*##The write-up at this stage contains lots of details and some repetitive sentences for the purpose of exploring the ideas, hence it is not necessarily written in the grant writing format and style. Once we are fine with the ideas I can change it for the grant. –Ying Oct 15 2013*

**1. Experiment setup**

sRNA-seq and RNA-seq will be performed from the shoots and roots of split-root plants at 0hr, 1hr, 2hr, 4hr, 6hr, 8hr and 10hr after transferring to heterogeneous-N treatment plates. In details:

1. At time 0, there will be one type of shoot and one type of root.
2. For the other time points, three treatments of shoots (Control N shoots=CNS, split root shoots=SRS, control KCl shoots=CKS) and four treatments of roots (Control N roots=CNR, split N roots=SNR, split KCl roots=SKR, and control KCl roots=CKR) were collected.
3. Control N shoots and Control N roots are from the same plants.
4. split root shoots, split N roots, and split KCl roots are from the same plants.
5. control KCl shoots and control KCl roots are from the same plants.
6. Three biological replicates per type/tissue/time point.

**2. Goal** The goal is to identify the traveling signals between shoots and roots, by a modeling approach integrating across time and space.

**3. Basic reasoning** The basic flow is source organ causes certain genes to express that can either result in hormones, small RNAs, or mRNAs/protein that are in turn transported to a target organ and cause changes there.

For each gene *g* under each treatment, we measured the RNA level in the shoots and the roots in 7 time points:

Ying, I think we should skip this part and go directly to the part where we explicitly identify and then model the intermediaries. So we will have TFs in shoot to intermediaries. Then intermediaries to targets in roots. It’s better mathematically because the number of intermediaries is relatively small.

*Shoot* Expression level of gene g E(*g,S*) *= {ES1, ES2, ES3, …, ES7}*

*Root* Expression level of gene g E(*g,R*) *= {ER1, ER2, ER3, …, ER7}*

E for expression; S for shoots; R for roots, subscript number for the time point.

Conceptually, if the shoot expression level of a gene *a* lag-correlates with the root expression level of a gene *b*, it possibly implies a shoot-to-root messenger generated downstream of A to affect B. Obviously, since the shoot-root traveling can only happen with in a plant, the above mentioned correlation should also be calculated within a plant, as bellows:

1. Corr (E(a,CNS),E (b,CNR))
2. Corr (E(a,CKS),E (b,CKR))
3. Corr (E(a,SRS),E (b,SNR))
4. Corr (E(a,SRS),E (b,SKR))

However, such simple correlation (or lag-correlation) approach will provide a big universe of possible Gene a-Gene b pairs, while lots of them will be generated just by chance. We address this problem in multiple ways described below.

**4. “targeted” correlation**

We proposed to focus on the “true” correlation based on what we know about the trafficking signals.

We know that the following three substrates could travel long distance:

1. Hormone (cytokinin reference[1])
2. sRNA ([2])
3. mRNA/protein ([3])

We will do the following analysis specifically for each type of long distance signal, within each experimental condition (homogeneous N, homogeneous KCl, heterogeneous):

1. Correlation of TFs to intermediary genes within source organs.

2. lagged correlation between intermediary genes in one organ to the predicted targets in the other organ.



***4.1 Hormones***

Because we can’t measure hormones directly, we will measure the genes that synthesize them.

Within the source organ (e.g. shoots), we will first find correlation between TF x and the hormone synthesis genes y:

Corr (E(x,S),E (y,S))

The AGRIS binding site information can be used to restrict the correlation. Such TF-gene pairs provide us information about how the long distance signals (hormone) is being regulated in the source organ.

Then, lagged correlation between the hormones synthesis gene y in the source organ will be calculated with known target z of specific hormone (Chory data) in the target organ:

Corr (E(y,S),E (z,R))

A strong correlation from such pair will indicate the hormone as traveling signal. To make sure the change of hormone responsive gene z is not due to local hormone synthesis, we also calculate

Corr (E(y,R),E (z,R))

and require a poor correlation is observed for the with-in target organ correlation.

***4.2 sRNA***

Within the source organ (e.g. shoots), we will first find correlation between TF x and the sRNA precursor genes (can be miRNA coding gene or known tasiRNA generating transcripts) y:

Corr (E(x,S),E (y,S))

The AGRIS binding site information can be used to restrict the correlation. Such TF-gene pairs provide us ideas about how the long distance signals (sRNA) is being regulated in the source organ.

Then, lagged correlation between the sRNA gene y in the source organ will be calculated with the abundance of the sRNA p, and predicted target q in the target organ:

Corr (E(y,S),E (p,R))

Corr (E(y,S),E (q,R))

A strong correlation here will indicate the sRNA as traveling signal.

To make sure the change of sRNA accumulation, and the target gene change is *not* due to local sRNA synthesis, we also calculate

Corr (E(y,R),E (p,R))

Corr (E(y,R),E (q,R))

and require a poor correlation is observed for the with-in target organ correlation.

***4.3 mRNA/protein***

There are reported cases that mRNA or protein can travel in the phloem from shoots to roots. To capture those, we again use lag correlation, but heavily reply on the known gene-to-gene interactions to help us filter out the spurious correlation.

Lagged correlation between any gene a in the source organ with any gene b in the target organ will be calculated, for example:

Corr (E(a,S),E (b,R))

A strong correlation of such pair will indicate the mRNA or protein of gene a as traveling signal.

We will further filter these gene a-b pairs with known gene-to-gene interaction based on literature, AGRIS binding site information and experimental/predicted protein-protein interaction from the multi-network knowledge database.

Within the source organ (e.g. shoots), we can also find correlation between TF x and the gene a:

Corr (E(x,S),E (a,S))

The AGRIS binding site information can be used to restrict this correlation. Such TF-gene pairs provide us ideas about how the long distance signals synthesis gene is being regulated in the source organ.

**5. Another filter to help identify changes caused by shoots-to-roots signals**

We will also use the inducible phloem-blocking line (pSUC2:iCalms) to get with-in root activity (without influence from the shoots). So, if there is a local (when the phloem blocker is deployed) expression pattern of some target gene g, then seeing that same expression pattern when there is no phloem blocker (i.e. wild type) would suggest that shoot-to-root transport is not causing the expression of gene g.

Conversely, if some gene’s expression changes in the wild type roots, but not in the roots of the phloem-blocked line, then a shoot-to-root transport of certain signal is the cause. At that point we can look at correlation of that gene’s expression with the transported elements within each experimental setting as well as across settings to see how robust that relationship is and what are the most likely correlating elements.

**6. Linear regression or non-linear terms.**

Wherever the word correlation is present from ***session 3-5***, we could use a linear regression and even introduce non-linear terms instead. However, the issue is that linear regression introduces many possible causal explanations for some target's behavior and therefore, even if a relationship is found, it might be caused simply by chance.

For this reason, it’s a better idea that we adopt these linear and non-linear regression approaches only when we have few possible influencers for a given target (*e.g.* under 15) from intermediaries that either are known to bind to a target or are highly correlated either positively or negatively with the target from the correlation (and lag-correlation) based method described in ***session 3-5***. If we happen to find more than 15 distant influencers for a given target, we can iterate on the regression, and take the highest 15 influencers based on their “weight” given by the regression and then redo the regression..

In detail, we will put experimental data into a model that uses both Stochastic Gradient Descent and Boosted ‘Regression Trees’ to learn from experimentally observed expression pattern of the influencers and targets, to determine which influencers are predictive of the expression level change of the targets over time. The two algorithms are complementary in that regression trees are easier to interpret, but stochastic gradient descent handles interactions better.

We will model the target x as a function of the its potential influencers (y1, y2, y3, …, yn) overtime, using the observed expression level over time of x and ys:

[(E(y1,S), E(y2,S), E(y3,S), …, E(yn,S), E(x,R)]

while E(*g,S*) *= {ES1, ES2, ES3, …, ES7},* g for genes, E for expression, s for shoots; r for roots, number for time points.

The central modeling problem consists of the use of different algorithms to find, for each condition over all influencers, a single set of coefficients Ci to each influencer yi so we can obtain equations of the form:

E(x,R)= Constant+C1 \* E(y1,S)+ C2 \* E(y2,S)+ … + Cn \* E(yn,S)

We will also analyze (using both algorithms) the product terms of two influencers to identify possible combined effects and interdependencies between two influencers. For example, for interaction between influencer y1 and y2:

E(x,R)= Constant+C1 \* E(y1,S)+ C2 \* E(y2,S)+ … + Cn \* E(yn,S)+ D1,2 \* E(y1,S)\*E(y2,S)

where the coefficients D describes the impact of the combined effect of y1 and y2. If two influencers are known to interact with each other (based on multinetwork knowledge for example), we will prioritize the test of the interaction between them.

***Validation and performance assessment:*** Model predictions will be validated by ten-fold cross-validation, i.e. by repeatedly training model parameters based on randomly selected 90% of the data and testing the predictions on the remaining 10%. We will derive estimates of prediction confidence and error, e.g. the relative error as (predicted-observed)/observed.

**7. How to use treatment information.**

We can determine which genes’ expressions depend on treatment settings and how. We can identify traveling signals that are triggered only by heterogeneous-N treatment, only by homogeneous KCl, or only by homogeneous N condition, etc.

Hence, the treatments information can be used to identify the traveling signal unique to systemic signal (differ between CKNO3 and spKNO3, or differ between CKCl and spKCl), or traveling signals triggered by the homogeneous-N supply (differ between CKNO3 and CKCl).