**EXERIMENTAL PLAN**

**Aim 1. Experimental Innovation: Generation of high through put, high confidence TF🡪target datasets.**

 1A. DEX expression data

 1B. TF-Dex-Chip

 1C. Time-series data for shoots

 1D.

**Aim 2. Computational Innovation: A Pipeline Machine Learning Approach to Finding Causal Links in regulatory networks using many kinds of genome-scale data**

**2. Computational novelty: Using a combination of different machine learning algorithms for analysis of steady state data, mutation data, and time-series data to generate predictive networks (Aim 2).**

**Aim 3. The Systems Biology Cycle: The feedback from analysis to subsequent experiment (Aim 3).**

 **Testing hypothese for TF interactions**

 **DEX-TF/TF**

 **DEX-TF1 in DEX-tf2 mutant background.**

**--------------------------------------------------------------------------------------------------------------------------**

**EXPERIMENTAL PLAN**

**Motivation & Significance**: Our ultimate goal is find a causal genetic network, effectively the circuit diagram underlying the regulation of genes in the N-assimilatory pathway. By analogy to an electrical network, such a gene regulatory network would enable us to infer the causal relationships between genes, thus suggesting targeted interventions on crops to reduce fertilizer usage with implications for human health, energy and environment. Importantly, this work will illustrate an experimental/informatics approach to the discovery of the causal network for any gene (or potentially any trait) of interest a wide range problems in biology and medicine.

**Aim 1. Experimental Innovation: Generation of high through-put, high-confidence TF🡪target data for hubs in N-assimilatory networks.**

***Rationale*:**  In this Aim, we will perturb the expression of TF hubs associated with N-assimilatory networks and examine alterations in genome-wide responses. This data will enable us to (i) identify TFs that control N-assimilatory genes in response to nitrogen sensing, (ii) identify biological processes and pathways that are coordinated with N-assimilation via these TFs, and (iii) produce high-confidence TF🡪target data to refine and train the machine learning approach to predict network states under untested conditions, the ultimate goal of Systems Biology.

**Approach**: Perturbations of TFs will be T-DNA mutants and/or overexpression of TFs in a transient expression system in protoplasts using dexamethasone (DEX) to induce nuclear localization of the TF. Nuclear localization will be performed in the presence or absence of cycloheximide. We will assay transcriptomic data (steady state and/or time series), as well as protein🡪DNA interaction data (ChIP-Seq) to enable us to identify direct and indirect targets of the TF. The data for each TF will be analyzed individually using gene networks to identify which biomodules are controlled by each TF. The data from all TFs will also fuel the computational pipeline developed in Aim 2 to find causal links in regulatory networks, and to generate predictive networks. We call our overall approach “Network Walking.”

**Identification and prioritization of TFs mediating inorganic and organic-N signaling.** To date, our regulatory network models have identified candidate TFs involved in regulating the N-assimilatory pathway genes in response to distinct N-signals [Krouk et al 2010][Gutierrez et al 2008]. Specific conditions for N-pre-treatments and treatments have uncovered nitrate regulation (growth in ammonium succinate, treatment with nitrate) or organic-N regulation (growth in 1mM nitrate, and treatment with 40mM ammonium/nitrate) of N-assimilatory genes [Wang 2004] [Gutierrez 2008]. The derived regulatory networks support a model in which inorganic-N (nitrate) signals induce TFs that activate genes involved in nitrate uptake, reduction, and assimilation into organic-N (Glu/Gln). These same genes are in turn feedback-repressed by the organic-N products of N-assimilation (Glu/Gln). (This may represent an “energy conservation” mechanism to conserve ATP & reducing equivalents, and carbon skeletons required for nitrate reduction and assimilation into organic-N, when levels of organic-N are abundant.) By contrast, organic-N signals (Glu/Gln) in turn activate TFs and target genes involved in converting “reactive” Gln into “inert” Asn, a more carbon-efficient N-transport amino acid used for N-transport/storage (e.g. to seeds) (see Fig 1).

To identify the TFs associated with this regulation, we used two types of network modeling approaches. TF hubs associated with nitrate-induction of N-assimilation were identified using time-series transcriptome data to drive a machine learning approach called State space modeling [Krouk 2010]. The derived network models – which were validated using out of sample data- were used to prioritize TFs induced “early” and at the top in the hierarchy for experimental testing (e.g. TFs HRS1, HHO1, HHO2, HHO3) (Table 1). To generate organic-N regulatory networks, we used steady state RNA data from plants treated with ammonium/nitrate, and used a Gln synthesis inhibitor (MSX), and Glu treatments [Gutierrez 2008]. TF hubs involved in organic-N signaling were identified using an Arabidopsis multinetwork in which TF🡪target edges were predicted based on correlation(>0.8) and representation of cis-elements [Gutierrez 2008]. Top scoring TF hubs (e.g. CCA1, GLK1, GLK2, WRKY1, bZip1) were identified, and selected for validation testing (e.g. CCA1) [Gutierrez 2008]. Surprisingly, the top-ranking TF hubs associated with nitrate-regulation (HRS1, HHO1, 2, 3) or organic-N regulation (GLK1 & GLK2) of the N-assimilatory pathway genes define 2 subclades of a single myb gene family, for which one member (HRS1) was identified to be involved in phosphate signaling [Liu 2009]. These results suggest that members of this TF family may be involved in coordinating plant responses to nutrient sensing of Nitrogen and Phosphate nutrient responses in plants. This hypothesis is supported by our preliminary results of HRS1 expression in transient assays, described in Aim 1B. The finding that network targets of a TF (HRS1) identified in our Dex-inducible transient assay system in protoplasts (e.g. activation of phosphate transport) recapitulate results uncovered in vivo (altered phosphate signaling in a transgenic 35S::HRS1 plant) [liu et al 2009] attests to the biological relevance of the rapid protoplast system. Table X lists the prioritized TF hubs for testing in Aim 1 for their role in regulating N-assimilation in response to inorganic (e.g. nitrate) or organic-N (Glu/Gln) sensing. In each of the subaims, we provide preliminary data using one of these TFs as a proof-of-principle example for the approach.

**Aim 1A. Generation of time-series transcriptome datasets for organic-N signaling.** Because causality moves forward in time, time-series experiments are a particularly promising source of structure for predictive network modeling. Our state-space modeling of time-series data from nitrate-treated plants, enabled us to successfully predict the direction of change in expresion levels under untested conditions with 76% accuracy (e.g. whether a gene in the network would go up or down) [Krouk 2010]. By contrast, our current network models for organic-N signaling are built using steady state transcriptome data [Gutierrez 2008], and are thus not able to predict network dynamics or responses under untested conditions, as we can for our nitrate-regulated networks. We will therefore generate fine-scale, kinetic transcriptome data (0,3,6,9,12,15,20,25,30 min) using conditions shown to elicit an organic-N response: growth on 1mM nitrate, treatment with 40mM ammonium/nitrate vs control KCl. This transcriptome data will be used to drive state-space modeling of regulatory networks, as described in [Krouk 2010] and will be combined with analysis of steady state and mutant data, as described in Aim 2 to impart dynamics and predictive power to our organic-N regulatory network models.

**Aim 1B. Network Walking: Generation of high through put, high confidence TF🡪target data for TF hubs.** To rapidly and reliably validate predicted network targets genome-wide for the TFs listed in Table 1, we will use the DEX-transient expression system (as also described in Progress). Following transient transfection of protoplasts with 35S::TF-GR constructs, DEX treatment will induce nuclear localization of the TF. Controls are -DEX. Prior to DEX treatment, treating with cycloheximide (CHX) will identify direct targets. Thus, +DEX+CHX will identify direct targets, while +DEX-CHX will identify direct and indirect targets. (DENNIS- WE WILL NEED AN ANALYSIS OF +/- CHX HERE TO CONVINCE REVIEWERS OF THIS [I sent Kranthi a way to do this]). Prediction of direct targets based on transcriptomic data (+/-DEX, +/-CHX) will also be confirmed using ChIP Seq (see Aim 1C). Prior to DEX-induction of TF nuclear localization, we will pre-treat the protoplasts with N-treatments for nitrate signaling (1mM nitrate) or organic-N signaling (40M ammonium/nitrate). This will test whether “pre-conditioning” plants with N-treatments affects the TF🡪target activation. If so, this could be due to induction of interacting TFs in response to N-treatment, and/or post-translational modification of the TF itself, as shown in the example below.TF interactions will also be explored in Aim 3.

As proof-of-principle for this approach, we tested HRS1 (At1g1330) a top scoring TF hub of our “learned” nitrate-regulatory network, which was previously shown to be involved in phosphate signaling [Liu 2009, Journal of Integrative Plant Biology]. We hypothesize that HRS1 may also be involved in nitrogen sensing, as TF HRS1 is induced by NO3- within 9 min of treatment and in the inferred network it is predicted to be the top most controlling and top most controlled gene [Krouk et al 2010]. [Gloria: please cut out the rest of this paragraph. Too much information.] To investigate the role of HRS1 in the modification of NO3- signal propagation throughout the learned N-regulatory gene networks, we validated its direct and indirect targets using the transient DEX system described more fully in the Progress Report. Briefly, using the transient expression system, a 35S::HRS1-GR construct is introduced into protoplasts, and the HRS1-GR fusion protein is transported into the nucleus only in the presence of Dexamethasone (DEX). We can therefore identify the genes activated by HRS1 following DEX treatment using ATH1 chip analysis. Our initial analysis of the genes associated with HRS1 activation by DEX, revealed four distinct clusters of genes that are influenced by a combination of HRS1 induction (by DEX treatment) and nitrate-status (Fig. X). For example, Cluster 4 genes are direct targets of HRS1 only in the *presence* of NO3-, are significantly overrepresented in the categories of phosphate transport (p-val 8.14 E-6) and cellular transport. Cluster 2 is comprised of genes that are induced by HRS1, but only in the *absence* of NO3-, and fall into functional categories such as RNA processing and cycling. Cluster 3 contains genes that induced only in the absence of HRS1 (e.g. HRS1 is a repressor) and NO3- and are involved in embryonal and plant development. Finally, Cluster 1 contains genes that are only induced by NO3-, not by HRS1, and is dominated by genes that fall into stress response functional categories. These distinct gene clusters reveal that the role of the TF hub HRS1 is very different depending on the N-status of the plant. Comparison of clusters 1 and 4 suggest that there are two different sets of N-responsive genes are either induced (Cluster 4) or repressed (Cluster 1) by HRS1, resulting in the activation of two different signaling pathways (phosphate transport and stress response, respectively). We have yet to complete the genome-wide analysis to reveal indirect targets of HRS1 (e.g. –CHX) in the presence and absence of NO3-, but anticipate changes in expression of N-sentinel genes as predicted in our networks. This approach (+/- DEX, +/-CHX, +/-N) will be used for all the prioritized TFs listed in Table 1, which also lists our progress for each of the TFs. Conditions for +/-N will be tailored to either nitrate responses (growth in ammonium succinate, treatment with nitrate) or organic-N responses (growth in 1mM nitrate, and treatment with 40mM ammonium/nitrate), as indicated in Table 1. Cluster analysis of transcriptomic data followed by GO-term analysis using the Biomaps function in VirtualPlant [Katari 2010], will identify genome-wide responses, and biological processes controlled by specific TFs. The normalized transcriptome data will also be fed into the pipeline of machine learning algoritms developed in Aim 2, to fuel the refinement of a causal network. [Gloria: cut to here. This paragraph needs only one thing: a description of what happened when we did the preliminary results on this gene +DEX and +/-CHX]

**Aim 1C. DEX- chip Seq (bZip1):**

**NEED SECTION HERE WITH PRELIMINARY RESULTS FROM bZIP1.**

**This will generate TF🡪target pairs based on protein-DNA interaction from chip seq.**

**Dennis- HOW WILL THIS DATA FEED INTO THE PIPELINE? [I think I describe this already: it causes a bias in the formation of the network]**

**Aim 1D. Perturbations of TFS in whole plants: T-DNA mutants and 35S::Over-expressors.**

The transient expression assays in Aims 1B & 1C, will enable us to overcome problems associated with functional redundancy in T-DNA mutants in TFs [Chen HW, Bandyopadhyay S, Shasha DE, Birnbaum KD, “[Predicting genome-wide redundancy using machine learning”.](http://www.ncbi.nlm.nih.gov/pubmed/21087504). BMC Evol Biol. 2010 Nov 18;10:357] [Cutler S and McCourt P (2005) Dude, Where’s my phenotype? Dealing with redundancy in Signaling Networks. Plant Physiology (2005) v. 138, pp558-9]. For example, because of functional redundancy of TFs in signaling pathways, a negative result with a T-DNA mutant in the TFs listed in Table I have been inconclusive. We failed to see a molecular phenotype for predicted network targets in the cca1 or glk1 single mutants, but observed molecular phenotypes in 35S::CCA1 [Gutierrez et al 2008] and one in a double mutant [glk1/glk2] (ADD FIGURE)? [I think we get the point. No need for more examples] Likewise, because of TF redundancy single mutants in bZip1 do not show a molecular phenotype for ASN1 expression [Obertello 200X][Baena Gonzalez], while transient expression confirms the edge [Baena Gonzales 200x]. Similarly, the single mutants of nitrate responsive myb family members HRS1, HHO1 and HHO2 also do not show molecular phenotypes, and we are currently testing the double mutants in collaboration with Dr. Krouk (INRA). Preliminary results show a mutant phenotype for hrs1/hho1 double mutants that is conditional on nitrate and phosphate conditions, as predicted by our DEX-perturbation studies of HRS1 (Aim 1B) and by HRS1 overexpression [Liu et al 2009].

[I don’t think you need this and it may be a problem for patentability] In rare cases, we have observed a molecular phenotype in the single T-DNA mutant of a TF hub. One such example is the master regulatory gene in the network TF WRKY1 (At2g04880), which is predicted to be a major hub in the N assimilatory network. The regulatory edge predictions are based on network analysis based on correlation (.0.8) and the presence of at least one CRE for the TF. This analysis predicts that WRKY1 influences the expression of 8 out of 15 sentinel genes from the N-assimilation network. In this network model, WRKY1 is predicted to act as a “toggle switch” to induce expression of genes involved in nitrate reduction and assimilation into Gln, while simultaneously repressing expression of genes involved in converting Gln to Asn (Fig. X). Therefore, WRKY1 is an excellent candidate for TF perturbation and genome-wide analysis. Our network predictions have been experimentally validated using both T-DNA plants and also the DEX inducible transient assay expression system in protoplasts (see Aim 1C). qRT-PCR on tissue from three independent WRKY1 T-DNA insertion mutants (SALK\_016954; SALK\_136009; SALK\_070989) (3 biological replications each) revealed decreased expression of NIA2 and NRT2.1 (targets of WRKY1 activation) and increased expression bZIP1🡪ASN1 (both targets of WRKY1 repression), as predicted by the network model (Fig. X), while the opposite expression patterns are found in the DEX system where WRKY1 is over-expressed (Aim 1C). These results support a model whereby WRKY1 acts as a toggle switch between nitrate assimilation into Gln, vs Gln conversion into Asn (for N-storage) depending organic-N status of the plant (Fig X). Preliminary studies with the DEX inducible system in combination with cyclohexamide treatment (e.g. to reveal direct targets) and qRT-PCR have revealed that NIA2, NRT2.1 and bZIP1 are direct targets of WRKY1. We are in the process of performing transient DEX experiments in protoplasts to determine indirect targets of WRKY1. This will be followed by genome-wide analysis and Network Walking to infer relationships between the direct and indirect targets of WRKY1 (as described in Aim 1B). These results will provide further insight into WRKY1 as a potential master regulatory hub of the N assimilatory network. For the TFs listed in Table I, we will perform transcriptomic analysis on T-DNAs only if Q-PCR shows mis-regulation of a target gene in the N-assimilation network. For TFs confirmed in the DEX- transient system, we will create transgenic plants with the Dex-inducible promoter. Stable transgenic constructs for TF candidates validated by transient assays in Aim 1B, will enable us to validate network targets in planta, and in distinct organs, or across distinct developmental time points.

**Aim 2. Computational Innovation: A Pipeline Machine Learning Approach to Finding Causal Links in regulatory networks using many kinds of genome-scale data**

***Rationale***: Our ultimate goal is to employ machine learning approaches to model a causal genetic network, effectively the circuit diagram underlying the regulation of genes in the N-assimilatory pathway. To date, we have used a machine learning approach to generate a predictive network model for nitrate control of N-assimilatory pathway based on time-series data from wild-type plants [Krouk 2010]. In this aim, we develop methods and approaches to refine and improve the predictive power of our networks, by feeding our network learning algorithms experimental data derived from plants (or cells) in which we have perturbed expression of the TF and monitored effects on target genes. The transcriptome data generated from these experiments will support either direct or indirect TF🡪target relationships, which is in some cases will also have experimental support at the level of TF🡪target binding.

To improve the predictive power of our networks, we will create a predictive network model that makes use of the three major kinds of transcriptomic data that are available to us, as generated in Aim 1: steady state data (e.g. N-treatments) and perturbation data (e.g. T-DNA or DEX data), as well as time-series data (e.g. expression over time). We will also have validation of TF🡪target relationships from ChIP seq. (Dennis- ChIP-seq is a fourth type of data….does this get put into the category perturbation data? Chip-seq shows what is bound to what, right? Under that assumption I guess this would also bias the determination of which edges are real. ) No single algorithm is best for machine learning using all three kinds of data, so we will use several algorithms together in a pipeline described below.

**Approach**: The causal network will consist of nodes that are genes and edges between genes that are labeled with coefficients. A positive coefficient corresponds to an inductive edge. A negative coefficient corresponds to a repressive edge. For a given target gene Z, these coefficients will be reflected in the form of an equation Z = c1\*A + c2\*B + c3\*C …. , (Dennis is the following correct?) where A, B, and C are normalized expression levels of transcription factor genes. They are not adjusted.????? So, if c1 is positive, then gene Z will tend to rise (increase in expression, because Z would be positive) as gene A rises (i.e., A to Z would be an inductive edge). If c2 were negative, then an increase in expression of B would cause a decrease in Z, all else being equal. Thus, B to Z would be a repressive edge.

Typically, machine learning algorithms to infer such equations will include a regularization factor that will limit the number of additive products on the right hand side to three or four, but this still means that even for 100 genes, our algorithms would have to find the proper 400 edges out of a possible 5,000 and assign coefficients to those edges. Regularization is a form of parsimony: we want to find the simplest model that explains the behavior. Simpler models tend to be more robust to noise because they avoid overfitting. In fact, as part of our quest for simplicity, we prefer “linear” models to quadratic ones (in which there would be terms of the form d\*B\*C, where B and C would be gene expressions). Surprisingly, the resulting models explain behavior well. A model explains behavior if it can predict the state of one gene given the state of other genes at the same or previous time points. (Dennis- I think we need to explain the rationale behind why mRNA can be used as a proxy for protein levels, that is our success in the State-space models. I don’t understand: this is a long-standing assumption/issue. mRNA is never a perfect proxy, but the fact that we can do anything with mRNA shows it can’t be completely wrong. Gab explained to me that your algorithm actually introduces “noise” to account for the conversion of TF mRNA to protein, and without that step, we would not have been so successful. Is this correct? No. It simply models the noise. I don’t think we need that level of detail here. If so, I think we should state it here explicitly---we cant expect the readers to remember this from our paper….or even to have read it) For example, we used this approach to successfully predict regulatory edges in a network generated from time-series data in nitrate-treated roots [Krouk et al 2010].

**Pipeline/Consensus of Machine Learning Algorithms**

One reasonable approach is to establish a pipeline (figure dfg.eps) that analyzes different kinds of data using different algorithms. In that figure, the MCZ algorithm (short for the Median Corrected Z-score method [Greenfield, A., Madar, A., Ostrer, H., Bonneau, R.: DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PloS one (2010]) is used for both the steady state and mutation data, followed by the DFG algorithm (short for Dynamic Factor Graph [Krouk,G.,Mirowski,P.,LeCun,Y.,Shasha,D.E.,Coruzzi,G.M.:Predictivenetworkmodeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biology 11(12), R123 (2010]) used for time-series data. The MCZ? algorithm assigns weights to certain edges that are then processed DFG algorithm. (DENNIS- I THINK YOU NEED A SHORT DESCRIPTION ABOUT HOW MCZ WORKS AND WHY IT IS GOOD…ALSO FOR DFG. WE CANT EXPECT THEM TO REMEMBER WHY DFG IS GOOD FROM OUR PAPER. I can do both, but let me wait until we figure out how much space we have. I don’t think we need much for our purposes here, viz. no figures.)

In preliminary work, done for a book we have written called “Network Inference in Molecular Biology”, [Jesse Lingeman and Dennis Shasha 2012 *Network Inference in Molecular Biology -- a hand-on framework*, Springer 2012], we show with data from the DREAM benchmark (Dialogue for Reverse Engineering Assessments and Methods) [Schaffter, T., Marbach, D., Floreano, D.: GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 27(16), 2263–2270 (2011)] that using the MCZ🡪DFG pipeline is better than using any single machine learning algorithm.

Alternatively, we can posit a collection of algorithms that each predicts edges and then use a “voting” approach to determine the best one. In that scheme, illustrated in the figure consensus.eps, in addition to the Median-Corrected Z-score and Dynamic Factor Graph algorithms, we use the NIR (Network Identification by Multiple Regression [Gardner,T.S.,DiBernardo,D.,Lorenz,D.,Collins,J.J.:Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Science 301(5629), 102–105

(2003)]) and CLR (Context Likelihood of Relatedness [Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S.: Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology 5(1), e8 (2007)]) on steady state data. We also use BANJO (Bayesian Inference with Java Objects [Yu,J.:AdvancestoBayesiannetworkinferenceforgeneratingcausalnetworksfromobservational biological data. Bioinformatics 20(18), 3594–3603 (2004)]), Time-Delay ARACNE [Zoppoli, P., Morganella, S., Ceccarelli, M.: TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinfor- matics (2010)], and Inferelator [Greenfield, A., Madar, A., Ostrer, H., Bonneau, R.: DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PloS one (2010)]. Because each machine learning algorithm ranks the TF🡪Target edges, we “weight” each vote for a TF🡪target edge depending on its rank in each of these programs. Highly ranked edges acquire a weight close to 1, and lowly-ranked edges acquire weights near 0, where the weight drops off exponentially. Once the edges are established- (BASED ON THE STEADY STATE DATA and mutation data?), one of the time-series based algorithms (e.g. Inferelator, Dynamic factor graphs, or Time Delay ARACNE, ?????) assigns weights to the edges of the reduced graph based on its dynamic view of the network.

**Preliminary Results**

To test the value of pipelines vs. weighted consensus views for predicting edges in a network, we compare the results of our previous Dynamic Factor Graph (DFG) Approach [Krouk,G.,Mirowski,P.,LeCun,Y.,Shasha,D.E.,Coruzzi,G.M.:Predictive networkmodeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biology 11(12), R123 (2010] built on time-series data, with a pipeline or consensus “weighted” approach to network inference. Our criterion for quality is how well the resulting network predicts out-of-sample data. In our previous paper [Krouk 2010], we showed that using the training data consisting of time-series data from the time of a perturbation, 3 minutes later, 6 minutes later, 9, 12, and 15, we were able to use a Dynamic Factor Graph Approach to predict the direction of expression change of 76 genes associated with the N-assimilation network between 15 and 20 minutes 74% of the time correctly. By contrast, a naïve trend forecasting method, which predicted the direction of expression change of genes between 15 and 20 minutes as being the same as between 12 and 15 minutes, was correct only 52% of the time, having a prediction accuracy marginally better than chance.

Using the new data from perturbation of TF function (using Dex) and the steady state data along with the time series data from wild-type plants, we can predict z% of the gene directions correctly. Using only the steady state data along with time series data, we can predict z’% of the gene directions correctly. Using the consensus approach, we obtain z’’% of the gene expression directions correctly.

**Intended Approach**

Based on our preliminary results, we will iterate the following steps: use a consensus approach on nitrogen treatment steady state data, nitrogen treatment time-series data, and TF perturbation experiments (e.g. DEX experiments or T-DNA mutants). This will result in a regulatory network of TFs targeting genes involved in N-assimilation.

(Dennis- We need to explain how we will define the subnetwork of genes involved in N-assimilation and the TFs that regulate them….OR are you intending to make a regulatory network using ALL TFs and ALL genes in the genome. We need to be clearer here. We do have a list of 57 target genes in the N-assimilatory pathway, including nitrate transporters…if that helps to anchor the network building. It will be a subnetwork consisting of the genes that could be relevant to nitrogen.).

That in turn will suggest new genes on which to try DEX experiments (e.g. putative hubs). That will suggest a new network and so on.

As usual, our criterion of goodness will be the ability to predict well on out-of-sample data, both missing time points and missing mutations. Ultimately, we would like to learn the model well enough that we choose the next DEX experiment to try based on which will be the hardest to predict.

**AIM 3.** **The feedback from analysis to subsequent experimentation: Testing new candidate TFs and TF interactions.**

Aim 3A. New candidate TFS:

Aim 3B. Interactions – Dex in mutant background

Aim 3C. Double expression vector

In order to co-express two TFs (TF1 and TF2) in the same cell, we resorted to use the MultiSite Gateway**®** Technology that allows to simultaneously clone multiple DNA fragments in a define order and orientation. To generate the expression construct, TF1 cDNA, TF2 cDNA and the “GR cassette”(that includes a terminator, 35S promoter and the GR sequence) were cloned in appropriate donor vectors and allowed to recombine with each other and with the destination vector pBob11. This results in the assembly of 2 consecutive expression cassettes in which each TF is fused to the GR sequence and the expression of each fusion is driven by a constitutive 35S promoter (FIGURE). Expressing both genes from the same vector has a few advantages: 1. it eliminates the problem of dealing with multiple vectors that could have different transfection efficiencies due to their size i.e. vectors containing small genes will be present in higher copy number than vectors containing long genes hence small genes might expressed at higher levels; 2. when multiple vectors are transfected, each cell is likely to contain a random number of copies of each vector while the population of transfected cells will be homogeneous in its transgenic content; 3. the expression vector harbors monomeric red fluorescent protein (mRFP) for positive fluorescence selection, therefore enabling the isolation of transfected protoplasts and avoiding confounding of the results by non-transformed cells. In addition, it will bypass the sorting of false positives due to fluorescence bleaching when multiple vectors with different fluorescence selections are used.