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

Dates:

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

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

* **Supplement Request**
	+ \*Summary of Proposed Work
	+ \*Justification for Supplement
	+ \*Budgets (NSF 1030s, budget justification and WHOI budget)
	+ Supplementary Documents (if any)

**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 with 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. 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 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 apply a classical physiological “split-root system” to the genomic era [Ruffel et al 2011]. This supplement relates to two computational approaches to analyze and integrate the large amount of data we have generated in this project. This data includes transcriptome data (mRNA and small RNAs) , SNP data, and … By contrast, the GWAS data has the potential to identify genes associated with NUE regardless of the. 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), and *vice versa*. Bridging the gap between Systems Biology and Genome Wide-Association Studies, 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 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” [Manolio et al 2009]. 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 [Nordberg 2005], yet most GWAS studies measure traits on under 100 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. 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 coming year, we will use 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 studies 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 X 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 y, (mean(x) – std(x)) > mean(all ecotypes)) or (2) that the mean for that trait plus the standard deviation for that gene is less than the global mean (i.e. (mean(x) + std(x)) < mean(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.

(Dennis: I like the idea, but this way we also eliminate the ecotypes that have the mediocre phenotype value (the ones very close to mean of all ecotypes), but have tight distribution within the ecotype. I am concern this might hurt the GWAS because we are eliminating the middle part of the continuous trait value, and reducing the ecotype numbers too much by removing the ones with small stds but mediocre mean value. Also I changed the origanl “genes” to “traits”, please see if it makes sense. ---Ying)

1. ***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 that include protein-protein, protein-DNA, miRNA-RNA connections [Gutierrez et al 2008], and 2. A systemic-N responsive network [Ruffel et al 2011] which will be improved using the identified genes from Aim S2 taking the multinetwork approach to connect the differentially regulated genes to their network partners by protein-protein, protein-DNA, miRNA-RNA connections [Gutierrez et al 2008], in addition to expression correlation. 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 (), [the equation is incomprehensible to me] 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).
2. ***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. So, for example if trait T1 and T2 exist as traits then so do 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. In the current year, we will start with these simple paired traits and three -way traits (Fig. 1A, x-axis). Based on these results, we may also explore whether parametric simulations are effective for improving GWAS results. (Dennis, what this means is that we may want to try alpha\*T1/T2 instead of T1/T2, while alpha is a range of parameters we test; Ulises thinks this might leads to more GWAS hits. Is “parametric simulation” a good word for this?)

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 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 rate) 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 space). 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.

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

(Dennis seems to be moving only on X-axis, I expand the logic he presents to both X-axis and Y-axis. ----Ying)

(Dennis, another question I have is that the method sounds a lot like Rank Product. Is there a difference between using sum and using product? For example, if using sum, SNP1 ranked as (2,4,3) and SNP2 ranked as (1,1,7) will have the same global rank, but they will have different rank if using product, in which case SNP2 will rank higher than SNP1. Of course we can find oppisite case that SNP1 and SNP2 will have the same global rank using product but different global rank if using sum. Does it matter? Just curious.)

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 traits we will randomly and independently permute the SNP rank corresponding to each GWAS, then compute the global ranks. This will give a 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 which mediate shoots-roots systemic signaling (Fig. 2)

**The data**: We have transcriptome data (RNAseq for mRNA and small RNA) from shoots and roots of split-root plants exposed to three N-treatments: Control-N roots (both root halves have adequate nitrogen); Control-KCl roots (both root halves are nitrogen deprived) and Heterogenous-N roots (one root ½ is in adequate nitrogen 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 3-way ANOVA model with nitrogen, split-condition, and time as described previously [Ruffel et al., 2011]. 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): [again, the equation is hard to read]



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



A higher ratio in either case suggests the gene is over-expressed in the foraging portion of the split-root system (spKNO3) 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 spKNO3 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 this product, reflecting the most deviation from the control values. (Dennis- how do we deal with replicates?) [it doesn’t matter too much as we just want genes that have low p-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 *PR*. 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:



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 [Ying add ref]. 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 miRNA, if the correlation between pre-miRNA and mature miRNA in different organs is high (e.g. pre-miRNA in shoots correlates with mature miRNA in roots), we infer transportation of miRNA or its precursor from shoots to roots (Fig. 2). We will use both simple correlation and lagged correlation, because if the time of transportation is much shorter than the time interval in our time series, correlation will reveal such relationship; if the time of transportation 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 split-root 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 in any species.

Finally, genes and miRNAs identified above can be used to construct a systemic-N responsive network taking a multinetwork approach to connect these differentially regulated genes to their network partners by protein-protein, protein-DNA, miRNA-RNA interaction [Gutierrez et al 2008], in addition to expression correlation. Such a network will fuel the GWAS in Aim S1.