***Aim 3. A modeling approach to identify the shoot-root long distance signals.***

***Rationale***. The goal of this aim is to identify the long-distance signals that travel from root-to-shoot, and shoot-to-root, in a root-shoot-root relay of systemic N-signaling. We will first use the space and time data from Aim 1 to identify the inter-organ correlation of causal and target gene pairs (between roots and shoots), which will be indicative of a long-distance signal. Specifically, we will identify gene pairs for which a causal gene in one organ (e.g. shoots) and target genes in the distal organ (e.g. roots), show a strong inter-organ correlation or lagged-correlation over time. We will then integrate such causal/target gene pairs into a predictive model, and use left-out data for *in silico* validation (Aim 3A). For long-distance signals suggested by the predictive modeling approach described in Aim 3A, we will focus on those relevant to the heterogeneous N-response, as identified in Aim 3B. Finally in Aim 3C, we will integrate the time/space data from Aim 1 and Aim 3A-B, with the phloem-specific data from Aim 2, using a Rank-Sum method to prioritize long-distance signals for testing in Aim 4.

**Aim 3A. To identify potential long-distance signals involved in response to a heterogeneous N-environment by integrating time and space data from Aim 1.**

**Known long-distance signals in plants.** Our first goal is to identify potential long-distance signals based on RNA data collected over space (organ) and time in Aim 1. We anticipate a long-distance N-signal to have the following properties: 1. A verified role in regulation of N-assimilation and 2. The ability to be transported across organs. For these reasons, the signal types we will focus our study are: hormones, small RNAs, and mRNA/Proteins (specifically TFs) (Fig. 6B). Hormones have been long known to travel across organs and they play a well-known role in the N-response [56]. In addition, we have previously shown a role for cytokinin in the signaling of systemic N-demand [1]. sRNAs (siRNA and miRNAs) are known to travel in the phloem [57]. Moreover, grafting experiments have shown that miRNAs act as long-distance signals of phosphate starvation [58]. We have also previously shown that miRNAs mediate nitrate control of root development [2,6] and

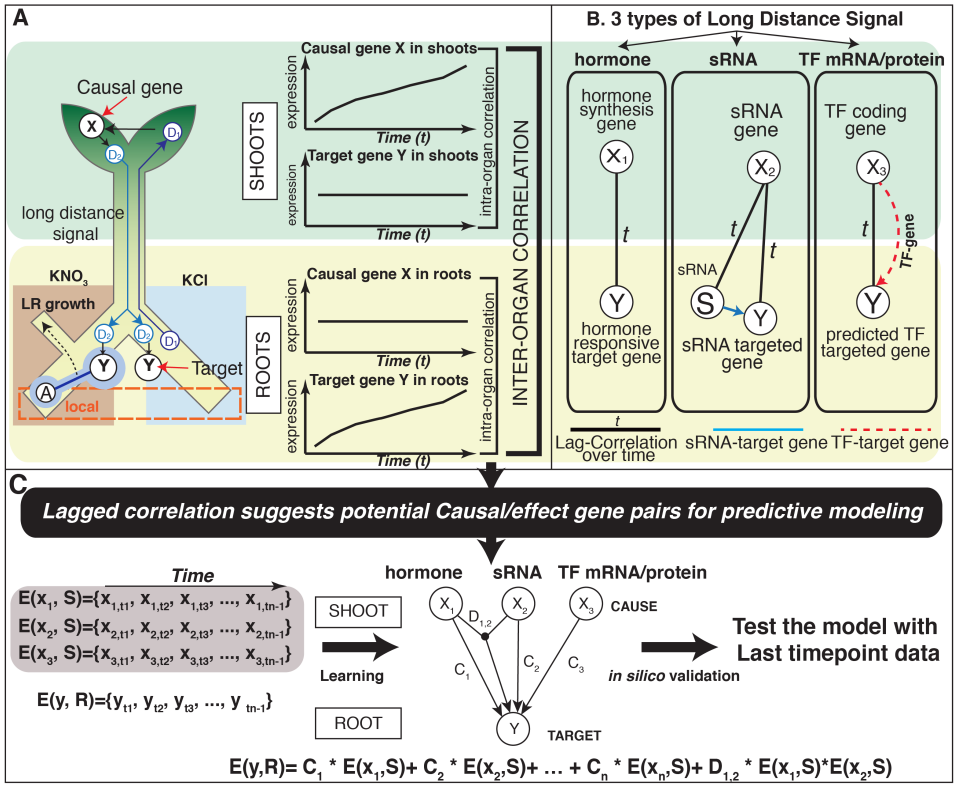


**Fig. 1 Modeling causal/target gene pairs will identify potential long-distance signals of systemic N-supply or demand.** (A) Inter-organ (lagged-)correlation identifies causal/target gene pairs (x and y) as indicators of (B) long-distance signals including hormones, sRNAs and TF mRNA/protein. For sRNA, note that the correlation of the sRNA precursor in the source organ and mature sRNA in the distal organ is also used to indicate the movement of the sRNA. (C) Modeling of causal/effector genes using time-series data and regression analysis identifies predictive models for systemic N-signals. Note: The example of shoot-to-root is shown here and the same logic is used for root-to-shoot signaling.

metabolism [3]. Thus, we will consider sRNAs as potential long-distance signals. Finally, mRNA and proteins especially transcription factors (TFs) are known to travel in the phloem [59]. Thus, we will consider the possibility that mRNAs or their encoded TFs may act as long-distance signals in root-shoot-root communication.

**Identifying causal/target "gene pair" read-outs as indicators of long-distance N-signaling.** To identify the potential long-distance signal(s) involved in systemic N-signaling, we will examine temporal expression of potential causal and target gene pairs across organ-types (Fig. 6A). Such gene pairs will be the molecular "read out" of a potential long-distance signal. For example, for hormone signals, the causal gene will be a hormone synthesis gene in one organ, and the target gene will be a hormone responsive gene [60] in the distal organ (Fig. 6B). For long-distance sRNA signals, the causal gene is a miRNA coding gene or siRNA generating locus, and the target will be computationally predicted or experimentally validated sRNA targets [61] (Fig. 6B). For traveling mRNA/proteins [62], we will specifically focus on TFs, as these regulators have been shown to travel in the phloem [63] to regulate developmental processes. For traveling TFs, the potential target gene can be identified using the TF-target database AGRIS [64] (Fig. 6B).

Next, ***correlation and lagged correlation*** between each potential causal gene *x* in the one organ (e.g. shoots), and each potential target gene *y* in the distal organ (e.g. roots) over time will be calculated, because such correlation indicates a long-distance signal downstream of *x* to affect *y*. If the expression (E) of a gene (g) in an organ (O) over time is denoted as $E(g,O) = \{E_{t1}, E_{t2}, E_{t3}, \ldots, E_{tn}\}$ (tn=time point). The correlation will be calculated as ***Corr(E(x,S), E(y,R))*** (S=shoots and R=roots) for shoot-to-root signal, or ***Corr(E(x,R), E(y,S))*** for root-to-shoot signal. Lagged correlation will also be calculated as the correlation between the shoot and root time-series data shifted in time relative to one another. We will evaluate a p-value of the correlation [65] and then compute a false discovery rate (FDR). A strong and significant (lag) correlation (FDR cutoff determined by a silhouette plot [66]) in such a causal/target gene pair over time, and across organ types, will be a strong indicator that the relevant signal indeed travels between one organ (e.g. shoots) and a distal organ (e.g. roots) (Fig. 6A). Meanwhile, to ensure the change of target gene y is not due to a local effect of the N-signal, we will also calculate intra-organ correlation of the causal/target gene pairs e.g. ***Corr (E(x,R), E(y,R)***) in the target organ, and consider cases where the intra-organ correlation << inter-organ correlation (Fig. 6A). This will identify potential gene pairs read outs that are indicative of the long-distance signal for the predictive modeling described below.

**Predictive modeling of the effect of long-distance signals on target genes**.
The above correlation analysis will suggest a restricted set of possible causal genes which effect long-distance regulation of a target gene. Since correlation alone doesn't support cause-effect relationship, and some genes can be regulated by multiple factors, we will next perform an integrated analysis across one to multiple (e.g. up to 15) causal genes to model their effect on the target gene using *Stochastic Gradient Descent* and *Boosted 'Regression Trees'* methods (Fig. 6C). These two machine-learning algorithms are complementary in that regression trees are easier to interpret, but stochastic gradient descent handles interactions better [67]. This analysis will allow us to:

- Predict cause and effect through a modeling approach that can be validated *in silico*.
- Test interactions between different causal genes on the same target gene.
- Integrate effects across the three major signal types (hormones, sRNAs, TFs).

The last point is important, because we have previously documented that interactions between hormones, sRNAs and TFs effect changes in root morphology in response to N-signals, so there is precedence for this type of signaling interaction [3,6].

In detail, we will use the observed dynamic expression levels of the causal genes $x_1, x_2, x_3, \ldots, x_n$ in the source organ (e.g. shoots) and the observed dynamic expression level of target

gene y in the recipient organ (e.g. roots), to model the target y as a function of its potential long-distance causal genes xs over time. The central modeling problem consists of the use of both algorithms to find a set of coefficients Ci to each causal gene $x_i$ so we can obtain equations of the form:

$$E(y,R) = Constant + C_1 \times E(X_1,S) + C_2 \times E(X_2,S) + \cdots + C_n \times E(X_n,S)$$

We will also analyze the product terms of two causal genes to identify possible combined effects and interdependencies between them. For example, the coefficients $D_{1,2}$ describes the impact of the interaction effect of $x_1$ and $x_2$ as:

$$E(y,R) = Constant + C_1 \times E(X_1,S) + C_2 \times E(X_2,S) + \cdots + C_n \times E(X_n,S) + D_{1,2} \times E(X_1,S) \times E(X_2,S)$$

Model predictions will be validated by a leave-out test, where one or more later time points are left out of the training data and then predicted using the model [29]. We will derive estimates of quality of the prediction by evaluating the mean square error (MSE) of the predictions [67] done using the model, compared with a prediction of no change in expression, and compared with a prediction of the continuation of a trend in the expression of each target gene (e.g. an auto-regression prediction). A low MSE indicates that the one or more causal genes in the learned model can be used to successfully predict the expression of the target gene y in a distal organ in later time-points. Biologically it means that the one or more causal genes plausibly influence the target gene cross-organ. The long-distance signals (hormone, sRNA, or TF) generated by these causal genes thus likely travel from one organ to the distal organ.

**Aim 3B. Identifying systemic signals relevant to the heterogeneous N-response.** The above machine-learning analysis will be used to infer the traveling signals in plants exposed to several distinct N-environments: 1. Plants exposed to a heterogeneous N-environment ($Sp.KNO_3$ and Sp.KCl), 2. Control plants exposed to a homogenous N-replete environment ($C.KNO_3$), or homogeneous N-deplete environment (C.KCl). We anticipate that many signals will be found to travel, and only a subset of them will be involved in long-distance systemic N-signaling. To identify those, we will do the following comparisons to distinguish between two hypothetical models for a heterogeneous N-response.

**Model 1: A unified "N-imbalance" signal for sensing heterogeneous N-supply**. In model one, we hypothesize that there is a unique systemic signal sensing the *nitrogen imbalance* that is only observed in plants exposed to heterogeneous N-environments (Fig. 7A). This "N-imbalance signal" in plants exposed to a heterogeneous N-environment, would trigger root growth on the N-replete side ($Sp.KNO_3$). To identify this signal, we will examine long-distance signals unique to $Sp.KNO_3$/Sp.KCl plants, but absent in $C.KNO_3$ and C.KCl plants, and subject it to Aim 3C for prioritizing for testing in Aim 4. While we will test such a formal possibility, the current data does not support a unified N-imbalance signal, since N-demand and N-supply are genetically distinct signals [1]. We showed using a cytokinin synthesis mutant, that the N-demand signal is dependent on cytokinin, while the N-supply signal is not [1], implying there are two separate systemic signals for N-supply and demand.

**Model 2: A specific combination of long-distance signals and local response genes triggers root responses to heterogeneous N-supply.** This model posits that the long-distance signals for N-supply ($S_1$ and $S_2$) or N-demand ($D_1$ and $D_2$) occur in plants exposed to *either* heterogeneous *or* homogeneous N-conditions (Fig. 7B). In this model, we posit that it is the *interaction* of these long-distance signals, with a response to local N-status in the heterogeneous roots (e.g. genes A or B in Fig. 7B) that triggers the root growth on the N-replete side (which has a unique $D_2$+ gene A combination), and represses root growth on the N-deplete side (which has a unique $S_2$+ gene B combination) (Fig. 7B). Therefore, we will identify the N-supply signal as the long-distance signals shared by C.$KNO_3$, Sp.$KNO_3$ and Sp.KCl, but absent in C.KCl. Similarly, we will identify the N-demand signal as the long-distance signals shared by C.KCl and split-root plants (Sp.KCl and Sp.$KNO_3$), but absent in C.$KNO_3$ plants (Fig. 7B). Next, we aim to detect the specific combination of long-distance signals and local-response genes (determined by ANOVA analysis in Aim 1) that triggers the root response to heterogeneous N-supply. For example, our studies suggest that cytokinin is the long distance signal for N-demand [1]. Thus, cytokinin traveling from shoots-to-roots may induce expression of a cytokinin-responsive gene Y in both root ½'s (Fig. 6A). Then, the interaction of gene Y product with gene A, which is responsive to local N-status, would specifically induce lateral root growth in the N-replete root ½ (Fig. 6A). To form hypothesis about such interactions, we will detect the interaction between each target gene Y of the potential N-supply or N-demand signal, with each local N-responsive gene A or B, using the multinetwork knowledge which documents all gene-to-gene interactions in Arabidopsis [26]. The Arabidopsis multinetwork includes all known protein-protein interactions, miRNA-RNA interactions, TF-DNA interactions, and literature-based interactions [26]. The potential N-supply/N-demand signals predicted to interact with a local N-response gene, will be used in Aim 3C to prioritize candidates for *in planta* testing. The above analyses will allow us to focus on the systemic signals (identified in Aim 3A) that are specifically relevant to the heterogeneous N-response, and to identify their potential interacting partners that respond to local N-signals.

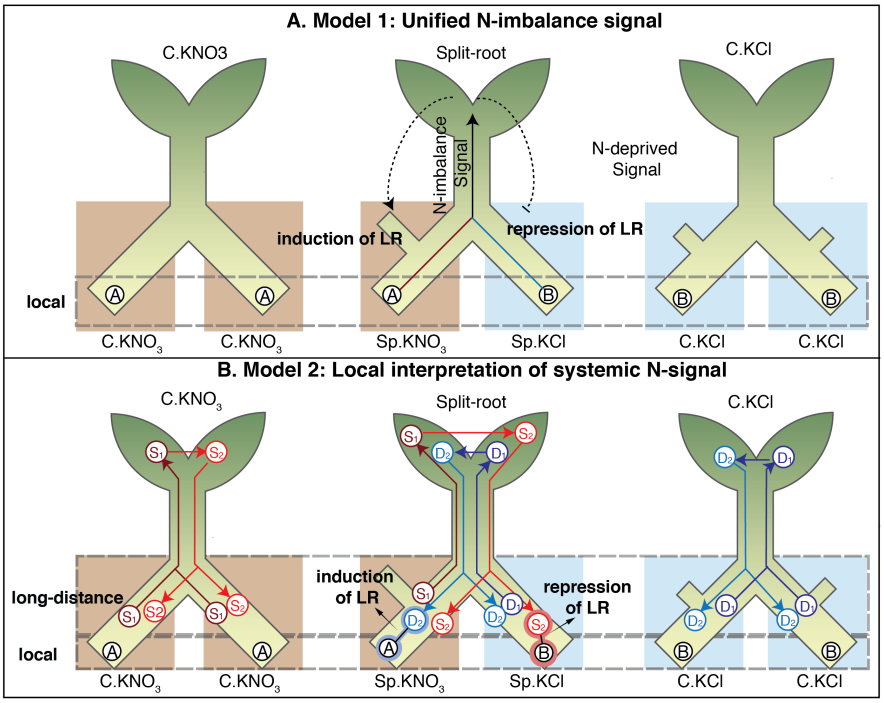


**Fig. 2. Models for root response to systemic signals of N-supply or demand.** A. Model 1 posits a unified signal for N-imbalance in Sp.$KNO_3$/Sp.KCl plants, compared to controls exposed to homogeneous N-environments (C.KCl or C.$KNO_3$). B. Model 2 posits that systemic signals of N-supply ($S_1$,$S_2$) and N-demand ($D_1$,$D_2$) are common to the heterogeneous-N and control plants, while the unique interaction of these systemic signals with local N-response genes (A or B) that causes induction of lateral roots in Sp.$KNO_3$ (D2 and A) and repression of lateral roots in Sp.KCl (S2 and B).

**Aim 3C. Prioritizing long-distance signals for experimental validation in mediating responses to a heterogeneous N-environment.**

In this aim, we prioritize the long-distance signals for testing by combining confidence and relevance measures from analysis in Aims 1-3, to rank signals for experimental validation (in Aim 4). This modeling combines the results of three analyses:

- In Aim 3A, we used correlation and machine learning to identify the potential traveling signals (technically their causal and target genes) using data from Aim 1.
- In Aim 3B, we determined the traveling signals that are specifically associated with a heterogeneous N-environment.
- In Aim 2, we physically capture the sRNAs and mRNAs traveling in the phloem.

Experimentally, Aim 1 and Aim 3 has the advantage of suggesting long-distance signals of a wide range of chemical nature, including hormones, sRNA/RNAs and protein. Aim 2 on the other hand, has the advantage that it physically captures the phloem-trafficking signal. Thus, integrating these complementary datasets will help us to prioritize signals for testing.

To do this, we will use a Rank-Sum method to rank all the potential long-distance signals based on their overall "goodness" determined by the two distinct experimental/computational approaches. Technically, the long-distance signals suggested by the modeling approaches using causal and effect gene pairs (Aim 3A), will be first filtered by their relevance to the heterogeneous N-response as described in Aim 3B, and then ranked based on the MSE of the regression model in Aim 3A (Table 1, (1)). The trafficking signals suggested by phloem-data in Aim 2 will be ranked by whether they are (2) responsive to nitrogen (measured by the p-val of the N-factor calculated in Aim 2), and (3) responsive to heterogeneous N-environment (measured by the p-val of the Het-factor calculated in Aim 2) (Table 1). The ranks of (1)-(3) will be summed up to produce a Sum-of-rank, generating a sorted global ranked order of long-distance signals whose effectiveness we can prioritize for testing in Aim 4 (Table 1).

To estimate the p-value and therefore the false discovery rate of the resulting sorted global rank, we will estimate the p-value of each global rank value by non-parametric reshuffling [65]. That is, for each of the 3 ranking factors, we will randomly and independently permute the signal rank, 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. (The null hypothesis would be that the three ranking methods are independent.) The signals will be further ranked by FDR and the top 5-10 will be subject to validation testing in Aim 4.

| ***Rank Annotation*** | ***Rank (1)*** | ***Rank (2)*** | ***Rank (3)*** | ***Rank Sum Analysis*** | |
|---|---|---|---|---|---|
| ***Source Data*** | Aim1. Time/space RNA data of Split-root system | Aim 2. Phloem-specific RNA capture | | | |
| ***Biological meaning*** | Likelihood to travel inter-organ | Response to Heterogeneous Environment | Response to N | | |
| ***Ranked by*** | Aim 3A. MSE of predictive model | Aim 2. *P-val* of Het factor in two–way ANOVA | Aim 2. *P-val* of N factor in two–way ANOVA | | |
| ***Rank Example*** | ***Rank (1)*** | ***Rank (2)*** | ***Rank (3)*** | ***Sum of Rank*** | ***Global Rank*** |
| ***Signal 1*** | **1** | **2** | **1** | **4** | **1** |
| ***Signal 2*** | **2** | **4** | **2** | **8** | **2** |

**Table 1. Rank-Sum method to prioritize candidate long-distance signals.**

**Expected Outcomes and alternative approaches for Aim 3**: Aim 3 will specifically test Hypothesis 2, to find the long-distance signal from modeling, and **Hypothesis 3. Root Response**: that specific combination of long-distance signals *and* the local N-response genes in roots is the trigger for root N-foraging in a heterogeneous N-environment. The modeling approach in Aim 3A will yield a list of potential shoot-to-root and root-to-shoot signals, whose relevance to N-Supply and N-Demand involved in heterogeneous N-response will be determined in Aim 3B. Spurious correlation could be a problem in such approaches, so we address this as follows: While the

currently proposed 7 time-points, combined with a focused space of genes of interest (e.g. hormone regulated or hormone biosynthesis genes) should allow us to overcome such problem, we will add 2 or 3 more times points if needed. Those will be added at time-points where sentinel genes to heterogeneous N-environment are shown to vary [1]. Such results from Aim 3A and B will be combined with the data from Aim 2 using Rank-sum method (Aim 3C) to provide a list of ranked signals for long-distance signaling of heterogeneous N-environments for *in planta* testing in Aim 4. Note, that Aim 1 detects both shoot-to-root signal and root–to-shoot signals, while Aim 2 only detects shoot-to-roots signal. Thus, the Rank Sum method will only be used to prioritize shoot-to-root signals for testing. For the putative root-to-shoot signals suggested by Aim 3 A&B analysis of the Aim 1 data, we will select 2-3 candidates for experimental testing in Aim 4, based on the modeling MSE and the biological relevance.