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

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 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/species; (ii) (technique) phylogeny and phenoclustering of ecotypes based on sequences and N-use traits, (iii) (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 traits with genes and with network edges \ using machine learning approaches (v) (validation) test in rapid cell-based assay as well as *in planta*. An approximate plan is as follows:

(i) **Data**: Genotype: Arabidopsis and rice (ecotypes) (http://1001genomes.org/); Phenotypes: N-responsive root parameters and 15N-labeled metabolites (n.b. root data for 96 Arabidopsis accessions collected). Fine-scale time-series: mRNA, small RNA, and proteome (n.b. time series RNA for Arabidopsis (Columbia) and Rice (Nipponbare japonica) collected.)

(ii) **Phylogeny and** **Phenoclusters**: Ecotypes will be clustered according to N-use traits using scaled measures of N-regulated root characters and 15N-labeled metabolites. Centroids of phenoclusters will be selected for network inference studies. Phylogenies based on ecotype sequences will be used identify clade or ecotype-specific gene characters. For species comparisons, we will exploit a phylogenomic matrix of 20 fully sequenced genomes spanning the major plant clades (Gymnosperms, Monocots, and Dicots (Asterid I & II, Rosids I and II) to identify species or clade specific gene attributes. [It seems to me Gloria that this should come before (i)]

(iii) **Network Inference:** (independently for selected ecotypes). Fine-scale time series data RNA (mRNA and small) from centroid ecotypes will be used to generate causal regulatory networks. Network inference techniques such as state-space analysis will be used to identify a first cut set of networks, one per ecotype.

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

(iv-b) **Use of Phylogenetic Analysis to Identify Distinct Network Edges** (across ecotypes/species): Conversely, when E and E’ have very different N-use 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 further support to the importance of an edge (g1, g2) in one ecotype where an edge (g1’, g2’) may be missing from another.

(iv) **Network-to-trait associations**. 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 that will be amenable for validation testing across a wide range of species.

**Future studies**. Our long-term goal is to extend the cross-species network inference approaches we develop across seed plant phylogeny. We will extend our network comparisons within the dicots (Arabidopsis & Medicago) and monocots (Rice and Corn) to identify clade-specific N-regulatory networks associated with N-use across seed plant phylogeny.