**500 word summary of research that support as an HHMI investigator would enable a selected applicant to carry out**. Evolutionary Genomics and Systems Biology

(BELOW Is >600 words….need to cut in to 500)

We will apply network inference approaches across both micro and macroevolutionary scales to identify how networks adapt and evolve relative to key agronomic traits. Arabidopsis (as a reference species) and rice (as a crop species) will be the initial targets to develop and validate both cross-ecotype and cross-species network inference approaches for the trait of N-use efficiency. Our long-term goal is to expand cross-species network inference approaches across seed plant phylogeny for any trait of interest.

This evolutionary systems biology program will initially integrate the following data and techniques: (i) (data) genome sequences, time-series RNA, short RNA, metabolomic, proteomic, and phenotypic trait data collected within each ecotype; (ii) (technique) construction of inferred regulatory networks independently within ecotypes based on time-series experiments and machine learning; (iii) (data/technique) orthology information between ecotypes and species leading to network inference comparisons to identify highly supported (conserved) edges and distinctive edges (unique to an ecotype or set of related ecotypes or a species), (iv) (technique) associate trait-to-gene or trait-to-edge predictions using machine learning approaches. (v) (validation) test in rapid cell-based assay as well as *in planta*. An approximate plan is as follows:

[Gloria: this numbering is not consistent with the above, e.g. network inference should move from (i) to (ii).]

(i) **Network Inference and Trait Adaptation** (independently for selected ecotypes). Phenotypic traits (e.g. N-responsive root characters and N15 labeled metabolites) will be used to phenocluster ecotypes according to distinct N-use strategies. Centroids of these clusters will be subjected to fine-scale time-series N-treatments. Network inference techniques will be used to identify a first cut set of networks, one per ecotype. [Gloria: network inference is a type of machine learning]

(ii) **Cross-Genome Network Inference to Find Common Edges** (across species/ecotypes). Comparisons of networks across species and ecotypes will be used to enhance the support of a predicted network edge (g1, g2) in a network for ecotype E, if another ecotype E’ has an edge (g1’, g2’) where g1’ is orthologous to g1 and g2’ is orthologous to g2.

(iii) **Use of Phylogenetic Analysis to Identify Distinct Edges**: Conversely, when E and E’ have very different phenotypes, we would expect the network of E to have edges missing from E’ and vice versa. Suppose phylogenetic analysis determines that the difference between g1 and g1’ and g2 and g2’ distinguish one clade from another. This will lend support to an edge (g1, g2) in one ecotype where an edge (g1’, g2’) may be missing from another.

(iv) **Connecting network nodes and edges with traits**. We will use two relationships (i) *ecotype-to-gene or edge* and (ii) e*cotype-to-traits*, to derive the relationship *genes-to-traits or edges-to-trait*. Because of combinatorial effects (e.g. several genes may be responsible for the trait) and the complexity of gene networks, we will use statistical and machine learning methods, including decision trees and support vector machines to generate trait-to-gene network predictions.

(v) **Validation**: Tests will occur in planta and using a rapid protoplast-based transient assay which will be amenable to validation testing across a wide range of species.

**[Gloria: I would drop this. Above is ambitious enough. You’ve already said that you will explore other traits]**

**Future Work**. **Network inference across plant phylogeny**. Our long-term goal is to extend the cross-species network inference approaches we develop and validate in the data rich models across seed plant phylogeny. To develop such approaches, we will focus network studies targeting 20 sequenced plant genomes spanning the major clades of seed plant phylogeny. We will experiment with methods to encode expression data into this phylogenetic matrix, and to perform cross-species network inference, using comparative network tools developed in the above project.

DENNIS – In this outline, how to you propose to associate the genes with N-use traits?

network edges and distinctive edges, and association of edges with N-traits

Cross-species time-series data will be used to enhance the predictive power of inferred regulatory networks in the models and their translation to crops. This will be done initially by comparing inferred networks derived based on time-series data collected across two distantly related species (e.g. Arabidopsis and Rice). We will further validate the network inference approach using Arabidopsis as the reference genome and a data-rich crop target (e.g. rice). The cross-species network inference approach will be embodied into a software platform to enable predictive network modeling, across species for which there is little genomic data.

(iv) **Phylogenomic Network Inference** (across clades). In this aim, evolutionary history across 20 species spanning the major plant clades (fig X) will be exploited to identify gene networks associated with evolution of shared derived traits by species within a clade (e.g. N-fixation in legumes). This should enable us to make trait-to-gene network predictions.

These three complementary aims build on each other and each involve a significant interplay between model, crop and biodiverse plant genomes. These studies will enable gene network-to-trait associations (aims i & ii) as well as trait-to-gene network associations (iii). The congruence of genes and gene networks identified by these complementary approaches (ecotype, species, and clade) comparisions will be examined. As proof-of-principle, we will develop these two approaches for the trait nitrogen-use. The tools and approaches developed will be embodied in a production quality Cross-species network inference platform that can be used across a wide variety of species and traits.

This project will bring to bear the following data and techniques: (i) (data) RNA expression, short RNA, possibly proteomic data, as well as phenotypic N-use traits (N-regulated root growth and N15-uptake/assimilation metabolites) within ecotypes for each species (e.g. initially, Arabidopsis and Rice); (ii) (technique) the construction of inferred networks independently within selected ecotypes based on time-series experiments and machine learning analysis; (iii) (data/technique) orthology information between ecotypes and across species leading to cross-species comparison to identify highly supported network edges and distinctive edges, and association of edges with N-traits (?), and (v) (community resource) the development of a production quality CSNI….

Need to add words about BigPlant.

(i) **Species Adaptation and network inference.** We will enhance the power of predictive regulatory networks by exploiting cross-ecotype comparisons. This will entail selecting ecotypes with distinct N-use traits by measuring N-induced changes in root architecture and uptake/metabolism of 15N-labeled inorganic N. Based on variation in these N-use traits, fine-scale time-series experiments for N-treatments will be conducted in a set of ecotypes (of Arabidopsis), to derive predictive networks for each, using machine learning methods. By using phylogenetic comparison of ecotypes (based on whole genome re-sequences), we can identify core as well as ecotype-specific edges within such regulatory networks. We will make network-to-phenome predictions by associating genes within these regulatory networks with the phenomic measurements of ecotypes related to nitrogen use efficiency including. Thus, this approach will uncover genes or network modules associated with N-traits.

ii) **Cross-species network inference**: We propose to develop a Cross Species Network Inference (CSNI) platform that will enable plant biologists to infer difficult-to-obtain relationships in gene or protein networks in less-studied species which we call Targets. CSNI will use edges in validated networks in well-studied species (Reference species), such as Arabidopsis, by integrating homology and easily collected data in the Target, such as gene chip expression measurements and RNA-Seq data. Inferred networks in Targets will provide testable hypotheses about their gene and protein interactions, as well as suggestions for future experiments, especially time-series experiments and genetic modifications. As a proof-of-principle, we will apply this Cross-Species Network Inference framework to predict metabolic and protein-protein interaction edges in rice where existing experimental data enables validation. We will then extend CSNI to other plant genomes, for which experimental data supporting network edges is not yet available. While CSNI is described with respect to plants, the framework and basic algorithms extend to any under-analyzed species. This work will achieve one of the main goals of Systems Biology – predicting network states under untested conditions.Finally, we will provide a biologist-friendly CSNI software platform (www.CrossSpecies.org) that will infer networks in a Target species, given experimentally validated networks in a Reference species, homology information and experimental data in the target species

iii) **Phylogenomic network inference**. Despite the considerable power in generating inferred regulatory networks, the approaches described in (i) rely heavily on the data rich reference models and crops, and their ecotypes. Thus, the CSNI approach alone will fail to uncover new genes and network modules that have evolved in under-studied species, and how such network structure differs across groups of species. For example, preliminary analysis of N-use traits in the 101 genera in the BigPlant tree has shown that N-fixation (a nitrogen-trait in wild relatives) has been lost in the cultivated species for several important clades. In parallel to (i) and (ii), in (iii) we propose to exploit the phylogenetic framework across seed plants to derive the genes and underlying networks associated with key clades, nodes and their associated traits that have evolved in nature. Our current studies in this area included the generation of a phylogeny of the seed plants based on full or partial sequences from 101 genera. We now propose to incorporate functional genomic data (gene expression, protein modification, etc) as well as phenomic traits within this a phylogenetic matrix, and to derive trait-to-gene networks. Initially, we will focus the development of tools and approaches using the 20 fully, or near fully sequenced plant genomes that span the major clades of seed plant diversity. This will involve developing approaches to encode functional genomic data into the matrix, and developing machine learning approaches to predict trait-to-gene predictions across and within species clades. We will use as proof-of-principle, we will collect functional genomic data on N-use across these 20 species (including mRNA, and small RNA), as well as other data-types, especially as they relate to the crops. We note that for our original matrix of 17 seed plant species we have uncovered a morphological data set containing 166 characters with 235 apomorphic states [4,9]. We have developed a machine learning method that enables us to make trait-to gene associations which exploit the power of comparisions across the entire phylogenetic tree.