Some Text and Equations for Our Grant

Our main goal in this supplement is to find genes that influence nitrogen use efficiency. This supplement focuses on computational analysis on our new experimental data, previous experiments, and known pre-cursors to omic interactions such as binding site information. Our methodology should apply both to new data types (e.g. mass spec data) as well as to the analysis of new functionality (i.e. non-nitrogen) on different species.

**GWAS**:

One problem with GWAS is that the set of SNPs can be much larger than the relatively small set of ecotypes that are available. For example, in Arabidopsis, there are <INSERT NUMBER HERE> SNPs. We have only <INSERT NUMBER HERE> ecotypes.

A second problem is that for some ecotypes, the phenotype of interest may be weakly present. For other ecotypes, the level of the phenotype may differ in different individuals. So the effective number of ecotypes is reduced.

A third problem is that the phenotypes may not be simple. For example instead of looking at lateral root length we may be interested in the ratio of <INSERT TRAIT HERE> to <INSERT OTHER TRAIT HERE>.

We address each of these problems as follows:

1. To reduce the set of genes upon which to do GWAS, we look for the genes that are most relevant to the treatments and phenotypes at hand. Further we expand that set to take into account genes that might interact with the directly relevant genes. <INSERT METHODOLOGY FOR REFINING GENE SET HERE INCLUDING MULTINETWORK, ONE-HOP ETC>
2. To address the issue of uninformative or widely variant ecotypes, we use an ecotype only if it has either of the following two properties: its mean value for the phenotype of interest less the standard deviation of that value is greater than the global mean for that phenotype (i.e. for gene g, (mean(pheno\_g) – std(pheno\_g)) > mean(pheno for all genes)) or that the mean for that gene plus the standard deviation for that gene is less than the global mean (i.e. (mean(pheno\_g) + std(pheno\_g)) < mean(pheno for all genes). 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.
3. To address the issue of the need for complex traits, we 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 (n\*(n-1)/2) \* 4 possible pairs. Further pairwise traits can be combined with other single traits to get traits like (T1-T2)\*T3, yielding ((n\*(n-1)/2) \* 4) \* (n-2) \* 4 three-way possibilities. Biological insight may cut down these possibilities to some set of interesting complex traits. In the current year, we will stop with pairwise traits.

As usual in GWAS studies, we will treat each simple and complex trait individually. For each particular trait, individual treatment will give us a set of p-values and false discovery rate. However there may be SNPS that rank highly in several of the GWAS analysis. We will assign a global rank to a SNP as the sum of the ranks of that SNP for each GWAS analysis over all traits of interest. Then we will sort the results by global rank. For example, suppose there are two SNPS S1 and S2 and three traits. For trait 1, S1 has rank 1, S2 has rank 2; similarly for trait 2; but for trait 3, S2 has rank 1 and S1 has rank 2. Then the global rank of S1 is 4, while the global rank of S2 is 5. In sorted order S1 is ranked 1 and S2 is ranked 2. This will give a sorted global ranked order of genes whose effectiveness we can study.

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, in each of M reshuffling we will randomly and independently permute the SNP identities corresponding to each trait, then compute the global ranks. This will give a p-value and false discovery rate for each global rank using the actual data.

**Split Root Analysis**

We have data from split root plants using both roots and shoots in three conditions: roots are both in nitrogen-rich soil (control-N); roots are both in nitrogen-poor soil (control-KCL); and one root is in nitrogen-rich soil and the other in nitrogen-poor soil (mixed). We also have four time points. What we would like to do is to determine which genes act very differently in the mixed setting from the other two settings and when.

Our strategy is simple, for each gene g and each time point t, we look at the mean ratio of the expression value of g in the nitrogen-rich portion of the mixed split root compared with the expression value of g in control-N (i.e. exp(g\_nitrogen-rich-in-mixed)/exp(g\_control-N)) and compare with the expression value of g in nitrogen-poor with g in the control-KCL (i.e. exp(g\_control-KCL)/exp(g\_nitrogen-poor-in-mixed)). These would be genes that are over-expressed in the foraging portion of the split-root system and under-expressed in the KCL portion of the split-root system relative to control. Of course, the converse is also interesting: under-expressed in the foraging portion relative to control-N and over-expressed in the KCL portion of the split-root system relative to control-KCL. The figure of merit is the product of these two ratios, i.e. exp(g\_nitrogen-rich-in-mixed)/exp(g\_control-N)) \* (i.e. exp(g\_control-KCL)/exp(g\_nitrogen-poor-in-mixed)). Here we rank genes based on the absolute value of this product, reflecting the most deviation from the control values. An alternate way to rank this is based on the p-value of this ratio which can be calculated as follows: for each time point and each gene take the 12 data points from the three replicates of control-N, nitrogen-rich mixed, nitrogen-poor mixed, and control-KCL. Permute the labels and recompute absolute value of the product of the two ratios. Do this M times (where M is something like 1000). Then see how often the absolute value of the computed product is greater than the absolute value of the product computed from the means with the proper labels.