely chain of events that lead to a certain regulatory outcome. This can include incorporation of binding information, chromatin contact links to confirm looping events, incorporation of temporal and spatial information, perturbation experiments of single or

multiple regulators, control for nonregulatory influences such as functional relationships in protein complexes or pathways, and foundational knowledge on the role of cofactors and mediators. Interactions between all components can also benefit from conditional tests, similar to TRIPOD. Also, integration of genetic information such as Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs) and Genome-Wide Association Studies (GWAS) data can help understand the causes of (e)GRN rewiring, especially in disease context. We believe that establishing causality is not only crucial for inferring correct interactions and eliminating false positives but also directly aids the interpretation of (e)GRNs and thus making them a more attractive model to address complex biological questions.

Third, the integration of additional data modalities allows complementing evidence from other omics and filling in regulatory blind spots and mechanisms that RNA- and ATAC-seq data cannot pick up [6]. The inclusion of spatial omics will enable us to investigate distribution of edges along cell type boundaries and trajectories [166], improve classification of subpopulations and cell states, and cell–cell communication. Existing chromatin contact data already helps to refine RE-TG interactions, and increased availability of Hi-C with kb or base-pair resolutions and single-cell Hi-C bear great potential to further improve this. DNA methylation was shown to influence TF binding and can deliver informative features for (e)GRN inference [167]. Parallel measures of proteins, e.g. derived from CITE-seq, can reveal post-translational regulatory effects [168, 169]. Measures of histone modifications such as H3K27me3, H3K4me1, and H3K27ac allow to characterize poised enhancers, add causality, and further complete the picture of gene regulation. Investigation of actively perturbed systems reduces dependency on natural variation for GRN inference and allows for direct investigation of downstream effects caused by regulators [157]. Also, omics of external origin such as microbiome-induced gut metabolomics contain regulatory cues and should not be neglected when studying regulation from a system-wide perspective [6]. Finally, the inclusion of spatial omics will enable us to investigate distribution of edges along cell type boundaries and trajectories [166], improve classification of subpopulations and cell states, and cell–cell communication. In conclusion, we believe that integrating many complementary omics has the potential to rapidly grow our knowledge and improve (e)GRN inference if it is done in a meaningful and mechanistically informed way.

Key Points

- Gene regulatory network inference greatly benefited from single-cell technologies and parallel measurements of complementary omics layers such as transcriptomics and chromatin accessibility data.
- GRN inference methods resolve TF-RE, RE-TG and TF-(RE)-TG interactions by relying on diverse TFBS detection, correlation, regression, probabilistic models and deep learning approaches, with some inference steps being streamlined in the community and others solved in very diverse ways.
- Future developments will likely focus on the incorporation of more omics layers, efficient integration of larger data resources and studies, quantitative modeling, incorporation of spatial and temporal dimensions as well as on *in silico* perturbation.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

Figures 1, 2, 5, and 6 were created with BioRender.com.

Funding

Jens Uwe Loers is supported by a Bijzonder Onderzoeksfonds (BOF) PhD scholarship (01D28520) and Vanessa Vermeirssen is granted a BOF starting grant (BOF/STA/201909/030) from Ghent University.

References

1. Ray-Jones H, Spivakov M. Transcriptional enhancers and their communication with gene promoters. *Cell Mol Life Sci* 2021;**78**:6453–85. https://doi.org/10.1007/s00018-021-03903-w.
2. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* 2021;**22**:108. https://doi.org/10.1186/s13059-021-02322-1.
3. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 2019;**20**:207–20. https://doi.org/10.1038/s41576-018-0089-8.
4. Kim S, Wysocka J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol Cell* 2023;**83**:373–92. https://doi.org/10.1016/j.molcel.2022.12.032.
5. Vandereyken K, Sifrim A, Thienpont B. *et al.* Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* 2023;**24**:494–515. https://doi.org/10.1038/s41576-023-00580-2.
6. Vandemoortele B, Vermeirssen V. Molecular systems biology approaches to investigate mechanisms of gut–brain communication in neurological diseases. *Eur J Neurol* 2023;**30**:3622–32. https://doi.org/10.1111/ene.15819.
7. Aibar S, González-Blas CB, Moerman T. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6. https://doi.org/10.1038/nmeth.4463.
8. Liu T, Ortiz JA, Taing L. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* 2011;**12**:R83. https://doi.org/10.1186/gb-2011-12-8-r83.
9. Bravo, González-Blas C, De Winter S, Hulselmans G. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;**20**:1355–67.
10. Deschildre J, Vandemoortele B, Loers JU. *et al.* Evaluation of single-sample network inference methods for precision oncology. *npj Syst Biol Appl* 2024;**10**:18. https://doi.org/10.1038/s41540-024-00340-w.
11. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. *Front Genet* 2019;**10**:535. https://doi.org/10.3389/fgene.2019.00535.
12. Hill SM, Heiser LM, Cokelaer T. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Methods* 2016;**13**:310–8. https://doi.org/10.1038/nmeth.3773.
13. Morgan D, Studham M, Tjärnberg A. *et al.* Perturbation-based gene regulatory network inference to unravel oncogenic mechanisms. *Sci Rep* 2020;**10**:14149. https://doi.org/10.1038/s41598-020-70941-y.
14. Aalto A, Viitasaari L, Ilmonen P. *et al.* Gene regulatory network inference from sparsely sampled noisy data. *Nat Commun* 2020;**11**:3493. https://doi.org/10.1038/s41467-020-17217-1.

15. Aygün N, Liang D, Crouse WL. *et al.* Inferring cell-type-specific causal gene regulatory networks during human neurogenesis. *Genome Biol* 2023;**24**:130. https://doi.org/10.1186/s13059-023-02959-0.
16. Hagemann-Jensen M, Ziegenhain C, Sandberg R. Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. *Nat Biotechnol* 2022;**40**:1452–7. https://doi.org/10.1038/s41587-022-01311-4.
17. Matuła K, Rivello F, Huck WTS. Single-cell analysis using droplet microfluidics. *Adv Biosyst* 2020;**4**:1900188. https://doi.org/10.1002/adbi.201900188.
18. Baysoy A, Bai Z, Satija R. *et al.* The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;**24**:695–713. https://doi.org/10.1038/s41580-023-00615-w.
19. Zhang Y, Xu S, Wen Z. *et al.* Sample-multiplexing approaches for single-cell sequencing. *Cell Mol Life Sci* 2022;**79**:466. https://doi.org/10.1007/s00018-022-04482-0.
20. Stoeckius M, Zheng S, Houck-Loomis B. *et al.* Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* 2018;**19**:224. https://doi.org/10.1186/s13059-018-1603-1.
21. Rosenberg AB, Roco CM, Muscat RA. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 2018;**360**:176–82. https://doi.org/10.1126/science.aam8999.
22. Buenrostro JD, Wu B, Chang HY. *et al.* ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 2015;**109**:21.29.1–9. https://doi.org/10.1002/0471142727.mb2129s109.
23. Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat Rev Genet* 2022;**24**:21–43. https://doi.org/10.1038/s41576-022-00509-1.
24. Cusanovich DA, Daza R, Adey A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 2015;**348**:910–4. https://doi.org/10.1126/science.aab1601.
25. O'Connell BL, Nichols RV, Pokholok D. *et al.* Atlas-scale single-cell chromatin accessibility using nanowell-based combinatorial indexing. *Genome Res* 2023;**33**:208–17. https://doi.org/10.1101/gr.276655.122.
26. Zhu C, Yu M, Huang H. *et al.* An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol* 2019;**26**:1063–70. https://doi.org/10.1038/s41594-019-0323-x.
27. Ma S, Zhang B, LaFave LM. *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 2020;**183**:1103–1116.e20. https://doi.org/10.1016/j.cell.2020.09.056.
28. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;**37**:1452–7. https://doi.org/10.1038/s41587-019-0290-0.
29. Battenberg K, Kelly ST, Ras RA. *et al.* A flexible cross-platform single-cell data processing pipeline. *Nat Commun* 2022;**13**:6847. https://doi.org/10.1038/s41467-022-34681-z.
30. Conesa A, Madrigal P, Tarazona S. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;**17**:13. https://doi.org/10.1186/s13059-016-0881-8.
31. Baek S, Lee I. Single-cell ATAC sequencing analysis: from data preprocessing to hypothesis generation. *Comput Struct Biotechnol J* 2020;**18**:1429–39. https://doi.org/10.1016/j.csbj.2020.06.012.
32. Yan F, Powell DR, Curtis DJ. *et al.* From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 2020;**21**:22. https://doi.org/10.1186/s13059-020-1929-3.
33. Sun S, Zhu J, Ma Y. *et al.* Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* 2019;**20**:269. https://doi.org/10.1186/s13059-019-1898-6.
34. Heumos L, Schaar AC, Lance C. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023;**24**:550–72. https://doi.org/10.1038/s41576-023-00586-w.
35. Kamimoto K, Stringa B, Hoffmann CM. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* 2023;**614**:742–51. https://doi.org/10.1038/s41586-022-05688-9.
36. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc Natl Acad Sci* 2019;**116**:27151–8. https://doi.org/10.1073/pnas.1911536116.
37. Faith JJ, Hayete B, Thaden JT. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 2007;**5**:e8. https://doi.org/10.1371/journal.pbio.0050008.
38. Margolin AA, Nemenman I, Basso K. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006;**7**:S7. https://doi.org/10.1186/1471-2105-7-S1-S7.
39. Huynh-Thu VA, Irrthum A, Wehenkel L. *et al.* Inferring regulatory networks from expression data using tree-based methods. *PloS One* 2010;**5**:e12776. https://doi.org/10.1371/journal.pone.0012776.
40. Bonneau R, Reiss DJ, Shannon P. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* 2006;**7**:R36. https://doi.org/10.1186/gb-2006-7-5-r36.
41. Vermeirssen V, Joshi A, Michoel T. *et al.* Transcription regulatory networks in Caenorhabditis elegans inferred through reverse engineering of gene expression profiles constitute biological hypotheses for metazoan development. *Mol Biosyst* 2009;**5**:1817–30. https://doi.org/10.1039/b908108a.
42. Roy S, Lagree S, Hou Z. *et al.* Integrated module and gene-specific regulatory inference implicates upstream Signaling networks. *PLoS Comput Biol* 2013;**9**:e1003252. https://doi.org/10.1371/journal.pcbi.1003252.
43. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* 2018;**19**:232. https://doi.org/10.1186/s12859-018-2217-z.
44. Chen J, Cheong C, Lan L. *et al.* DeepDRIM: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell RNA-seq data. *Brief Bioinform* 2021;**22**:bbab325. https://doi.org/10.1093/bib/bbab325.
45. Papili Gao N, Ud-Dean SMM, Gandrillon O. *et al.* SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 2018;**34**:258–66. https://doi.org/10.1093/bioinformatics/btx575.
46. Chan TE, Stumpf MPH, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems* 2017;**5**:251–267.e3. https://doi.org/10.1016/j.cels.2017.08.014.
47. Matsumoto H, Kiryu H, Furusawa C. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* 2017;**33**:2314–21. https://doi.org/10.1093/bioinformatics/btx194.

48. Mahmood O, Gibbs CS, Bonneau R. *et al.* A Variational inference approach to single-cell gene regulatory network inference using probabilistic matrix factorization. *Genome Biol* 2024;**25**:88. https://doi.org/10.1186/s13059-024-03226-6.
49. Badia-i-Mompel P, Vélez Santiago J, Braunger J. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* 2022;**2**:vbac016. https://doi.org/10.1093/bioadv/vbac016.
50. Müller-Dott S, Tsirvouli E, Vazquez M. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res* 2023;**51**:10934–49. https://doi.org/10.1093/nar/gkad841.
51. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol* 2018;**19**:196. https://doi.org/10.1186/s13059-018-1575-1.
52. The DREAM5 Consortium, Marbach D, Costello JC. *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* 2012;**9**:796–804. https://doi.org/10.1038/nmeth.2016.
53. Vermeirssen V, De Clercq I, Van Parys T. *et al.* Arabidopsis ensemble reverse engineered gene regulatory network discloses interconnected transcription factors in oxidative stress. *Plant Cell* 2014;**26**:4656–79. https://doi.org/10.1105/tpc.114.131417.
54. Pratapa A, Jalihal AP, Law JN. *et al.* Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;**17**:147–54. https://doi.org/10.1038/s41592-019-0690-6.
55. Cantini L, Zakeri P, Hernandez C. *et al.* Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. *Nat Commun* 2021;**12**:124. https://doi.org/10.1038/s41467-020-20430-7.
56. Nguyen H, Shrestha S, Tran D. *et al.* A comprehensive survey of tools and software for active subnetwork identification. *Front Genet* 2019;**10**:155. https://doi.org/10.3389/fgene.2019.00155.
57. Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics. *Nat Methods* 2019;**16**:381–6. https://doi.org/10.1038/s41592-019-0372-4.
58. McCalla SG, Fotuhi Siahpirani A, Li J. *et al.* Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data. *G3 Genes|Genomes|Genetics* 2023;**13**:jkad004. https://doi.org/10.1093/g3journal/jkad004.
59. Kang Y, Thieffry D, Cantini L. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. *Front Genet* 2021;**12**:617282. https://doi.org/10.3389/fgene.2021.617282.
60. Xue L, Wu Y, Lin Y. Dissecting and improving gene regulatory network inference using single-cell transcriptome data. *Genome Res* 2023;**33**:1609–21. https://doi.org/10.1101/gr.277488.122.
61. Sonawane AR, DeMeo DL, Quackenbush J. *et al.* Constructing gene regulatory networks using epigenetic data. *npj Syst Biol Appl* 2021;**7**:45. https://doi.org/10.1038/s41540-021-00208-3.
62. Mariani L, Weinand K, Gisselbrecht SS. *et al.* MEDEA: analysis of transcription factor binding motifs in accessible chromatin. *Genome Res* 2020;**30**:736–48. https://doi.org/10.1101/gr.260877.120.
63. Corces MR, Granja JM, Shams S. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* 2018;**362**:eaav1898. https://doi.org/10.1126/science.aav1898.
64. Stuart T, Srivastava A, Madad S. *et al.* Single-cell chromatin state analysis with Signac. *Nat Methods* 2021;**18**:1333–41. https://doi.org/10.1038/s41592-021-01282-5.
65. Pliner HA, Packer JS, McFaline-Figueroa JL. *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell* 2018;**71**:858–871.e8. https://doi.org/10.1016/j.molcel.2018.06.044.
66. Kiani K, Sanford EM, Goyal Y. *et al.* Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. *Mol Syst Biol* 2022;**18**:e10979. https://doi.org/10.15252/msb.202210979.
67. Chereji RV, Eriksson PR, Ocampo J. *et al.* Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation. *Genome Res* 2019;**29**:1985–95. https://doi.org/10.1101/gr.249326.119.
68. Siahpirani AF, Roy S. A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res* 2017;**45**:e21. https://doi.org/10.1093/nar/gkw1160.
69. Jiang Y, Harigaya Y, Zhang Z. *et al.* Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions. *Cell Systems* 2022;**13**:737–751.e4. https://doi.org/10.1016/j.cels.2022.08.004.
70. Duren Z, Chang F, Naqing F. *et al.* Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with scREG. *Genome Biol* 2022;**23**:114. https://doi.org/10.1186/s13059-022-02682-2.
71. Kamal A, Arnold C, Claringbould A. *et al.* GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Mol Syst Biol* 2023;**19**:e11627. https://doi.org/10.15252/msb.202311627.
72. Xu Q, Georgiou G, Frölich S. *et al.* ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic Acids Res* 2021;**49**:7966–85. https://doi.org/10.1093/nar/gkab598.
73. Kartha VK, Duarte FM, Hu Y. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genomics* 2022;**2**:100166. https://doi.org/10.1016/j.xgen.2022.100166.
74. Osorio D, Capasso A, Eckhardt SG. *et al.* Population-level comparisons of gene regulatory networks modeled on high-throughput single-cell transcriptomics data. *Nat Comput Sci* 2024;**4**:237–50. https://doi.org/10.1038/s43588-024-00597-5.
75. Trimbour R, Deutschmann IM, Cantini L. Molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS. *Bioinformatics* 2024;**40**:btae143. https://doi.org/10.1093/bioinformatics/btae143.
76. Chen H, Lareau C, Andreani T. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol* 2019;**20**:241. https://doi.org/10.1186/s13059-019-1854-5.
77. Granja JM, Corces MR, Pierce SE. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* 2021;**53**:403–11. https://doi.org/10.1038/s41588-021-00790-6.
78. Bravo González-Blas C, Minnoye L, Papasokrati D. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods* 2019;**16**:397–400. https://doi.org/10.1038/s41592-019-0367-1.
79. Baran Y, Bercovich A, Sebe-Pedros A. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol* 2019;**20**:206. https://doi.org/10.1186/s13059-019-1812-2.
80. Persad S, Choo Z-N, Dien C. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat Biotechnol* 2023;**41**:1746–57. https://doi.org/10.1038/s41587-023-01716-9.

81. Xu Y, McCord RP. Diagonal integration of multimodal single-cell data: potential pitfalls and paths forward. *Nat Commun* 2022;**13**:3505. https://doi.org/10.1038/s41467-022-31104-x.
82. Chen H, Ryu J, Vinyard ME. *et al.* SIMBA: single-cell embedding along with features. *Nat Methods* 2023;**21**:1–11. https://doi.org/10.1038/s41592-023-01899-8.
83. 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. Science. *Advances* 2022;**8**:eabl7393. https://doi.org/10.1126/sciadv.abl7393.
84. Fleck JS, Jansen SMJ, Wollny D. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* 2023;**621**:365–372. 1–8. https://doi.org/10.1038/s41586-022-05279-8.
85. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;**40**:1458–66. https://doi.org/10.1038/s41587-022-01284-4.
86. He Z, Hu S, Chen Y. *et al.* Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nat Biotechnol* 2024;**42**:1–12. https://doi.org/10.1038/s41587-023-02040-y.
87. Hao Y, Stuart T, Kowalski MH. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;**42**:293–304. https://doi.org/10.1038/s41587-023-01767-y.
88. Hammal F, de Langen P, Bergon A. *et al.* ReMap 2022: a database of human, mouse, drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* 2022;**50**:D316–25. https://doi.org/10.1093/nar/gkab996.
89. Davis CA, Hitz BC, Sloan CA. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;**46**:D794–801. https://doi.org/10.1093/nar/gkx1081.
90. Mulero Hernández J, Fernández-Breis JT. Analysis of the landscape of human enhancer sequences in biological databases. *Comput Struct Biotechnol J* 2022;**20**:2728–44. https://doi.org/10.1016/j.csbj.2022.05.045.
91. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* 2019;**48**:gkz980. https://doi.org/10.1093/nar/gkz980.
92. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* 2017;**6**:e21856. https://doi.org/10.7554/eLife.21856.
93. Kaya-Okur HS, Wu SJ, Codomo CA. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* 2019;**10**:1930. https://doi.org/10.1038/s41467-019-09982-5.
94. Fu Z, Jiang S, Sun Y. *et al.* Cut&tag: a powerful epigenetic tool for chromatin profiling. *Epigenetics* 2024;**19**:2293411. https://doi.org/10.1080/15592294.2023.2293411.
95. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&tag profiles histone modifications and transcription factors in complex tissues. *Nat Biotechnol* 2021;**39**:825–35. https://doi.org/10.1038/s41587-021-00869-9.
96. Pantier R, Chhatbar K, Alston G. *et al.* High-throughput sequencing SELEX for the determination of DNA-binding protein specificities *in vitro*. *STAR Protocols* 2022;**3**:101490. https://doi.org/10.1016/j.xpro.2022.101490.
97. Xu J, Gao J, Ni P. *et al.* Less-is-more: selecting transcription factor binding regions informative for motif inference. *Nucleic Acids Res* 2024;**52**:e20. https://doi.org/10.1093/nar/gkad1240.
98. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;**27**:1017–8. https://doi.org/10.1093/bioinformatics/btr064.
99. Santana-Garcia W, Castro-Mondragon JA, Padilla-Gálvez M. *et al.* RSAT 2022: regulatory sequence analysis tools. *Nucleic Acids Res* 2022;**50**:W670–6. https://doi.org/10.1093/nar/gkac312.
100. Heinz S, Benner C, Spann N. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89. https://doi.org/10.1016/j.molcel.2010.05.004.
101. Schep A, University S. *motifmatchr: Fast Motif Matching in R*2023. https://doi.org/10.18129/B9.bioc.motifmatchr (19th August 2024, date last accessed).
102. Bruse N, van Heeringen SJ. GimmeMotifs: an analysis framework for transcription factor motif analysis. *bioRxiv* 2018. https://doi.org/10.1101/474403.
103. Georgakopoulos-Soares I, Deng C, Agarwal V. *et al.* Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nat Commun* 2023;**14**:2333. https://doi.org/10.1038/s41467-023-37960-5.
104. Inukai S, Kock KH, Bulyk ML. Transcription factor–DNA binding: beyond binding site motifs. *Curr Opin Genet Dev* 2017;**43**:110–9. https://doi.org/10.1016/j.gde.2017.02.007.
105. Schep AN, Wu B, Buenrostro JD. *et al.* chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 2017;**14**:975–8. https://doi.org/10.1038/nmeth.4401.
106. Tognon M, Giugno R, Pinello L. A survey on algorithms to characterize transcription factor binding sites. *Brief Bioinform* 2023;**24**:bbad156. https://doi.org/10.1093/bib/bbad156.
107. Yu X, Buck MJ. Pioneer factors and their in vitro identification methods. *Mol Genet Genomics* 2020;**295**:825–35. https://doi.org/10.1007/s00438-020-01675-9.
108. Bentsen M, Goymann P, Schultheis H. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* 2020;**11**:4267. https://doi.org/10.1038/s41467-020-18035-1.
109. Sherwood RI, Hashimoto T, O'Donnell CW. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* 2014;**32**:171–8. https://doi.org/10.1038/nbt.2798.
110. Piper J, Elze MC, Cauchy P. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res* 2013;**41**:e201. https://doi.org/10.1093/nar/gkt850.
111. Li Z, Schulz MH, Look T. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* 2019;**20**:45. https://doi.org/10.1186/s13059-019-1642-2.
112. Jiang J, Lyu P, Li J. *et al.* IReNA: integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles. *iScience* 2022;**25**:105359. https://doi.org/10.1016/j.isci.2022.105359.
113. Wang L, Trasanidis N, Wu T. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multi-omics. *Nat Methods* 2023;**20**:1368–1378. https://doi.org/10.1038/s41592-023-01971-3.
114. Cazares TA, Rizvi FW, Iyer B. *et al.* maxATAC: genome-scale transcription-factor binding prediction from ATAC-seq with deep neural networks. *PLoS Comput Biol* 2023;**19**:e1010863. https://doi.org/10.1371/journal.pcbi.1010863.
115. Kshirsagar M, Yuan H, Ferres JL. *et al.* BindVAE: Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin. *Genome Biol* 2022;**23**:174. https://doi.org/10.1186/s13059-022-02723-w.
116. Li Z, Nagai JS, Kuppe C. *et al.* scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference.

Bioinformatics. *Advances* 2023;**3**:vbad003. https://doi.org/10.1093/bioadv/vbad003.
117. Popp AP, Hettich J, Gebhardt JCM. Altering transcription factor binding reveals comprehensive transcriptional kinetics of a basic gene. *Nucleic Acids Res* 2021;**49**:6249–66. https://doi.org/10.1093/nar/gkab443.
118. Li G, Sun T, Chang H. *et al.* Chromatin interaction analysis with updated ChIA-PET tool (V3). *Genes (Basel)* 2019;**10**:554. https://doi.org/10.3390/genes10070554.
119. Schoenfelder S, Javierre B-M, Furlan-Magaril M. *et al.* Promoter capture hi-C: high-resolution, genome-wide profiling of promoter interactions. *J Vis Exp* 2018;**136**:e57320. https://doi.org/10.3791/57320.
120. Arrastia MV, Jachowicz JW, Ollikainen N. *et al.* Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nat Biotechnol* 2022;**40**:64–73. https://doi.org/10.1038/s41587-021-00998-1.
121. Dixon JR, Selvaraj S, Yue F. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80. https://doi.org/10.1038/nature11082.
122. Rao SSP, Huntley MH, Durand NC. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80. https://doi.org/10.1016/j.cell.2014.11.021.
123. Javierre BM, Burren OS, Wilder SP. *et al.* Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* 2016;**167**:1369–1384.e19. https://doi.org/10.1016/j.cell.2016.09.037.
124. McArthur E, Capra JA. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am J Hum Genet* 2021;**108**:269–83. https://doi.org/10.1016/j.ajhg.2021.01.001.
125. Downes DJ, Smith AL, Karpinska MA. *et al.* Capture-C: a modular and flexible approach for high-resolution chromosome conformation capture. *Nat Protoc* 2022;**17**:445–75. https://doi.org/10.1038/s41596-021-00651-w.
126. Tran A, Yang P, Yang JYH. *et al.* scREMOTE: using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model. *NAR Genomics and Bioinformatics* 2022;**4**:lqac023. https://doi.org/10.1093/nargab/lqac023.
127. Reiff SB, Schroeder AJ, Kırlı K. *et al.* The 4D Nucleome data portal as a resource for searching and visualizing curated nucleomics data. *Nat Commun* 2022;**13**:2365. https://doi.org/10.1038/s41467-022-29697-4.
128. Starks RR, Biswas A, Jain A. *et al.* Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics Chromatin* 2019;**12**:16. https://doi.org/10.1186/s13072-019-0260-2.
129. Jo H, Kim T, Chun Y. *et al.* A compendium of chromatin contact maps reflecting regulation by chromatin remodelers in budding yeast. *Nat Commun* 2021;**12**:6380. https://doi.org/10.1038/s41467-021-26629-6.
130. Sakaue S, Weinand K, Isaac S. *et al.* Tissue-specific enhancer–gene maps from multimodal single-cell data identify causal disease alleles. *Nat Genet* 2024;**56**:615–26. https://doi.org/10.1038/s41588-024-01682-1.
131. Wang H, Huang B, Wang J. Predict long-range enhancer regulation based on protein–protein interactions between transcription factors. *Nucleic Acids Res* 2021;**49**:10347–68. https://doi.org/10.1093/nar/gkab841.
132. Lettice LA, Heaney SJH, Purdie LA. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 2003;**12**:1725–35. https://doi.org/10.1093/hmg/ddg180.
133. Jin S, Zhang L, Nie Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol* 2020;**21**:25. https://doi.org/10.1186/s13059-020-1932-8.
134. Duren Z, Chen X, Jiang R. *et al.* Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci* 2017;**114**:E4914–23. https://doi.org/10.1073/pnas.1704553114.
135. Fang R, Preissl S, Li Y. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun* 2021;**12**:1337. https://doi.org/10.1038/s41467-021-21583-9.
136. Zhang K, Zemke NR, Armand EJ. *et al.* A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat Methods* 2024;21:1–11. https://doi.org/10.1038/s41592-023-02139-9.
137. Glass K, Huttenhower C, Quackenbush J. *et al.* Passing messages between biological networks to refine predicted interactions. *PloS One* 2013;**8**:e64832. https://doi.org/10.1371/journal.pone.0064832.
138. Liu Y, Shi N, Regev A. *et al.* Integrated regulatory models for inference of subtype-specific susceptibilities in glioblastoma. *Mol Syst Biol* 2020;**16**:e9506. https://doi.org/10.15252/msb.20209506.
139. Song Q, Ruffalo M, Bar-Joseph Z. Using single cell atlas data to reconstruct regulatory networks. *Nucleic Acids Res* 2023;**51**:e38. https://doi.org/10.1093/nar/gkad053.
140. Yuan Q, Duren Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nat Biotechnol* 2024;**42**:1–11. https://doi.org/10.1038/s41587-024-02182-7.
141. La Manno G, Soldatov R, Zeisel A. *et al.* RNA velocity of single cells. *Nature* 2018;**560**:494–8. https://doi.org/10.1038/s41586-018-0414-6.
142. Bergen V, Lange M, Peidli S. *et al.* Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 2020;**38**:1408–14. https://doi.org/10.1038/s41587-020-0591-3.
143. Gorin G, Fang M, Chari T. *et al.* RNA velocity unraveled. *PLoS Comput Biol* 2022;**18**:e1010492. https://doi.org/10.1371/journal.pcbi.1010492.
144. Weiler P, Lange M, Klein M. *et al.* CellRank 2: unified fate mapping in multiview single-cell data. *Nat Methods* 2024;**21**:1196–205. https://doi.org/10.1038/s41592-024-02303-9.
145. Zhang S, Pyne S, Pietrzak S. *et al.* Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nat Commun* 2023;**14**:3064. https://doi.org/10.1038/s41467-023-38637-9.
146. Burdziak C, Zhao CJ, Haviv D. *et al.* scKINETICS: inference of regulatory velocity with single-cell transcriptomics data. *Bioinformatics* 2023;**39**:i394–403. https://doi.org/10.1093/bioinformatics/btad267.
147. Merrill CB, Montgomery AB, Pabon MA. *et al.* Harnessing changes in open chromatin determined by ATAC-seq to generate insulin-responsive reporter constructs. *BMC Genomics* 2022;**23**:399. https://doi.org/10.1186/s12864-022-08637-y.
148. Duren Z, Chen X, Xin J. *et al.* Time course regulatory analysis based on paired expression and chromatin accessibility data. *Genome Res* 2020;**30**:622–34. https://doi.org/10.1101/gr.257063.119.

149. Ramasamy S, Aljahani A, Karpinska MA. *et al.* The mediator complex regulates enhancer-promoter interactions. *Nat Struct Mol Biol* 2023;**30**:991–1000. https://doi.org/10.1038/s41594-023-01027-2.
150. Yang R, Das A, Gao VR. *et al.* Epiphany: predicting hi-C contact maps from 1D epigenomic signals. *Genome Biol* 2023;**24**:134. https://doi.org/10.1186/s13059-023-02934-9.
151. Guilliams M, Bonnardel J, Haest B. *et al.* Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell* 2022;**185**:379–396.e38. https://doi.org/10.1016/j.cell.2021.12.018.
152. Aizarani N, Saviano A, . *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* 2019;**572**:199–204. https://doi.org/10.1038/s41586-019-1373-2.
153. Badia-i-Mompel P, Wessels L, Müller-Dott S. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023;**24**:739–54. https://doi.org/10.1038/s41576-023-00618-5.
154. Fouché A, Zinovyev A. Omics data integration in computational biology viewed through the prism of machine learning paradigms. *Front Bioinform* 2023;**3**:1191961. https://doi.org/10.3389/fbinf.2023.1191961.
155. Song D, Li K, Ge X. *et al.* ClusterDE: a post-clustering differential expression (DE) method robust to false-positive inflation caused by double dipping. *bioRxiv* 2023. https://doi.org/10.21203/rs.3.rs-3211191/v1.
156. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 2016;**13**:966–7. https://doi.org/10.1038/nmeth.4077.
157. Datlinger P, Rendeiro AF, Schmidl C. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* 2017;**14**:297–301. https://doi.org/10.1038/nmeth.4177.
158. Nießl C, Herrmann M, Wiedemann C. *et al.* Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *WIREs Data Mining and Knowledge Discovery* 2022;**12**:e1441. https://doi.org/10.1002/widm.1441.
159. Williams RM, Candido-Ferreira I, Repapi E. *et al.* Reconstruction of the global neural crest gene regulatory network In vivo. *Dev Cell* 2019;**51**:255–276.e7. https://doi.org/10.1016/j.devcel.2019.10.003.
160. Kim D, Tran A, Kim HJ. *et al.* Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *npj Syst Biol Appl* 2023;**9**:1–13. https://doi.org/10.1038/s41540-023-00312-6.
161. Felce C, Gorin G, Pachter L. A biophysical model for ATAC-seq data analysis. *bioRxiv* 2024. https://doi.org/10.1101/2024.01.25.577262.
162. Browaeys R, Gilis J, Sang-Aram C. *et al.* MultiNicheNet: a flexible framework for differential cell-cell communication analysis from multi-sample multi-condition single-cell transcriptomics data. *bioRxiv* 2023. https://doi.org/10.1101/2023.06.13.544751.
163. Kyaw W, Chai RC, Khoo WH. *et al.* ENTRAIN: integrating trajectory inference and gene regulatory networks with spatial data to co-localize the receptor–ligand interactions that specify cell fate. *Bioinformatics* 2023;**39**:btad765. https://doi.org/10.1093/bioinformatics/btad765.
164. Lin Y, Wu T-Y, Chen X. *et al.* Data integration and inference of gene regulation using single-cell temporal multimodal data with scTIE. *Genome Res* 2024;**34**:119–33. https://doi.org/10.1101/gr.277960.123.
165. Cui H, Wang C, Maan H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1–11. https://doi.org/10.1038/s41592-024-02201-0.
166. Zhang Z, Han J, Song L. *et al.* CeSpGRN: inferring cell-specific gene regulatory networks from single cell multi-omics and spatial data. *bioRxiv* 2023. https://doi.org/10.1101/2022.03.03.482887.
167. Héberlé É, Bardet AF. Sensitivity of transcription factors to DNA methylation. *Essays Biochem* 2019;**63**:727–41. https://doi.org/10.1042/EBC20190033.
168. Chen AF, Parks B, Kathiria AS. *et al.* NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat Methods* 2022;**19**:547–53. https://doi.org/10.1038/s41592-022-01461-y.
169. Stoeckius M, Hafemeister C, Stephenson W. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**:865–8. https://doi.org/10.1038/nmeth.4380.