Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

Computational prediction of gene regulatory networks in plant growth and development

Samiul Haque[1,3], Jabeen S Ahmad[2,3], Natalie M Clark[2], Cranos M Williams[1,5] and Rosangela Sozzani[2,4]

Plants integrate a wide range of cellular, developmental, and environmental signals to regulate complex patterns of gene expression. Recent advances in genomic technologies enable differential gene expression analysis at a systems level, allowing for improved inference of the network of regulatory interactions between genes. These gene regulatory networks, or GRNs, are used to visualize the causal regulatory relationships between regulators and their downstream target genes. Accordingly, these GRNs can represent spatial, temporal, and/or environmental regulations and can identify functional genes. This review summarizes recent computational approaches applied to different types of gene expression data to infer GRNs in the context of plant growth and development. Three stages of GRN inference are described: first, data collection and analysis based on the dataset type; second, network inference application based on data availability and proposed hypotheses; and third, validation based on *in silico, in vivo*, and *in planta* methods. In addition, this review relates data collection strategies to biological questions, organizes inference algorithms based on statistical methods and data types, discusses experimental design considerations, and provides guidelines for GRN inference with an emphasis on the benefits of integrative approaches, especially when *a priori* information is limited. Finally, this review concludes that computational frameworks integrating large-scale heterogeneous datasets are needed for a more accurate (e.g. fewer false interactions), detailed (e.g. discrimination between direct versus indirect interactions), and comprehensive (e.g. genetic regulation under various conditions and spatial locations) inference of GRNs.

Addresses
[1] Electrical and Computer Engineering, North Carolina State University, Raleigh, USA
[2] Plant and Microbial Biology, North Carolina State University, Raleigh, USA

Corresponding authors: Williams, Cranos M (cmwilli5@ncsu.edu), Sozzani, Rosangela (ross_sozzani@ncsu.edu)
[3] These authors contributed equally to this work.
[4] Postal address: Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, 27695, USA.
[5] Postal address: Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, 27606, USA.

Current Opinion in Plant Biology 2019, **47**:96–105

This review comes from a themed issue on **Growth and development**

Edited by **Adrienne H K Roeder** and **C Jill Harrison**

For a complete overview see the Issue and the Editorial

Available online 14th November 2018

https://doi.org/10.1016/j.pbi.2018.10.005

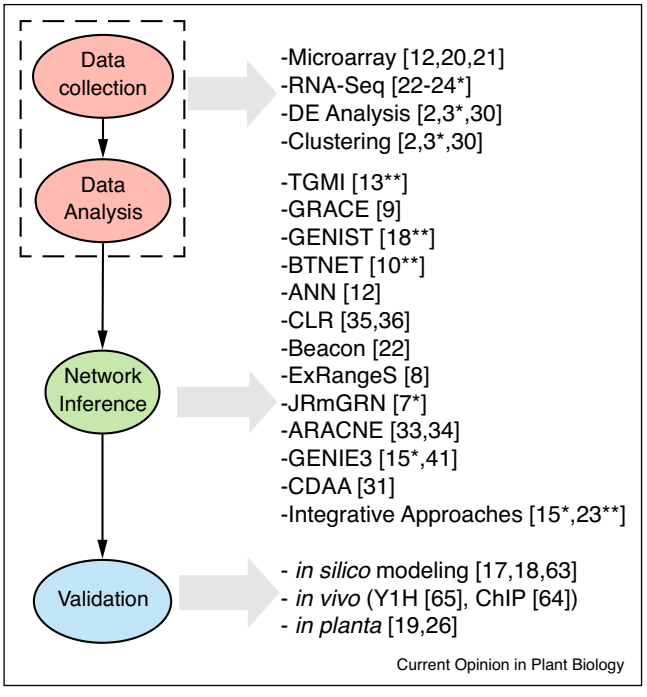
Introduction

An important question in plant biology is how interactions among genes across different cells, tissues, and organs influence plant growth and developmental processes such as germination, flowering, regeneration, and stress response. A systematic understanding of these developmental aspects will ultimately contribute to increases in plant resiliency, sustainable agriculture, and food security [1]. Advancements in high-throughput technology enable collection and analysis of genome-wide expression data at a systems level [2,3•], which is then used to form a gene regulatory network (GRN). A GRN is a graphical or mathematical representation that conveys the causal relationships among genes in an organism [2]. The GRN can take the form of a graph (where nodes are genes and edges represent relationships), a logical relationship (e.g. Boolean network), or a mathematical equation (e.g. algebraic or ordinary differential equations) that captures the change in expression over time [2]. GRNs are an effective tool for identifying genes with essential biological functions or that are involved in a specific pathway or process [4,5]. Inference methods that take transcriptional data and reveal connections between genes [6,7•,8,9,10••,11,12,13••,14] have been employed to construct GRNs and find important genes and regulatory relationships involved in plant growth and developmental processes such as cell wall synthesis [15•], regeneration [16], and root hair growth [17]. These inference methods utilize spatial, temporal, and perturbation (which includes

changes in environmental condition, stress, or mutation) gene expression data. Inferred GRNs allow one to identify key regulators through hub genes or network motif analysis for downstream validation. Validation provides evidence of gene regulatory relationships as well as functional relevance, particularly when the connection between gene expression data and the subsequent molecular or physiological phenotype is unclear. Accordingly, identified regulatory interactions can be validated *in silico* (e.g. ODE analysis), *in vivo* (e.g. ChIP assays), and/or *in planta* (e.g. mutant phenotype analysis) [14,17,18••,19].

In this review, we examine recent approaches for predicting GRNs in the context of plant growth and development (Figure 1). First, we detail how data collection and analysis depend on the biological question of interest with an overall focus on the role of these steps in GRN construction. Then, we highlight different GRN inference methods used to answer persistent questions in plant growth and development. To select the appropriate inference method based on collected and/or available data, we group methods according to the types of data they require (Figure 2). We note considerations for good experimental design and examine ways to improve GRN inference through integrative approaches. Finally, we discuss how an inferred GRN can be validated *in silico*, *in vivo*, and/or *in planta*. We conclude that gene expression data across different cells, tissues, organs, and environmental conditions can provide an enhanced understanding of the spatiotemporal dynamics of gene regulation. Moreover, incorporation of additional types of data such as transcription factor binding sites and protein-DNA binding can improve inference accuracy. Thus, future research endeavors should focus on integrating these heterogeneous datasets as well as multiple GRN inference methods to capture a more detailed view of transcriptional regulation.

Figure 1

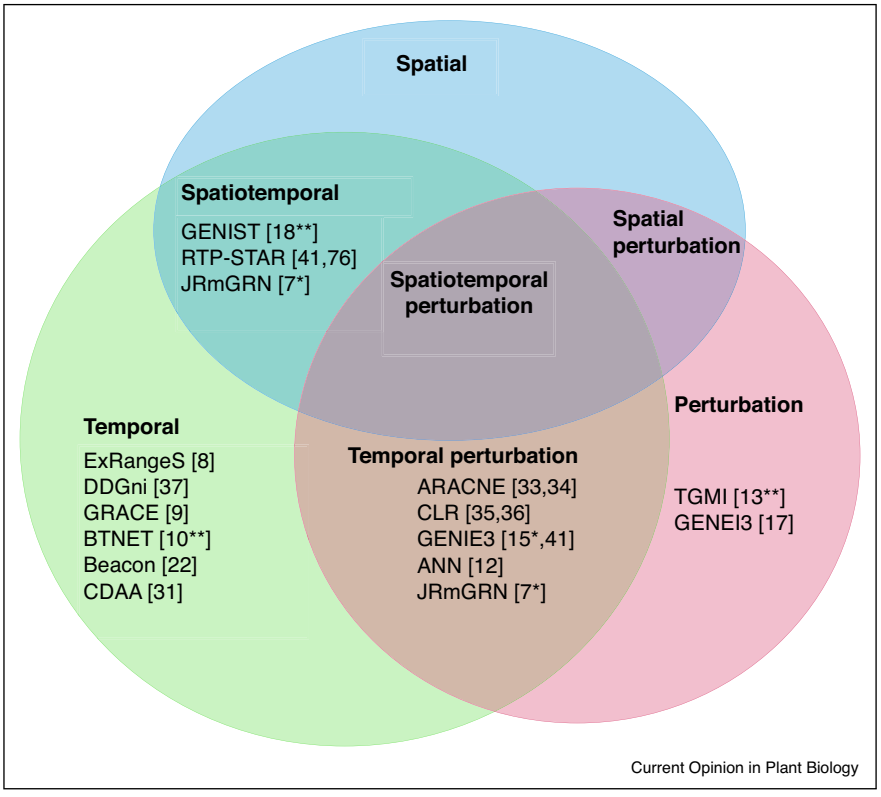


Stages of GRN inference. We organize GRN inference into a workflow of three broadly defined stages: data collection and analysis, network inference, and validation. Within this workflow, we list examples of techniques, tools, and methods that can be used during each stage of GRN inference. GRN inference begins with data collection and analysis. Data collection and data analysis are diagrammed separately as they use different tools but are color-coded together to indicate that they represent one stage of our workflow. This stage identifies differentially expressed genes, which will be used to infer a GRN. The next stage is network inference, which involves using computational methods to model a GRN including integrative approaches. The modeled GRN will then be validated in stage three using a combination of *in silico*, *in vivo*, and *in planta* methods to confirm whether the predicted model behaves as expected.

Data collection and analysis

GRN inference requires quantitative data that capture gene expression over space (e.g. cells, tissues, and organs), time, and/or under various environmental or stress conditions (later referred to as perturbation data). Generally, the biological question of interest and the researcher's hypothesis drive experimental specifications for gene expression data collection (e.g. temporal, spatial, perturbation, or a combination). Gene expression data are typically collected using high-throughput methods such as microarrays [12,20,21] and/or RNA-sequencing (RNA-Seq) [22,23••,24•]. Spatial gene expression data are collected from different cell types (including single and population of cells), tissues, or organs [25]. Temporal data are appropriate when studying changes in gene expression during stages of plant growth and development (e.g. seeds [11,22], leaves [26], and root elongation [18••]) or during the steps in metabolic processes (e.g. proanthocyanidins [27], lignin biosynthesis [13••], and phenylpropanoid, carbohydrate, fatty acid, and terpene pathways [28]). Perturbation data focus on changes in expression between a normal condition and an environmental change or mutation. Examples include gene expression data from biotic and abiotic stresses, such as nitrogen deficiency [12], osmotic stress [26], and over accumulation of phosphate under zinc-deficient conditions [29]. Transcriptomic data may fall into two or more categories such as spatiotemporal, spatial perturbation, temporal perturbation, and spatiotemporal perturbation (Figure 2). Data from these different combinations are especially valuable for understanding plant growth and development in specific tissues, under specific conditions, or both. For example, to understand the metabolic pathway for lignin biosynthesis, Gunasekara *et al.* obtained spatial perturbation data from *Arabidopsis thaliana* hypocotyledonous stem tissue during a shorter day [13••]. Overall, the biological question and corresponding

Figure 2



Types of gene expression data with recently applied algorithms. The Venn diagram shows different types of gene expression data and the corresponding algorithms applied to the data type for GRN inference. Temporal data refer to gene expression data collected over a period of time to study growth and development (e.g. time course or time series data). Spatial data refer to gene expression data collected from different cell types, tissues, and/or organs. Perturbation data refer to gene expression data for changes in environmental condition, stress, or mutation. It appears that the most recent GRN inference methods utilize temporal data either alone or in combination with another data type.

collected data must be considered and logically related to provide the appropriate biological context for the GRN.

High-throughput data collection provides information about hundreds to thousands of genes. However, not all genes play a relevant or important regulatory role with respect to the biological question or hypothesis. To isolate those genes of interest, differential expression analysis and clustering is necessary [2,3•,30]. To that end, statistical algorithms are applied to identify differentially expressed (DE) genes (i.e. genes with significant changes in expression) in gene expression data after which clustering is used to group together genes with similar expression patterns. For example, Koryachko *et al.* filtered out DE genes in *A. thaliana* under iron deficiency before applying their inference algorithm (Cluster and Differential Alignment Algorithm — CDAA) to build the resulting network [31]. In another example, de Luis Balaguer *et al.* used spatial gene expression data to cluster the genes and then applied dynamic Bayesian network (DBN) inference to identify relationships between genes within the resulting clusters [18••]. Specific approaches for DE analysis and clustering are extensively discussed in Refs. [2,3•,30]. Focusing on a subset of genes decreases computational time for inference algorithms while a lower number of interactions in the inferred GRN makes it feasible for *in vivo* and *in planta* validation of these connections. Given the advantages, DE analysis and clustering is useful and should be implemented when transcriptome data is used for network inference.

Network inference

Similar to data collection, the selection of a network inference algorithm depends on the biological question as well as the type of data available (Figure 2). Network inference algorithms have been applied to answer questions about metabolic pathways [13••,27,28], biotic and abiotic stress [12,29], and other aspects of plant growth and development. For the purposes of this review, these inference algorithms have been sorted into three categories: correlation and mutual information (MI), probabilistic graphical models, and machine learning. Correlation and MI methods mostly use temporal data to infer gene interactions. Correlation methods employ different correlation metrics and assume

that co-expression is an indicator of coregulation. Some popular correlation metrics, such as Pearson's correlation coefficient [32], more accurately infer regulations when relationships among genes are linear (i.e. change in regulator expression alters the expression of its targets proportionally). For example, Liesecke *et al.* used different correlation coefficients for constructing a global co-expression network of *A. thaliana* to recover genes involved in the same metabolic pathways [28]. Unlike Pearson's correlation, Spearman's rank correlation coefficient [32] and MI capture nonlinear relationships (i.e. a target's gene expression pattern is not directly proportional to its regulator's expression) between genes [2,32]. In one example, Montes *et al.* used the MI inference algorithm ARACNE [33] to construct GRNs in *A. thaliana* [34]. They recapitulated known networks involved in ground tissue patterning and stem cell maintenance. In another example, Xiong *et al.* unveiled important genes involved in maize seed development [35] using Context Likelihood of Relatedness (CLR) [36]. Their analysis identified 15 different network communities or clusters including one contributing to seed phenotype and another responsible for nutrient transport. A final MI method, TGMI identified regulatory transcription factors (TFs) in lignin and pigment biosynthesis pathways in *A. thaliana* using perturbation data [13••]. Unlike the other methods, TGMI calculates MI score among three genes instead of two. One major limitation of these methods is that they cannot assign direction of regulation when regulator genes are unknown or when time delay between gene expression is not considered [3•]. Thus, these methods may not be suitable in applications where causality is important (e.g. regulators of a hub gene).

Recently, improvements have been made to correlation and MI methods to consider time delays between expressions of different genes as well as uneven sampling and sparseness (i.e. having few time samples over a large period of time), which is typically seen in time series gene expression data. Among these methods is ExRangeS, which calculates rate of change in gene expression levels and assigns more importance to time points where transcriptional changes are high [8]. Koryachko *et al.* proposed the CDAA method, which inferred a GRN from time course data and identified novel regulators involved in iron homeostasis by scoring the alignment of differential expression over time [14]. Another algorithm, DDGni, captured the time-delayed regulation patterns by introducing gaps in gene expression profiles to align regulators with expression targets [37]. DDGni has been successfully used on yeast cell cycle data and could be applied to plants. Overall, ExRangeS, DDGni, and CDAA capture additional information by processing the time series data strategically and by considering the time delay between expression of two genes. In this way, these algorithms produce a directed network that reflects causality between regulators and potential targets.

Correlation and MI methods assume that gene expression is deterministically controlled by upstream regulators. However, recent studies reveal that gene expression patterns have some degree of random variation across individual cells [38,39]. Probabilistic graphical models address this issue by treating gene expression as random variables with a certain probability distribution that depends on time, space, environment, or a combination of these. Thus, this approach is suitable for understanding dynamic regulation of plants in different cell types under various environmental conditions. Additionally, in a probabilistic framework, *a priori* information can be incorporated as prior probability distribution. In one example, de Luis Balaguer *et al.* identified novel *A. thaliana* root stem cell regulators by developing a dynamic Bayesian network (DBN)-based algorithm called GENIST [18••]. DBNs model gene expression as a probabilistic function of itself and its regulator(s) at previous time points. GENIST clusters the gene expression pattern using spatial data from *A. thaliana* root stem cell data and uses temporal data to infer important regulatory connections for stem cell maintenance. In another case, Deng *et al.* introduced a Gaussian graphical model-based inference algorithm called JRmGRN [7•], which allowed for network inference over different tissues and environmental conditions. JRmGRN used temporal gene expression data from several tissues (spatiotemporal data) or conditions (temporal perturbation data). This algorithm identified hub genes that are common in GRNs under different light conditions (low and high red) in *A. thaliana*. It also found hub genes that were specific to these two conditions. One drawback of probabilistic methods is that they often require data with high spatial and temporal resolution (i.e. high number of time points compared to other methods) to estimate the probability distribution. Nevertheless, probabilistic approaches provide a strong mathematical framework to understand gene regulation.

Another type of inference method gaining significant attention is machine learning [22]. Machine learning can be further divided into unsupervised, supervised, and semi-supervised approaches depending on whether *a priori* interactions are incorporated into the inference method [40]. Unsupervised approaches do not incorporate *a priori* interaction data and are suitable when knowledge about regulatory relationships is limited. Examples of unsupervised approaches include the regression tree inference method GENIE3 [17,41], BTNET [10••], and fused LASSO [42]. Shibata *et al.* used GENIE3 to infer a GRN for root hair growth in *A. thaliana* [17], while BTNET and fused LASSO successfully identified developmental regulators in animals and could be applied to plants [10••,42]. In general, unsupervised approaches are less prone to overfitting (i.e. where a model fits the data and fails to extrapolate) and can be applied to multiple types of gene expression data (e.g. GENIE3 is applied to temporal [15•,41] and perturbation

[17] data). However, unsupervised methods cannot directly incorporate known regulatory connections into the inference framework. Additionally, as unsupervised approaches have fewer constraints, they are prone to yielding results with reduced mechanistic significance. On the other hand, when multiple regulators are known *a priori* and finding new targets of a key regulator is the major focus, supervised methods are more suitable for GRN inference. Examples of supervised machine learning methods include Beacon, which was used to predict a GRN for specific embryonic developmental stages in *A. thaliana* [22]. Beacon used temporal gene expression data along with known regulatory relationships to predict 521 genes downstream of their genes of interest. Jiang *et al.* used an artificial neural network (ANN) for GRN analysis of maize under nitrogen treatment [12]. They used microarray datasets from two different studies and generated an ANN which identified influential genes in the nitrogen use efficiency system. As the performance of supervised learning is dependent on *a priori* data, possible effects of overfitting must be considered before applying these approaches to situations with small training datasets. Machine learning algorithms that use both labeled (i.e. gene expression data of known regulators and downstream targets) and unlabeled (i.e. gene expression data without *a priori* data on regulation) data are referred to as semi-supervised approaches [40]. A semi-supervised method GRACE used ensembles of Markov Random Fields [43] and recovered cell cycle control mechanisms in *A. thaliana* [9]. Semi-supervised methods require much fewer *a priori* data for training than supervised methods, but they are often less accurate [40].

Experimental design for network inference

The accuracy of an inference algorithm is highly dependent on the experimental design. Determining optimal experimental specifications, however, remains a complex problem. Changes in experimental specifications can contribute to noise, sparseness, variation, and biases in data. The minimum data requirements for accurate network inference are usually unknown beforehand, leading to experimental specifications that are determined empirically or based on prior knowledge or experiences as well as availability in resources. Nevertheless, there have been efforts to provide information for adequate experimental design specifications using *in silico* and known GRN models for which the amount, quality, and type (i.e. perturbation or steady state) of experimental data can be controlled systematically [44–48]. Though analysis related to plant GRNs are still lacking, insights from existing approaches involving *in silico* and model GRNS can be applied to plant related problems. Accordingly, the DREAM challenge initiative provides a set of gold standard benchmark networks and datasets on which researchers can evaluate the performance (accuracy or precision of inferred GRN) of their inference algorithm [49–52]. These networks have been used to relate the performance of different inference algorithms to the types of experimental data used (i.e. steady state, time series, knockout, and overexpression). The DREAM3 competition provided three *in silico* networks, each containing a different number of interactions and asked participants to develop novel inference algorithms for GRN reconstruction. The best algorithm in this competition, which used a combination of ODE and Gaussian noise models [53], utilized additional information from knockout datasets along with time series and steady state data [50]. The DREAM5 challenge benchmark dataset contained simulated data from an *in silico* network and experimental data for *Escherichia coli* and *Saccharomyces cerevisiae* GRNs. These GRNs were significantly larger and more complicated (i.e. different network motifs and higher number of interactions) than the DREAM3 networks. Marbach *et al.* examined the performance of 35 GRN inference algorithms submitted to DREAM5 network inference competition and found that methods using TF perturbation data had improved accuracy in predicting downstream targets [51]. Overall, the studies suggest that incorporating knockout experiments in the experimental design can produce improved accuracy in GRN inference [50,51].

In addition to determining the type of experimental data to use, another important experimental design consideration is the total number of time points and/or biological replicates collected. The number of collected samples is often constrained by the amount of resources. One example in yeast research, that could also apply to plant GRNs modeling and inference, is given by Markdahl *et al.* Using an *in silico* model of the yeast cell-cycle network, the authors demonstrated that increasing the number of samples improved the accuracy of the inferred GRN using a linear regression-based method (i.e. expression of each gene is a linear response to its regulators' expression) [44]. Importantly, after 15 or more time samples, it plateaued, indicating that adding more time samples did not improve inference accuracy. This example shows that having more time samples does not necessarily guarantee improved inference. Inference accuracy can increase if new data provide additional information regarding gene interactions. Consequently, additional information can come from more biological replicates or a different type of experiment, including diverse time course transcriptional data. Moreover, for high-throughput gene expression acquisition, there is often a tradeoff between the number of sampled time points and biological replicates due to experimental budgets and limited resources [54]. In view of this, Geier *et al.* investigated data requirements for reconstructing a synthetic DBN. By analyzing 100 different synthetic networks (each containing 30 genes), they demonstrated that sampling fewer time points with more biological replicates significantly improved inference accuracy compared to data that covered more time points with fewer biological replicates per time point [45].

One way of determining the best experimental design is to frame it as an optimization problem where an objective function (an algebraic function designed to reflect the tradeoffs between experimental cost and performance of an inference algorithm) is minimized (or maximized) [46–48]. This objective function takes experimental design costs and inference accuracy as input and outputs a number that is minimized (or maximized) by modifying the experimental specifications. For example, using synthetic GRNs, Dehghannasiri *et al.* developed an optimization-based framework to prioritize potential experiments that reduced the uncertainty in predicting correct edges [46]. In another example, Ud-dean *et al.* developed the REDuction of UnCertain Edges (REDUCE) algorithm to find the optimum number of knockout experiments required to infer the GRN of *E. coli* under ideal conditions [47]. A comprehensive review of similar optimizations of experimental design can be found in Ref. [48]. Optimization approaches work well with *in silico* models and address the data requirements for inference algorithms. However, in practical applications, to determine good experimental design, the optimization framework can be applied to parts of the GRN which have been experimentally validated.

Analysis of existing GRN inference approaches can provide additional insights for choosing the number of time samples and biological replicates. Studies related to plant growth and development have used different experimental specifications depending on the biological question as well as the requirements of the inference algorithm. Among the algorithms (discussed in this paper) that inferred GRNs from time course data in plants, the number of time points varied from seven to 15 with two to four biological replicates per time point [6,14,23••,37]. Experiments with more time points collected fewer biological replicates per time point. Two methods requiring specific perturbation experiments used six knockout lines with three biological replicates each [17] and six time points with two biological replicates each (under two different conditions) [7•]. Though none of these works presented strong mathematical justification for experimental design, they all used a similar amount of data for similar biological questions, indicating that there may be some consensus on the minimum number of data points required for GRN inference. Given that these methods successfully predicted genes involved in plant growth and development, their experimental specifications could provide guidelines for future research.

Integrative network inference approaches

To obtain a more comprehensive and accurate GRN, it is recommended to combine multiple inference methods and/or multiple, different data types [3•,55,56]. In existing studies both the combination of network inference methods and different data types have been referred to as consensus approaches [51,57,58]. However, as the term consensus approach can have different meanings depending on the field of study [51,57,58], for the remainder of this review, these processes will be referred to as integrative approaches.

The use of multiple datasets is a useful integrative approach as the accuracy of GRN inference can be improved by incorporating additional gene expression datasets [3•,55,56]. For example, Shahan *et al.* created a combined dataset from 46 RNA-seq libraries of strawberry flower and fruit development. They randomly sampled this dataset 1000 times and applied the Weighted Gene Co-expression Network Analysis (WCGNA) [59] to each subset. Their resulting GRN from this analysis gave important insights for iron transport during fruit development and identified potential floral meristem and receptacle meristem regulators [57]. Considering datasets beyond the transcriptome is also important as these data alone cannot provide a complete picture of a plant's molecular state. This method of combining datasets on different molecular scales is called a multiomics approach [55,56]. For example, Walley *et al.* used a combination of RNA, protein, and phosphoprotein data to create an expression atlas for maize [60]. Their integrated dataset proved to be better than using only one type of dataset as they discovered that nearly 85% of maize hub genes were not shared in gene networks derived from RNA or protein data alone [60]. In another example, researchers used transcripts, proteins, and metabolites to understand changes in gene expression under different environmental conditions [61]. Their aggregated model helped to confirm which genes were responsible for distinguishing variations among grape cultivars. The model also helped to reveal complex variation that was not apparent from a single source of data [61]. Additional information about GRN inference using multiple datasets can be found in Refs. [3•,55,56].

Another type of integrative approach involves using multiple inference methods on the same dataset to obtain multiple GRNs. Each GRN produced is then aggregated into one final GRN. For example, Taylor-Teeples *et al.* developed a GRN in *A. thaliana* for secondary wall synthesis [15•] using four unsupervised (GENIE3, Inferelator, TIGRESS, and ANOVerence) and one supervised (SIRENE) method. By incorporating multiple inference algorithms, they identified many novel regulators and gained a better understanding of the developmental regulation of xylem cell differentiation. Similarly, Redeker *et al.* identified key genes and their regulatory interactions underlying seed development metabolisms in low phytic acid soybean using an integrated network from five different methods (ARACNE, GENIE3, LARS, partial correlation, CLR) [23••]. An integrative approach using multiple methods can exploit the advantages of different types of inference algorithms and often results in higher

accuracy in identifying biologically relevant regulatory relationships [2,15•,23••,42]. Marbach *et al.* demonstrated the benefits of an integrative approach by testing the accuracy of 35 different inference algorithms using three *in silico* and well-known GRNs. Their evaluation showed that a combination of multiple approaches constantly outperformed any single method in reconstructing the GRNs [51]. Thus, an integrative approach is strongly recommended to overcome the shortcomings of any individual method while developing an accurate GRN.

Validation of inferred GRNs

Once a GRN is inferred, hub genes need to be identified. These genes play important functional roles in plant growth and development as they often encode regulators essential for biological processes [2] and are good candidates for mutant experiments. Hub genes can be identified by the number of interactions (i.e. outdegree or indegree) each gene in the network has with a larger number of interactions indicating greater importance. Genes can also be ranked by the number of network motifs (common sub networks that appear repeatedly in the same GRN or across different GRNs) to which they belong [19].

After identifying the important genes, methods that test how these genes and their regulatory connections affect growth and development are needed. *In silico* models provide opportunities to validate the network by comparing the model output with experimental data. In addition, these models predict gene expression based on the inferred GRN and can be used to simulate the influence of novel perturbations (e.g. knockdown or overexpression of specific genes) in the network. Ordinary Differential Equations (ODE), which mathematically model changes in gene expression over time, are used frequently for modeling gene expression dynamics. For example, de Luis Balaguer *et al.* constructed ODE equations for eight key transcription factors involved in *A. thaliana* root stem cell regulation to simulate expression in wildtype and mutants [18••]. In their simulations, comparison of the steady states of the TFs regulated by a specific root stem cell regulator indicated that the TFs are functionally interconnected [18••]. Similarly, Shibata *et al.* presented ODE equations for two regulators (GTL1 and RSL4) controlling root hair growth [17]. Through *in silico* experiments, they predicted that GTL1 and RSL4 together function in a negative-feedback loop for consistent root growth by maintaining steady state levels of expression [17]. Besides ODE, other mathematical functions for modeling gene expression are available and extensively discussed in Refs. [62,63].

Inferred GRNs and their associated dynamic models produce *in silico* results, which should then be verified *in vivo*. With *in vivo* validation, the inferred interactions are characterized under the appropriate biological context [19]. For small or moderate sized networks, inferred GRNs and their *in silico* simulations can be experimentally validated using qRT-PCR analysis [14,17,18••]. While qRT-PCR quantifies changes in gene expression, it cannot inform whether the regulation is indirect or due to direct binding to the promoter region of target genes. Chromatin ImmunoPrecipitation (ChIP) [64] and Yeast 1 Hybrid (Y1H) [65] methods can be used to determine direct TF-DNA binding. Additionally, examination of plant phenotypes, such as quantitative measurements of root length, leaf size, and chlorophyll content are important to connect the changes in gene expression predicted by the network to a developmental phenotype [19,26]. While validation using *in silico* methods provides support for an inferred GRN, validation using *in vivo* and *in planta* methods are necessary to connect the changes in gene expression predicted by the GRN with a gene developmental function.

Conclusion

Computational inference of GRNs is advantageous for understanding gene regulation in plant growth and development. Accuracy of the inference algorithm is dependent on the availability and type of experimental data collected. However, constructing good experimental design for GRN inference remains a challenging task and can vary based on the biological problem. Recent inference methods focus on improving accuracy of GRNs and identification of hub genes through integrative approaches that aggregate multiple computational methods. GRN inference can be further improved by incorporating additional datasets beyond gene expression data for a comprehensive understanding of the plant's molecular state. Specifically, techniques such as ATAC-Seq [66–68], Yeast 1 Hybrid (Y1H) [65,69], and DAP-Seq [70] can be used to elicit detailed information about chromatin accessibility, protein–DNA interactions, and binding sites for TFs. Another promising data collection method uses single-cell technology, where a single cell is studied to detect coregulated genes and/or infer genes at the top of a regulation hierarchy [71,72]. This technology allows for examination of changes in gene expression that otherwise may not be observed among a population of cells [71,72]. Single-cell genomic assays can be used for integrated predictions of GRNs and applied to biological questions (e.g. understanding individual cell response to disease, nutrient deficiency, or environmental stress) where spatial, temporal, or perturbation data are desired. Computational inference methods of GRNs from single-cell data have already been introduced (PIDC and SCENIC) [71,73]. PIDC provides a network inference algorithm for single-cell gene expression data and compares its GRN inference performance using known GRNs for the human megakaryocyte and erythroid differentiation pathways [73]. PIDC successfully inferred the known GRN but worked best when using an aggregate of gene expression data from multiple single cells. SCENIC successfully identified known TFs in mouse brain cells while also providing information about the mechanism driving cellular heterogeneity [71]. In plants, single-cell technology could provide

greater insight on transcriptional regulation. For example, the technology can identify the underlying networks controlling the developmental trajectory of a tissue at the cellular level [74].

Furthermore, integrating heterogeneous data from different sources (e.g. RNA-Seq, ChIP-seq, or ATAC-Seq) provides a more comprehensive GRN as including information about interactions beyond gene expression improves the accuracy of any prediction and reduces the prevalence of false positives. Recent computational methods such as NECorr [25] and Beacon [22] are valuable for their ability to integrate heterogeneous data. Integration of time series experiments among different cell types and under different environmental conditions allows for improved overall understanding of gene regulation. Such massive amounts of data from various sources also opens up an opportunity for employing advanced machine learning algorithms (e.g. SVM [22], ANN [12], and deep neural networks [75]) for GRN inference. Future research endeavors should focus on building machine learning methods for large scale GRN inference that can incorporate heterogeneous data.

Finally, for many biological researchers, lack of knowledge and experience in mathematics, statistics, and programming limits the use of computational prediction methods. Education in computational biology, bioinformatics, and other cross-subject courses as well as collaborative efforts between biologists, data scientists, and computer scientists can bridge the gap. In addition, a continued development of user-friendly interfaces, like TuxNet [76] and Seidr [77], that are easy to use, access, and understand will increase the use of computational methods in GRN inference. These tools allow users to input gene expression data, then select a method of analysis for GRN inference (TuxNet) [76] or receive an aggregate GRN from a variety of algorithms that can be analyzed for further inference (Seidr) [77]. By utilizing these novel tools for computational prediction, the next generation of scientists will be able to gain key insights on plant regulatory mechanisms in a cost-effective manner with substantial supporting data.

Acknowledgements

Funding to NMC and RS is provided by the National Science Foundation (NSF) (CAREER MCB 1453130) (to RS). Funding to JSA and RS is provided by the NSF and the Biotechnology and Biological Sciences Research Council (BBSRC) (NSF MCB 1517058) (to RS). Funding to SH and CMW is provided by the NSF (MCB 1247427) (to CMW).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hu H, Scheben A: *Advances in Integrating Genomics and Bioinformatics in the Plant Breeding Pipeline*. 2018 http://dx.doi.org/10.3390/agriculture8060075.

2. Li Y, Pearl SA, Jackson SA: **Gene networks in plant biology: approaches in reconstruction and analysis**. *Trends Plant Sci* 2015, **20**:664-675.

3. • Banf M, Rhee SY: **Computational inference of gene regulatory networks: approaches, limitations and opportunities**. *Biochim Biophys Acta - Gene Regul Mech* 2017, **1860**:41-52.
This review provides an in-depth overview of computational methods and algorithms for GRN inference. It also highlights advantages and limitations of various approaches.

4. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R: **Gene regulatory network inference: data integration in dynamic models-a review**. *BioSystems* 2009, **96**:86-103.

5. Krouk G, Lingeman J, Colon AM, Coruzzi G, Shasha D: **Gene regulatory networks in plants: learning causality from time and perturbation**. *Genome Biol* 2013, **14**:123.

6. de Luis Balaguer MA, Fisher AP, Clark NM, Fernandez-Espinosa MG, Möller BK, Weijers D, Lohmann JU, Williams C, Lorenzo O, Sozzani R: **Predicting gene regulatory networks by combining spatial and temporal gene expression data in *Arabidopsis* root stem cells**. *Proc Natl Acad Sci U S A* 2017, **114**:E7632-E7640.

7. • Deng W, Zhang K, Liu S, Zhao P, Xu S, Wei H: **JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions**. *Bioinformatics* 2018, **34(20)**:3470-3478 http://dx.doi.org/10.1093/bioinformatics/bty354.
This paper introduced an algorithm to jointly infer multiple GRNs in different tissues or under different conditions. JRmGRN was able to identify common hub and condition-specific genes in*A. thaliana* under two different light conditions.

8. Desai JS, Sartor RC, Lawas LM, Jagadish SVK, Doherty CJ: **Improving gene regulatory network inference by incorporating rates of transcriptional changes**. *Sci Rep* 2017, **7**:1-12.

9. Banf M, Rhee SY: **Enhancing gene regulatory network inference through data integration with markov random fields**. *Sci Rep* 2017, **7**:1-13.

10. •• Park S, Kim JM, Shin W, Han SW, Jeon M, Jang HJ, Jang I-S, Kang J: **BTNET: boosted tree based gene regulatory network inference algorithm using time-course measurement data**. *BMC Syst Biol* 2018, **122**:20.
While it has not been used in plants, this paper describes machine learning algorithm BTNET, which shows great promise and outperformed many well-known unsupervised machine learning tools. This algorithm was applied to infer GRN from antidepressant-treated human neuroblastoma cell gene expression data.

11. Gupta C, Krishnan A, Schneider A, Denbow C, Collakova E, Wolinski P, Pereira A: **SANe: the seed active network for discovering transcriptional regulatory programs of seed development**. *bioRxiv* 2018.

12. Jiang L, Ball G, Hodgman C, Coules A, Zhao H, Lu C: **Analysis of gene regulatory networks of maize in response to nitrogen**. *Genes (Basel)* 2018, **9**:1-20.

13. •• Gunasekara C, Zhang K, Deng W, Brown L, Wei H: **TGMI: an efficient algorithm for identifying pathway regulators through evaluation of triple-gene mutual interaction**. *Nucleic Acids Res* 2018, **46**.
This study describes mutual information-based algorithm TGMI, which evaluated a three-gene block using perturbation data. TGMI identified many regulatory transcription factors in lignin and pigment biosynthesis pathways in*A. thaliana*.

14. Koryachko A, Matthiadis A, Muhammad D, Foret J, Brady SM, Ducoste JJ, Tuck J, Long TA, Williams C: **Clustering and differential alignment algorithm: identification of early stage regulators in the Arabidopsis thaliana iron deficiency response**. *PLoS One* 2015, **10**:1-21.

15. • Taylor-Teeples M, Lin L, De Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R *et al.*: **An Arabidopsis gene regulatory network for secondary cell wall synthesis**. *Nature* 2015, **517**:571-575.

This study used a consensus approach with four unsupervised methods and one supervised method for inferring a GRN for secondary cell wall synthesis in*A. thaliana*. The inferred network was used to validate a new hypothesis about secondary cell wall gene regulation under abiotic stress. They showed that abiotic stress can co-opt the xylem regulatory network.

16. Ikeuchi M, Shibata M, Rymen B, Iwase A, Bgman AM, Watt L, Coleman D, Favero DS, Takahashi T, Ahnert SE *et al.*: **A gene regulatory network for cellular reprogramming in plant regeneration**. *Plant Cell Physiol* 2018, **59**:765-777.

17. Shibata M, Breuer C, Kawamura A, Clark NM, Rymen B, Braidwood L, Morohashi K, Busch W, Benfey PN, Sozzani R *et al.*: **GTL1 and DF1 regulate root hair growth through transcriptional repression of *ROOT HAIR DEFECTIVE 6-LIKE 4* in *Arabidopsis***. *Development* 2018, **145** dev159707.

18. •• de Luis Balaguer MA, Fisher AP, Clark NM, Fernandez-Espinosa MG, Möller BK, Weijers D, Lohmann JU, Williams C, Lorenzo O, Sozzani R: **Predicting gene regulatory networks by combining spatial and temporal gene expression data in *Arabidopsis* root stem cells**. *Proc Natl Acad Sci U S A* 2017, **114**: E7632-E7640.

This paper describes a dynamic Bayesian network-based algorithm called GENIST, which used spatiotemporal data to predict a GRN in the root stem cells of*A. thaliana*. GENIST outperformed many well-known probabilistic graphical models.

19. Gaudinier A, Brady SM: **Mapping transcriptional networks in plants: data-driven discovery of novel biological mechanisms**. *Annu Rev Plant Biol* 2016, **67**:575-594.

20. Valentim FL, Van Mourik S, Posé D, Kim MC, Schmid M, Van Ham RCHJ, Busscher M, Sanchez-Perez GF, Molenaar J, Angenent GC *et al.*: **A quantitative and dynamic model of the arabidopsis flowering time gene regulatory network**. *PLoS One* 2015, **10**:1-18.

21. Bonneau R, Lingeman J, Colon AM, Coruzzi G, Shasha D, Bonneau R, Bonneau R, Facciotti M, Reiss D, Schmid A *et al.*: **Learning biological networks: from modules to dynamics**. *Nat Chem Biol* 2008, **4**:658-664.

22. Ni Y, Aghamirzaie D, Elmarakeby H, Collakova E, Li S, Grene R, Heath LS: **A machine learning approach to predict gene regulatory networks in seed development in Arabidopsis**. *Front Plant Sci* 2016, **7**.

23. •• Redekar N, Pilot G, Raboy V, Li S, Saghai Maroof MA: **Inference of transcription regulatory network in low phytic acid soybean seeds**. *Front Plant Sci* 2017, **8**:1-14.

This study discovered key genes and their regulatory interactions underlying seed development metabolism in low phytic acid soybean. Five different unsupervised network inference methods were used to create a consensus GRN. The authors used a diverse set of computational methods maximizing the capability of capturing different types of interactions.

24. • Thorne T: **Approximate inference of gene regulatory network models from RNA-Seq time series data**. *Thorne BMC Bioinform* 2018, **19**:1-13.

While it has not been used in plants, this methodology paper describes methods to improve GRN inference from RNA-Seq temporal data that involves sparse regression-based methods.

25. Liseron-Monfils CV, Olson A, Ware D: **NECorr, a tool to rank gene importance in biological processes using molecular networks and transcriptome data**. *bioRxiv* 2018.

26. Van den Broeck L, Dubois M, Vermeersch M, Storme V, Matsui M, Inzé D: **From network to phenotype: the dynamic wiring of an Arabidopsis transcriptional network induced by osmotic stress**. *Mol Syst Biol* 2017, **13**:961.

27. Nishiyama S, Onoue N, Kono A, Sato A, Yonemori K, Tao R: **Characterization of a gene regulatory network underlying astringency loss in persimmon fruit**. *Planta* 2018, **247**:733-743.

28. Liesecke F, Daudu D, de Bernonville RD, Besseau S, Clastre M, Courdavault V, De Craene J-O, Crèche J, Giglioli-Guivarc'h N, Glévarec G *et al.*: **Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks**. *Sci Rep* 2018, **8**10885 http://dx.doi.org/10.1038/s41598-018-29077-3.

29. Pal S, Kisko M, Dubos C, Lacombe B, Berthomieu P, Krouk G, Rouached H: **TransDetect identifies a new regulatory module controlling phosphate accumulation**. *Plant Physiol* 2017, **175** pp.00568.2017.

30. Thompson D, Regev A, Roy S: *Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution*. 2015.

31. Koryachko A, Matthiadis A, Ducoste JJ, Tuck J, Long TA, Williams C: **Computational approaches to identify regulators of plant stress response using high-throughput gene expression data**. *Curr Plant Biol* 2015, **3-4**:20-29.

32. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ: **Co-expression tools for plant biology: opportunities for hypothesis generation and caveats**. *Plant Cell Environ* 2009, **32**:1633-1651.

33. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**. *BMC Bioinform* 2006, **7**:1-15.

34. Chávez Montes RA, Coello G, González-Aguilera KL, Marsch-Martínez N, de Folter S, Alvarez-Buylla ER: **ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks**. *BMC Plant Biol* 2014, **14**:97.

35. Xiong W, Wang C, Zhang X, Yang Q, Shao R, Lai J, Du C: **Highly interwoven communities of a gene regulatory network unveil topologically important genes for maize seed development**. *Plant J* 2017, **92**:1143-1156.

36. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles**. *PLoS Biol* 2007, **5**:0054-0066.

37. Yalamanchili HK, Yan B, Li MJ, Qin J, Zhao Z, Chin FYL, Wang J: **DDGni: dynamic delay gene-network inference from high-temporal data using gapped local alignment**. *Bioinformatics* 2014, **30**:377-383.

38. Efroni I, Ip PL, Nawy T, Mello A, Birnbaum KD: **Quantification of cell identity from single-cell gene expression profiles**. *Genome Biol* 2015, **16**:1-12.

39. Shapiro E, Biezuner T, Linnarsson S: **Single-cell sequencing-based technologies will revolutionize whole-organism science**. *Nat Rev Genet* 2013, **14**:618-630.

40. Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA: **Supervised, semi-supervised and unsupervised inference of gene regulatory networks**. *Brief Bioinform* 2014, **15**:195-211.

41. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: **Inferring regulatory networks from expression data using tree-based methods**. *PLoS One* 2010, **5**:1-10.

42. Omranian N, Eloundou-Mbebi JMO, Mueller-Roeber B, Nikoloski Z: **Gene regulatory network inference using fused LASSO on multiple data sets**. *Sci Rep* 2016, **6**:20533.

43. Li SZ: *Markov Random Field Modeling in Image Analysis*. London: Springer; 2009.

44. Markdahl J, Colombo N, Thunberg J, Goncalves J: *Experimental Design Trade-Offs for Gene Regulatory Network Inference: An In Silico Study of the Yeast Saccharomyces cerevisiae Cell Cycle*. 2017.

45. Geier F, Timmer J, Fleck C: **Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge**. *BMC Syst Biol* 2007, **1**:11.

46. Dehghannasiri R, Yoon B-J, Dougherty ER: **Optimal experimental design for gene regulatory networks in the presence of uncertainty**. *IEEE/ACM Trans Comput Biol Bioinform* 2015, **12**:938-950.

47. Ud-Dean SMM, Gunawan R: **Optimal design of gene knockout experiments for gene regulatory network inference**. *Bioinformatics* 2016, **32**:875-883.

48. Sverchkov Y, Craven M: **A review of active learning approaches to experimental design for uncovering biological networks**. *PLoS Comput Biol* 2017, **13**:1-26.

49. Marbach D, Schaffter T, Floreano D, Prill RJ, Stolovitzky G: *The DREAM4 In-Silico Network Challenge*. . Training 2009.

50. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G: **Revealing strengths and weaknesses of methods for gene network inference**. *Proc Natl Acad Sci U S A* 2010, **107**:6286-6291.

51. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Aderhold A *et al.*: **Wisdom of crowds for robust gene network inference**. *Nat Methods* 2012, **9**:796-804.

52. Meyer P, Cokelaer T, Chandran D, Kim K, Loh P-R, Tucker G, Lipson M, Berger B, Kreutz C, Raue A *et al.*: **Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach**. *BMC Syst Biol* 2014, **8**:13.

53. Yip KY, Alexander RP, Yan KK, Gerstein M: **Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data**. *PLoS One* 2010, **5**.

54. Sefer E, Kleyman M, Bar-Joseph Z: **Tradeoffs between dense and replicate sampling strategies for high-throughput time series experiments**. *Cell Syst* 2016, **3**:35-42.

55. Kim M, Tagkopoulos I: **Data integration and predictive modeling methods for multi-omics datasets**. *Mol Omics* 2018, **14**.

56. Li Y, Wu F-X, Ngom A: **A review on machine learning principles for multi-view biological data integration**. *Brief Bioinform* 2016, **19** bbw113.

57. Shahan RM, Zawora C, Wight HM, Sittmann J, Wang W, Mount SM, Liu Z: **Consensus co-expression networks provide insights into wild strawberry flower and fruit development**. *Plant Physiol* 2018, **178** pp.00086.2018.

58. Taylor-Teeples M, Lin L, De Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R *et al.*: **An Arabidopsis gene regulatory network for secondary cell wall synthesis**. *Nature* 2015, **517**:571-575.

59. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinform* 2008, **9**.

60. Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, Urich MA, Nery JR, Smith LG, Schnable JC, Ecker JR *et al.*: **Integration of omic networks in a developmental atlas of maize**. *Science (80-)* 2016, **353**:814-818.

61. Ghan R, Van Sluyter SC, Hochberg U, Degu A, Hopper DW, Tillet RL, Schlauch KA, Haynes PA, Fait A, Cramer GR: **Five omic technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (Vitis vinifera L.) cultivars**. *BMC Genom* 2015, **16**:1-26.

62. Rué P, Garcia-Ojalvo J: **Modeling gene expression in time and space**. *Annu Rev Biophys* 2013, **42**:605-627.

63. Panse C, Kshirsagar M: **Survey on modelling methods applicable to gene regulatory network**. *Int J Bioinform Biosci* 2013, **3**:13-23.

64. Raha D, Hong M, Snyder M: **ChIP-Seq: a method for global identification of regulatory elements in the genome**. *Curr Protoc Mol Biol* 2010:1-14. Chapter 21.

65. Reece-Hoyes JS, Marian Walhout AJ: **Yeast one-hybrid assays: a historical and technical perspective**. *Methods* 2012, **57**:441-447.

66. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position**. *Nat Methods* 2013, **10**:1213-1218.

67. Buenrostro J, Wu B, Chang H, Greenleaf W: *HHS Public Access*. 2016 http://dx.doi.org/10.1002/0471142727.mb2129s.ATAC-seq.

68. Bajic M, Maher KA, Deal RB: *Plant Chromatin Dynamics* 2018, **vol. 1675**183-201.

69. Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, Kadreppa S, Pesyna C, Dekker J, Myers CL, Walhout AJM: **Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping**. *Nat Methods* 2011, **8**:1059-1068.

70. Bartlett A, O'Malley RC, Huang SSC, Galli M, Nery JR, Gallavotti A, Ecker JR: **Mapping genome-wide transcription-factor binding sites using DAP-seq**. *Nat Protoc* 2017, **12**:1659-1672.

71. Fiers MWEJ, Minnoye L, Aibar S, Bravo González-Blas C, Kalender Atak Z, Aerts S: **Mapping gene regulatory networks from single-cell omics data**. *Brief Funct Genom* 2018, **17(4)**:246-254.

72. Yuan Y, Lee HT, Hu H, Scheben A, Edwards D: **Single-cell genomic analysis in plants**. *Genes (Basel)* 2018, **9**.

73. Chan TE, Stumpf MPH, Babtie AC: **Gene regulatory network inference from single-cell data using multivariate information measures**. *Cell Syst* 2017, **5**:251-267.e3.

74. Wendrich JR, Möller BK, Li S, Saiga S, Sozzani R, Benfey PN, De Rybel B, Weijers D: **Framework for gradual progression of cell ontogeny in the *Arabidopsis* root meristem**. *Proc Natl Acad Sci U S A* 2017, **114**:E8922-E8929.

75. Angermueller C, Pärnamaa T, Parts L, Stegle O: **Deep learning for computational biology**. *Mol Syst Biol* 2016, **12**:878.

76. de Luis Balaguer MA, Spurney RJ, Clark NM, Fisher AP, Sozzani R: **TuxNet: a simple interface to process RNA sequencing data and infer gene regulatory networks**. *bioRxiv* 2018 http://dx.doi.org/10.1101/337147.

77. Schiffthaler B, Serrano A, Delhomme N, Street NR: **Seidr: a toolkit for calculation of crowd networks**. *bioRxiv* 2018 http://dx.doi.org/10.1101/250696.