OXFORD

*Briefings in Bioinformatics*, 20(3), 2019, 1021–1031

**doi: 10.1093/bib/bbx152**
Advance Access Publication Date: 27 November 2017
Software Review

# Construction of a hierarchical gene regulatory network centered around a transcription factor

Hairong Wei

Corresponding author: Hairong Wei, State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin, Heilongjiang 150040, China and School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA. Tel.: +1-906-487-1473; Fax: +1-906-487-2915; E-mail: hairong@mtu.edu



## Abstract

We have modified a multitude of transcription factors (TFs) in numerous plant species and some animal species, and obtained transgenic lines that exhibit phenotypic alterations. Whenever we observe phenotypic changes in a TF's transgenic lines, we are always eager to identify its target genes, collaborative regulators and even upstream high hierarchical regulators. This issue can be addressed by establishing a multilayered hierarchical gene regulatory network (ML-hGRN) centered around a given TF. In this article, a practical approach for constructing an ML-hGRN centered on a TF using a combined approach of top-down and bottom-up network construction methods is described. Strategies for constructing ML-hGRNs are vitally important, as these networks provide key information to advance our understanding of how biological processes are regulated.

**Key words:** backward elimination random forest; bottom-up graphic Gaussian model algorithm; multilayered hierarchical gene regulatory network; top-down graphic Gaussian Model algorithm; transcription factor

## Background

A multitude of transgenic lines of various transcription factors (TFs) have been developed in various plant and some animal species. Overexpression or suppression of a TF of interest often leads to phenotypic changes; some of these changes are readily visible, whereas others, such as anatomical changes, are subtle and discernible only via microscopy. In such circumstances, we can endeavor to identify target genes of the TF of interest by examining genome-wide expression data generated either from its stable transgenic lines or from a transient expression system. While elucidation of target genes is likely to be the primary goal, there may also be interest in determining collaborative regulators associated with a particular TF as well as its upstream regulators. By identifying and annotating the genes surrounding a TF in a hierarchical gene regulatory network (hGRN), we can significantly improve our understanding of why the altered expression of a particular TF results in the phenotypic changes observed in the transgenic plants.

We recently developed a top-down graphic Gaussian model (GGM) algorithm [1], a bottom-up GGM algorithm [2, 3] and backward elimination random forest algorithm (BWERF) [4]. The top-down GGM algorithm enables the construction of a two-layered hGRN mediated by a TF, whereas the bottom-up GGM and BWERF [4] allow for the construction of a multilayered-hGRN (ML-hGRN) that operates immediately above a pathway or biological process. Experimental validation of the ML-hGRN derived from both the top-down and bottom-up GGM algorithms suggests that both approaches are highly accurate [1, 3]. When used jointly, these methods enable us to reconstruct a ML-hGRN centered around a TF using gene expression data generated from properly designed experiments. In this review, we will examine the principles underlying each of these algorithms, discuss strategies for using the algorithms and describe methods for generating data required for accurate reconstruction of a ML-hGRN centered around a TF.

## Top-down GGM algorithm

The top-down GGM algorithm can be used to infer one or two layers of ML-hGRNs using high-throughput gene expression

**Hairong Wei** is an Associate Professor at Michigan Technological University. His research is primarily focused on construction of gene networks that include but are not limited to hierarchical regulatory networks, collaborative networks and coexpression networks. He proposed the concept of collaborative networks and developed multiple algorithms for not only constructing collaborative networks but also for decomposing collaborative network into subnetworks, and each subnetwork regulates a biological process or a complex trait.

**Submitted:** 27 June 2017; **Received (in revised form):** 11 October 2017

data generated either from stable transgenic lines or from a transient expression system in which TF expression is perturbed.

The top-down GGM algorithm builds an ML-hGRN in two steps: (1) identification of TF-responsive genes from differentially expressed genes (DEGs), and (2) identification of TF-responsive DEGs that have causal relationships with the given TF using the triple gene model shown above.

## 1. Identification of TF-responsive genes from DEGs

Using high-throughput gene expression data generated either from transgenic lines (stable overexpression or suppression) or from a transient system, DEGs can be identified by comparing data from one line with that from the corresponding control using appropriate methods. These methods include edgeR [5], DEseq [6], vst-limma [6, 7] and voom-limma [7] for RNA sequencing (RNA-seq) data, and Rank Product [8], Weighted Average Difference [9] and Moderated T-Statistics [7] for microarray data. Next, TF-responsive DEGs can be evaluated and potentially used as inputs to build the layer immediately below the TF in the ML-hGRN. Eliminating DEGs that are not closely responsive to the TF's overexpression or downregulation will decrease noise that reduces the accuracy of the triple gene model. To eliminate irrelevant DEGs, we used Fisher's exact test in conjugation with a condition probability method for DEG reduction in our top-down GGM algorithm [1]. Briefly, the significance of dependence between each DEG and the TF was examined using Fisher's exact test [10, 11] with a null hypothesis of independence between the TF and DEG. When the TF and a target were dependent, we continued to test which type of the DEG's response concordance was significant using the following hypotheses: $H_0$ = one of the four types of responses is significant (null hypothesis) and $H_a$ = none of the four types of responses is significant (alternative hypothesis). The four types of response concordance are defined as follows: concurrence (Type I), one-level lag (Type II), two-level lag (Type III) and three-level lag (Type IV) [1]. Concurrence describes a phenomenon in which a DEG closely follows the overexpressed TF of interest, and their relative expression values are always in the same discretized levels. One-level lag indicates that the DEG lags by one level regarding its relative expression levels compared with those of the overexpressed TF. If any of the four types of concordance has a *P*-value exceeding the significance level, then the type with the largest *P*-value is considered the most significant. The details for testing the four types of concordance using Pearson's chi-square test are provided in our earlier publication [1]. Although both Fisher's exact test and probability methods can help reduce noise, the efficiency of the probability method is limited when the sample size is small. In that circumstance, the probability test for concordance between the TF and DEGs can be omitted without significantly affecting the outcome; this is the case because the use of Fisher's exact test appears to be sufficient for eliminating the majority of irrelevant genes.

## 2. Identification of TF-interfering genes using the triple gene model

### *Triple gene model and identification of interference*

Triple gene model uses a first-order correlation to evaluate causal relationships among three genes (Figure 1A). Given a pair of candidate target genes, x and y, we can examine whether the presence of the TF, z, makes their coordination tighter or looser. We first calculate Spearman's partial correlation coefficient between x and y in the absence of z using the formula $r_{xy|z} = pcor(x, y|z)$, which represents the correlation between x and y after the effect of z on the genes is removed. We then compare $r_{xy|z}$ with the x and y correlation coefficient $r_{xy}$, where $r_{xy}$ is Spearman's rank correlation coefficient (rho) for variables x and y [12] when the effect of z is present. If z interferes with x and y, there should be a significant difference between $r_{xy}$ and $r_{xy|z}$. Based on this assumption, we implement the following procedure to determine whether z has causal relationships with x and y.

A. If $r_{xy}$ is significant and $r_{xy|z} = 0$ (insignificant) or vice versa, then z interferes with x and y.
B. If both $r_{xy}$ and $r_{xy|z}$ are significant, we need to test the significance of the difference $d = r_{xy} - r_{xy|z}$. The standard error of the difference can be calculated using the multivariate delta method as described previously [13]. The Z score can be calculated using the formula $Z = d/SED$, which follows a normal distribution. If the *P*-value is significant, we can then conclude that z interferes with x and y.
C. When z interferes with x and y, $z \rightarrow x$ and $z \rightarrow y$ are recorded.

### *Top-down GGM algorithm*

Expression data for the identified TF-responsive genes can be used as inputs for causality analysis using the triple gene model to infer the next one or two layers, namely, the second and first layers, which contain the direct and indirect target genes of the given TF, respectively (Figure 1A). For each combination of three genes, which includes a TF in an upper layer and two TF-responsive candidate genes in the lower layer, if the TF significantly interferes with the two TF-responsive genes, then the TF is said to control the two genes and their regulatory relationships (edges) are recorded. When all combinations of the TF and two TF-responsive candidate target genes have been evaluated, the interference frequency between the TF and each TF-responsive candidate target gene is calculated; the TF-responsive candidate target genes having the highest frequency are retained in the second layer. As not all TF-responsive candidate target genes retained at the second layer are TFs, the next layer (namely, the first layer) is generated in top-down fashion only from each TF present in the second layer (direct targets) by recursively calling the top-down GGM algorithm, as shown in Figure 1A. To generate the first layer from a TF present in the second layer, the responsive genes for this TF are identified from the remaining DEGs, excluding the genes retained in the second layer. The core algorithms of triple gene interference and top-down GGM are shown below:

**Core algorithm of triple gene interference for top-down (TGI-TD)**

1. procedure
2. Input: Expression profiles of (Z, G), where G are candidate targets of a regulatory gene Z
3. for each pair of $(X_i, Y_j)$, where $(X_i, Y_j) \in G$ do
4. if $r_{xy}$ is significant and $r_{xy|z} = 0$ (insignificant) or vice versa
5. Z interfere $X_i$ and $Y_j$, then
6. $Z \rightarrow X_i$ and $Z \rightarrow Y_j$;
7. else if both $r_{xy}$ and $r_{xy|z}$ are significant, then
8. if $d = r_{xy} - r_{xy|z}$ is significant after being tested with multivariate delta method, then
9. Z interfere $X_i$ and $Y_j$, then
10. $(E_{Z \rightarrow X_i}$ *and* $E_{Z \rightarrow Y_i}) \rightarrow N_z$, where $N_z$ hosts edges between Z and inferred targets
11. $N_{Z \rightarrow X_i}$ ++; $N_{Z \rightarrow Y_i}$ ++; the number of each edge increases 1
12. Else
13. discard ($X_i$ and $Y_j$), go to step 2

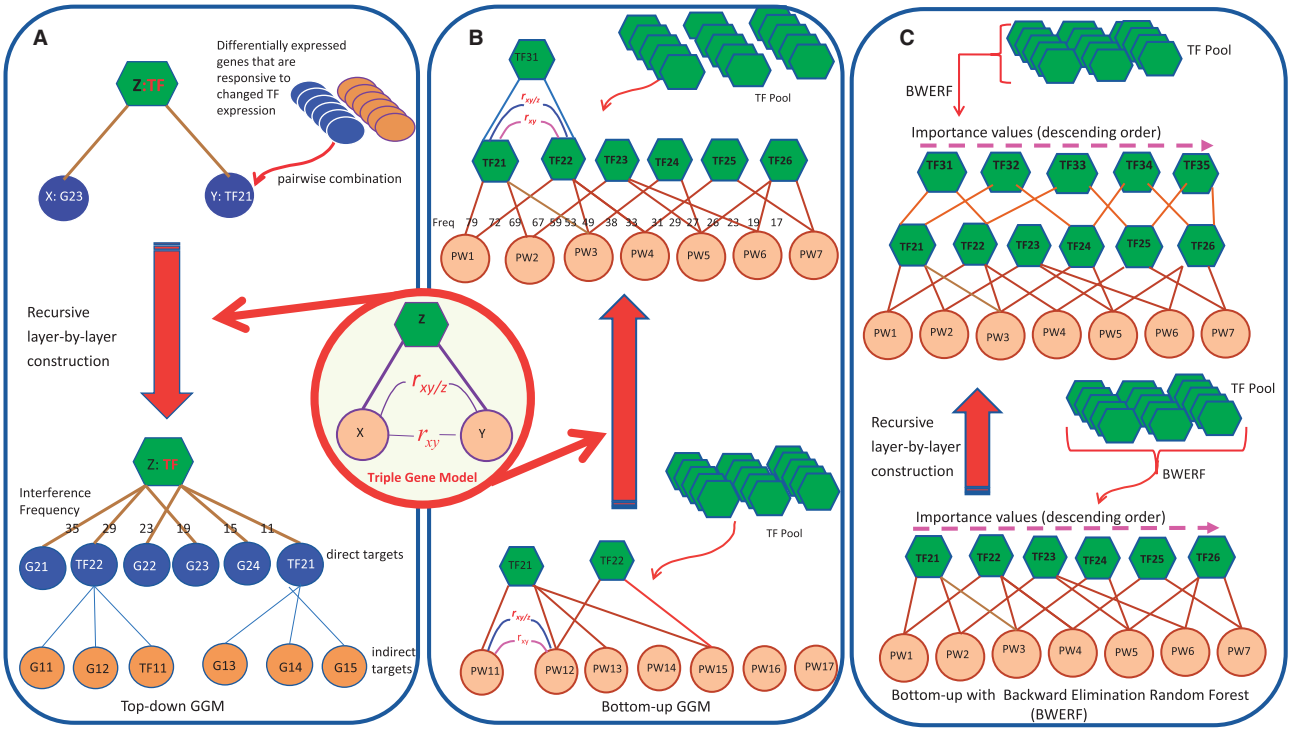


**Figure 1.** Triple gene model and how to use it for constructing an ML-hGRN. (**A**) Triple gene model for top-down GGM algorithm; (**B**) triple gene model for bottom-up GGM algorithm; (**C**) BWERF for bottom-up network work construction.

14. else if $r_{xy} = 0$ (insignificant) and $r_{xy|z} = 0$ (insignificant)
15. discard ($X_i$ and $Y_j$), go to step 2
16. end for
17. output: (1) $G_Z^I$ sorted by interference frequency of Z; (2) $N_z$ network contains interfered edges
18. end procedure

**Top-down GGM algorithm**

1. procedure
2. Input: $D$ data of n samples from tissues where Z (a TF) is overexpressed;
3. do $DEG \leftarrow D$; DEG hosts differentially expressed genes (DEGs);
4. if $DEG \neq 0$ **do**
5. $G_Z^r \leftarrow DEG$; where $G_Z^r$ holds responsive genes to Z (by Fisher Exact test &&/| | concordance test)
6. else
7. end procedure
8. if $G_Z^r \neq 0$ **do**
9. $G_Z^i \leftarrow$TGI-TD (Z, $G_Z^r$); where $G_Z^i$ holds genes sorted by interfered frequency of Z in descending order;
10. $N_z \leftarrow$TGI-TD (Z, $G_Z^r$);
11. $G_Z^i(k) \leftarrow$ sig $(G_Z^I)$; # keep most significant k genes to current layer immediately below the TF's
12. else
13. end procedure
14. for a new $Z'$ in $G_Z^i(k)$ && $DEG' = DEG - G_Z^i(k)$ do
15. Step 3–10 to build next layer
16. end of for
17. end procedure

***An example to illustrate use of the top-down GGM algorithm***

The top-down GGM algorithm was first applied to RNA-seq data sets generated from a time course experiment (7, 12 and 25 h after the perturbation of TF) in which the TF, *SND1*, was transiently overexpressed in protoplasts isolated from developing xylem. There were three biological replicates at each time point. The blank plasmid (i.e. no *SND1*) was used to transfect protoplasts to generate control samples for each time point. Using the edgeR algorithm, we identified 178 DEGs across the three time points. The expression profiles of these 178 DEGs and that of *SND1* were used as inputs for building the ML-hGRN regulated by SND1. The third layer was generated by extending from two TF genes located in the second layer. In total, 16 genes selected from Layer 2 and 16 from Layer 3 were validated by Chromatin Immunoprecipitation-Polymerase Chain Reaction (ChIP-PCR) using an antibody against SND1. All 16 genes from Layer 2 were true targets of SND1, whereas only 1 of 16 genes in Layer 3 was a true target of SND1 indicating that 97% (31 of 32) of the relationships inferred by the top-down GGM algorithm were correct. This ML-hGRN governed by SND1 can be found in our previous publication [1].

## Bottom-up network construction approaches

### Bottom-up GGM algorithm

Initially, genes involved in a metabolic pathway or biological process (as defined by gene ontology) were used as inputs for the bottom layer, whereas all TFs or differentially expressed TFs were used as inputs for constructing the upper layers [2, 3]. The method by which genes involved in a biological process of interest are obtained depends largely on the experiments from which the gene expression data were generated. DEGs from experiments involving two or more treatments can be easily identified through multiple pairwise comparisons of data from different conditions or treatments, and biological processes of interest can be identified via gene ontology enrichment analyses of DEGs. When a single experiment includes multiple time

points, each with more than one treatment condition, the enriched biological processes identified at each time point can be summed. The importance of biological processes that are enriched at only one time point and the importance of those that are enriched across multiple time points need to be weighed properly based on experimental goal and present knowledge. The identification of enriched biological processes is further complicated when compendium data sets pooled from multiple experiments are used. These issues are beyond the scope of this review, however, and will not be discussed further.

One efficient why to construct a ML-hGRN above a bottom layer of genes is to use the triple gene model, as illustrated in Figure 1B. Each evaluation involves three genes, including two genes from the current bottom layer and a candidate TF, which may be chosen from all TFs or the differentially expressed TF pool. These three genes are then used to build the next layer above. When a TF significantly interferes with the two bottom-layer genes (false discovery rate (FDR)-corrected $P$-value $< 0.05$), the regulatory relationship (edge) between the TF and each gene (edge) is recorded. After all combinations are evaluated, the interference frequency of each TF for all pathway genes is calculated and sorted. TFs with the highest interference frequencies are retained in the second layer; the number retained can be determined empirically or by a weighted sparse canonical correlation analysis method [2]. At this point, the second layer will be used as the bottom layer, and the remaining TFs will be used as inputs to obtain the third layer by repeating the aforementioned procedure. We can construct an ML-hGRN in a layer-by-layer fashion until the expected number of layers is obtained or the program terminates automatically when there are no more layers that can be built. The core algorithms of triple gene interference and bottom-up GGM are shown below:

**Core algorithm of triple gene interference for bottom-up (TGI-BU)**

1. **procedure**
2. **Input**: (Z, B), where Z contains a number of TFs while B contains genes of current bottom layer
3. **foreach** combination of $(X_i, Y_j)$, where $(X_i, Y_j) \in B$, test all genes in Z **do**
4. **if** $r_{xy}$ is significant and $r_{xy|z} = 0$ (insignificant) or vice versa
5. $Z_k$ interfere $X_i$ and $Y_j$, **then**
6. $Z_k \rightarrow X_i$ and $Z_k \rightarrow Y_j$;
7. **else if** both $r_{xy}$ and $r_{xy|z}$ are significant, **then**
8. **if** $d = r_{xy} - r_{xy|z}$ is significant after being tested with multivariate delta method, **then**
9. $Z_k$ interfere $X_i$ and $Y_j$, **then**
10. $(E_{Z_k \rightarrow X_i}$ *and* $E_{Z_k \rightarrow Y_i}) \rightarrow N_z$, where $N_z$ contains edges between Z and inferred targets
11. $N_{Z \rightarrow X_i}$ ++; $N_{Z \rightarrow Y_i}$ ++; the number of each edge increases 1
12. **else**
13. discard ($X_i$ and $Y_j$), go to step 2
14. **else if** $r_{xy} = 0$ (insignificant) and $r_{xy|z} = 0$ (insignificant)
15. discard ($X_i$ and $Y_j$), go to step 2
16. **end foreach**
17. **output:** (1) $Z^i$sorted by interference frequency on all bottom-layered genes in B;
18. (2) $N_z$ network contains interfered edges
19. **end procedure**

**Bottom-up GGM algorithm**

1. procedure
2. Input: $D$ high-throughput expression data of n samples
3. do $DEG \leftarrow D$; where DEG holds differentially expressed genes;
4. if $DEG \neq 0$ **do**
5. $G_b \leftarrow DEG$; where $G_b$-genes involved in a biological process/pathway of interest
6. Z$\leftarrow$ *All TFs or TFs in DEG*
7. else
8. end procedure
9. if $G_b \neq 0$ && $Z \neq 0$ do
10. $Z^i \leftarrow$TGI-BU (Z, $G_b$); where $Z^i$ holds TFs sorted by their interfered frequency
11. $N_z \leftarrow$TGI-BU (Z, $G_b$);
12. $Z^i(k) \leftarrow sig(Z^i)$; # keep most significant genes to current layer immediately above bottom layer
13. else
14. end procedure
15. foreach a new $Z' = Z - Z^i(k)$ and a new $G_b' = Z^i(k)$ do
16. Step 3–14 to build next layer
17. end of foreach
18. end procedure

The bottom-up GGM algorithm has been used to construct an ML-hGRN using RNA-seq data from nine independent poplar stable transgenic lines of *Ptr-miR397* (under the control of the cauliflower mosaic virus promoter, CaMV 35 S) and three wild-type lines [3]. The nine transgenic lines were selected to include lines with high, medium and low *Ptr-miR397a* expression to maximize its variation. In total, 459 DEGs were identified by comparing the transgenic lines with the wild-type controls. The bottom layer included 17 differentially expressed laccase (LAC) genes and four peroxidase genes that are postulated to function in lignin polymerization. Additionally, 1208 TFs plus *Ptr-miR397a* were used as regulatory genes for bottom-up network construction. A three-layer hGRN was obtained. The causal relationships between *Ptr-miR397a* and the LAC genes were captured using the bottom-up GGM algorithm. Of 17 LAC genes, 13 were identified as direct targets of *Ptr-miR397a*. Five of these LAC genes were randomly selected for experimental validation, and all were proven to be true targets of ptr-*miR397* [3].

## Using BWERF

In addition to the bottom-up GGM algorithm, a BWERF algorithm [4] has been developed for constructing the ML-hGRN operating above a biological pathway or process (Figure 1C). For each pathway/process gene in the bottom layer, the BWERF algorithm uses a random forest model not only to calculate the importance values for all TFs, as they relate to genes in the bottom layer, but also to remove some of the least important TFs. For example, the least important 10% of TFs are excluded in each round of modeling. In each new round, importance values are updated and ranked; this procedure, termed BWERF, continues until only one TF is retained. Afterward, the importance value for each TF, as it relates to all pathway genes is aggregated and fitted to a Gaussian mixture model to clarify TF retention for the second layer immediately above the bottom layer. The TFs retained in the secondary layer are then set as the new bottom layer to infer the third layer, and this process is repeated until the expected number of layers is obtained or the program terminates automatically.

BWERF improves the accuracy for constructing ML-hGRNs because it uses backward elimination to exclude noise-causing genes and aggregates the individual importance values for determining TF retention. We validated the BWERF algorithm by using it to construct ML-hGRNs operating above a mouse pluripotency maintenance pathway and an *Arabidopsis*

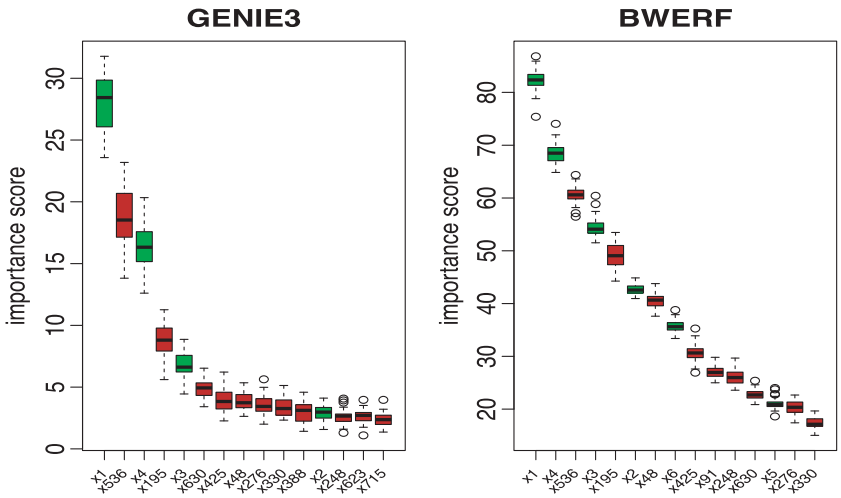


**Figure 2.** The effect of recursive elimination of noise variables (genes) on the final rankings of true variables when a simulate data set containing 5 positive variables and 1000 noise variables was used. Only the top 15 variables (X axis) emerged from 1000 variables that were subjected to GENIE3 and the BWERF algorithms are shown. Green bars: true positives, red: noise variables. The boxplots were based on 30 rounds of BWERF with an elimination rate of 10%.

lignocellulosic biosynthesis pathway. Like GENIE3, the champion of the DREAM4 in silico network challenge (http://dreamchallenges.org/project/dream4-in-silico-network-challenge/), BWERF also uses random forest, but has backward elimination process to augment the emergence of true TFs, as the noise variables are eliminated. The pronounced effect of BWERF is fully displayed in Figure 2, where the rankings of the true regulatory genes with medium-level regulatory strengths were significantly elevated. Compared with the bottom-up GGM algorithm, BWERF can construct ML-hGRNs with significantly reduced edges that enables biologists to choose edges more easily for experimental validation. For example, the two ML-hGRNs built above the lignocellulosic pathway using bottom-up GGM and BWERF and the same input data set have 576 and 88 edges, respectively, and the average connectivities in the two networks are 6.9 and 1.8 edges, respectively. A flowchart of BWERF is shown in the original publication [4] to facilitate the understanding of detailed procedures of this algorithm.

## The use of top-down and bottom-up approaches to construct an ML-hGRN centered around a TF

When overexpression or suppression of a TF leads to phenotypic changes, we can obtain an ML-hGRN centered around the TF by sequentially using the top-down GGM algorithm and bottom-up approaches, namely, the bottom-up GGM or BWERF algorithms. The framework and detailed procedures for constructing ML-hGRNs are illustrated in Figure 3. As illustrated in Steps 1 and 5 (Figure 3), two RNA-seq (or two microarray experiments) are required: one that yields data for the top-down GGM algorithm and the other to provide data for the bottom-up approaches (Figure 3). In some situations, however, data produced from a single experiment can be used for both the top-down and bottom-up approaches.

### Experiments to generate data for the top-down GGM algorithm

Two different experimental systems can be used to generate the RNA-seq or microarray data for the top-down GGM algorithm (Step 1, Figure 3). The first is a transient expression system in which plasmids carrying the TF are infiltrated or injected into newly plant cells or tissues that include protoplasts [1, 14], leaves [15], cotyledonary explants [16], young cuttings [17] and roots [18], or animal cell lines or tissues using, for example, electroporation [19, 20] and a stearylated cell-penetrating peptide (CPP)-based transfection methods [21]. Infiltrated/transfected cells or tissues can be harvested in a short time series. As plant protoplasts usually survive for 48 h, transfected protoplasts can be harvested at 6, 12, 24 and 36 h; at least three biological replicates should be used for each time point to ensure the quality of DEGs. To increase the accuracy of DEG identification, protoplasts infiltrated with a blank plasmid vector are used as controls and should be harvested at the same time points (i.e. 6, 12, 24 and 36 h). For plant tissues, the peak of transient expression generally appears at 72–120 h; therefore, harvesting infiltrated tissues at 48, 72, 96 and 120 h is suggested. DEGs identified at multiple time points can be merged and used collectively as inputs for the top-down GGM algorithm.

The second experimental system for producing data for the top-down GGM algorithm involves taking advantage of existing stable transgenic lines. If the trans-TF is under the control of a promoter other than the constitutive 35 S promoter, then temporal changes in the expression of the TF should be investigated. When a TF has dramatically varying expression levels within a day or a short developmental period, a few or even a single transgenic line can be used and harvested in a time series. The goal is to generate gene expression profiles in which the trans-TF varies in its expression throughout the time course of experiment. However, if a TF is under the control of the constitutive 35 S promoter or does not exhibit dramatic changes in expression over a period of time, more independent stable transgenic lines of this TF with different expression levels are anticipated. Under this circumstance, plant materials may not be harvested from multiple transgenic lines in a time series; instead, one sample is harvested from each line at the same time to generate a 'pseudo-profile' in which the TF has varied expression levels across different lines. As demonstrated previously [3], a data set was generated from nine independent *Ptr-miR397a* transgenic lines with varied *Ptr-miR397* expression levels as well as three wild-type lines to enable construction of an ML-hGRN.

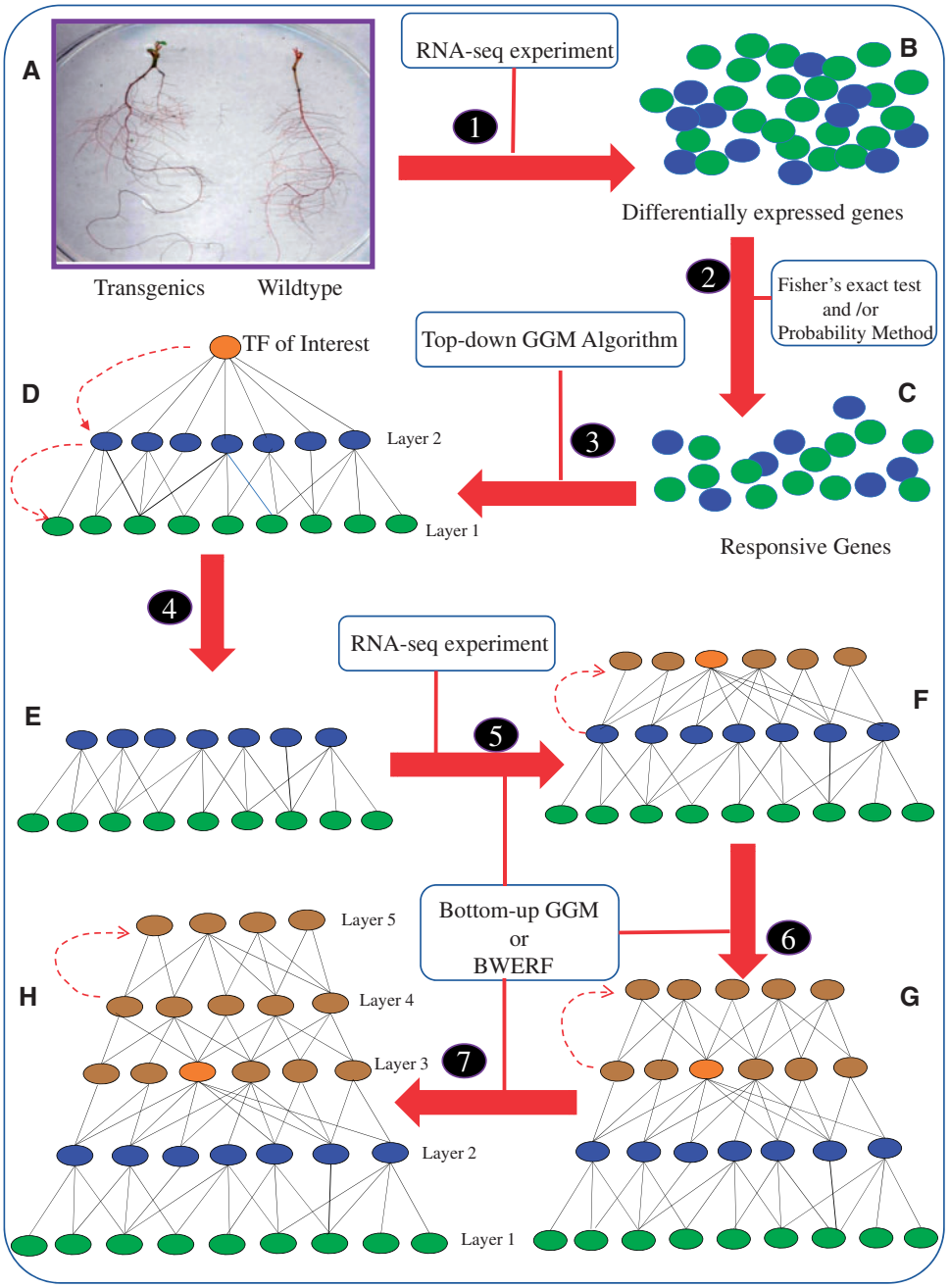


**Figure 3.** Illustration of using the top-down GGM algorithm followed by bottom-up approaches, bottom-up GGM or BWERF, to construct ML-hGRN.

Conclusive evidence regarding which of these experimental systems yields a ML-hGRN with more authentic regulatory relationships has not yet been obtained.

## Experiments to generate data for the bottom-up GGM algorithm

As shown in Figure 3, once we obtain a TF-mediated two-layered hGRN using the top-down GGM algorithm, then Layer 2 (Figure 3D), which contains the predicted direct target genes of the TF, can be used as the bottom layer for a bottom-up approach. This approach involves the use of either the bottom-up GGM or BWERF algorithm and, as a result, transforms the existing two-layered hGRN into an ML-hGRN. The data sets needed for bottom-up construction require careful consideration. One possibility is to use the same data that was used for the top-down GGM algorithm; one concern with this practice, however, is that strong relationships between the TF and direct target genes in Layer 2 may overwhelm relationships between other regulators and Layer 2 genes, making it difficult to identify additional regulators that are at the same or a higher hierarchical level as the TF. Although, at present, we cannot assess the 'overwhelming effect' of a TF, we have used data produced from nine *Ptr-miR397a*-over-expressing lines to build an ML-hGRN; when the direct target genes, LAC family members, of *Ptr-miR397a* were used as bottom layer, and 1208 TFs plus *Ptr-miR397* were used as inputs for the bottom-up GGM algorithm, we obtained a four-layered hGRN containing both *Ptr-miR397a* (in the second layer) and multiple

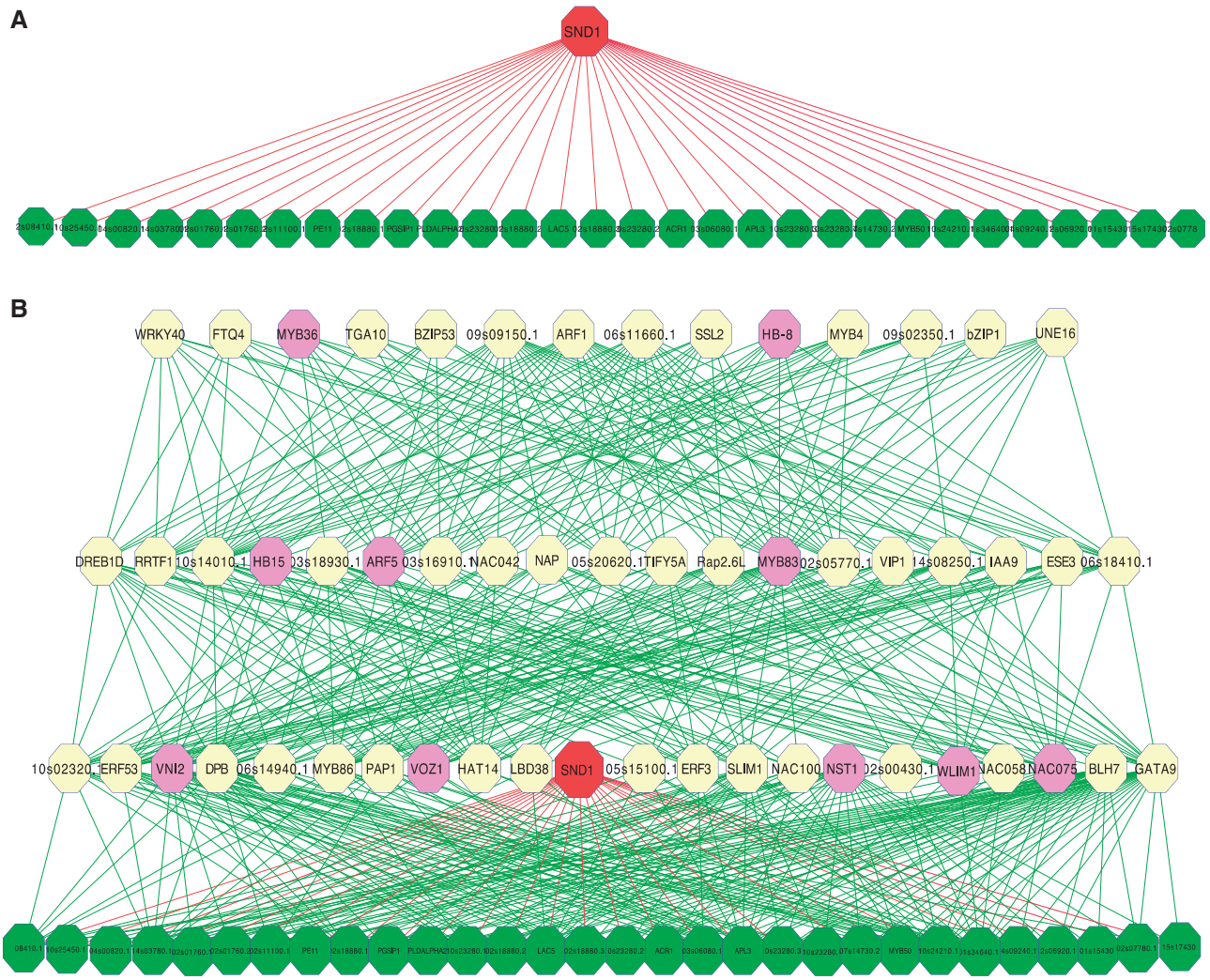


**Figure 4.** The ML-hGRN built using top-down GGM algorithm followed by bottom-up GGM algorithm. The input data set was generated from protoplasts of the developing xylem of *P. trichocarpa* through transient overexpression of SND1 gene and RNA-seq technology. The data set (GSE49911) is available in Gene Expression Omnibus (GEO), NCBI. Green nodes: bottom-layer genes; purple node: known positive regulators of wood formation.

well-known wood formation regulatory genes [3], including *PtrMYB42*, *PtrMYB48*, *PtrMYB 52*, *PtrMYB021*, *VND1* and *NST1* [22]. This example illustrates that data generated for the top-down GGM algorithm can, in fact, be used again for the bottom-up GGM algorithm to construct a multilayered hGRN that contains an aggregate of functionally collaborative TFs.

A second possibility is to generate a new data set specifically for bottom-up construction using a RNA-seq experiment (Figure 3, Step 5), which should be carried out under the same or similar experimental conditions as used for RNA-seq experiment at Step 1 (Figure 3). The use of the same or similar experimental conditions guarantees that regulatory relationships existing in the lower and upper layers of the ML-hGRN built with two data sets are authentic for that condition. When designing an RNA-seq or microarray experiment for bottom-up network construction, we need to consider if the gene profiles produced have sufficient length for recognizing or differentiating regulatory interactive events. This is a complicated issue that has been discussed in the literature [23, 24]; the answer is contingent on many variables that include considerations of how the data will be used and what kinds of methods or algorithms will be used. For the bottom-up method, we have previously used data sets with 12 samples (nine independent transgenic lines plus three control samples) [3] or 24 samples of six time points (two treatments versus two controls at each time point) [25]. In the future, it will be fascinating to investigate whether use of one data set for both the top-down and bottom-up approaches produces different results from use of data sets from two separate experiments.

Additional guidelines of experimental design for producing interactive gene expression profiles are elaborated in our previous review paper [26], and interested readers may consult this review article and our previous publications [1–4, 27] for more details.

## Demonstration of the feasibility of top-down and bottom-up network construction algorithms with real gene expression data sets

### *Construction of an ML-hGRN centered on SND1*

The aforementioned data set (Accession #: **GSE49911,** Gene Expression Omnibus, NCBI) resulting from SND1 transient overexpression in xylem-derived protoplasts of *Populus trichocarpa* was used for top-down GGM algorithm. The inputs include the expression profiles of SND1 and 178 DEGs, and only one layer

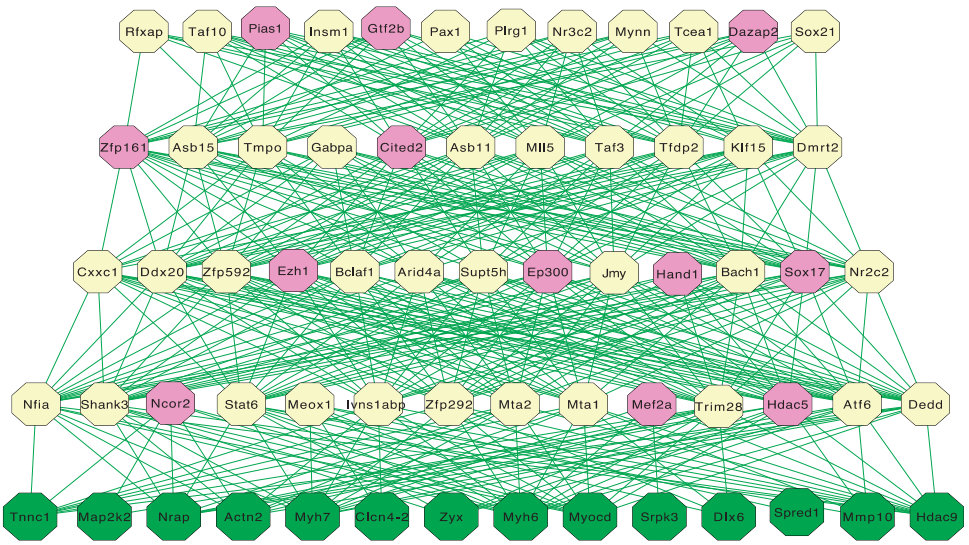


**Figure 5.** The ML-hGRN that regulates mouse heart development. The network was built using bottom GGM algorithm with the bottom-layer genes being selected from existing literature and evidenced to participate the cardiac muscle development. Green nodes: bottom-layer genes; purple node: known positive regulators of heart development.

immediately below SND1 was built (Figure 4A). Then, the 30 inferred targets genes of SND1 with the highest interference frequency were used as the bottom layer and the expression profiles of 1284 TFs were used as inputs for upper layers to construct a four-layered hGRN using bottom-up GGM algorithm (Figure 4B). Note that the SND1 was built into the network again during bottom-up network construction process (Figure 4B). The SND1 responsive genes (Supplementary Table S1), interfered genes (Supplementary Table S2) and network (Supplementary Table S3) are all provided in the Supplemental Excel File. The purple nodes in Figure 4B contain those genes that are evidenced by existing literature to regulate wood formation. These genes include *MYB36* [28], *ARF5* [29], *HB-8* [30], *MYB83* [31], *NAC075* [32], *WLIM1* [33], *NST1* [22], *VOZ1* [34] and *VNI2* [35]. The regulatory roles of some other genes in the ML-GRN may be implicated by their functions. For example, IAA3, a homolog of IAA9 controls wood formation in aspen trees [36]. All these facts suggest the viability of the combined approach.

### *Construction of an ML-hGRN controlling heart development*

The usefulness of the developed network construction approaches is also demonstrated by constructing an ML-hGRN governing heart development using a compendium microarray data set that was described earlier in our publication [37]. Briefly, the data set comprises 172 microarrays pooled from nine independent experiments with the following Gene Expression Omnibus (GEO) accession IDs: GSE11291, 15078, 19875, 29145, 30495, 3440, 38754, 5500 and 7781. First, 14 genes involved in cardiac muscle development were identified from existing literature [38–40] and used as bottom-layered terminal genes. Then, the profiles of these 14 genes and 1675 TFs were used as inputs to build an ML-hGRN. The resulting network is shown in Figure 5 and Supplementary Table S4. In this network, 12 regulatory genes including *Ep300* [41], *HDAC9* [42], *NCOR2* [43], *HAND1* [44], *SOX17* [45], *Ezh1* [46], *ZFP161*[47], *Cited2* [48], *GTF2b* [49], *Mef2a* [38], *Pias1* [50] and *Dazap2* [51] were found to be positive regulators of heart development. Their annotations and interference frequencies are shown in Supplementary Table S5. In addition to those known regulators that are explicitly evidenced to control heart development, some other genes in the network have the functions that indicate they may play a role in regulating heart development. For example, Pax1 is implicated to control the formation of heart outflow tract [52], Nr2C2 participates in muscle differentiation [53] and Dedd controls cell cycle and organ size [54].

## ML-hGRNS can provide new biological insights

The ML-hGRN that is best characterized in *Escherichia coli* has three layers. The top layer contains seven global, pleiotropic regulators that regulate 51% of operons in various metabolic pathways [55, 56]. These global regulators play a role in monitoring energetic and nutritional states, sensing environmental cues and regulating gene transcription by modulating DNA topology of the cells. The middle layer contains 105 TFs that have more local and dedicated regulation than the global regulators, and the bottom layer contains 749 terminal genes that include structural and functional genes. Studies in yeast also demonstrate that global regulators respond to various cellular cues [57], environmental factors and stresses [58]. In plants, there is also evidence for the existence of high hierarchical regulators. For example, *HY5* is defined as a high hierarchical regulator of transcriptional networks for photomorphogenesis [59, 60], and PsnSHN2 is a high hierarchical activator that coordinately regulates secondary wall formation through selective upregulation/downregulation of its downstream TFs that control secondary wall formation [61]. Mutations of high hierarchical regulators in some species have been linked to historical or milestone evolutionary events [62, 63]. Middle-level genes in the ML-hGRNs of five species, including *E. coli*, yeast, rat, mouse and humans, show consistent tendencies for collaboration [64], indicating that middle-level regulators tend to regulate terminal genes in a collaborative manner. In addition, middle-level regulators can receive 'commands' from multiple high hierarchical regulators and pass them down to terminal genes, exerting more specific regulation on structural or functional genes. Such a topology also provides opportunity for some midlevel genes to serve as 'control bottlenecks' in the gene hierarchy [57].

It is clear that the establishment of ML-hGRNs is critically important to further understanding of the regulation of biological processes. I have demonstrated the feasibility of combined

top-down and bottom-up approaches in both previous publications [1–4, 27] and this review article using real gene expression data. The ML-hGRNs constructed can provide information for the following purposes: (1) acquisition of an aggregation of most of TFs functionally associated with the terminal genes in the bottom layer; (2) creation of a causal map of molecular regulatory interactions; (2) generation of hypotheses for perturbation experimental design and ChIP-seq experiments; (3) selection of biomarkers within a network; (4) comparative network analysis for knowledge discovery; (5) identification of high hierarchical regulators, middle-level specific regulators and terminal genes as well as top-down 'chains-of-command' that can guide genetic engineering efforts.

## Comparison with other network construction methods

Currently, there are no methods that have been developed and tailored specifically for the construction of ML-hGRNs except those described above. Though many methods have been developed for building GRNs, these are not suited for constructing ML-hGRNs. Methods for constructing GRNs can be classified into correlation-, mutual information- and probability-based methods. Correlation-based methods include ParCorA [65] and GGM [66]. ParCorA uses zero second-order partial correlation as a measure of interaction between any triple genes, while GGM uses a novel shrinkage covariance estimator to calculate the optimal shrinkage intensity to infer gene association. Mutual information-based methods include RN [67], ARACNE [68], MRNET [69], C3NET [70] and MI3 [71]. RN, ARACNE, MRNET and C3NET use mutual information to evaluate dependency relationships between each pair of genes followed by additional filtering strategies to refine networks; MI3 uses mutual information to evaluate dependency relationships among three genes, that is two regulators and one target. This is different from our triple gene model, as our three genes consist of one regulator and two target genes. Probability-based methods, such as Bayesian network, are generally computationally intensive because learning graphic structures is nondeterministic polynomial time (NP)-hard problem [70]. Moreover, probability-based methods are generally less efficient than mutual information-based methods [71].

It is well established that the detection of causal patterns is more effective in a trivariate setting than in a bivariate context [72]. Both our top-down and bottom-up GGM algorithms lay the GGM on the top of a biological triple gene model to infer ML-hGRNs. Because the triple gene model is based on the well-accepted theory that two coexpressed target genes are more likely to be under the control of the same regulatory mechanism, it aligns well with general biological regulatory models, thereby enhancing the power and accuracy of network inference. Compared with ParCorA, we introduced the delta method to examine interference between one regulator and two target genes for each combination of triple genes. In addition, BWERF was added and the backward elimination method in BWERF method makes BWERF more efficient than GENIE3 [73] in recognizing authentic TFs that regulate target genes.

## Conclusion

As more TF overexpression lines with discernible phenotypic changes become available in various plant or animal species, we can take advantage of these lines and design small-scale experiments to acquire data for constructing an ML-hGRN centered around each trans-TF. In many circumstances, stable overexpression lines can be also superseded by a transient expression system. A transient expression system may generate perfect perturbed data for capturing causal relationships, but it is usually much more difficult to manipulate and produce replicable expression across multiple samples. With an awareness of the availability of top-down and bottom-up approaches and the corresponding data requirements, one can scale up routine comparative RNA-seq or microarray experiments (e.g. treatments versus controls) conducted on a TF of interest by including a few more time points and/or samples, which will ensure that the resulting gene expression data contain the prevailing regulatory interaction profiles for associating molecular entities, and such data can then be used to construct an ML-hGRN centered around the TF of interest. Such a strategy will not significantly increase experimental costs but will enable us to obtain layered hierarchical networks for identifying high hierarchical regulators, middle-level specific regulators and terminal genes as well as top-down 'chains-of-command' that can guide genetic engineering studies, leading to the discovery of novel biological knowledge that can eventually bridge genotypes and phenotypes.

**Key Points**

- We need to go far beyond treatment versus control experiments that can only provide a large number of DEGs. Methods and algorithms for constructing ML-hGRNs have become available using slightly larger-scale experiments.
- The construction of ML-hGRN can be initialized from a TF of interest. The resulting network can not only advance our understanding of regulatory roles and functions of each TF but also identify new regulators including those that are collaborative or located at higher hierarchic levels.
- The experimental validation of the direct target layer of a TF using ChIP-PCR before implementing bottom-up network construction methods will increase overall accuracy of the ML-hGRN.
- Although not specifically emphasized in this review, the construction of ML-hGRN can also start with a bottom layer that contains genes from a biological process or a pathway, leading to an ML-hGRN that governs a biological process or pathway.

## Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

## Funding

This work was supported by grant from US National Science Foundation Advances in Biological Informatics (Grant#: DBI-1458130) to H.W. and an open fund from China's State Key Laboratory of Tree Genetics and Breeding to H.W.

## References

1. Lin YC, Li W, Sun YH. SND1 transcription factor-directed quantitative functional hierarchical genetic regulatory

network in wood formation in *Populus trichocarpa*. *Plant Cell* 2013;**25**(11): 4324–41.
2. Kumari S, Deng, W Gunasekara, C, *et al*. Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes. *BMC Bioinformatics* 2016;**17**:132.
3. Lu S, Li Q, Wei H, *et al*. Ptr-miR397a is a negative regulator of laccase genes affecting lignin content in *Populus trichocarpa*. *Proc Natl Acad Sci USA* 2013;**110**(26):10848–53.
4. Deng W, Zhang K, Busov V, *et al*. Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways. *PLoS One* 2017;**12**(2):e0171532.
5. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**(1):139–40.
6. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**(10):R106.
7. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**(1):1.
8. Breitling R, Armengaud P, Amtmann A, *et al*. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* 2004;**573**(1–3):83–92.
9. Kadota K, Nakai Y, Shimizu K. A weighted average difference method for detecting differentially expressed genes from microarray data. *Algorithms Mol Biol* 2008;**3**:8.
10. Fisher RA. On the interpretation of x-square from contingency tables, and the calculation of P. *J R Stat Soc* 1922;**85**(1):87–94.
11. Fisher RA. *Statistical Methods for Research Workers*. New York: Harper–Biological Monographs and Manuals, No. 5, p. 356.
12. Kumari S, Nie J, Chen HS, *et al*. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One* 2012;**7**(11):e50411.
13. MacKinnon DP, Lockwood CM, Hoffman JM, *et al*. A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods* 2002;**7**(1):83–104.
14. Katagiri F. Transient expression in protoplasts. *CSH Protoc* 2007;**2007**(2):pdb.prot4692.
15. Pitzschke A, Persak H. Poinsettia protoplasts—a simple, robust and efficient system for transient gene expression studies. *Plant Methods* 2012;**8**(1):14.
16. Nanasato Y, Konagaya KI, Okuzaki A, *et al*. Improvement of Agrobacterium-mediated transformation of cucumber (Cucumis sativus L.) by combination of vacuum infiltration and co-cultivation on filter paper wicks. *Plant Biotechnol Rep* 2013;**7**(3):267–76.
17. Takata N, Eriksson ME. A simple and efficient transient transformation for hybrid aspen (*Populus tremula* x *P. tremuloides*). *Plant Methods* 2012;**8**(1):30.
18. Zhong L, Zhang Y, Liu H, *et al*. Agrobacterium-mediated transient expression via root absorption in flowering Chinese cabbage. *Springerplus* 2016;**5**(1):1825.
19. Zu Y, Huang Lu SY, *et al*. Size specific transfection to mammalian cells by micropillar array electroporation. *Sci Rep* 2016;**6**:38661.
20. Chang DC. Experimental strategies in efficient transfection of mammalian cells. Electroporation. *Methods Mol Biol* 1997;**62**: 307–18.
21. Karro K, Mannik Mannik TA, *et al*. DNA transfer into animal cells using stearylated CPP based transfection reagent. *Methods Mol Biol* 2015;**1324**:435–45.
22. Ye Z-H, Zhong R. Molecular control of wood formation in trees. *J Exp Bot* 2015;**66**(14):4119–31.
23. Zien A, Fluck J, Zimmer R, *et al*. Microarrays: how many do you need?. *J Comput Biol* 2003;**10**(3-4):653–67.,
24. Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol* 2002;**3**(5): research0022.1–10.
25. Wei H, Yordanov YS, Georgieva T, *et al*. Nitrogen deprivation promotes Populus root growth through global transcriptome reprogramming and activation of hierarchical genetic networks. *New Phytol* 2013;**200**(2):483–97.
26. Yang C, Wei H. Designing microarray and RNA-Seq experiments for greater systems biology discovery in modern plant genomics. *Mol Plant* 2015;**8**(2):196–206.
27. Wei H, Yordanov Y, Kumari S, *et al*. Genetic networks involved in poplar root response to low nitrogen. *Plant Signal Behav* 2013;**8**(11):e27211.
28. Kamiya T, Borghi M, Wang P, *et al*. The MYB36 transcription factor orchestrates Casparian strip formation. *Proc Natl Acad Sci USA* 2015;**112**(33):10533–8.
29. Przemeck GK, Mattsson J, Hardtke CS, *et al*. Studies on the role of the Arabidopsis gene MONOPTEROS in vascular development and plant cell axialization. *Planta* 1996;**200**(2): 229–37.
30. Baima S, Possenti M, Matteucci A, *et al*. The *Arabidopsis* ATHB-8 HD-zip protein acts as a differentiation-promoting transcription factor of the vascular meristems. *Plant Physiol* 2001; **126**(2):643–55.
31. Ko JH, Jeon HW, Kim WC, *et al*. The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis. *Ann Bot* 2014;**114**(6): 1099–107.
32. Endo H, Yamaguchi M, Tamura T, *et al*. Multiple classes of transcription factors regulate the expression of vascular-related NAC-domain7, a master switch of xylem vessel differentiation. *Plant Cell Physiol* 2015;**56**(2):242–54.
33. Andersson-Gunneras S, Mellerowicz EJ, Love J, *et al*. Biosynthesis of cellulose-enriched tension wood in Populus: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J* 2006; **45**(2):144–65.
34. Mitsuda N, Hisabori T, Takeyasu K, *et al*. VOZ; isolation and characterization of novel vascular plant transcription factors with a one-zinc finger from *Arabidopsis thaliana*. *Plant Cell Physiol* 2004;**45**(7):845–54.
35. Yamaguchi M, Ohtani M, Mitsuda N, *et al*. VND-INTERACTING2, a NAC domain transcription factor, negatively regulates xylem vessel formation in *Arabidopsis*. *Plant Cell* 2010;**22**(4):1249–63.
36. Nilsson J, Karlberg A, Antti H, *et al*. Dissecting the molecular basis of the regulation of wood formation by auxin in hybrid aspen. *Plant Cell* 2008;**20**(4):843–55.
37. Li J, Wei H, Zhao PX. DeGNServer: deciphering genome-scale gene networks through high performance reverse engineering analysis. *Biomed Res Int* 2013;**2013**:856325.
38. Black BL. Myocyte enhancer factor 2 transcription factors in heart development and disease. In: RP Harvey (ed). *Heart Development and Regeneration*. Oxford: Academic Press, 2010, 673–99.
39. Paris J, Virtanen C, Lu Z, *et al*. Identification of MEF2-regulated genes during muscle differentiation. *Physiol Genomics* 2004; **20**(1):143–51.
40. Risebro CA, Searles RG, Melville AA, *et al*. Prox1 maintains muscle structure and growth in the developing heart. *Development* 2009;**136**(3):495–505.

41. Shikama N, Lutz W, Kretzschmar R, *et al*. Essential function of p300 acetyltransferase activity in heart, lung and small intestine formation. *Embo J* 2003;**22**(19):5175–85.
42. Chang S, McKinsey TA, Zhang CL, *et al*. Histone deacetylases 5 and 9 govern responsiveness of the heart to a subset of stress signals and play redundant roles in heart development. *Mol Cell Biol* 2004;**24**(19):8467–76.
43. Jepsen K, Gleiberman AS, Shi C, *et al*. Cooperative regulation in development by SMRT and FOXP1. *Genes Dev* 2008;**22**(6): 740–5.
44. McFadden DG, Barbosa AC, Richardson JA, *et al*. The Hand1 and Hand2 transcription factors regulate expansion of the embryonic cardiac ventricles in a gene dosage-dependent manner. *Development* 2004;**132**(1):189–201.
45. Pfister S, Jones VJ, Power M, *et al*. Sox17-dependent gene expression and early heart and gut development in Sox17-deficient mouse embryos. *Int J Dev Biol* 2011;**55**(1):45–58.
46. Ai S, Yu X, Li Y, *et al*. Divergent requirements for EZH1 in heart development versus regeneration. *Circ Res* 2017;**121**(2): 106–12.
47. Li Y, Klena NT, Gabriel GC, *et al*. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. *Nature* 2015;**521**(7553):520–4.
48. Bamforth SD, Braganca J, Eloranta JJ, *et al*. Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. *Nat Genet* 2001;**29**(4):469–74.
49. Sayed D, Yang Z, He M, *et al*. Acute targeting of general transcription factor IIB restricts cardiac hypertrophy via selective inhibition of gene transcription. *Circ Heart Fail* 2015;**8**(1): 138–48.
50. Constanzo JD, Deng M, Rindhe S, *et al*. Pias1 is essential for erythroid and vascular development in the mouse embryo. *Dev Biol* 2016;**415**(1):98–110.
51. Cohen-Barak O, Yi Z, Hagiwara N, *et al*. Sox6 regulation of cardiac myocyte development. *Nucleic Acids Res* 2003;**31**(20):5941–8.
52. Monsoro-Burq AH. PAX transcription factors in neural crest development. *Semin Cell Dev Biol* 2015;**44**:87–96.
53. Kim SY, Shim J-H, Chun E, *et al*. TGF-beta-activated kinase 1 (TAK1) and apoptosis signal-regulating kinase 1 (ASK1) interact with the promyogenic receptor Cdo to promote myogenic differentiation via activation of p38MAPK pathway. *J Biol Chem* 2012;**287**(5):11602–15.
54. Arai S, Miyake K, Voit R, *et al*. Death-effector domain-containing protein DEDD is an inhibitor of mitotic Cdk1/cyclin B1. *Proc Natl Acad Sci USA* 2007;**104**(7):2289–94.
55. Martınez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* 2003;**6**(5):482–9.
56. Martinez-Antonio A. *Escherichia coli* transcriptional regulatory network. *Netw Biol* 2011;**1**(1):21–33.
57. Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci USA* 2006; **103**(40):14724–31.
58. Balazsi G, Barabasi AL, Oltvai ZN. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc Natl Acad Sci USA* 2005;**102**(22): 7841–6.
59. Nawkar GM, Kang CH, Maibam P, *et al*. HY5, a positive regulator of light signaling, negatively controls the unfolded protein response in Arabidopsis. *Proc Natl Acad Sci USA* 2017;**114**(8): 2084–9.
60. Lee J, He K, Stolc V, *et al*. Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development. *Plant Cell* 2007;**19**(3): 731–49.
61. Liu Y, Wei M, Hou C, *et al*. Functional characterization of populus PsnSHN2 in coordinated regulation of secondary wall components in tobacco. *Sci Rep* 2017;**7**(1):42.
62. Crombach A, Hogeweg P. Evolution of evolvability in gene regulatory networks. *PLoS Comput Biol* 2008;**4**(7):e1000112.
63. Peter IS, Davidson EH. Evolution of gene regulatory networks controlling body plan development. *Cell* 2011;**144**(6): 970–85.
64. Bhardwaj N, Yan KK, Gerstein MB. Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proc Natl Acad Sci USA* 2010;**107**(15):6841–6.
65. de la Fuente A, Bing N, Hoeschele I, *et al*. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 2004;**20**(18):3565–74.
66. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 2005;**21**(6):754–64.
67. Butte A, Kohane I. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Proc Pac Symp Biocomput* 2000;**5**:415–26.
68. 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**(Suppl 1):S7.
69. Meyer PE, Lafitte F, Bontempi G. Minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* 2008;**9**:461.
70. Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. *BMC Syst Biol* 2010;**4**:132.
71. Luo W, Hankenson KD, Woolf PJ. Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC Bioinformatics* 2008;**9**:467.
72. Schäfer J, Strimmer K. Learning large-scale graphical Gaussian models from genomic data. In: JF Mendes (ed). Proceedings of "Science of Complex Networks: from Biology to the Internet and WWW" (CNET 2004). Aveiro, PT: The American Institute of Physics, 2005.
73. Huynh-Thu VA, Irrthum A, Wehenkel L, *et al*. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;**5**(9):e12776.