**NSF Arabidopsis 2010 Genome Grant (MCB-0929338)**: "Nitrogen networks in plants"

Dates: BECCA FILL IN HERE????

PI: Gloria M. Coruzzi (NYU Center for Genomics and Systems Biology)

Co-PI: Dennis Shasha (NYU Courant Institute of Mathematical Sciences)

* **Supplement Request**
	+ \*Justification for Supplement- BELOW
	+ \*Summary of Proposed Work- BELOW
	+ \*Budgets (NSF 1030s)- BECCA
	+ Budget justification - BECCA
	+ WHOI budget- BECCA
	+ Supplementary Documents (if any) (NONE)

**Justification for Supplement:** We are requesting a one-year supplement to our existing NSF Grant MCB-0929338 “Nitrogen Networks in Plants” to accomplish two computational goals that are necessary to analyze and integrate the large datasets we have developed under the initial funding of this project. In our first supplemental Aim 1 (S1), we will develop and test new computational methods to integrate our Network Analysis and Genome-Wide Association mapping Studies (GWAS). Combining the NUE trait information with SNP analysis refined by gene network information based on gene transcriptional regulation and other previously known interactions (post transcriptional, translational, etc) will improve our identification of candidate NUE genes. Importantly, developing methods to integrate genomic datasets across Networks and GWAS should be of broad interest and have broad applications in the general field of Systems Biology. In our Supplemental Aim 2 (S2), we will test/develop computational approaches to interrogate -omic data across both time and space. Specifically, we seek to understand how Nitrogen (N) sensed in roots coordinates a systemic N-response via root-shoot-root long distance signaling in a dynamic fashion. Developing such computational methods will enable us to identify mechanisms by which plants respond to environmental signals as an integrated system. Again, our methods can be used beyond our study of NUE in the field of dynamic and spatial regulatory network modeling. Developing such computational approaches is critical to analyze the large amount of genomic (RNA-seq) and phenotype data (NUE traits) generated in this project. This data integration is crucial to derive testable hypotheses of genes controlling systems-wide use of nitrogen in higher plants, the ultimate goal of our project.

**Summary of Proposed Work:** This NSF project aims to determine the mechanisms by which plants respond systemically to a changing nitrogen (N) nutrient environment. To do this, we adapted a classical physiological “split-root system” to analysis approaches of the genomic era ([*1*](#_ENREF_1)). This supplement relates to two computational approaches to analyze and integrate the large amount of data we have generated using the split-root approach in this project. This includes transcriptome data (mRNA and small RNAs), and phenotype traits (NUE and root architecture) for 80 ecotypes. Our aim in this supplement to combine transcriptome data with GWAS/SNPs information has the potential to identify genes associated with NUE including, but not limited to transcriptional changes. With this supplement, we will explore whether and how network analysis of –omic data can be used to refine outcomes of our GWAS studies across different levels of control (e.g. transcriptional, post-translational, etc). Our studies should bridge the gap between Systems Biology and Genome Wide-Association Studies, and should have impact on both fields.

**Focus Areas of Supplement Request:** This work will include development of computational approaches for:

* S1. Using “network knowledge” to refine identification of genes involved in complex phenotypes (GWAS).
* S2. Analyzing regulatory connections across time *and* space (e.g. organs)

**S1. Using “network knowledge” to refine the identification of genes involved in complex phenotypes (GWAS).** GWAS is a potentially powerful approach to identify genes underlying traits. However, GWAS studies across animals and plants have failed to uncover genes underlying complex traits, referred to as ‘missing heritability” ([*2*](#_ENREF_2)). Several features contribute to this problem.

*One problem* that limits the power of GWAS to identify genes underlying complex traits is that the set of SNPs is usually much larger than the set of ecotypes examined, leading to ambiguity in the possible causes of ecotypic differences. In Arabidopsis, for example, there are >250K SNPs that are commonly used ([*3*](#_ENREF_3)), yet most GWAS studies measure traits on at most hundreds of ecotypes. In our study, where we measured traits involved in systemic N-signaling and NUE, we examined 80 ecotypes – due to the laborious nature of the phenotypic measurements using the split-root system. However, even if we could examine traits in thousands of ecotypes, there would still be a large discrepancy with the number of SNPs. One approach we will explore to reduce this discrepancy is to refine our SNP search space using “network knowledge”, as described below.

*A second problem* is that for some ecotypes, the phenotype of interest may be weakly supported (low heritability). For example, the level of the phenotype may differ widely in different individuals within the same population (e.g. large variation). So the *effective* number of ecotypes is further reduced. Because GWAS methods were initially developed for humans, the problem of variation within genetically identical individuals within an ecotype population is a problem present in all model systems. We propose to address this issue in this supplement.

*A third problem* is that the phenotypes may not be simple. For example, in our case instead of looking at lateral root length (of split-root plants), we may be interested in the *ratio* of lateral roots to primary roots. Indeed, we may not know which combination of root traits will help us find significant underlying genes in GWAS. Thus, in this supplementary year, we will develop methods to systematically test combinatorial traits in GWAS analysis.

During this supplementary year, we will develop computational approaches that should address these problems, and improve candidate gene identification as described below. Briefly, we will attempt to increase the power of GWAS by combining two forces: 1. Refining the “gene space” using “network knowledge” and 2. Develop methods to systematically explore the combinatorial phenotypic traits, as shown in Fig. 1. The specific steps we will take are outlined below.

1. ***Refining Ecotype selection***. GWAS approaches were initially developed in humans – where trait measurements correspond to a single individual. However, in Arabidopsis and other biological models, we measure traits of genetically identical individuals within an ecotype, yet standard GWAS approaches use the mean value of the trait without considering the between-individual variation. To address the issue of ***uninformative or widely variant phenotypes***, we will refine our selection to include an ecotype e in our GWAS analysis only if it has either of the following two properties: (1) its mean value for the trait of interest less the standard deviation of that value is greater than the global mean for that traits across all ecotypes (i.e. for trait T, (mean(e) – std(e)) > mean(expression of all ecotypes)) or (2) that the mean value for that trait plus the standard deviation for that value is less than the global mean (i.e. (mean(e) + std(e)) < mean(expression of all ecotypes). The first set are statistically speaking positive species for that phenotype and the second set are negative ones. This eliminates ambiguous ecotypes to give us clearer input.
2. ***Refining the “gene space” of GWAS searches***: Another issue with most GWAS studies is that the set of possible SNPs is too high when considering the whole genome. To address this, we will use a principled approach to focus GWAS analysis on sets of genes that are most relevant to the NUE phenotypes at hand. We will refine our GWAS mapping studies by focusing on N-regulated genes that are connected in networks (Fig. 1A, y-axis) including: 1. N-regulated gene network and 2. A systemic-N responsive network ([*1*](#_ENREF_1)), that include protein-protein, protein-DNA, miRNA-RNA connections (*[4](#_ENREF_4" \o "Gutierrez, 2008 #205)*). Each of these networks employ the Arabidopsis Multinetwork we developed and implemented in the VirtualPlant software platform ([*5*](#_ENREF_5)), to connect the differentially regulated genes to their network partners by protein-protein, protein-DNA, miRNA-RNA connections, in addition to expression correlation connections ([*4*](#_ENREF_4)). Using a “one-hop” network approach, we can capture additional transcriptional, post-transcriptional and translational regulators linked to these N-regulated gene networks, expanding the scope beyond genes that respond transcriptionally to a N-treatment. Preliminary analysis suggests that this approach has promise. In a GWAS mapping study of a complex NUE trait (), a test of SNPs in 123 genes responsive to systemic N-signaling was able to uncover a highly significant GWAS hit in Chromosome 2 (FDR<0.05) that was missed in a comparable whole genome analysis (Fig. 1B).

1. ***A systematic exploration of complex traits***: To address the need for a systematic test of ***complex traits***, we will introduce the notion of ***trait lattice*** in which simple traits are combined into complex traits based on ratios, differences, sums, and products (Fig. 1A, x-axis). For example, if trait T1 and T2 exist as individual traits then so do combinatorial traits including T1+T2, T1-T2, T1\*T2, T1/T2. If there are n traits, this results in  possible pairs. Further pairwise traits can be combined with other single traits to get traits like (T1-T2)\*T3, yielding three-way possibilities. Biological insight may cut down these possibilities to some set of interesting complex traits. (Fig. 1A, x-axis).

Step 2 and 3 leads to the construction of the GWAS space (Fig. 1A), where a series of GWAS are performed with combinations of all gene spaces (X-axis) and all single and combinatorial traits (Y-axis). As usual in GWAS studies, we will treat each simple and complex trait and gene space individually (the circles 1 to 9 in Fig. 1A). For each particular trait, individual gene space will give us a set of p-values (and false discovery rates) for a set of SNPs. However, there may be SNPs that rank highly in several of the GWAS analysis. To detect those SNPs, we will assign a global rank to a SNP x as the sum of the ranks of that SNP for each GWAS analysis over all GWAS performed (SR(x)). Then, we will sort the SNPs by their sum of ranks (SR). For example (Table 1), suppose there are two SNPs S1 and S2 and three GWAS (with different traits and gene spaces). For GWAS 1, S1 has rank 1, S2 has rank 2; similarly for GWAS 2; but for GWAS 3, S2 has rank 1 and S1 has rank 2. Then the SR(S1) is 4, while the SR(S2) is 5. In sorted SR order, S1 is ranked 1 and S2 is ranked 2. This will give a sorted global ranked order of SNPs whose effectiveness we can study.

Table 1. Example of calculation of Sum of Rank (SR) and global rank for SNPs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SNP rank | GWAS1 | GWAS2 | GWAS3 | Sum of Rank (SR) | Global Rank |
| S1 | 1 | 1 | 2 | 4 | 1 |
| S2 | 2 | 2 | 1 | 5 | 2 |

To estimate the p-value and therefore the false discovery rate of the resulting sorted global rank, we must acknowledge the fact that the traits may be correlated with one another, either by construction (e.g. a gene positively influencing T1 will also positively influence T1 + T2) or by nature (e.g. lateral root length is related to lateral root density). As there is no principled way to disentangle these co-dependencies, we will estimate the p-value of each global rank value by non-parametric reshuffling. That is, for each of M GWAS tests, we will randomly and independently permute the SNP rank corresponding to each GWAS, then compute the simulated global ranks. This will give an estimated p-value and false discovery rate for each true global rank using the actual data.

**S2. Modeling gene networks across both time *and* space (e.g. organs).** The goal of this aim is to identify the regulatory relationships between genes that respond to a systemic N-signal over space and time. We will test two complementary approaches to identify such tempo-spatial regulatory relationships: 1. Integration of time-and-space data, and 2. Time-Lagged correlation to identify long distance traveling miRNAs that mediate shoot-root systemic N-signaling (Fig. 2).

**The data**: We have measured transcriptome data from shoots and roots of split-root plants exposed to three N-treatments: Control-N roots (both root halves are nitrogen replete); Control-KCl roots (both root halves are nitrogen-deprived) and Heterogeneous-N roots (one root ½ is in nitrogen-replete and the other is nitrogen-deprived). We have mRNA and small RNA data from four time points (0h, 2h, 4h, 8h) following these N-treatments. We will develop two new approaches to identify genes involved in systemic-N signaling based on their distinct behavior over space (root vs. shoot) and time, as follows.

1. **Integrating time and Space**. To determine which genes act very differently in the heterogeneous N-environment compared to the homogeneous-N controls and the timing of such gene regulation, we will explore a **new method to improve statistical power** in addition to the ANOVA model with nitrogen and split-condition described previously ([*1*](#_ENREF_1)). Our new strategy is simple yet powerful because it makes a weak assumption on the data type, as described in detail below: for each gene *g* and each time point *t*, we calculate the mean ratio of the expression value of *g* (denoted E(g)) in the nitrogen-rich portion of the heterogeneous N-treatment (Sp.KNO3) compared with the expression value of *g* in homogeneous control-N treatment (C.KNO3) in roots:

Similarly, the e­xpression value of *g* in nitrogen-poor patch in heterogeneous environment (Sp.KCl) with *g* in the homogeneous control-KCl environment (C.KCl) is calculated as:



A higher ratio in either case suggests the gene is over-expressed in the N-foraging portion of the split-root system (Sp.KNO3) and under-expressed in the KCl portion of the split-root system (Sp.KCl) relative to controls. Of course, the converse is also interesting: under-expressed in the sp.KNO3 portion relative to control-N and over-expressed in the Sp.KCl relative to control-KCl. The figure of merit is the product of the ratios (PR):



Next, we rank genes based on the absolute value of PR, reflecting the most deviation from the control values. Alternatively, we can calculate the p-value of PR as follows: for each time point *t* and each gene *g*, take the 12 gene expression measurements from the three replicates of four treatment conditions (C.KNO3, Sp.KNO3, Sp.KCl, and C.KCl), permute the labels and recompute the absolute value of PR M times (where M = 1000). We can then just count the frequency at which simulated PR value is equal to or greater than the absolute value of the product computed for the true PR (x(PR)) and calculate the significance *pval*=(x(PR))/1000. A large number of permutation tests will provide better estimates of the P-value. We can then rank the genes by the *pval*. In both ranking methods, the top ranked genes on both ends are the ones likely associated with the root foraging ability under heterogeneous N-environment trigger by systemic N-signaling at time *t*.

The above equations can be also applied to the shoots, in with case the product of ratio can be calculated as:



Where E(g,SplitRoot) is the expression of gene *g* in the shoots of the plants where roots are exposed to the heterogeneous-N treatment. The PR in shoots will be used to identify shoot genes associated with the systemic response to heterogeneous N-environment at time *t* as described above.

**2. Time-lagged correlation.** Here, we use time to study the role of miRNA in long distance systemic-N signaling, as they are important developmental regulators that are able to travel from shoot-to-root ([*6-10*](#_ENREF_6)). We will mine the time-series mRNA and sRNA expression data to identify putative miRNAs traveling between shoots-to-roots based on the cross-organ correlation between precursor miRNA and mature miRNA over time. We will first identify miRNAs that show significant variance across the time-series (Fig. 2). For such miRNAs, if the correlation between pre-miRNA in one organ and mature miRNA in the other organ is high (e.g. pre-miRNA in shoots correlates with mature miRNA in roots), we will infer transport of a miRNA or its precursor from one organ to the other (Fig. 2). We will use both simple correlation and lagged correlation, because if the time of inter-organ transport is much shorter than the time interval in our time-series, correlation will reveal such relationship; if the time of transport between organs is comparable to our time-interval, lagged correlation will be used. In our study, we will focus on traveling miRNAs that differentially accumulate in the heterogeneous N-treated plants compared to the control (*e.g.* Sp.KNO3 vs. C.KNO3), because they possibly mediate systemic N-signaling. However, such methods of detecting traveling miRNAs based on correlation or lagged correlation across organs can be applied as a general method for mathematically detecting long distance traveling miRNAs for any signal in any species.

Finally, genes and miRNAs identified above can be used to construct a systemic-N responsive network to refine the GWAS in Aim S1. To do this, we will query the identified genes and miRNAs against the Multinetwork (*[4](#_ENREF_4" \o "Gutierrez, 2008 #205)*) to construct the connections among these genes, and between these genes and their interaction partners, via protein-protein, protein-DNA, miRNA-RNA interactions ([*4*](#_ENREF_4)), in addition to correlation connections. A software platform, VirtualPlant, was developed in our lab and is being improved and expanded to support such network analysis for our lab and for the research community ([*5*](#_ENREF_5)).



**Fig. 1. GWAS refined by “network knowledge” and combinatorial traits uncovers candidate NUE genes.**

******

**Fig. 2 Detecting traveling miRNAs based on analysis of space and time data.**

***Reference***

1. S. Ruffel *et al.*, Nitrogen economics of root foraging: transitive closure of the nitrate-cytokinin relay and distinct systemic signaling for N supply vs. demand. *Proc Natl Acad Sci U S A* **108**, 18524 (Nov 8, 2011).

2. T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature* **461**, 747 (Oct 8, 2009).

3. M. Nordborg *et al.*, The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol* **3**, e196 (Jul, 2005).

4. R. A. Gutierrez *et al.*, Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1. *Proc Natl Acad Sci U S A* **105**, 4939 (Mar 25, 2008).

5. M. S. Katari *et al.*, VirtualPlant: a software platform to support systems biology research. *Plant Physiol* **152**, 500 (Feb, 2010).

6. C. A. Brosnan, O. Voinnet, Cell-to-cell and long-distance siRNA movement in plants: mechanisms and biological implications. *Curr Opin Plant Biol* **14**, 580 (Oct, 2011).

7. A. Buhtz, J. Pieritz, F. Springer, J. Kehr, Phloem small RNAs, nutrient stress responses, and systemic mobility. *BMC Plant Biol* **10**, 64 (2010).

8. K. Furuta, R. Lichtenberger, Y. Helariutta, The role of mobile small RNA species during root growth and development. *Curr Opin Cell Biol* **24**, 211 (Apr, 2012).

9. A. Molnar *et al.*, Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. *Science* **328**, 872 (May 14, 2010).

10. B. C. Yoo *et al.*, A systemic small RNA signaling system in plants. *Plant Cell* **16**, 1979 (Aug, 2004).