**Gene regulatory networks in plants: Learning causality from time.**

**Gabriel Krouk3, Jesse Lingeman 2, Gloria Coruzzi1, Dennis Shasha2**

**Gab, as this has been a tremendous amount of work for Jesse, would you mind making him co-first author. Gloria and I should be co-senior authors**

**1** Center for Genomics and Systems Biology, Department of Biology, New York University, 12 Waverly Place, New York, 10003, USA

**2** Courant Institute of Mathematical Sciences, New York University, New York, NY, 10003, USA

**3** Institut de Biologie Intégrative des Plantes, UMR 5004, Biochimie et Physiologie Moléculaire des Plantes, Agro-M/CNRS/INRA/SupAgro/UM2, Montpellier, France

**Abstract*.***

The ultimate goal of Systems Biology is to generate models that can predict how a system will react to untested conditions or genetic perturbations. This area of research is particularly relevant to plants where such predictive models can be useful for interventions in agriculture, and to engineer plants improved in their response to environmental change. Due to their sessile mode of life, plants are subject to drastic variations in their environment that lead to rapid adaptation of molecular behaviors of gene regulatory networks. A critical question is understanding the cascade of gene responses that underlie a plant’s adaptation to environmental perturbations. In the present review we: i) describe experimental approaches to understand such dynamic and causal gene relationships in plants using time-series (kinetic) and other data, ii) review the analytical approaches used to infer causality in the Gene Regulatory Network (GRNs) from these types of genomic data, iii) suggest best practices in experimental design and analytical approaches for future efforts of this kind, iv) review methods for high through put validation of gene regulatory networks in plants.

**Introduction: After we figure out the rought outline, please ask Sandrine to read to correct English problems. I’ll do a final pass of course, but in the meantime, please try to bring it up to the standards of your normal papers.**

Gene Regulatory Networks are key to understand the living cell behavior. Very successful studies have been performed in prokaryotes. These models are able to predict genome-wide variations in untested environmental conditions and regulatory genes involved [what is “involved” doing here?] [[1](#_ENREF_1), [2](#_ENREF_2)]. [I think this general discussion of prokaryotic vs. eukaryotic cells doesn’t need to be here. Isn’t it general knowledge? Let’s quickly get to the point. So cut from here] However, this is rather limited when comes multi-cellular organisms or even eukaryotic cells as compared to prokaryotes. Indeed, prokaryotic sequence-specific DNA-binding transcription factors often recognize binding sites longer than 12 base-pairs (bp) [[3](#_ENREF_3), [4](#_ENREF_4)], whereas binding sites for eukaryotic transcription factors are usually shorter, 5 to 10 bp long [[5](#_ENREF_5)]. Some hypotheses propose, that a combinatorial mechanism composed of factors that recognize short sequences is probably a more efficient way (requires a reduced number of factors) for selectively regulating the expression of tens of thousands of genes, than a mechanism based upon factors that are each dedicated to control a small number of genes and operate through longer target sites (as in prokaryotes). This is also why eukaryotes networks are probably several order of complexity above prokaryotes ones.

Moreover, high throughput transcriptomic techniques recently allowed investigating the transcriptomic state under many environmental conditions. Network inference has become a scientific challenge that tends to predict targets controlled by a particular TF. The union of all the interactions of this kind (TF 🡪 Target), is called herein GRN. Many inference techniques have been developed that use only gene expression data [[5-8](#_ENREF_5)], or several layers of information [[2](#_ENREF_2), [9-12](#_ENREF_9)]. These approaches helped to deepen our knowledge in biology in several different organisms; including into Arabidopsis. This is what we are proposing to review herein in the first part of this work. We particularly emphasize the role of kinetic data to understand GRNs. In a second part of this work we propose *in silico* results in order to determine best practices to minimize cost of experiments when maximizing the power of currently available machine algorithms. This article focusses on inference of causality in genomics, but its techniques apply to any setting (whether in plants or other species) in which elements may singly or collectively affect others. [cut to here] The article consists of three parts: 1. A review of efforts to use time series and other data to infer regulatory edges, showing the kinds of biological insights that can be obtained. 2. A description and a categorization of the experimental methods that are used. 3. An in-silico exploration using the DREAM framework to determine questions of practical interest, notably how to plan experiments to gain the maximum insight from each experiment.

**I) Successful stories in the field of plant science to learn GRN and their behavior over time.**

Different kinds of systems approaches are used to model GRN in plants. They basically can be classified into 2 different kinds. Here we exemplify these 2 kinds

**“Bottom-up” approaches. [Since we use the term “prior” for already known knowledge, can we use “Strong prior” vs. “Weak prior” approaches?]**

“Bottom-up” approaches are grounded into a very strong previous knowledge of the gene regulatory networks [[13](#_ENREF_13)] such as for instance: auxin signaling [[14](#_ENREF_14)], circadian clock [[15](#_ENREF_15), [16](#_ENREF_16)], or flower development [[17](#_ENREF_17)]. [Cut from here] This previous knowledge is formalized (using ODE systems, Boolean model), which are numerically solved and studied. Outputs of the models are then compared to experimental data in order determine their predictive power. When predictive, the models can be used to explore *in silico* GRN behavior in untested conditions and determine overall system properties/architecture.

This kind of investigations led to stinking [ca pue? Pouquoi?] results exemplified herein. [To here] For auxin signaling, Vernoux et al., 2001 [[14](#_ENREF_14)] built a model based on previous knowledge of the AUX/IAA-ARF transcription factors network and Yeat-2-hybrid experiments drafting the possibility of interactions between the protein partners. This model demonstrated that the GRN built (Resolution of ODE) display a strong buffering capacity that is revealed *in planta* in the shoot apical meristem by using DII-VENUS as a reporter of the input of signaling pathway and DR5 as an output.

Circadian clock is also a well-determined gene regulatory system that consists in interlocked transcription factors feedback loops [[18](#_ENREF_18), [19](#_ENREF_19)]. GRN modeling of this particular system has been successful in determining its over-time evolution and the critical components involved in some key features of the oscillations. For instance, in Pokhilko et al, (2010) [[18](#_ENREF_18)], the GRN model was key to reveal/predict the role of PRR5 as a night inhibitor of the LHY/CCA1 expression including a role of in the control of the phase of the of morning gene expression. In the same work these GNR generated hypothesis was validated by matching *prr5* mutants behavior to gene expression predicted by the model [[18](#_ENREF_18)]. This bottom-up approach is another example of the key role of GNR modeling, and probing of evolution over time, in solving complex gene expression systems.

Flower development (ABC model) is a textbook example of conserve GRN that control the fate of cells into sepals, petals, stamens, and carpels [[20](#_ENREF_20)]. A successful approach using a discrete network model (gene expression in coded into discrete values) has been able to simulate the cell-fate determination during floral organ primordial formation in Arabidopsis [[17](#_ENREF_17)]. This model demonstrated again that this particular GRN dynamically converge towards steady states in gene expression that defines the different cell fates into flower composing organs. These steady states (or basin of attraction) can be seen as an equilibrium states obtained independent of the gene expression initial values. This shows that this GRN as buffering capacities that channel gene expression behavior towards a dedicated state (make a particular organ) [[17](#_ENREF_17)].

These few examples of successful bottom-up approaches all demonstrate that connections in GRNs confer robust emerging properties dedicated to developmental or environmental adaptations.

 **“Top-down” approaches.**

In bottom-up approaches (above) the key prerequisite is to gain enough knowledge of the modeled GRNs in order to capture its intrinsic behaviors see above or Middleton et al, (2012) for a good review [[13](#_ENREF_13)]. However for many systems this initial knowledge still need to be discovered.

“Top-down” approaches use computers algorithms in order to infer potential connections/causality in GRNs from Omics datasets. In general many techniques are used to infer GRNs in the field of systems biology (for reviews see [[1](#_ENREF_1), [21](#_ENREF_21)]). These techniques tend to be very successful in simpler systems such as bacteria. Striking example is the model built from multi level dataset (including transcriptomic data and CRE inference) of the gene regulatory programs that lead *Halobacterium salinarum* response to environmental cues. The model has been built *de novo* by machine learning procedure based on 72 transcription factors responding to 9 environmental factors. The same model was able to predict correct gene response (80% of the genome) in 147 untested conditions [[2](#_ENREF_2)]. In our opinion this demonstrates that “top-down” approaches can be very successful and has to be scaled up in order to reach the same predictive power in eukaryotic multi-cellular system [[22](#_ENREF_22)]. However in plant science, since this eukaryotic system is far more complex than yeast or bacteria, we can say that the field of GRN *de novo* learning if far less advanced.

[Because I discuss this later, please cut from here] In the plant field the 3 most popular/used top-down approaches methods are i) classical correlations networks, ii) Graphical Gaussian model (GGM) (partial correlation), iii) State-space modeling or combinations of the above. Each of them has different power for defining causality in networks. [To here]

Correlations networks are very popular in the plant science field [[23](#_ENREF_23)]. For instance coupled to other level of integrations they where able to highlight key features of GRNs in plants. For instance the use of the Multinetwork [[24](#_ENREF_24), [25](#_ENREF_25)], that collate (among others integration levels): putative TF🡪Cis-Regulatory-Elements (CRE), protein-protein, miRNA--| mRNA interactions highlighted important Nitrogen/circadian clock connections. Indeed, together with correlation networks the Multinetwork was able to highlight a key role of Nitrogen (N) regulation of CCA1 and *vice versa,* the role of CCA1 in the control of N assimilation through the direct (validated by Chromatin-IP) transcriptional regulation of the Asparagine Synthase 1 gene [[26](#_ENREF_26)]. The same work revealed that bZIP1 could be a regulator of ASN1, which was experimentally validated by a different lab [[27](#_ENREF_27)]. Moreover, correlation network approaches were strikingly successful in identifying 2 genes (a myo-inositol-1-phosphate synthase, and a Kelch-domain protein) correlating with biomass accumulation in plants [[28](#_ENREF_28)]. Their individual role was further supported by association mapping study that demonstrated coherent allelic diversity at their locus [[28](#_ENREF_28)].

[Again, I discuss this, so cut from here] GGM can be viewed as “partial correlation networks”. Partial correlation is a technique devoted to infer causality in steady correlations. Mathematically if A, B, and C correlate together, partial correlation correlates A and B by “subtracting” the correlation due to A and C. Practically, partial correlation is the correlation between the residuals resulting from the linear regression of A with C, and of B with C. GGM has been successfully developed [[29](#_ENREF_29)] and applied to plant GRN [[30](#_ENREF_30), [31](#_ENREF_31)]. [to here] In Ingkasuwan et al., time-series are used to identify gene regulated across diurnal cycle. Then a sub-networks of starch metabolism involved genes together with regulated TFs were subjected to a GGM modeling. This model were challenged and validated by the study of regulators mutants that displayed starch granule defects in plastids [[30](#_ENREF_30)].

State-Space modeling is one modern machine learning technique that is devoted to detect causality in networks by solving ODE on artificially noised dataset (avoid over-fitting see part xxx). In plant science this technique has been applied to probe GRNs involved in leave senescence [[32](#_ENREF_32)] as well as or GRNs involved in regulating early transcriptional response to NO3- [[33](#_ENREF_33)]. Breeze et al, (2011) provided a high-resolution temporal picture of the ageing leaf transcriptome. Authors demonstrate that different programs are deployed across time that significantly involved particular TF families and CREs. In order to draft some new relationships between TFs in a set of clusters they use SSM to propose a GRN-model that correctly predicts the influence of ANAC092 and proposes several new connections that still need to be validated [[32](#_ENREF_32)]. In Krouk et al, (2010) a high resolution time course in response to NO3- also demonstrates different cluster of regulations that tends to be lighten up sequentially involving different Gene Ontology categories known to prepare plant to reduce NO3-. A subset of TFs and N transport and assimilation genes have been modeled in order to propose a GRN that explain NO3- signal propagation. The modeled has been challenge *in silico* as well as experimentally. *In silico* validation demonstrated that the model trained on the beginning of the kinetic is able to predict gene expression modulation on later time point (not used to train the model). Experimental validation consisted in studying the effect of over-expressing a predicted hub (SPL9 TF) on NO3- response of other NO3- regulated genes. Indeed, SPL9 over-expression modified the regulation of NIR gene, NIA2, but also of many NO3- regulated TFs [[33](#_ENREF_33)]. This brings the question of network adaptation to genetic perturbations. Indeed, TF regulation

**II) Analytical approaches used to infer causality in the Gene Regulatory Network (a mathematical point of view).**

Inferring a causal link is useful in many applications in plant biology, from genomics to ecology. If some A can cause some B to take on a high value (where A could be a gene in our context, a hormone, or a species in ecology), then preventing B from taking such a value can be done by removing some B, by removing some A or by interfering with the link from A to B. Conversely, making B achieve a higher value can be done by adding more B, adding more A, or enhancing the efficiency of the link from A to B. Commonly, causal relationships in biology may involve several elements A1, ..., Ak influencing some B, sometimes positively and sometimes negatively. The influences can be "linear" in which each element has either a positive or negative weight (or coefficient) or "non-linear" in which case the elements work synergistically. An example of synergy would be a dependency of B on the product of the concentrations of A3 and A7.

**Dennis, I think this (yellowed part below) indeed overlap very much with the description of the papers I did.**

**I was thinking that it could be more Math-or-computer-oriented where you could categorize the different algorithms used and math methods… Something more synthetic maybe… not a one by one paper…? Can we categorize ODE, decision tree, etc… something like that.**

**Can you have your student write this? It can be largely reduced since we are already too long.**

**Gab: Ok. There is no math framework to characterize them, but at least I can escape the paper-by-paper approach.**

**The approaches to network inference fall into the following categories (we include references to papers that use them). Virtually all approaches deteriorate as the size of networks becomes larger, some more than others. Fortunately, biology tends to be modular, so large analyses can be broken down into smaller ones and then recombined [Middleton et al.** Vernoux, T., et al. (2011). The auxin signalling network translates dynamic input into robust patterning at the shoot apex. Mol. Syst. Biol. 7: 508]

**Use** ordinary differential equations, often based on mass action, yielding equations of the form mRNA concentration = Transcription rate - Decay rate. Such approaches work especially well for small networks such as auxin networks. [Middleton et al and Yuan et al] An issue with the mass action approach is that it assumes that different inputs interact in a multiplicative manner (product of concentration of each component) whereas the interaction is likely more complex in biological as opposed to chemical settings.

An alternative is to use a Boolean approach which allows other logical relationships [Middleton et al and Espinosa-Soto]. Logic gates are based on thresholds, e.g. an AND gate will fire if the minimum input reaches a certain threshold, thus permitting non-linear relationships. These tend to work on smaller networks than the differential equation approaches.

Correlation techniques are automatic techniques to try to find single source-target relationships. To try to isolate the effects of one gene on another, many researchers make use of partial correlations (Schaefer and Strimmer, Ingkasuwan et al.) present an analysis of Graphical Gaussian Models. These models assume a Gaussian noise distribution and try to infer partial correlations (gene X influences gene Y while holding the effects of other genes constant). Partial correlations can be computed indirectly by computing regressions and then computing the correlations among the residuals. Such analyses cannot be done without heuristic approximations because the number of experiments (e.g. microarrays) is always far less than the number of genes. For small circuits the heuristics work well.

Mutual information (Carrera et al) still seeks pairwise relationships among variables (as do correlation methods) but without assumptions of linear dependencies. The advantage is that it can be used for large scale networks, because it looks just at linear relationships, but does not try to compute the weight of one gene on another in predicting the target’s expression value.

Some techniques analyze how changes can cause divergent behavior over time (Mendoza-Parra et al). The idea is that genes are in some steady state before some perturbation occurs and the technique follows the genes that change first, that change second, and so on to try to guess causality. This is the qualitative idea behind the differential equation approaches.

To discover physical connections directly (but without necessarily knowing the effect of those connections) much research (e.g. Brady et al) measure protein-protein interactions and protein-DNA interactions using chromotin IP and other such techniques. They try to derive a network based on follow-up experiments using qPCR. They don’t use time series experiments at all.

Finally, there are techniques that integrate many different data types (

Greenfield et al. [Alex Greenfield, Christoph Hafemeister, and Richard Bonneau
**Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks** Bioinformatics first published online March 21, 2013 doi:10.1093/bioinformatics/btt099 ]). Thus, if some target gene Z has possible connections from X and Y but not from W based on physical experiments, then only X and Y will be considered in the subsequent analysis. For steady state experiments, such a method might use mutual information methods. The kinetic inference algorithm uses these potential edges to derive an ordinary differential equation model that may combine linear and non-linear terms. The result is a set of equations that estimate the change in transcription of a target gene based on transcriptional levels of other genes using time series data. Such machine learning approaches that make strong use of parsimony to try to avoid overfitting; even when doing such corrections, some genes may see their weights reduced more or less depending on previous structural knowledge.

Finally, other work suggests trying lots of methods together (Marbach et al) showing empirically that a combination of strategies (wisdom of crowds) often lead to the best outcomes.

[cut as much of the rest of the yellow stuff as you think you need to]

In many cases, however, we simply lack sufficient data to explore all possible synergies. Suppose for example we wanted to explore the effects of all pairs of genes. The most straightforward way to do that would be to over-express or knock out every pair. This would require something like 300 million manipulations. Thus, methods often work in two phases whose first phase consists of finding a good-fitting linear model and whose second phase consists of exploring the synergies among elements that have large positive or negative weights in the linear models. A sort of pre-first phase is to cluster expression patterns in order to create "super-nodes" that can then be analyzed.

Regardless of the analysis that follows, experimental approaches to finding such causal links may entail performing:

A) "Steady state" experiments under multiple different conditions to detect associations between A and other elements. Such associations are bi-directional but may acquire directionality if it is known that, for example, A is an element that can change other elements (in the genomic context A could be a "transcription factor") and B is not.

B) Experiments that increase the quantity of some A to see which other elements are either enhanced (quantity increases) or repressed (quantity decreases).

C) Experiments that decrease the quantity of A or even knock it out (A goes to 0) may also reveal something about the influence of A.

D) Experiments over a closely spaced time course to enable inferences of the form "the state of A at time t may influence B at time t+1."

New table:

Method Expressiveness Scalability

Correlation Low High (thousands of genes)

Partial Correlation Medium Medium (up to 100 genes using heuristics)

Differential equations and linear regression Medium Medium

Non-linear regression High Low (up to 25 genes)

Boolean High Low (up to 25 genes)

Here would be the caption:

This table shows the tradeoff between expressiveness (the number of factors that can be applied to determine gene expression)

and the size of the analyzed network. Small networks can be handled by methods that are highly expressive. Combining

several small network modules holds the potential to analyze a large network.

**III) *In silico* probing of best practices in machine learning procedures.**

In these experiments, we have focused on algorithms that did well (define well?) in the DREAM competition (present the competition for newbies). The software we provide with this paper is extendible, so other algorithms (including ones referred to in the related work) can be incorporated to improve the results. Our intent here is not to advocate any particular set of algorithms but rather to show the kinds of results one can achieve.

The benefit of using simulated data is that we can change the size of the network, the noise level, the number of time points to take, the number of replicates, and the algorithms all while being able to judge the quality of inferences because the underlying network is known. Using the tool we analyze two questions that can be useful to experimentalists trying to construct causal networks:

Given a fixed number of tests that can be done, where each test is some kind of whole genome expression readout, should we prefer tests at more time points or more replicates per time point? (The time points in both cases are spread out evenly over 1008 simulation time units, so when there are fewer time points, the interval between them is more than when there are many.) Practically speaking given an experimental budget, can we characterize data in such a way that we can determine whether to use more replicates or more time points assuming the data is varying over time?

How much gain do we get from steady state data, from mutation data, from both put together?

We answer these questions in both low and high noise contexts as well as for 20 gene as well as 100 gene networks. Low noise means … [*Need to specify]* We use two algorithms that did very well in the in the DREAM competitions – Inferelator (though we use the latest version discussed above) for the time series and mutational aspects and GENIE3 for the steady state portions of the experiment. GENIE3 [ref to Lingeman/Shasha book] is an algorithm that uses an ensemble of regression trees to infer networks from steady state data, thus giving useful priors for the Inferelator.

Our figure of merit is to get the best precision-recall possible. That is we measure “Area Under Precision Recall” curve. Interpret this as follows: lay out edges in descending order of their score (given by whatever algorithm one uses). When all real edges are found, the curve reaches the right side of the graph. At any given time, we evaluate the precision. If we now look at the area under the curve, we get a figure of merit. The ideal area under the curve would be 1 represented by figure Perfect [*figure to come]* If the proportion of real edges is small compared to the number of possible edges (say 5%), then a random choice of edges would lead to an area under the curve of 5% [*figure to come].*

1. *figure:* Time Series only

three histograms corresponding to low noise for different tradeoffs

of time points vs. replicates

ditto for middle noise

ditto for high noise

Caption: In all noise contexts, using many time points with relatively few replicates is an advantage for small networks.

2. *figure* Time series + Steady State

same as above

Caption: When steady state data is included, then ….

3. *figure* Time Series + Steady State + Mutant

Caption: When both steady state and mutant data are included…

The story changes when we go to 100 genes. First, the quality of the network goes way down. Second, it is much more useful to have more replicates even if at the cost of fewer time points.

*Gab: There will be three more figures for the 100 gene networks.*

Now we look at real data. When looking at real data, we cannot ask for precision and recall because there is not gold standard. However, we can look at out-of-sample data. That is, we train on a subset of the data (from earlier time points) and predict on a later time point.

For the German data, we capture the percent correctly predicted from

the Inferelator and by naive extrapolation (i.e. if gene g went up from

t-1 to t then it will go up again between t and t+1). *We might need one figure but I don’t think so.*

We conclude with advice for experimentalists. Prior information is very useful. How to trade off replicates with time points.

In the case of DREAM, For small 20 gene networks, 2 replicates and 15 time points is the best approach as you can see from fewerreps20genePR.pdf The overall Area Under Precision Recall Curve is 0.23. By contrast, 5 replicates and 5 time points gives only 0.13. That is under medium noise conditions. When the noise is high, AUPR goes down to 0.15 without steady state and to 0.17 using the best algorithm for that data (GENIE 3) when using steady state. So for small networks, using more time points seems to be a good idea. When we add in steady state data, that makes the area under precision recall to rise to 0.25. So not so much.

**IV) Experimental validations of inferred GRNs.**

GRNs modeling are providing very strong hypothesis that need to be experimentally probed/validated. So far experimental techniques used to validate TF🡪target relationships are i) chromatin immune-precipitation (Chip) followed by genome wide techniques (deep-seq or Arrays) or the use of Glucocorticoid receptor fusion to the TF of interest.

The first approach identifies TF binding to the DNA when the second detects TF action on target transcription. It is interesting to note that binding does not warrant gene activation and the other way around. This has been clearly demonstrated in Arabidopsis for a couple of transcription factor [[34](#_ENREF_34), [35](#_ENREF_35)]. These techniques can be time consuming because transgenic lines need to be produced in order to tag the studied TF with GFP, GR (or any other specific tag) in order to proceed with Chip-Seq or DEX activation [[34](#_ENREF_34), [35](#_ENREF_35)]. Recently, protoplast system and fluorescent activated cell sorting have been employed together, in order to scale up the validation procedure of GRN in a plant system.

Briefly, plant protoplasts are transformed with plasmid harboring a selective marker (Red Fluorescent Protein- RFP), together with the over-expression of a studied TF fused to GR (glucocorticoid receptor (from rat)). Fluorescent Activated Cell Sorting (FACS) sorts only successful transformation events. This technique allows to i) overproduce the studied TF but ii) controls its massive entrance in the nucleus by dexamethasone (DEX) treatment. This permits to co-treat protoplast with DEX and cycloheximide (CHX:Translation inhibitor). This rapid technique (named TARGET for: Transient Assay Reporting Genome-wide Effects of Transcription factors) allows the TF targets identification in less than a week of experimentation, opening perscpectives towrds high throughput investigations/validations of GRNs in plants [[36](#_ENREF_36)].

**References:**

1. Bonneau R: **Learning biological networks: from modules to dynamics.** *Nat Chem Biol* 2008, **4:**658-664.

2. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, et al: **A predictive model for transcriptional control of physiology in a free living cell.** *Cell* 2007, **131:**1354-1365.

3. Robison K, McGuire AM, Church GM: **A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome.** *J Mol Biol* 1998, **284:**241-254.

4. Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12.** *Nucleic Acids Res* 2004, **32:**D303-306.

5. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, et al: **Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes.** *Science* 2000, **290:**2105-2110.

6. De Smet R, Marchal K: **Advantages and limitations of current network inference methods.** *Nat Rev Microbiol* 2010, **8:**717-729.

7. 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.

8. Carrera J, Rodrigo G, Jaramillo A, Elena SF: **Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions.** *Genome Biol* 2009, **10:**R96.

9. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks.** *Nat Biotechnol* 2003, **21:**1337-1342.

10. Reiss DJ, Baliga NS, Bonneau R: **Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks.** *BMC Bioinformatics* 2006, **7:**280.

11. Lemmens K, De Bie T, Dhollander T, De Keersmaecker SC, Thijs IM, Schoofs G, De Weerdt A, De Moor B, Vanderleyden J, Collado-Vides J, et al: **DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli.** *Genome Biol* 2009, **10:**R27.

12. Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M: **Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks.** *Genome Res* 2012, **22:**1334-1349.

13. Middleton AM, Farcot E, Owen MR, Vernoux T: **Modeling regulatory networks to understand plant development: small is beautiful.** *Plant Cell* 2012, **24:**3876-3891.

14. Vernoux T, Brunoud G, Farcot E, Morin V, Van den Daele H, Legrand J, Oliva M, Das P, Larrieu A, Wells D, et al: **The auxin signalling network translates dynamic input into robust patterning at the shoot apex.** *Mol Syst Biol* 2011, **7:**508.

15. Pokhilko A, Fernandez AP, Edwards KD, Southern MM, Halliday KJ, Millar AJ: **The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops.** *Mol Syst Biol* 2012, **8:**574.

16. Salazar JD, Saithong T, Brown PE, Foreman J, Locke JC, Halliday KJ, Carre IA, Rand DA, Millar AJ: **Prediction of photoperiodic regulators from quantitative gene circuit models.** *Cell* 2009, **139:**1170-1179.

17. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER: **A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles.** *Plant Cell* 2004, **16:**2923-2939.

18. Pokhilko A, Hodge SK, Stratford K, Knox K, Edwards KD, Thomson AW, Mizuno T, Millar AJ: **Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model.** *Mol Syst Biol* 2010, **6:**416.

19. Pruneda-Paz JL, Kay SA: **An expanding universe of circadian networks in higher plants.** *Trends Plant Sci* 2010, **15:**259-265.

20. Coen ES, Meyerowitz EM: **The war of the whorls: genetic interactions controlling flower development.** *Nature* 1991, **353:**31-37.

21. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G: **Wisdom of crowds for robust gene network inference.** *Nat Methods* 2012, **9:**796-804.

22. Ruffel S, Krouk G, Coruzzi GM: **A systems view of responses to nutritional cues in Arabidopsis: toward a paradigm shift for predictive network modeling.** *Plant Physiol* 2010, **152:**445-452.

23. Mao L, Van Hemert JL, Dash S, Dickerson JA: **Arabidopsis gene co-expression network and its functional modules.** *BMC Bioinformatics* 2009, **10:**346.

24. Gutierrez RA, Lejay LV, Dean A, Chiaromonte F, Shasha DE, Coruzzi GM: **Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis.** *Genome Biol* 2007, **8:**R7.

25. Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, Thompson LP, Cabello JM, Davidson RS, Goldberg AP, Shasha DE, et al: **VirtualPlant: a software platform to support systems biology research.** *Plant Physiol* 2010, **152:**500-515.

26. Gutierrez RA, Stokes TL, Thum K, Xu X, Obertello M, Katari MS, Tanurdzic M, Dean A, Nero DC, McClung CR, Coruzzi GM: **Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1.** *Proc Natl Acad Sci U S A* 2008, **105:**4939-4944.

27. Baena-Gonzalez E, Rolland F, Thevelein JM, Sheen J: **A central integrator of transcription networks in plant stress and energy signalling.** *Nature* 2007, **448:**938-942.

28. Sulpice R, Pyl ET, Ishihara H, Trenkamp S, Steinfath M, Witucka-Wall H, Gibon Y, Usadel B, Poree F, Piques MC, et al: **Starch as a major integrator in the regulation of plant growth.** *Proc Natl Acad Sci U S A* 2009, **106:**10348-10353.

29. Carrera J, Rodrigo G, Jaramillo A: **Model-based redesign of global transcription regulation.** *Nucleic Acids Res* 2009, **37:**e38.

30. Ingkasuwan P, Netrphan S, Prasitwattanaseree S, Tanticharoen M, Bhumiratana S, Meechai A, Chaijaruwanich J, Takahashi H, Cheevadhanarak S: **Inferring transcriptional gene regulation network of starch metabolism in Arabidopsis thaliana leaves using graphical Gaussian model.** *BMC Syst Biol* 2012, **6:**100.

31. Ma S, Gong Q, Bohnert HJ: **An Arabidopsis gene network based on the graphical Gaussian model.** *Genome Res* 2007, **17:**1614-1625.

32. Breeze E, Harrison E, McHattie S, Hughes L, Hickman R, Hill C, Kiddle S, Kim YS, Penfold CA, Jenkins D, et al: **High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation.** *Plant Cell* 2011, **23:**873-894.

33. Krouk G, Mirowski P, LeCun Y, Shasha D, Coruzzi G: **Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate.** *Genome Biol* 2010.

34. Monke G, Seifert M, Keilwagen J, Mohr M, Grosse I, Hahnel U, Junker A, Weisshaar B, Conrad U, Baumlein H, Altschmied L: **Toward the identification and regulation of the Arabidopsis thaliana ABI3 regulon.** *Nucleic Acids Res* 2012, **40:**8240-8254.

35. Zheng Y, Ren N, Wang H, Stromberg AJ, Perry SE: **Global identification of targets of the Arabidopsis MADS domain protein AGAMOUS-Like15.** *Plant Cell* 2009, **21:**2563-2577.

36. Bargmann BO, Marshall-Colon A, Efroni I, Ruffel S, Birnbaum KD, Coruzzi GM, Krouk G: **TARGET: A Transient Transformation System for Genome-wide Transcription Factor Target Discovery.** *Mol Plant* 2013.

**FIGURE LEGENDS**

**Figure 1**

**Figure 2**

**Figure 3**