**Research Strategy**

**1. Significance**

In eukaryotes, gene regulation involves a complex interplay between transcription factors (TFs) and epigenetic regulators, which set the chromatin stage for TFs to activate or repress target genes (Fig. 1) [1]. This multilayer control allows living organisms to respond to environmental and developmental cues through coordinated expression of thousands of genes which can be abstracted as gene networks [2]. Gene network dysfunction has been implicated in many disease states [3, 4]. Prominent examples include cancer [5–7], inflammation [8] and T cell differentiation [9, 10]. Gene network inference is the task of identifying causal influences between regulator genes and regulated genes in a network to best explain the observed gene expression patterns [3]. Current popular gene network inference approaches usually use TFs as regulators [10–13], but don’t include the important layer of epigenetic regulation (Fig. 1), despite evidence that dysfunction in epigenetic control could lead to severe developmental defects and diseases [14, 15].

The exponential accumulation of epigenomics data in recent years (e.g. in ENCODE [16]) has led to recent efforts to incorporate epigenomic data into gene network studies [17–19]. These algorithms use histone profiles to improve network analysis. However, the molecular regulators of histone profiles - the epigenetic regulators - and their causal influence on the genomic targets have been so far largely overlooked. This ***EpiNet*** proposal will complement the current effort of bridging epigenomics data with transcriptome inference, by specifically testing the hypothesis that **incorporating epigenetic regulators as predictors will expand and refine predictive network models.** I will start with Arabidopsis SET Domain Group 8 (SDG8), a H3K36 methyltransferase homologous to human SETD2 (Huntingtin-interacting protein), to test this hypothesis. Arabidopsis is a good system for this purpose, because i) Arabidopsis SDG8 has specific target loci, unlike its yeast homology SET2 that targets the whole genome and ii) Unlike its human homolog, the Arabidopsis *sdg8* mutant is viable, which allows *in vivo* target identification. My preliminary genome-wide study has identified functional targets and bound targets of SDG8, which allows using this target loci information as prior knowledge to inform network inference. Starting with SDG8, I will also determine the functional targets of other epigenetic regulators, to complement the effort of identifying genomic targets for chromatin regulators in higher eukaryotes [20] and to facilitate network inference.

This proposal will link epigenetic modifiers with predictive network modeling, which will be informative for both fields. For the field of predictive network modeling, this proposal will offer a tool to incorporate a yet missing layer of gene regulation to allow more accurate modeling of biological systems. For the field of epigenetics, this proposal will provide a method to view epigenetic modifiers in the context of gene regulatory networks. The concept and methodology of *EpiNet* can be applied to study crucial biological questions across plant and animal phyla.

**2. Innovation.**

As mentioned above, some recent efforts have been made to integrate epigenomics data into network construction. Web-based informatics tools have been developed to construct gene networks based on shared histone marks among co-regulated genes [18], or to infer multi-level networks by incorporating histone profiles, TF-binding and gene expression data [17]. Histone profiles have also been used to derive prior knowledge to inform and improve the reconstruction of gene networks in yeast [21]. These studies demonstrate the power of incorporating epigenomics data into gene network analysis but overlook the causal influence of epigenetic regulators on the expression level of their genomic targets. My proposal is that epigenetic regulators should be allowed the same “predictor” roles as TF regulators in gene network modeling. This hypothesis is based on the knowledge that epigenetic regulators affect the gene expression level of their genomic targets [[22, 23] and my preliminary study]. By including epigenetic regulators in gene network models, I expect the resultant gene network will be better at predicting the “true” biological states, and this hypothesis is borne out by my preliminary results discussed below. When required, the current network inference tools will also be modified to treat epigenetic regulators as a special class of predictors, *e.g.* as an interactor with a TF regulator. Thus, the incorporation of epigenetic modifiers as potential regulators in gene network models, and their interactions with TFs, is an unexplored area of predictive network modeling that this proposal will test.

**3. Approaches**

**Overview:** This ***EpiNet*** proposal will specifically test the hypothesis that **incorporating epigenetic regulators as predictors will expand and improve predictive network modeling.** To do this, I will use a combination of experimental and computational approaches in three specific aims: In **Aim 1**, I will generate a fine scale time-course transcriptome data with a specific focus on one master epigenetic regulator, SDG8, to provide perturbation data for predictive network modeling in Aim 2. Next, I will test the performance of predictive network modeling with and without epigenetic regulators as predictors (**Aim 2A**), and with and without prior knowledge of genomic targets of SDG8 (**Aim 2B**). My preliminary data supports the idea that epigenetic regulators improve the accuracy of predictive network modeling. The gene network generated in Aim 2 will be validated, expanded and refined by experimentally testing genomic targets of additional epigenetic regulators implicated in the network in **Aim 3** (Fig. 2). Aim 1 and Aim 2A will be performed during my mentored phase, while Aim 2B and Aim 3 will be performed during my independent phase.

**Background**:

**Network inference:** One goal of systems biology is to be able to infer Gene Regulatory Networks [spell out all acronyms except DNA and RNA] with sufficient accuracy to predict *in silico* their behavior in response to future perturbations and untested conditions. Current efforts in this direction involve mining experimental perturbation data, *e.g.* time course transcriptome data, to infer causative relationships between regulatory genes and regulated targets [2] [10–13]. The basic logic behind network inference is quite simple: If expression level of gene A is perturbed, and subsequently the expression level of gene B is observed to change, then gene A may be causally upstream of gene B. Dynamic Bayesian Networks have become a popular choice for this task because they have the power to deal with experimental noise, missing data and hidden variables [24]. State-space modeling (SSM) is a subclass of dynamic Bayesian network learning method, and a modern machine learning technique, which has been applied to reverse engineer GRNs from time-series expression data [25, 26]. In this proposal, I will primarily use Dynamic Factor Graph approach (DFG) [12], a form of SSM, to infer gene networks.

**Epigenetic Regulators:** In eukaryotes, gene regulation involves a complex interplay between transcription factors and epigenetic regulators, which pack genomic DNA and poise genes for activation or repression (Fig. 1). Ever since the crystal structure of nucleosomes was published in 1997 [27], researchers have been hypothesizing that the N-terminal histone tail could be modified to regulate chromatin configuration. Experimental evidence accumulated in the last 20 years indeed proves that histone modification, a type of epigenetic control, plays an important role in modulating gene expression in the context of development and response to external stimulus [15, 28–32]. Histone modification maintains differentiated cell status, records memory of stress [31] and transmits traits across generations [33]. Histone modification is maintained by specific protein families. For example, histone methylation is mediated by histone methyltransferases (HMT) - SET domain containing proteins [34], and histone acetylation is maintained by histone acetyltransferases (HAC) and histone deacetylases (HDAC). In this proposal, I will study the role of epigenetic regulators – specifically histone modifiers – in gene regulatory networks.

**HMT SDG8:** My studies will initially focus on SDG8 (AT1G77300), a H3K36 methyltransferase. SDG8 is ideally suited for the proposed study because i) it mediates genome-wide responses to environmental signals and ii) it displays target specificity. ***A) SDG8 as master regulator of signal responsive GRNs***: mutations in *sdg8* have been implicated in a wide range of developmental and environmental responses from flowering to nutrient sensing [35–46]. For example, 90% of the genetic response to the phytohormone Brassinosteroid [44] is disrupted in an *sdg8* mutant. This allows ease in generating perturbation data to probe SDG8 function in GRN. ***B) Genomic targets of SDG8:*** SDG8 is well conserved among yeast (SET2)[47], plants, and humans (SETD2, Huntingtin-interacting protein). Yeast SET2 is the only H3K36 methyltransferase in the genome, by contrast, Arabidopsis *sdg8* mutant has drasctically lowered but residual levels of H3K36 di or tri-methylation. This indicates a scenario where there is more than one functioning H3K36 methyltransferase in the genome, and some level of target specificity for the various H3K36 HMTs. Using Arabidopsis as the model system, I identified genomic targets of SDG8 *in vivo.* This assay is possible in Arabidopsis as the mutant in *sdg8* is developmentally impaired but viable and fertile, unlike its mammalian homologs. My own studies on the role of SDG8 in sucrose and light signaling [Li et al., unpublished] shows that SDG8 targets ~700 genomic targets with specific biological functions (e.g. metabolism & photosynthesis) for H3K36 methylation to elevate their mRNA levels [Fig 3]. This latter finding suggests that the validated genomic targets of SDG8 can be used as priors to inform predictive gene network modeling. While the rich dataset I have accumulated on SDG8 will be used to seed the proposed study in Aim 1, I will expand my studies to include other epigenetic regulators in Aim 2 and 3. The approach and preliminary data for each specific aim are detailed below.

**Specific Aim 1: Perturbation data: Generate fine-scale time-series transcriptome data to learn regulatory networks (mentored phase).**

**Rationale:** Because*causality moves forward in time*, fine-scale time-course gene expression data provides an extremely valuable dataset to support the inference of causal relationships between regulators and regulated genes [12]. Thegoal of this aim is to generate a perturbation time-course transcriptome where the data-rich HMT SDG8 likely serves as a master regulator. SDG8 expression is induced by sucrose supply by 2-10 fold in dark-grown Arabidopsis seedlings [NASCARRAY database: NRID5273, 5275, 5277; NRID5279, 5281, 5283] and [48–50]. This observation agrees with the notion that SDG8 mediates a sucrose response [51] [Li et al unpublished]. Therefore, I will generate a fine-scale time-course transcriptome profile in response to sucrose nutrient sensing.

**Experimental Approaches:** Wild-type Arabidopsis plants will be grown in a hydroponic system, which facilitates rapid and instant treatment with sucrose. 7-day-old dark-grown Arabidopsis seedlings will be treated with 10mM sucrose and compared to controls [48, 52]. Both sucrose treated plants and control plants will be harvested, in triplicates, at 10 time points for RNA-seq (Fig. 4). To capture the best dynamic pattern of SDG8 mRNA level, the time points will be determined based on a pilot experiment to measure the temporal induction of SDG8 mRNA by qRT-PCR. My preliminary data suggests that a 2hr treatment of sucrose induces 1.5 fold increase in mRNA level of SDG8, therefore, the time points for qRT-PCR test will be 5min, 20min, 60min, 90min, 2hr, 4hr, 8hr and 12hr. [Hmm. Do we want to look at some targets of SDG8 too to help determine these points] This pilot study will help determine the range and intervals of time points for RNA-seq for best time resolution. I aim to perform reliable quantitative comparison of gene expression across time points in an efficient and economical way. To satisfy these demands, I will use a block design for multiplexed RNA-Seq libraries as shown in Fig. 4 [53]. Total RNA will be extracted and barcoded RNA-seq libraries will be prepared according to [54]. The RNA-seq libraries will be sequenced on an Illumina HiSeq platform at the NYU genomics core. **Bioinformatics analysis:** The Illumina HiSeq raw reads will be aligned to Arabidopsis TAIR 10 genome and gene counts will be generated using the TopHat-Cufflinks suite [55, 56]. The gene counts will be normalized by Quantile normalization, to render the expression measures across time points comparable [57]. Finally ANOVA, Linear model [58] and other appropriate statistical tests will be conducted in R to determine the genes responding to sugar treatments over time. These sugar-responsive genes form the universe for predictive network modeling in Aim 2.

**Innovation:** The transcriptome response of Arabidopsis to sucrose-treatments has been measured [50, 51, 59–61]. However, a detailed time-course transcriptome describing the fine-scale kinetics of gene response to sucrose – which is required to derive predictive network models is still lacking. [Doesn’t seem that fine as a time scale, but ok] The proposed research will fill this information gap, providing a temporal insight of the genetic response to a supply of energy molecules, and to fuel GRN analysis as proposed in Aim 2.

**Outcomes and alternative approaches:** This aim will generate a fine-scale time-course transcriptome profile in response to sucrose by RNA-seq, to provide gene expression datasets for predictive network modeling in Aim 2. Our lab has successfully profiled fine-scale time-course transcriptome using the Arabidopsis system [12](Coruzzi, unpublished). I also have extensive experience in preparing RNA-seq libraries and RNA-seq data analysis, so I do not expect major difficulties in performing the proposed experiments. A challenge could be to figure out the optimal experimental conditions to capture the dynamic range of SDG8 and target genes. As described above, pilot experiments by qRT-PCR will help determine the optimal time points, developmental stages and growth conditions.

**Specific Aim 2: Assessing the influence of epigenetic regulators in predictive network modeling.   
Rationale:** In this aim, I will infer gene networks from the fine-scale time-series data generated in Aim 1, using a machine learning approach – Dynamic Factor Graph (DFG), a form of state space modeling (SSM). Previously, SSM-based studies took transcription factors as predictors for gene expression change, while ignoring the influence of epigenetic regulators (Karlebach and Shamir, 2008; Krouk et al., 2010). In this aim, I will construct GRNs with epigenetic modifiers as well as transcription factors as potential regulating gene to predict GRN behavior in untested conditions. This will be performed in two sub-aims as described below.

**A. Incorporating epigenetic modifiers as putative regulators in network models (mentored phase).** In this sub-aim, I will use DFG to reconstruct a gene network from the time series transcriptome data generated in Aim 1, either with or without epigenetic regulators as predictors, to assess the influence of epigenetic regulators on the nutrient responsive gene network.

**Modeling approaches:** I will be trained in using the DFG approach to infer GRNs from time course expression data. DFG is a version of state space modeling (SSM) that synthesizes dynamic Bayesian network and Markovian approaches [12]. In a Markov chain, each gene’s activity at time *t+1 is* dependent *only* on the state of regulatory genes (“predictors”) at previous time point *t*, regardless of the trajectory that led to the state at time *t* [2]. Mathematically, this can be described by ordinary differential equations, as below:

where the mRNA concentration for a given gene *i* is ( being a kinetic constant which allows modeling of mRNA degradation), and the rate of change of mRNA concentration of gene *i* is determined by function describing the weighted influence of the several predictors (e.g. transcription factors) represented by vector . Hmm. I don’t love this formulation. Yi(t) should have a coefficient that models mRNA degradation and maybe it should be something yi(t+1) – yi(t) = gi(y(t)) – tau \* yi(t)

However, measurement of large-scale biological data is noisy by nature. To address this problem, DFG assumes that the measured gene expression level is generated from an underlying, hidden state of “true” mRNA expression as with a Gaussian noise term (Fig. 5A). DFG infers function *f,* which describes how true expression levels of predictor genes at time t influence the change in true gene expression level of gene *i (* ) from time t to time t+1 (see the equation in Fig. 5A). I will train the model with the first 9 time points to solve the best function *f* to explain the expression pattern of each regulated gene *i* (Fig. 5B). To validate the “trained” model, I will use the learnt function *f* to predict the change of expression level of gene *i* in the last time point. I will then compare it to the experimentally observed value to test the predictive power of the learnt model (Fig. 5C). Finally, if *N* predictor genes influence the gene expression level of gene *i*, an inferred causal regulatory edge can be drawn between these *N* predictor genes and gene *i*. The union of such edges results in an overall GRN. [This sentence should go earlier when justifying this approach]Compared with other network inference approaches [11, 62, 63], DFG had a slight improvement in accuracy, and had a better signal-to-noise ratio using the same data [12].

To test the hypothesis that adding epigenetic regulators (such as SDG8) will improve predictive network modeling, I will first identify epigenetic regulators among the sugar responsive genes from the time-series data, which comprise the gene universe for predictive gene network modeling. This will be based on published annotation of epigenetic regulators including but not limited to HMT, HAT and HDAC gene families [34, 65]. These epigenetic regulators in addition to TFs [64] will be used as predictors to create a learned gene network model and GRN (Fig. 2). As control, I will also run DFG with *only* transcription factors as predictors [64] (Fig. 2), as previously described in [12], which will generate gene network models and a resultant GRN. These two network models will be compared in their network properties, subnetwork(s) of key genes, success rate in predicting left-out time-point, and *in vivo* validation with published transgenic experimental data if available (to estimate precision and recall), to test if epigenetic regulators help yield a better gene network**.**

**Preliminary analysis:** A fine-scale time-course transcriptome data from nitrogen treatment was generated to test the feasibility of my proposed approach. This fine-scale time series data, containing 10 time points (0, 5, 10, 15, 20, 30, 45, 60, 90, and 120 minutes) and two conditions (Nitrogen treated vs. controls), was analyzed by fitting a temporal trend for each condition using a cubic spline in the LIMMA package in R [58]. A *F*-test for differences in the coefficients of the two trends identified 2,173 genes that change in response to nitrogen treatment over time. These 2,173 genes comprise the gene universe for the predictive network modeling. In this universe, there are 161 TFs based on current Arabidopsis TF annotation in AGRIS database [64]. Importantly, this N-response gene universe also contains ten epigenetic regulators, including four HDACs (HDA8, HDT1, HDT3, and HDT4) [65], five HMTs (SDG19/SUVH3, SDG39/ASHR2, SDG15/ATXR5 and two putative HMTs) [66], and one RNA-directed DNA methylase (RDM1). With the help of co-mentor Dr. Dennis Shasha, the first 9 time points of the time-series expression data of these 2,173 N-response genes were used to train a DFG model as described in [12]. For run No.1, only the 161 TFs were used as the predictors in the modeling (“TF-DFG”); for run No. 2, TFs and the 10 epigenetic regulators were used as predictors in the modeling (“TF-EpiReg-DFG”) (Table 1). A highly stringent p-value cutoff (p-val<0.001) was applied to derive GRNs from the learnt models, *i.e.* if a predictor gene has high-confidence influence on a regulated gene by p-value < 0.001, this predictor gene and the regulated genes are connected by a inferred causal regulatory edge.

Significantly, the “TF-EpiReg-DFG” network contains 1,408 *de novo* edges linking epigenetic regulators to putative target genes, compared to the “TF-DFG” network. Importantly, when I focus on regulatory hubs inthe “TF-EpiReg-DFG” network, three epigenetic regulators are among the top 20 hubbiest regulatory genes: 1) a putative histone methyltransferase AT3G26850 (ranked 4th); 2) histone deacetylase HDT4 (ranked 5th); and 3) SET domain containing protein coded by AT3G56570 (rank 7th). Unfortunately, the genome wide targets of any these histone modifiers are not known. Nonetheless, this preliminary result supports a central role of epigenetic regulators in the nutrient responsive GRN. For example, all the epigenetic regulators and their regulated genes comprise a subnetwork of 453 genes, covering 45.85% of the nitrogen responsive GRN. This epigenetic-regulator based subnetwork uncovers interesting insights on how epigenetic regulators and TFs interact to regulate key metabolism genes in the nitrogen assimilation pathway (Fig. 6). Finally, allowing even a few epigenetic regulators as predictors in DFG increases the success rate of predicting left-out time points by 0.4%-9%.

These preliminary results thus support the feasibility of the approaches to incorporate epigenetic regulators in predictive network modeling using DFG. For my proposed study, I will apply the DFG machine learning approach to reconstruct GRNs from the sugar time-series transcriptome data from Aim 1, with only TFs as predictors, or with TFs and epigenetic regulators transcriptionally regulated in the sugar time-course (*e.g.* SDG8) as predictors, to derive different GRNs for comparison.

**B. Use experimentally derived SDG8->target information as “priors” to inform predictive network modeling (independent phase).** Experimentally confirmed regulatory edges between TFs and their target genes can be used as biological prior knowledge to inform and improve network modeling [67][19, 67]. In this subaim, I will use experimental confirmed genomic targets of SDG8 as prior knowledge, to inform predictive network modeling.

**Modeling approaches:** Typically, the algorithm for DFG starts with an initial *f* matrix representing the influence of the predictors (e.g. in my study the TFs and epigenetic regulators) on the regulated genes (Fig. 5B), which will be optimized by the end of the machine learning procedure. If no prior knowledge is used, this initial matrix is seeded with normal distributed pseudorandom numbers. When prior knowledge is used, a matrix of 1 and -1 representing the prior knowledge of the regulatory relationship between predictors and target genes (e.g. SDG8 and its genomic targets) will be overlaid on top of the random matrix as the initial matrix. I have profiled genomic targets of SDG8 through multiple *–omics* approaches using *sdg8* mutant and transgenic SDG8 epitope-tagged lines (see preliminary results below). This information will be provided as priors in the form of an initial seeding matrix for DFG, while epigenetic regulators and TFs will be used as predictors to create a third learnt model and GRN from the time-series data from Aim 1 (Fig. 2). The resulting GRN model will be compared to the ones constructed without prior knowledge (Fig. 2) in their network properties, including their success rate in accurately predicting gene expression states in left-out time-point data, and *in vivo* validation with published transgenic experimental data if available, to test if the epigenetic regulator priors help yield greater accuracy of prediction in the gene network**.**

**Preliminary data:** ***Identification of SDG8 genome-wide targets:*** A fast-neutron deletion mutant in *sdg8*, *sdg8-5*, was isolated previously in our lab in a positive genetic screen seeking molecular components involved in integrating light and sugar signaling [51]. Overall, the *sdg8-5* deletion mutant provides an ideal system to study genomic targets of HMTs. To identify its genome-wide targets, I performed histone ChIP-seq assays and detected a sharp and specific decrease of H3K36me3 marks in 4,060 genes in the *sdg8-5* mutant compared to WT (Fig. 3B & C) (FDR<0.05, fold change>2-fold by SICER [68]). By contrast, the rest of the genome displays a wild-type amount of H3K36me3 marks. These results demonstrate that SDG8 functions to target a specific set of genes for histone methylation. These 4,060 genes comprise the “functional targets” of SDG8, which are not dependent on developmental stages (*e.g.* juvenile *vs.* reproductive). To probe direct binding of SDG8 to its functional targets, I created a HA-tagged SDG8 transgenic line (*h*SDG8) (Fig. 3A). Using the *h*SDG8 transgenic line, 2,267 genes were detected to directly bind to SDG8 (FDR<0.01 by SICER), which contains 728 of the 4,060 functional target genes. These 728 genes are thus “high confidence direct targets” of SDG8 (Fig. 3A), since they are both bound by SDG8 and their associated histones are methylated by SDG8. Additionally, as a group these specific targets of SDG8 have functional specificity related to energy metabolism (Fig. 3E) and cis-regulatory binding motif specificity (Fig. 3A). This indicates that a targeting mechanism exists to enable SDG8 to specifically recognize and regulate genes involved in energy metabolism by epigenetic control. Importantly for network inference, SDG8 is required for elevating the mRNA level of its genomic targets (Fig. 3D). Therefore, I will use the 728 genomic targets of SDG8 as prior knowledge to inform the DFG modeling of sucrose time-series data as described above in **modeling approaches**.

**Innovation:** In traditional gene network inference approaches, transcription factors have been typically used as predictors for gene expression change [10–13]. However, to date these have not incorporated other regulatory molecules, like epigenetic regulators. In this aim, I am going to add epigenetic regulators in addition to TFs, as predictors in network inference, to study the function of epigenetic regulators in GRNs. In addition, biological databases such as TF-DNA binding databases [64] and KEGG metabolic pathway data [69] have been used as prior knowledge to inform network learning (Werhli and Husmeier, 2007). Recently, correlation data derived from histone modification profiles measured by ChIP-Seq was used as epigenetic priors to inform dynamic Bayesian network modeling, which improved the sensitivity and accuracy of the gene network model [19]. However, experimentally confirmed genomic targets of epigenetic regulators have not previously been tested as prior knowledge for network inference. This is partially because genomic targets of epigenetic regulators are not as well documented and studied as TF targets are, given the relatively short history of such effort [20, 22, 70]. Here, I aim to explore the innovative idea of using experimentally confirmed genomic targets of epigenetic regulators as prior knowledge to inform network inference.

**Outcomes and alternative approaches for Aim 2:** In Aim 2 I will be able to improve predictive network modeling by incorporating epigenetic regulators as predictors (Aim 2A) and genomic targets of epigenetic regulators as priors (Aim 2B) using network inference methods such as Dynamic Factor Graphs. I will assess if epigenetic data increases the performance of network inference using “left-out” data. My preliminary data supports this hypothesis. Importantly, this new and improved method of predictive network modeling will produce GRNs with epigenetic regulators as regulatory hubs, and reveal new biology on the epigenetic and transcriptional regulation of gene networks in response to environmental stimulus (for example see Fig. 6).

The Coruzzi lab and Dr. Dennis Shasha (co-mentor of this K99 proposal) have successfully implemented DFG for reconstructing causal GRNs from time series data as published in [12], and we have already tested the feasibility of including epigenetic regulators predictors as described above in preliminary results in Aim 2A, so I don’t expect major problem in implementing the proposed modeling approach. One possible scenario is that epigenetic regulators might have different behavior from TF regulators in network modeling. After I get trained in machine learning and developed better understanding of the scripts of DFG, I might modify DFG scripts to treat epigenetic regulators as a special class of predictors *e.g.* as an interactor with a TF regulator, or give SDG8 targets a special status *e.g*. altered response to TF influence. In the unlikely scenario that DFG approach is not suitable for the proposed study of priors in Aim 2B, I will consider other network modeling tools such as Inferelator [11], developed by Bonneau Lab in the Center for Genomics and Systems Biology at NYU, as an alternative.

**Specific Aim 3: Expansion and refinement of the epigenomic-informed GRN (independent phase).**

**Rationale**: In Aim 3, I will experimentally identify genome-wide targets of more epigenetic regulators besides SDG8 by ChIP-Seq, to validate and refine the DFG network generated in Aim 2. This iteration cycle (Fig. 2) will help refine the network prediction and expand our experimental target data on epigenetic regulators.

**Experimental and modeling Approaches:** As shown by my preliminary study, 16 epigenetic regulators (including HMTs, HATs and HDACs) are significantly regulated by a 2hr supply of sucrose at a FDR-adjusted p-value < 0.05 (Table 2). Therefore, I expect that Aim 2 will likely uncover multiple epigenetic regulators (including SDG8) that regulate the sucrose-responsive GRN. I will pick 3-5 epigenetic regulators to experimentally determine their genome-wide functional targets, by comparing histone profiles between WT and mutants using histone ChIP-Seq, as described in preliminary data in Aim 2B. I will prioritize epigenetic regulators for target identification as follows: i) Epigenetic regulators in the GRN from Aim 2 will be ranked by their out-degree, which measures how many genes are regulated by each of these epigenetic regulators. The epigenetic regulators ranked higher will be prioritized. ii) Epigenetic regulators with reported substrate information, *e.g.* H3K27m or H3K9Ac, will be prioritized, to speed up the functional target identification by histone ChIP-Seq. iii) Epigenetic regulators with available homozygous mutants from ABRC Arabidopsis seed stock center (abrc.osu.edu) will be prioritized. From this list of 3-5 prioritized genes, I will also pick 1-2 highly influential and interesting epigenetic regulators to create epitope-tagged transgenic lines as described in preliminary data in Aim 2B. This will enable me to determine their directly bound genomic targets by epitope-tag ChIP-Seq in addition to determining their functional targets by histone ChIP-Seq. The growth conditions for histone ChIP-Seq and epitope-tag ChIP-Seq will be the same as Aim 1, as the genomic targets of epigenetic regulators can be developmental-context dependent (REF).

***Validation***: The experimentally identified genomic targets of these 3-5 epigenetic regulators will be compared to their predicted targets in the GRN in Aim 2, [Will you be modifying these targets? I think that would be cool.] to serve as an independent experimental validation of network prediction. The significance of overlap between the experimentally validated targets and network predicted targets will be estimated using hypergeometric distribution. ***Expansion and refinement:*** Most importantly, the experimental validated genomic targets will provide new prior knowledge to fuel another round of predictive network modeling in Aim 2B, which will hopefully result in a more refined GRN by incorporating more biological information. This more refined GRN will predict new epigenetic regulators and targets for validation test in Aim 3 (Fig. 2). This iterative loop will gradually refine the GRN for generating new biological hypothesis.

**Innovation:** Numerous publications have reported genome-wide targets of TFs [71, 72]. By contrast, genomic targets of epigenetic regulators are less understood. In recent years, reports on global targets of epigenetic regulators have started to accumulate [73–77][1, 20]. Since binding doesn’t imply function [1, 78], it is equally important to identify functional targets of epigenetic regulators [23, 70] as well as bound targets. In this proposal, I took advantage of my preliminary data on genomic functional and bound targets of SDG8, to model its influence as a regulatory protein in a GRN. From this starting point, I will expand my research to identify genomic functional targets and bound targets of other epigenetic regulators, to help build the knowledgebase of genomic targets for epigenetic regulators.

**Outcomes and alternative plans:** Aim 3 will generate experimental data on genomic targets of epigenetic regulators in addition to SDG8 (e.g. HMTs, HATS, and HDACs as seen in Table 2), to validate and refine the GRNs in an iterative manner. As an alternative approach to expand the *EpiNet* study during the 3-year independent research phase, a shift of focus to other epigenetic regulators may lead to the discovery of GRNs that are responsive to other stimuli like abiotic/biotic stress, depending on the nature of these epigenetic regulators.

**Summary and Perspective:** This *EpiNet* proposal will complement the current effort of linking epigenomics data to GRN modeling, by specifically testing the hypothesis that incorporating epigenetic regulators as predictors will expand and improve predictive network modeling. The epigenetic regulation mechanism is largely conserved across kingdoms, therefore, a possible future direction would be to expand my research to animals, to test the conservation and uniqueness of epigenetic controlled GRNs in plants *vs.* animals. Also maybe other kinds of regulators such as hormones or whatever?

**Reference**

1. Van Steensel B: **Mapping of genetic and epigenetic regulatory networks using microarrays**. *Nat Genet* 2005, **37**:S18–S24.

2. Karlebach G, Shamir R: **Modelling and analysis of gene regulatory networks**. *Nat Rev Mol Cell Biol* 2008, **9**:770–780.

3. De la Fuente A: **From “differential expression” to “differential networking” - identification of dysfunctional regulatory networks in diseases**. *Trends Genet TIG* 2010, **26**:326–333.

4. Ideker T, Sharan R: **Protein networks in disease**. *Genome Res* 2008, **18**:644–652.

5. Wachi S, Yoneda K, Wu R: **Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues**. *Bioinformatics* 2005, **21**:4205–4208.

6. Jonsson PF, Bates PA: **Global topological features of cancer proteins in the human interactome**. *Bioinformatics* 2006, **22**:2291–2297.

7. Pujana MA, Han J-DJ, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, ElShamy WM, Rual J-F, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Solé X, Hernández P, Lázaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, et al.: **Network modeling links breast cancer susceptibility and centrosome dysfunction**. *Nat Genet* 2007, **39**:1338–1349.

8. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF, Program I and HR to ILSCR: **A network-based analysis of systemic inflammation in humans**. *Nature* 2005, **437**:1032–1037.

9. Elo LL, Järvenpää H, Orešič M, Lahesmaa R, Aittokallio T: **Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process**. *Bioinformatics* 2007, **23**:2096–2103.

10. Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, Wu C, Karwacz K, Xiao S, Jorgolli M, Gennert D, Satija R, Shakya A, Lu DY, Trombetta JJ, Pillai MR, Ratcliffe PJ, Coleman ML, Bix M, Tantin D, Park H, Kuchroo VK, Regev A: **Dynamic regulatory network controlling TH17 cell differentiation**. *Nature* 2013, **496**:461–468.

11. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo**. *Genome Biol* 2006, **7**:R36.

12. Krouk G, Mirowski P, LeCun Y, Shasha DE, Coruzzi GM: **Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate**. *Genome Biol* 2010, **11**:R123.

13. Nachman I, Regev A, Friedman N: **Inferring quantitative models of regulatory networks from expression data**. *Bioinformatics* 2004, **20**(suppl 1):i248–i256.

14. Greer EL, Shi Y: **Histone methylation: a dynamic mark in health, disease and inheritance**. *Nat Rev Genet* 2012, **13**:343–357.

15. Nimura K, Ura K, Kaneda Y: **Histone methyltransferases: regulation of transcription and contribution to human disease**. *J Mol Med* 2010, **88**:1213–1220.

16. Weinstock GM: **ENCODE: More genomic empowerment**. *Genome Res* 2007, **17**:667–668.

17. Guan D, Shao J, Deng Y, Wang P, Zhao Z, Liang Y, Wang J, Yan B: **CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data**. *Bioinforma Oxf Engl* 2014.

18. Wang LY, Wang P, Li MJ, Qin J, Wang X, Zhang MQ, Wang J: **EpiRegNet: Constructing epigenetic regulatory network from high throughput gene expression data for humans**. *Epigenetics* 2011, **6**:1505–1512.

19. Zheng J, Chaturvedi I, Rajapakse JC: **Integration of Epigenetic Data in Bayesian Network Modeling of Gene Regulatory Network**. In *Pattern Recognit Bioinforma*. Edited by Loog M, Wessels L, Reinders MJT, Ridder D de. Springer Berlin Heidelberg; 2011:87–96. [*Lecture Notes in Computer Science*, vol. 7036]

20. Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M, Durham T, Zhang X, Donaghey J, Epstein CB, Regev A, Bernstein BE: **Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells**. *Cell* 2011, **147**:1628–1639.

21. Zheng Y, Ren N, Wang H, Stromberg AJ, Perry SE: **Global Identification of Targets of the Arabidopsis MADS Domain Protein AGAMOUS-Like15**. *PLANT CELL ONLINE* 2009, **21**:2563–2577.

22. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K: **Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions**. *Genes Dev* 2006, **20**:1123–1136.

23. Robyr D, Suka Y, Xenarios I, Kurdistani SK, Wang A, Suka N, Grunstein M: **Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases**. *Cell* 2002, **109**:437–446.

24. Murphy K, Mian S: *Modelling Gene Expression Data Using Dynamic Bayesian Networks*. 1999.

25. Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL: **A Bayesian approach to reconstructing genetic regulatory networks with hidden factors**. *Bioinformatics* 2005, **21**:349–356.

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

27. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 Å resolution**. *Nature* 1997, **389**:251–260.

28. Brusslan JA, Rus Alvarez-Canterbury AM, Nair NU, Rice JC, Hitchler MJ, Pellegrini M: **Genome-Wide Evaluation of Histone Methylation Changes Associated with Leaf Senescence in Arabidopsis**. *PLoS ONE* 2012, **7**:e33151.

29. Maze I, Covington HE, Dietz DM, LaPlant Q, Renthal W, Russo SJ, Mechanic M, Mouzon E, Neve RL, Haggarty SJ, Ren Y, Sampath SC, Hurd YL, Greengard P, Tarakhovsky A, Schaefer A, Nestler EJ: **Essential Role of the Histone Methyltransferase G9a in Cocaine-Induced Plasticity**. *Science* 2010, **327**:213–216.

30. Van Dijk K, Ding Y, Malkaram S, Riethoven J-J, Liu R, Yang J, Laczko P, Chen H, Xia Y, Ladunga I: **Dynamic changes in genome-wide histone H3 lysine 4 methylation patterns in response to dehydration stress in Arabidopsis thaliana**. *BMC Plant Biol* 2010, **10**:238.

31. Trollope AF, Gutièrrez-Mecinas M, Mifsud KR, Collins A, Saunderson EA, Reul JMHM: **Stress, epigenetic control of gene expression and memory formation**. *Exp Neurol* 2012, **233**:3–11.

32. Yuan L, Liu X, Luo M, Yang S, Wu K: **Involvement of Histone Modifications in Plant Abiotic Stress Responses**. *J Integr Plant Biol* 2013.

33. Gaydos LJ, Wang W, Strome S: **H3K27me and PRC2 transmit a memory of repression across generations and during development**. *Science* 2014, **345**:1515–1518.

34. Alvarez-Venegas R, Avramova Z: **SET-domain proteins of the Su(var)3-9, E(z) and Trithorax families**. *Gene* 2002, **285**:25–37.

35. Berr A, Ménard R, Heitz T, Shen W-H: **Chromatin modification and remodelling: a regulatory landscape for the control of Arabidopsis defence responses upon pathogen attack**. *Cell Microbiol* 2012, **14**:829–839.

36. Cazzonelli CI, Cuttriss AJ, Cossetto SB, Pye W, Crisp P, Whelan J, Finnegan EJ, Turnbull C, Pogson BJ: **Regulation of Carotenoid Composition and Shoot Branching in Arabidopsis by a Chromatin Modifying Histone Methyltransferase, SDG8**. *PLANT CELL ONLINE* 2009, **21**:39–53.

37. Cazzonelli CI, Roberts AC, Carmody ME, Pogson BJ: **Transcriptional Control of SET DOMAIN GROUP 8 and CAROTENOID ISOMERASE during Arabidopsis Development**. *Mol Plant* 2009, **3**:174–191.

38. Dong G, Ma D-P, Li J: **The histone methyltransferase SDG8 regulates shoot branching in Arabidopsis**. *Biochem Biophys Res Commun* 2008, **373**:659–664.

39. Hoppmann V, Thorstensen T, Kristiansen PE, Veiseth SV, Rahman MA, Finne K, Aalen RB, Aasland R: **The CW domain, a new histone recognition module in chromatin proteins**. *EMBO J* 2011, **30**:1939–1952.

40. Kim SY: **Establishment of the Vernalization-Responsive, Winter-Annual Habit in Arabidopsis Requires a Putative Histone H3 Methyl Transferase**. *PLANT CELL ONLINE* 2005, **17**:3301–3310.

41. De-La-PeñA C, Rangel-Cano A, Alvarez-Venegas R: **Regulation of disease-responsive genes mediated by epigenetic factors: interaction of Arabidopsis-Pseudomonas**. *Mol Plant Pathol* 2012, **13**:388–398.

42. Palma K, Thorgrimsen S, Malinovsky FG, Fiil BK, Nielsen HB, Brodersen P, Hofius D, Petersen M, Mundy J: **Autoimmunity in Arabidopsis acd11 Is Mediated by Epigenetic Regulation of an Immune Receptor**. *PLoS Pathog* 2010, **6**:e1001137.

43. Soppe WJ, Bentsink L, Koornneef M: **The early-flowering mutant efs is involved in the autonomous promotion pathway of Arabidopsis thaliana**. *Development* 1999, **126**:4763–4770.

44. Wang X, Chen J, Xie Z, Liu S, Nolan T, Ye H, Zhang M, Guo H, Schnable PS, Li Z, Yin Y: **Histone Lysine Methyltransferase SDG8 Is Involved in Brassinosteroid-Regulated Gene Expression in Arabidopsis thaliana**. *Mol Plant* 2014, **7**:1303–1315.

45. Xu L, Zhao Z, Dong A, Soubigou-Taconnat L, Renou J-P, Steinmetz A, Shen W-H: **Di- and Tri- but Not Monomethylation on Histone H3 Lysine 36 Marks Active Transcription of Genes Involved in Flowering Time Regulation and Other Processes in Arabidopsis thaliana**. *Mol Cell Biol* 2007, **28**:1348–1360.

46. Zhao Z, Yu Y, Meyer D, Wu C, Shen W-H: **Prevention of early flowering by expression of FLOWERING LOCUS C requires methylation of histone H3 K36**. *Nat Cell Biol* 2005, **7**:1156–1160.

47. Springer NM: **Comparative Analysis of SET Domain Proteins in Maize and Arabidopsis Reveals Multiple Duplications Preceding the Divergence of Monocots and Dicots**. *PLANT Physiol* 2003, **132**:907–925.

48. Rogers LA, Dubos C, Cullis IF, Surman C, Poole M, Willment J, Mansfield SD, Campbell MM: **Light, the circadian clock, and sugar perception in the control of lignin biosynthesis**. *J Exp Bot* 2005, **56**:1651–1663.

49. Rogers LA, Dubos C, Surman C, Willment J, Cullis IF, Mansfield SD, Campbell MM: **Comparison of lignin deposition in three ectopic lignification mutants**. *New Phytol* 2005, **168**:123–140.

50. Stokes ME, Chattopadhyay A, Wilkins O, Nambara E, Campbell MM: **Interplay between sucrose and folate modulates auxin signaling in Arabidopsis**. *Plant Physiol* 2013, **162**:1552–1565.

51. Thum KE, Shin MJ, Gutiérrez RA, Mukherjee I, Katari MS, Nero D, Shasha D, Coruzzi GM: **An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis**. *BMC Syst Biol* 2008, **2**:31.

52. Romano JM, Dubos C, Prouse MB, Wilkins O, Hong H, Poole M, Kang K-Y, Li E, Douglas CJ, Western TL, Mansfield SD, Campbell MM: **AtMYB61, an R2R3-MYB transcription factor, functions as a pleiotropic regulator via a small gene network**. *New Phytol* 2012, **195**:774–786.

53. Auer PL, Doerge RW: **Statistical Design and Analysis of RNA Sequencing Data**. *Genetics* 2010, **185**:405–416.

54. Wang L, Si Y, Dedow LK, Shao Y, Liu P, Brutnell TP: **A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq**. *PLoS ONE* 2011, **6**:e26426.

55. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**:R36.

56. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**:511–515.

57. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC Bioinformatics* 2010, **11**:94.

58. Carey VJ, Huber W, Irizarry RA, Dudoit S: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. *Volume 746718470*. New York: Springer; 2005.

59. Gutiérrez 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.

60. Loreti E, Poggi A, Novi G, Alpi A, Perata P: **A Genome-Wide Analysis of the Effects of Sucrose on Gene Expression in Arabidopsis Seedlings under Anoxia**. *Plant Physiol* 2005, **137**:1130–1138.

61. Palenchar PM, Kouranov A, Lejay LV, Coruzzi GM: **Genome-wide patterns of carbon and nitrogen regulation of gene expression validate the combined carbon and nitrogen (CN)-signaling hypothesis in plants**. *Genome Biol* 2004, **5**:R91.

62. Shimamura T, Imoto S, Yamaguchi R, Fujita A, Nagasaki M, Miyano S: **Recursive regularization for inferring gene networks from time-course gene expression profiles**. *BMC Syst Biol* 2009, **3**:41.

63. Wang Y, Joshi T, Zhang X-S, Xu D, Chen L: **Inferring gene regulatory networks from multiple microarray datasets**. *Bioinformatics* 2006, **22**:2413–2420.

64. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E: **AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks**. *Plant Physiol* 2006, **140**:818–829.

65. Pandey R, Müller A, Napoli CA, Selinger DA, Pikaard CS, Richards EJ, Bender J, Mount DW, Jorgensen RA: **Analysis of histone acetyltransferase and histone deacetylase families of Arabidopsis thaliana suggests functional diversification of chromatin modification among multicellular eukaryotes**. *Nucleic Acids Res* 2002, **30**:5036–5055.

66. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, et al.: **InterPro in 2011: new developments in the family and domain prediction database**. *Nucleic Acids Res* 2012, **40**(Database issue):D306–312.

67. Werhli A, Husmeier D: **Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge**. *Stat Appl Genet Mol Biol* 2007, **6**.

68. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: **A clustering approach for identification of enriched domains from histone modification ChIP-Seq data**. *Bioinformatics* 2009, **25**:1952–1958.

69. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 1999, **27**:29–34.

70. Ekwall K: **Genome-wide analysis of HDAC function**. *Trends Genet TIG* 2005, **21**:608–615.

71. Gaudinier A, Zhang L, Reece-Hoyes JS, Taylor-Teeples M, Pu L, Liu Z, Breton G, Pruneda-Paz JL, Kim D, Kay SA, Walhout AJM, Ware D, Brady SM: **Enhanced Y1H assays for Arabidopsis**. *Nat Methods* 2011, **8**:1053–1055.

72. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B: **JASPAR: an open‐access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, **32**(suppl 1):D91–D94.

73. Chruscicki A, Macdonald VE, Young BP, Loewen CJR, Howe LJ: **Critical determinants for chromatin binding by Saccharomyces cerevisiae Yng1 exist outside of the plant homeodomain finger**. *Genetics* 2010, **185**:469–477.

74. Greil F, Kraan I van der, Delrow J, Smothers JF, Wit E de, Bussemaker HJ, Driel R van, Henikoff S, Steensel B van: **Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location**. *Genes Dev* 2003, **17**:2825–2838.

75. Lopez-Atalaya JP, Ito S, Valor LM, Benito E, Barco A: **Genomic targets, and histone acetylation and gene expression profiling of neural HDAC inhibition**. *Nucleic Acids Res* 2013, **41**:8072–8084.

76. Luo C, Durgin BG, Watanabe N, Lam E: **Defining the Functional Network of Epigenetic Regulators in Arabidopsis thaliana**. *Mol Plant* 2009, **2**:661–674.

77. Taverna SD, Ilin S, Rogers RS, Tanny JC, Lavender H, Li H, Baker L, Boyle J, Blair LP, Chait BT, Patel DJ, Aitchison JD, Tackett AJ, Allis CD: **Yng1 PHD finger binding to H3 trimethylated at K4 promotes NuA3 HAT activity at K14 of H3 and transcription at a subset of targeted ORFs**. *Mol Cell* 2006, **24**:785–796.

78. Biggin MD: **Animal Transcription Networks as Highly Connected, Quantitative Continua**. *Dev Cell* 2011, **21**:611–626.