**PROJECT SUMMARY**

1. **Senior Personnel PI**: Gloria Coruzzi (NYU Biology)

**CoPIs**: Dennis Shasha (NYU Computer Science), Alessia Para Gallio (NYU Biology).

**Senior Personnel:** Manpreet Katari (multinetworks).

**Collaborators**: Ulises Rosas (morphometrics), Gabriel Krouk (N-physiology).

**2. Intellectual merit of proposed project (Project Summary)**

In now classic experiments on plant nutrition from 1962, Murashige and Skoog showed that specific combinations of N (nitrate) with P (phosphate) and K (potassium) could lead to an increase in biomass under low N – “for unknown reasons” {Murashige, 1962 #46}. This observation has inspired us to explore the molecular underpinnings of this “NPK effect” - by testing the hypothesis that the enhancement of biomass under low-N conditions is the result of the synergy of NPK nutrients acting at the *signaling* level. In this proposal, we aim to identify the regulatory networks underlying this “NPK signal interaction effect” by combining genomic, phenomic, and network inference approaches. These studies will uncover the genome-wide basis for the NPK enhancing effect on biomass. We seek to discover: **1.** The NPK conditions to optimize biomass. **2.** The morphological traits that are early predictors of high biomass. **3.** The molecular markers for high biomass and the genetic circuits that control them. **4.** Genes whose overexpression or disruption could lead to an increase in biomass production. Our experimental and analytical strategy is the result of a highly successful collaboration between biologists and computer scientists, and involves an iterative cycle of experimentation and computation, as follows:

**Aim 1. Generation of an NPK nutrient-to-phenotype matrix**. We will determine the effects of a systematic matrix of NPK combinations on plant growth, using morphometric analysis of seedlings, and biomass quantification of mature plants. Correlation of these datasets will be used to i) identify early morphometric predictors of biomass, and ii) select three NPK:phenotype states for comparative analysis: (1) High-N:High biomass, (2) Low-N:Low biomass, and (3) Low-N:High biomass -the case study .

**Aim 2*.* The NPK nutriome: Identification of early markers of biomass and nutrient-responsive pathways**. Using transient treatments and a developmental series, we will generate transcriptomes (“nutriomes”) from the matrix of NPK treatments from Aim 1. Correlation and possibly regression analysis will allow us to uncover early molecular markers for morphometric indicators and biomass, and to identify metabolic and cellular pathways (e.g. N, C-metabolism, photosynthesis, etc.) whose regulation is correlated with biomass production.

**Aim 3. Inference of NPK regulatory networks: Time-series omics and state-space modeling** We will generate fine-scale time-series transcriptome data from the three NPK:phenotype states determined in Aim 1 and use *machine-learning approaches* (e.g. state-space modeling) to derive causal TF regulatory networks controlling the genes and the metabolic pathways identified in Aim 2. While we have successfully used time-series data to derive causal networks {Krouk, 2010 #95}, the novel aspect of this study will be to integrate/compare such networks across multiple nutriome states and to generate predictive regulatory networks that identify central integrators of the NPK effect on quantified phenotypes (e.g. morphometrics and biomass) The result will be a *gene-to-phenotype* predictive regulatory network.

**Aim 4. Functional validation of regulatory NPK network predictions**. We will identify candidate regulators from our inferred network *in silico* and test them *in vivo* using mutants and transgenics – in which we will monitor molecular, trancriptome and phenotypic changes. The results of these validations will correct our network as well as generate new hypotheses. The idea is to iterate the systems biology cycle of experimentation and computation.

**Justification for NSF Network and Regulation goals**

This research addresses a number of the goals of the Networks and Regulation cluster:

1. Fundamental research about how cells integrate environmental signals (e.g. NPK nutrients) with their internal genetic & metabolic programs to regulate physiology & development. (Aim 1 & 2)
2. Uncovering mechanisms of signal transduction, metabolic pathways and networks. (Aim 2 & 3)
3. Development of quantitative, predictive theories of cellular function through iterative cycles of theory and experiment. (Aim 3 & 4)

**3. Broader impacts of proposed research**

1. **Applications to Agriculture**: Modification N-use efficiency in plants.
2. **Development of informatic tools**: Development of regulatory network inference approaches.
3. **Training in Systems Biology**: Postdocs & students are trained in Systems Biology by co-mentorship between biologists (Coruzzi) and Math/Computer scientists (Shasha) from The Courant Institute.
4. **Collaborations:** This project involves collaborations related to computational phenotyping (Ulises Rosas), and predictive network modeling (Shasha/LeCun).

# -------------------------------------------------------------------------------------------------------------------------------

# RESULTS OF PRIOR SUPPORT

# This application is most closely related to NSF Arabidopsis 2010: Nitrogen Networks in Plants. In this awarded NSF 2010 grant, we developed a systems approach to identify regulatory networks that coordinate regulation of plant metabolism, growth and development in response to nitrogen (N) signals. In previous cycles of this 2010 grant (NSF-IOB0519985), we generated transcriptome datasets and analyzed them in the context of an Arabidopsis multinetwork (Gutierrez et al 2007). Further network analysis and refinement, identified the first regulatory networks and components (TFs and miRNAs) that regulate metabolism (Gutierrez et al 2008, PNAS) and development (Gifford et al 2008, PNAS) in response to sensing N-signals. Our goals for the current cycle of this NSF 2010 grant (NSF MCB-0929338), exploit new technology (deep-transcriptomics) and new biology discovered during the previous cycle, namely the prevalence of the “RNA world” in controlling plant adaptation to the environment. We are using deep-Seq of RNA to determine the role of RNA-based network modules in mediating N-responses (Aim 1). We are combining this, with two experimental approaches to identify mechanisms by which N-regulates changes in root development. The first, is a unique experimental set up (split-root) designed to identify and test components involved in adapting root growth in response to local vs. systemic N-signals (Aim 2) (Ruffel et al, 2011). The second, addresses how these RNA-based network motifs and modules evolve across micro-evolutionary time, to enable populations to adapt to changes in N-nutrient acquisition in natural environments (Aim 3). To integrate the exhaustive data, models and pipeline analysis tools will be developed to encompass RNA data from wild-type, mutants and ecotypes (Aim 4). The data, approaches and tools developed in this project will enable a systems approach to answer a wide range of problems in plant biology. Below are highlights relevant to this application:

**Nitrogen Regulatory Networks I: Adaptation to changing N-environments.**

***Relevant Publication***: Ruffel *et al*., (2011) *“Nitrogen-economics of root foraging: Transitive closure of the nitrate-cytokinin relay and new systemic signals for N-supply vs. demand”.* ***PNAS*** *(In Press).*

As sessile organisms, root plasticity enables plants to forage for and acquire nutrients in a fluctuating underground environment. Here, we use genetic and genomic approaches in a “split-root” framework - in which physically isolated root systems of the same plant are challenged with different nitrogen (N) environments- to investigate how systemic signaling affects genome-wide reprogramming and root development. The integration of transcriptome and root phenotypes, enables us to identify new mechanisms underlying “N-economy” (i.e., N-supply and demand) of plants as a system. Under nitrate-limited conditions, plant roots adopt an “active-foraging strategy”, characterized by lateral root outgrowth and a shared pattern of transcriptome reprogramming, in response to either local or distal nitrate deprivation. By contrast, in nitrate-replete conditions, plant roots adopt a “dormant strategy”, characterized by a repression of lateral root outgrowth, and a shared pattern of transcriptome reprogramming, in response to either local or distal nitrate supply. Sentinel genes responding to systemic N-signaling identified by genome-wide comparisons of heterogeneous vs. homogeneous split-root N-treatments, were used to probe systemic N-responses in Arabidopsis mutants impaired in nitrate reduction, hormone synthesis, and also in decapitated plants. This combined analysis identified genetically distinct systemic signals underlying plant N-economy: (i) N-supply: a newly identified long-distance systemic signal triggered by nitrate sensing, and (ii) N-demand: experimental support for the transitive closure of a previously inferred nitrate-cytokinin shoot-root relay system that reports the nitrate demand of the whole plant, promoting a compensatory root growth in nitrate-rich patches of heterogeneous soil.

**Nitrogen Regulatory Networks II: Adaptation to N-hormone interactions.**

***Relevant Publication****: Ristova et al., (2011) “Nitrogen-hormone interactions mediate root plasticity” (In preparation).* The goal of this study is to gain insight into the systemic integration of nitrogen (N) and hormone (H) signals that affect root developmental plasticity through regulation of gene expression. We are using a combinatorial approach of five binary (+/-) signals (NO3-, NH4+, IAA, CK and ABA), which yields 32 (25) combinations, and examine their effect on root architecture and on gene expression. Our goal is to combine these two datasets - root phenotype and transcriptome - to create a genome-to-phenotype network. We have used two approaches to quantitate the root phenotypes: i) Single root trait measurements (e.g. lateral root number, size of primary root, etc) quantified using Optimas6 software (REF), and Fig. 2) and ii) A landmark-based morphometric approach called AAMToolbox (REF), a MATLAB plugin (see Fig. 3). In the AAMT toolbox morphometric method, we designated 20 landmarks to quantify the overall root system architecture. This landmark-based morphometrics approach, used in combination with Principal Component Analysis (PCA), identified four Principal Components that captured more than 90% of the variation, some of which were largely driven by particular hormone treatments (e.g. auxin) (Fig. 3). In our goal to associate genes with root traits, we have built *gene-to-phenotype* networks. We have initial results from integration of phenotype data (individual traits) and gene expression (Fig. 4), and will expand this analysis using PCs as nodes. To make the gene-to-trait network, we correlated the average gene expression with average trait values, using Pearson correlation and a cutoff of 0.7. We then used this correlation matrix to create *gene-to-root trait* network as displayed using Cytoscape (Fig. 4). This gene-to-phenotype network contains 280 unique genes associated with individual root traits. GO term analysis showed significant overrepresentation of ‘*response to auxin stimuli*’, including list of 15 genes known to affect root development and auxin signaling. To include the signals into the trait-to-gene network, we draw an edge if the gene is significantly controlled by a signal(s), as captured by the 5-way ANOVA, and a p-value cutoff of 0.001. If a gene is also significantly controlled by two signals in interaction, we represent this as a black box in the network (*not shown*). Future work will focus on identification of selected candidate genes for which interaction between the signals is highly correlated with trait(s) and a high p-value. The final step will be validation of candidate genes that modulate N/H interactions, through reverse genetic approaches.

**Networks III: Dynamic transcriptomes and Predictive Network Modeling**: The ultimate goal of systems biology is to predict how network states change under untested conditions or in response to modifications [1] (Gutierrez et al 2005). Our first step towards this goal, was the creation of an Arabidopsis multi-network where the “edges” connecting gene “nodes” are supported by metabolic, protein, RNA connections ([www.virtualplant.org](http://www.virtualplant.org)) (Katari et al 2010). By querying this multinetwork with microarray data, the subnetworks generate testable hypotheses. Validated examples include a role for the central clock gene CCA1 as a hub of an organic-N regulated network (Gutierrez et al 2008) and a miR-TF motif involved in the N-regulation of lateral root outgrowth (Gifford et al 2008). Most recently, we generated a High Resolution Dynamic Transcriptome (HRDT) and used a machine learning approach, “state space modeling”, to generate predictive network models for N-sensing which were validated *in silico* (using left out data), and experimentally (Krouk et al, 2010). In Aim 3, we describe the use of this machine-learning method in detail, as it applies to this current NSF grant application.

***(SUZAN PLEASE CHECK IF THESE GRANTS LIST NSF 2010 funding)***

***Publications and Products******:***

Widiez T, El-Kafafi ES, Girin T, Berr A, Ruffel S, Krouk G, Vayssières A, Shen WH, Coruzzi G, Gojon A, and Lepetit M (2011) “HNI9/AtIWS1-mediated systemic repression of root NO3- uptake is associated with changes in histone methylation” (2011). ***Proc. Natl. Acad. Sci. USA*** *(In Press)*

Ruffel S, Krouk G, Shasha D,Birnbaum KD, and Coruzzi GM (2011) “Nitrogen-economics of root foraging: Transitive closure of the nitrate-cytokinin relay and new systemic signals for N-supply vs. demand." ***Proc. Natl. Acad. Sci.*** (Conditionally accepted, pending revisions).

Krouk G, Ruffel S, Gutiérrez RA, Gojon A, Crawford NM, Coruzzi GM and Lacombe B. (2011) “A framework integrating plant growth with hormones and nutrients.” ***Trends in Plant Science***, 16 (4) 178-182.

Krouk, G, Mirowski, P, LeCun, Y, Shasha, D and Coruzzi, G. (2010) “Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate.” ***Genome Biology*** 11 (12), R123

Krouk G, Crawford NM, Coruzzi GM, Tsay YF (2010) “Nitrate signaling: adaptation to fluctuating environments.” ***Curr Opinion in Plant Biol***. (3): 266-73.

Ruffel S, Krouk G, Coruzzi GM (2010). “A Systems View of Responses to Nutritional Cues in Arabidopsis: Towards a Paradigm Shift for Predictive Network Modeling.” ***Plant Physiol***. 152(2): 445-52.

Krouk, G., Tranchina, D., Lejay, L., Cruikshank, A., Shasha, D., Coruzzi, G., and Gutierrez, R. (2009). “A systems approach uncovers restrictions for signal interactions regulating genome-wide responses to nutritional cues in Arabidopsis”. ***PloS Comp Biol.***Mar 5(3):e1000326. (Highly Accessed).

Gutierrez, R.A.; Coruzzi, G.M., Eds (2009) ***Plant Systems Biology***, Annual Plant Reviews; Blackwell Publishing: Oxford, UK, 2009, Vol. 35. 360 pages.

Coruzzi GM, Burga A, Katari MS, and Gutierrez RA (2009) Systems Biology: Principles and Applications in Plant Research. In *Plant Systems Biology*, Annual Plant Reviews; Blackwell Publishing: Oxford, UK, 2009, Vol. 35. Pgs 3-31. Book Chapter.

Gifford, M.L., Dean, A., Gutierrez, R.A., Coruzzi, G.M. and Birnbaum, K.D. (2008) “Cell-specific nitrogen responses mediate developmental plasticity”. ***Proc. Natl. Acad. Sci. USA***: 105, pp. 803-8. *Cited as a "must read" Factor 6 in the Faculty of 1000.*

Gutiérrez, R.A., Stokes, T.L., Thum, K., Xu, X., Obertello, M., Katari, M.S., Tanurdzic, M., Dean, A., Nero, D.C., McClung, C.R. and Coruzzi, G.M. (2008). “Systems approach identifies an organic nitrogen-responsive gene network regulated by the master clock gene CCA1”. ***Proc. Natl. Acad Sci* USA**:105, pp. 4939-44. *Cited as a "recommended" Factor 3 in the Faculty of 1000.*

**Shasha, D** is **Series Editor** of a two volume series “**Systems Biology” I & II**: on Genomics, Networks, Models, & Applications; Edited by I. Rigoutsos and G Stephanopoulos, Oxford University Press.

**Dissemination of Microarray Data:** Our microarray is disseminated to the community via our 2010 website: <http://coruzzilab.bio.nyu.edu/home/n2010>**.** We also deposit in journals and repositories including ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>); GEO (<http://www.ncbi.nlm.nih.gov/geo/>) & NASC.

**Commercial Application of N2010 project: Patents Filed.**

Serial No. 60/918,443; “Methods of affecting plant growth with microRNA” Filed: March 16, 2007.

Serial No. 60/919,818; “Methods of affecting nitrogen assimilation in plants” Filed: March 23, 2007.

**Training in Systems Biology:** Post-docs and students in this project are trained in Systems Biology by co-mentorship by faculty in Biology (Coruzzi/Birnbaum/Crawford) and Math & Computer Science (Shasha/Tranchina). **Post-docs trained:** Miriam Gifford (current Asst. Prof, U Warwick), Mariana Obertello, Gabriel Krouk, & Sandrine Ruffel. **PhDs**: Damion Nero (African American), Julie Young (African American), **MS**: Marcela Soruco (Hispanic) current in a PhD program at Brown & Alexis Dean (Science Education). **Undergrads**: David Hersh (Beckman Scholar), Giovanni Bonomo, Maricela Castillo (Hispanic), Harinder Singh, Brijesh Penugonda & David Iaea. Three UGs presented posters at NYU Undergraduate Research Symposium (April 2009) and completed Honors Theses. D. Hersh’s thesis won the university-wide Borgman prize. Four are in Medical School: Hersh (NYU), Bonomo (Stonybrook), Castillo (Cornell), Singh (SUNY). This grant supported a UCSD senior to perform research under the BISP199 program. UCSD also hosted a visiting professor from Nanjing Agricultural University (Dr. Yali Zhang).

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**RESEARCH PLAN**

**BACKGROUND, SIGNIFICANCE, AND MOTIVATION**

A key goal in modern agriculture is to engineer plants to produce higher yield (biomass or seed) with less N input. The benefits would be a reduction of ground water contamination by nitrates, and a decrease in energy and costs associated with fertilizer production. Our inspiration to tackle this issue comes from a now classic paper by Murashige and Skoog (1962) “A revised medium for rapid growth and bioassays with tobacco tissue cultures” {Murashige, 1962 #46}. This paper is often cited because it introduced the “MS growth medium” to the plant community; however, it also holds important observations concerning the plant nutrient interactions. Murahige and Skoog observed that culturing tobacco explants ?? explants?? on low N and specific concentrations of P and K in the medium achieved biomass levels comparable to high N (Fig. 4A). At one point, they state: “It may be noted that under the conditions of this experiment, the increase in K resulted in a marked increase in yield at the low N levels*- for unknown reasons*. In this experiment, the high phosphate level […] depressed growth *except* in the case of high K and low N media.”

More recent studies cited below suggested that the “unknown” molecular underpinnings of the interplay between N, P and K may result from the synergy of NPK nutrients acting at the *signaling* level. The hypothesis is supported by multi-level analysis that indicates how N, P and K are acting not only as nutrients *per se,* but also as *signals* that interact both at the cellular and at the systemic level to influence each other’s uptake and usage {Amtmann, 2005 #77; Armengaud, 2009 #80; Hirai, 2004 #89; Ho, 2010 #91; Tsay, 2011 #90}. The emerging concept of an existing crosstalk among these key nutrients NPK, suggests the intriguing hypothesis that the reprogramming of plant growth, development, and metabolism towards high-yield crops, may potentially be achieved as a result of the optimization of the signaling pathways that are involved in nutrient sensing and signaling. An understanding of the molecular mechanisms underlying the “NPK effect”, could lead to the creation of crops that require a reduced amount of fertilizers with a consequent reduction of the health and ecological impact {Kant, 2011 #88}.

In previous studies, we showed that by manipulating target genes in the N-assimilatory pathway resulted in an increase in biomass (Oliveira et al, 2002) or N-transport to seed (Lam et al., 2003). Our findings have led to gene technology transfer into crop plants, which are now in field-trials with several US and European AgBiotech companies. This proposal aims to identify regulatory genes, enzymes and pathways that are the main players in regulating biomass production under distinct NPK regimes that promote high biomass under low N conditions.

Our approach involves quantifying plant growth phenotypes using morphometrics and biomass under a systematic matrix of NPK treatments (Fig. 5) (Aim 1), thus identifying the NPK states that lead to high biomass under low N conditions. We next identify the transcriptional responses that are linked to the early markers for biomass in specific NPK conditions that optimize N-use (Aim 2). Our ultimate goal is to use predictive regulatory network models to identify genes and pathways that support optimal plant growth under specific NPK treatments of lowN-high biomass conditions (Aim 3). Manipulating these genes in transgenic plants (Aim 4) may contribute to the cause of sustainable agriculture.

**Aim 1. The nutrient-to-phenome matrix: correlation of NPK nutrient combinations to phenotypes.**

***Rationale***: Interactions among the nutrients NPK have been shown to affect growth and gene regulation, but the underlying mechanisms are unknown. In this aim, we will create and analyze a nutrient-to-phenome matrix to i) identify early morphometric seedling markers as biomass predictors, ii) correlate NPK matrix combinations with morphometric measures and biomass yields, and iii) select NPK matrix treatments for transcriptome analysis in Aim 2 and 3. As for most plants, Arabidopsis biomass is sensitive to N-limitation (REFs). We will exploit this feature to investigate the effect of a macronutrient matrix of low, intermediate and high concentrations of nitrogen (N), combined with low and/or high concentrations of potassium (K) and phosphorus (P), to select NPK matrix combinations that result in 1) High-N:High biomass, 2) Low-N:Low biomass, and 3) Low-N:High biomass for further analysis in Aims 2 and 3 (Fig. 4).

***Approach***: To begin, we will grow Arabidopsis seedlings on a complete matrix of NPK combinations (Fig. 4), and quantify the attributes that define root architecture (e.g. primary root length and lateral root number) and shoot morphology (e.g. leaf area and leaf number). In parallel, we will use the AAMT toolbox, a landmark-based morphometric method to quantify plant allometry (for roots and shoots) (REF) and apply Principal Component Analysis (PCA). For each NPK combination, we will also measure the biomass of adult plants (shoots and roots), as well as other N-use parameters (e.g. chlorophyll). The integrated analysis of these datasets (e.g. NPK matrix, morphometrics, biomass) will allow us to identify NPK treatments and growth strategies that represent adaptive responses to N-use efficiency. By examining the correlation between plant allometry and biomass, we will determine whether morphometric markers present in seedling early development could be exploited to predict biomass.

***Pilot Experiment:*** Arabidopsis (Col-0) seeds were germinated on N-free MS media (Murashige and Skoog, 1962), as well as on MS media containing increasing concentration of N (nitrate), to establish plant growth parameters on “low”, “intermediate” and “high” levels of inorganic N. Nitrate was used as the sole source of N, as it was previously shown that nitric nutrition is more effective than ammonium (NH4+) or ammonium/nitrate nutrition for sustaining Arabidopsis growth and biomass production (M’rah Helali, Nebli et al. 2010) Seedlings were grown for 6 weeks in short-day conditions (8 light-16 dark), to increase vegetative growth (biomass) and suppress flowering. The plants were grown on increasing concentrations of KNO3 (0.05–20mM), and fresh weight was determined as an index of biomass production (REF). The KNO3 concentrations that yielded low, intermediate and high amount of biomass but did not result in a nutritional stress (e.g. chlorosis, necrosis, anthocyanin accumulation, etc.) will be combined with P and K in the NPK matrix shown in Fig. 4, and described below.

**Aim 1A**. **Generation of an NPK matrix treatment: phenotype dataset**. To generate testable biological hypotheses for how N, P and K signaling interplay to regulate plant growth and biomass, we will assess the effect of a complete matrix of NPK treatments that represent all combinations of low, intermediate and high N (as derived from our pilot experiment), with low versus high P and/or K (Fig. 5B). As N is a growth limiting nutrient/signal, we will test both “low” and “intermediate” concentrations of N to determine the optimal concentration of nitrate that will allow us to observe the PK growth-enhancing effect on Arabidopsis, originally observed by Murashige and Skoog in tobacco pith (Fig. 5A) (REF). For media formulations, we will use the software tool ARS MEDIA (REF) that assists the design of experiments where multiple ion concentrations can be varied independently while keeping the proper balance of the other mineral components in the proportions determined for MS medium (REF). Arabidopsis seedlings will be grown for 9 days on vertical plates containing complete MS medium with ammonium succinate as N-source. The use of ammonium succinate in the initial growth media insures that the plants sense the nitrate as a “signal” in the subsequent NPK treatments, without being subjected to N-starvation, which would induce a stress response (Crawford, other refs). The seedlings will then be transferred to N-, P- and K-free medium, long enough to develop N, P and K deficiencies that can be detected at the molecular level. To determine the appropriate depletion time for each nutrient, sentinel genes for N, P, K starvation will be monitored in shoots and roots of plants moved to NPK free medium for 24, 48, 72 and 96 hours (REF). A suitable depletion time will be defined as the time when the expression of the starvation marker(s) increases at least 2-fold upon nutrient depletion (REF). Upon depletion, the plants will be exposed to the complete matrix of NPK combinations for 4 days, and then subjected to the detailed phenotypic quantitation described in Aim 1B. The same experimental design will also be applied to hydroponically grow plants, in order to measure root and shoot biomass production (as described in REF, Fig. X) after 6-week resupply of the selected NPK combination. The integration and analyses of these data (matrix, phenotype, biomass) will allow us to capture the variation in phenotype due to different NPK treatments and distinguish alterative growth strategies at seedling and mature stages as driven by NPK nutrient signaling.

**Aim 1B. Quantifying the NPK matrix effect on plant phenotype and biomass.** To acquire a systems-wide view of the growth strategies that are driven by the matrix of NPK nutrient/signals, we will quantify plant growth phenotype using individual trait quantification, as well as using a newer landmark-based geometric morphometric method (AAMT Toolbox) coupled with Principal Component Analysis to characterize root and shoot architecture (REFs).

***Phenotype I: Unidimensional morphometrics***. Root parameters will measured for a large number of characters including primary root length, lateral root density, total root length, etc. under all NPK treatments using the Optimas6 software for processing of digital images (REF). For shoots, we will measure basic leaf dimension parameters such as leaf length, width and area as well as leaf number.

***Phenotype II: Multivariate landmark-based morphometrics***. This analysis will be carried out using the AAMT toolbox plugin for Matlab that applies multivariant statistical analysis methods (REF). This method - originally used for face recognition - has been employed to quantify changes in shoot architecture in Arabidopsis (REF). More recently, our lab has shown that this landmark-based approach can be used to capture the geometry of roots and then convert these morphometric measurements into Principal Components (REF) (Fig. 3). When applying AAMT to object recognition, one places primary and secondary landmarks on the object to be analyzed at recognizable features. (Mention shoot and flower paper as ref.) In the case of Arabidopsis root morphometrics, the root architecture is captured by a total of 20 landmarks that include 6 primary landmarks (e.g. the root-hypocotyl junction) and 14 secondary landmarks that are regularly spaced between the primary landmarks by the AAMT software (Fig. 6) (REF for AAMT). Differences in root geometry between seedlings can be described by the difference in coordinates of corresponding landmarks. The resulting sets of coordinates are aligned (Procrustes Alignment), and subjected to principal component analysis by the AAMT software (REF). The first Principal Component (PC) accounts for as much of the variability in the data as possible, and each succeeding PC accounts for the remaining variability. The top-ranked PCs provides a quantitative measure of variation in root or shoot morphology across different NPK matrix conditions. This will allow us to assess the *quantitative allometry* of root and/or shoot differences and to synthesize the relationship between shoots and roots in a unified statistical model (Fig. 6B). This approach has been successfully used in our laboratory and others to identify new meaningful underlying variables of phenotypes (REF) (Fig. 3) (Ulises paper).

***Phenotype III: Biomass***. To quantify the effect of varying concentrations of NPK on biomass production at later developmental stages, we will measure the fresh weight of the rosettes and roots of 6 week-old plants to evaluate biomass partitioning for each NPK treatment (Fig. 7A). We will also measure soluble sugar and protein content, because of their close relationship with fresh weight (REF). For the shoot, we will also measure chlorophyll content which is an index of nitrogen use (REF). Next, we will select the NPK combinations according to the amount of biomass produced, and the level of N supplied in the growth medium, to represent distinct N-usage strategies (Fig. 7A). After measuring root and shoot biomass, we will calculate the mean for each partition -as well as for the whole plant- for each NPK combination in order to pick out the following NPK:phenotype states: 1) High-N:High biomass, 2) Low-N:Low biomass, and 3) Low-N:High biomass. Because many NPK combinations may give a Low-N:High biomass combination, we will opt for a low N combination with the lowest amount of PK nutrients. To determine that two combinations both give high biomass, we will do a non-parametric t test on the biomasses of plants resulting from each combination. If their biomasses differ by a p-value of at least 0.3, then we consider them statistically identical.

**Aim 1C. Correlation between biomass and morphometric** **analysis for early prediction of biomass production.** For eachNPK combination, the morphometric analysis will provide a quantitative measure of the effect of the matrix on the morphology of young seedlings, while biomass will represent the result of prolonged growth in the same conditions. We will use ***stochastic gradient descent*** to establish relationships between the root and shoot traits, Principal Components derived from morphometrics, and the biomass measurements, in order to identify early predictors of biomass (Fig. 7B). That is, for total biomass as well as root and shoot separately, stochastic gradient descent [refs] will identify the values of the coefficients (a1, a2, and a3 as well as b1 through b??) of a linear equation of the form

Total biomass = a1\*PC1 + a2\*PC2 + a3\*PC3 + b1\*PrimaryRootLength + b2\*LateralRootLength + c1\* Leaf area + c2 Leaf main axis+ …

Shoot biomass (?) = a1\*PC1 + a2\*PC2 + a3\*PC3 + b1\*Leaf area + b2\*Leaf main axis + b3\*Leaf number+…

Root biomass (?) = a1\*PC1 + a2\*PC2 + a3\*PC3 + b1\*PrimaryRootLength + b2\*LateralRootLength + b3\*

We will use **stochastic gradient descent** with a regularization term (to reduce the tendency to overfit) and a training coefficient **eta** (which regulates the magnitude of changes to coefficients). To discover the proper values of **eta**, we will use 10-fold cross-validation with different parameter settings. In each “fold” of cross-validation, we withhold 10% of the plants from the training set, and then evaluate our results on the left-out plants (the test plants). Different folds differ based on which plants are chosen as test and which as training. The regularization and training coefficients that offer the best results on cross-validation, will then be used on all the data to determine the coefficient values. To evaluate the confidence interval of the coefficients, we will sample the plants with replacement to determine the range of values of each coefficient. The net result of all this analysis will be the identification of the features (whether principal component or classic root/lateral root measurement) that are the best early predictors of biomass (Fig. 7B).

**Aim 2. The nutriome: Identification of early markers of biomass and NPK-responsive pathways**

***Rationale:*** In this aim, we build on the phenotype analysis from Aim 1 to explore the molecular basis of the effect(s) of the NPK matrix on biomass using transcriptome analysis. We propose to uncover the transcriptional responses to NPK signaling (e.g. the “nutriome”) in the roots and shoots of young seedlings and of plants at later developmental stages. By correlating the NPK effect on transcriptome across these developmental stages, we aim to identify early gene markers that correlate with early morphological traits associated with biomass production in mature plants (Fig. 8A). On a practical level, the discovery of such molecular predictors of biomass at the seedling stage, will represent an invaluable tool for genetic screens and field studies that aim to isolate high-yield crops. We will also analyze transcriptome data using Gene Ontology Analysis tools to identify the metabolic and cellular processes underlying biomass production (Fig. 8B). The studies in this Aim will identify the target genes and pathways that will be investigated in our time-series analysis in Aim 3, where we propose to predict the TF networks controlling the expression of the early markers as well as the NPK nutriome-responsive metabolic and cellular pathways.

***Approach:*** To generate the NPK “nutriome” datasets, we will perform transcriptome analysis on roots vs. shoots of Arabidopsis seedlings transiently treated with the complete NPK matrix from Aim 1A. For the developmental series, we will restrict our transcriptome analysis to the 3 NPK:phenotype states selected in Aim 1B, to determine how and when the “nutriome” landscapes reach equilibrium in the two organ systems. We will then integrate these nutriome datasets with morphometric parameters (from Aim 1), using parametric and non-parametric correlation. The approach will allow us to identify early molecular sentinels of biomass production and metabolic/cellular pathways that, in turn, correlate with changes in biomass. The results of this analysis will show how the execution of a specific genetic program can shape a number of cellular machines - metabolic or signaling pathways - to support growth in relation to N-availability.

**Aim 2A. The Seedling Nutriome: Identifying early molecular predictors of biomass.** In this subaim, we will measure genome-wide expression changes that arise in shoots and roots of seedlings transiently treated with all the combinations of the NPK matrix. This complete NPK matrix dataset will provide a transcriptional baseline to identify the genes that are highly induced or repressed in each NPK combination. Using these data, we will proceed to establish an association between gene expression and seedling traits that are early markers for biomass (see Aim 1B). To gather transcriptomic data, Arabidopsis seedlings will be grown and transiently depleted of NPK, under conditions determined in Aim 1. Following deprivation, seedlings will be subject to a 2-hour treatment with all NPK matrix combinations and mRNA will be extracted for transcriptome analysis from shoots and roots. As controls, seedlings that are nutrient starved, but not resupplied with any NPK combination, will be used. Trancriptome analysis will be performed using either ATH1 chips, or deep-sequencing. Both methods are currently used in our laboratory. As costs of RNA-Seq decrease –especially using pooled bar-coded samples-, we anticipate fully switching to deep-sequencing for the increased coverage of the genome. We will integrate transcriptome analysis with the root and shoot traits measured across treatments, using parametric and non-parametric correlation and correction for multiple testing {Storey, 2003 #2}. This analysis will identify genes that are significantly correlated or anti-correlated (>0.8 or <-0.8) with the morphometric root and shoot parameters. In addition, we will be able to pinpoint genetic markers for biomass production as the genes that will be found to correlate with specific morphometric root and shoot traits indicated as early predictor of biomass in Aim 1C (Fig. 8A).

**Aim 2B. The Developmental Nutriome: Identifying metabolic and cellular pathways associated with selected NPK:biomass states.** To identify the metabolic and cellular processes that correlate with the NPK effect on biomass, we will generate a developmental series of root and shoot transcriptomes for the selected NPK:phenotype states from aim 1: 1) High-N:High biomass, 2) Low-N:Low biomass, and 3) Low-N:High biomass. mRNA samples will be extracted from root and sample leaves of 2-, 4- and 6-week old plants, grown hydroponically on the three specific NPK combinations associated with the above-mentioned states to be analyzed by ATH1 chips or RNA-seq. Using correlation with correction for multiple testing as above, we will find genetic markers for biomass at each developmental time point. We expect to find both positive genetic markers (i.e. genes that are induced during growth on the NPK combination(s) that results in high biomass) and negative ones. We will use the genes that are consistently up-regulated for high biomass plants as positive genetic markers and conversely for negative markers. These positive and negative markers will serve as sentinel target genes for regulatory networks determined in Aim 3.

**Aim 2C. Identification of metabolic and cellular pathways correlated with biomass production.** We next aim to uncover the metabolic and cellular pathways that correlate with three selected NPK:Biomass states: 1) High-N:High biomass, 2) Low-N:Low biomass, and 3) Low-N:High biomass. Toward this goal, we will exploit GO term analysis tools like BioMaps and Sungear from our software platform VirtualPlant(www.virtualplant.org) {Katari, 2010 #97} to uncover significantly represented biological processes associated with biomass production over developmental time (Fig. 8A). Analogous to our correlation analysis for genetic markers (Aim 2A), we will mine our NPK nutriome matrix dataset for biological terms (GO or MIPS) that are over- and under-represented compared to their representation in the entire genome (Fig. 8B). The significance will be determined by a probability measure based on a hyper-geometric distribution. The data is presented with p-values in table format or as a directed acyclic graph enabling a view of the hierarchy of GO connections, as reported for these functions {Katari, 2010 #97}.

**Aim 3. Inference of NPK regulatory networks: Time-series omics and state-space modeling.**

***Rationale.*** The goal of this aim is to generate a predictive regulatory network model to identify the transcription factors (TFs) that control early genetic predictors of biomass (Aim 2A), as well as the TFs associated with the metabolic pathways that correlate with biomass (Aim 2B). Whereas it is possible to create models and draw association networks based on correlation and other data-driven relationships, we will pursue the creation of a causal network to determine which upstream transcription factors will have the greatest influence on gene markers for biomass. We will test this network model by its ability to predict the values of assays (e.g. expression levels) on conditions that were not used in training. Constructing the model in the form of a causal network will aid the experimental design by indicating the gene modifications that may optimize the performance of a species to some useful end, such as increasing biomass production. Because causality moves forward in time, time-series experiments are a particularly promising source of network structure, and we propose to examine NPK regulatory networks from a dynamic point of view, at very early stages of the NPK signaling cascade(s). Below, we describe our methodology in some detail, explain prior predictive modeling results {Krouk, 2010 #95}, and illustrate how we will integrate the methodology based on expression data with other existing information (e.g. validated transcription factor binding sites and structurally based contact-binding sites).

***Approach.*** We will generate high-resolution time-course transcriptomic datasets from plants transiently treated with the NPK combinations selected in Aim 2: 1) High-N:High biomass, 2) Low-N:Low biomass, 3) Low-N:High biomass. To define the scale of the time-series, we will first use Q-PCR to monitor the expression of the early gene markers of biomass identified in Aim 2 on a fine-grain time scale, to pinpoint the smallest time intervals in which the expression of these genes varies significantly. This profiling will indicate which time-points to choose for the genome-wide expression experiments that will be analyzed using a state-space machine learning algorithm (need ref for state space method REF). We will then construct a causal model based on the transcriptome data, and will test it using “leave-out-last” validation. That is, we will create a model with all but the last time point and then predict expression changes from the penultimate to the last time point as done in Krouk 2010. Such a test will allow us to estimate the predictive accuracy of our final model. We will use the resulting causal network to identify genes that have the most influence over early markers of biomass, and for metabolic pathways associated with biomass. The identified early, upstream master regulators of these genes and pathways will be prioritized for functional testing in Aim 4.

**Aim 3A. Generation of High Resolution Dynamic Transcriptome (HRDT) data.** Arabidopsis seedlings grown on complete MS media for 12 days, will be transferred to NPK-free MS medium for 24 hours, and then treated with the 3 selected NPK combinations that result in the three N-Biomass states 1) High-N:High biomass, 2) Low-N:Low biomass, 3) Low-N: High biomass (as determined in Aim 2). RNA will be collected at 0, 3, 6, 9, 12, 15, 20, 25, 40, 45, and 60 min., following transfer to NPK media. “T0” (harvest time zero, before treatment) will be used as control. To select a subset of time-points for transcriptome analysis, the early molecular markers for biomass from Aim 2, will be monitored by QPCR (Fig. X). Based on these results, selected RNA samples (in biological duplicates) will be analyzed using Affymetrix ATH1 Chips or by Illumina Deep-sequencing, depending on costs at the time of assays. We will determine which genes vary significantly over the course of the time series for at least one of the three NPK combinations chosen in Aim 2. To determine significant variation, we will take each gene and determine its variance and consider all genes that exceed a threshold. The value of that threshold will be determined by the quality of the leave-out-last test described below. DENNIS- can you clear up confusion in this section? (WHAT METHOD WILL BE USED TO DETERMINE GENES THAT ARE REGULATED BY NPK OVER TIME???????) (Gloria- IT seems to me that someone deleted the method that will be used to determine which genes are significantly regulated as a function of treatment and time????? *No. I changed this so there is no more statistics. Just the factor of 2. Please reread.*) The genes that will be found to be NPK responsive in this time-series transcriptome data will make up the nodes in the causal network generated in Aim 3B.

**Aim 3B. Predicting NPK regulatory networks using time-series data and “State Space” analysis: A machine learning approach. *State space method.*** We will use the NPK time-series transcriptome data generated in Aim 3A to support the creation of a predictive regulatory network that controls sentinel genes and pathways associated with biomass. The experimental approach using high-resolution time-series was previously exploited in our laboratory to monitor transcriptome responses to nitrate treatment {Krouk, 2010 #45}. In order to build a regulatory network that could predict TF-target interactions, we used a machine learning method, “State-Space” modeling to generate predictions for regulatory networks {Mirowski, 2009 #47}. The State-Space model synthesizes Bayesian and Markovian approaches in which each gene’s expression value at a time t is assumed to depend directly only on the state of potentially all the genes at the previous time point and indirectly on values from previous time points {Mirowski, 2009 #47;Murphy, 1999 #65}.

In the “State Space” model depicted in Fig. 10A, each node represents the values of all gene expression at a particular time point. Typical values of all gene expression are depicted as a heat map in Fig. 9. The goal of this approach is to learn the function that determines the change in expression of a target gene as a linear (or if needed non-linear) combination of the expression of a relatively small number of transcription factors (typically up to three or four). As applied to our problem, the set of all genes at time t is modeled by a “latent” (i.e. hidden) variable (denoted Z(t)) from which noisy and sometimes missing observations Y(t) are made. Latent variables are represented by large red circles, and observed variables by large black circles in Fig. 10A. The relationship between latent and observed variables is the identity function h with added Gaussian noise (represented by a black square in Fig. 10A). An unknown function f (represented by a red square in Fig. 10A) relates the values of latent variables Z(t) and Z(t+1) (for all t) corresponding to consecutive time measurements as a Markov chain. The dynamical function f factors in both transcription factors and their target genes (e.g. other TFs or target genes), as shown in Fig. 9. Learning the function *f* corresponds to finding parameters of *f* that minimize the prediction error, while penalizing functions that are excessively complex (i.e. require many transcription factors to determine the change in expression of a target). The state-space method uses an iterative procedure that attempt to learn the dynamical relationship between latent gene expression variables z(t) while maintaining the latent variables z(t) as close as possible to the observed Affymetrix measures y(t). The algorithm consists in a) minimizing the sum of quadratic errors of the dynamical and the observation models with respect to the latent variables Z by using gradient descent on the latent variables {Mirowski, 2009 #47} (this is the inference step); and in b) minimizing the sum of quadratic errors of the dynamical model using conjugate gradient, LARS {Efron, 2004 #67} or Elastic Nets {Zou, 2005 #88} optimization on the parameters of *f* (this is the learning step). During the learning step, sparse gene regulation networks are obtained by penalizing dense solutions using L1-norm regularization, which amounts to adding a λ-weighted penalty to the dynamical error term, as in the LASSO initially described by {Tibshirani, 1996 #117}.

***State-space validation.*** In our previous work, to test the ability of the “State Space” approach to generate a predictive regulatory network, we built a regulatory network using Arabidopsis time-series data up to 15 minutes (training set: 0, 3, 6, 9, 12, 15 min) and used the resulting network to predict the direction of gene change (up regulation or down regulation) from 15 min to 20 min (Fig. 10) (Krouk 2010). Our State Space predictions of gene regulation were correct for 74% of the genes in a small network of 76 genes (Fig. 10B). As a basis for comparison, the "naive trend forecast" that predicted the direction of change from 15 to 20 min to be in the same direction as the movement from 12 to 15 min, was correct for only 52% of the genes, just slightly better than random (Fig. 10C) p-value < 0.006. This “State Space” model can also be used to predict the “most influential TFs” in the network (e.g. the one that is predicted to influence the most genes in the network), and to generate a time-dependent regulatory network model for the control of N-assimilatory pathway genes {Krouk, 2010 #95}.

When compared with other network inference approaches {Bonneau, 2007 #118;Bonneau, 2006 #119;Shimamura, 2009 #121;Wang, 2006 #176} our state-space method in Krouk et al 2011, showed an improvement in accuracy and had a better signal to noise ratio when compared on the same data. Further, the method reduces the importance of initial parameters by using random starting points and bootstrapping, thus offering a principled way to deal with uncertainty and avoid over-fitting in microarray measurements. Further, our method easily allows the addition of “hints” in the form of known transcription factor-binding relationships. Finally, the state space method combined with biclustering generalizes to larger networks. In a network of 550 genes including the original 76, we ran the biclustering algorithm CMonkey {Reiss, 2006 #268} using default settings on the non-TF genes (the biclustering algorithm makes use of metabolic interactions, the Arabidopsis Prolinks file, as well as gene expression.) This resulted in a reduced network consisting of 67 TF genes and 63 biclusters among the 483 non-TF genes. On this network of 67 + 63 = 130 supernodes, the state space method predicted the direction (expression up or down) accurately for 67.7% of all supernodes in the 15-20 minute time point compared with 51.9% for the trend forecast prediction (Fig. 10BC) {Krouk, 2010 #95}.

**Aim 3C. Applying state-space method to a combined analysis of three NPK states.** We will use the time-series on the three different NPK states from Aim 2; 1) High-N:High Biomass, 2) LowN-Low Biomass, 3) Low N-High biomass, synergistically in our state-space modeling. This means that we will infer the causal function ***f***, using the experiments and replicates from all three NPK states combined. This is possible, because the state space model seeks gene-to-gene causality. If, for example, a transcription factor (tf1) induces the target g2, then we should observe this relationship even if tf1 (and therefore g2) are repressed in one NPK condition, and induced in another. Given the large amount of data that we will gather per time-point, we expect our networks to have better predictive power than those that were generated by {Krouk, 2010 #45}, which used only one N-treatment condition. We will use all possible prior information (e.g. validated TF-target pairs from Transfac (REF)) to bootstrap our approach.

The precise edge topology of this causal network will be assessed experimentally in Aim 4,by probing the relationship between specific interacting components. Forward (mutants) and reverse (overexpressors) genetics will enable us to examine the consequences of altering the levels of a selected TF on its putative targets. If predicted changes are shown to be false, we will remove the predicted TF-target edge from the network, and rerun the state space analysis without that edge, hence improving (by validating or correcting) the causal network.

**Aim 4. Functional validation of regulatory NPK network predictions.**

***Rationale.*** In this Aim, we will validate the regulatory interactions predicted from the causal networks that we generated in Aim 3. We will functionally validate the most influential TFs associated with the early molecular markers for biomass. The validation will be conducted based on a prioritized scheme and will exploit reverse and forward genetics, as well as a rapid transient expression system in root and mesophyll protoplast. The results of this aim will generate two important outcomes as they will: (i) validate the key role of TF predicted to orchestrate the genome-wide responses that influence biomass especially under low nitrogen conditions, and (ii) test the causal relationships predicted by the network by seeding a genome-scale validation of TF-targets relationships and defining new interacting regulatory partners on a systems-wide level.

***Approach***. To prioritize the TF genes for experimental validation, we will define an “influential” TF based on a) *functional category* (GO term classification); b) *network connectivity* (the TF is predicted to control several downstream targets); c) *hierarchical position* in the causal network (the TF is predicted to control a number of downstream target but is not a target itself or has few upstream regulators) and d) *temporal induction* (the expression of the TF is found to change early upon treatment), as described in Aim 4A. Functional validation, in Aim 4B, will use both forward and reverse genetics, as well as a rapid, transient assay inducible root protoplast system. As the ultimate goal is to understand the molecular underpinning of the “NPK effect” on biomass production, we will conduct our experiments on the mutants/transgenics in the 3 selected NPK combinations that determine the N: biomass phenotype states 1) High-N: High biomass, 2) Low-N: Low biomass, 3) Low-N: High biomass (Aim 4C). This approach will begin to test TF-targets predictions based on current network models and we will use the results to generate new, refined ones. The results of these analyses will enable us to address a big challenge in network biology, which is to decipher molecular cascades that lead to interactions of network motifs and compensatory changes in a network in response to genetic perturbation.

**Aim 4A. Prioritizing the most influential TFs for validation studies.** Based on our previous success in generating predictive regulatory networks {Krouk, 2010 #3}, we anticipate that the machine learning approach from Aim 3 is likely to return a large number of TFs as regulatory nodes of the causal network. To prioritize these TFs for testing, we will analyze the network properties (connectivity, hubbiness, etc.) as well as use results from the previous aims (Fig. 11A). Specifically, in Aim 2C, we will identify the functional categories that are associated with biomass production. For each module, we will focus on the hubs where a TF is connected to multiple early molecular markers of biomass identified in Aim 2, and for which the modality of the TF-target correlation displays the same sign as the marker-biomass correlation (Fig. 11A). For example, an ‘influential’ transcription factor should induce the targets that are positively associated with the predictors of biomass, and/or repress the markers that are negatively associated with the predictors of biomass. Next, we will consider the hierarchical position of the TF in the overall network (or in the module), and choose the TFs predicted to be controlled by a small number of upstream regulators (Fig.11A). Last, we will prioritize the TFs whose expression will be found to change the earliest in our fine-scale time course from Aim 3 (Fig. 11A). This will give us a priority among genes.

**Aim 4B. Functional validation by forward-and reverse-genetics and network refinement.** We will perform functional analysis of TFs in mutant backgrounds, as well as conduct constitutive and conditional overexpression studies for the most influential TFs (see below). Both approaches will be complementary for confirming the TF-target interaction. In particular, overexpression studies will allow us to overcome the issues that might occur because of gene and signal redundancy that often complicate the interpretation of the molecular phenotype. The use of a rapid, transient, protoplast-based inducible system, will also guide us in distinguishing direct from indirect interactions genome-wide, the analysis of which will drive network improvement and refinement.

To decipher the role of the most influential TFs in biomass production, we will first analyze the molecular and biomass phenotype of Arabidopsis T-DNA mutant lines for the selected TFs that will be obtained from the available collections of insertional mutants (SALK, SAIL, etc.) (REFS and web sites) (Fig.11B); for each TF, we will genotype and study at least two alleles per locus. As the selected TFs are predicted to control the expression of a set of genes representing early molecular markers of biomass production, we will first perform expression Q-PCR analysis to monitor the RNA levels of the putative target genes, and we will compare the T-DNA mutant vs. wild-type for two informative time-points (inferred from the expression time-series in Aim 3) when the expression of the TFs as well as the targets is predicted to change the most after induction with the selected NPK combinations. If a TF is found to be required for triggering the molecular phenotype associated with a specific NPK combination (e.g. the change in expression of the predicted targets), we will also perform morphometric analysis on the mutants, to quantify the effect of knocking out that particular TF on the early predictor of biomass in young seedlings and eventually, on biomass production in adult plants.

If the selected prioritized TFs belong to a large gene family whose members have redundant functions, we might not be able to fully quantify the consequences of deleting just one member. We will then address the role a TF using conditional overexpression expression using a rapid transient expression system in protoplasts (Fig. 11B), as described in Bargman and Birnbaum, . Briefly, a fusion between the TF and the glucocorticoid receptor (GR) will be expressed in protoplasts, and the timing of the binding of the TF to the promoter of the target genes will be controlled by applying DEX (dexamethasone), which mediates the translocation of the TF-GR fusion to the nucleus (Dex REF). Cycloheximide (CHX), an inhibitor to protein synthesis, will be added to distinguish direct vs. indirect targets and the use of NPK pretreatments (e.g. prior to DEX induction of the TF) will allow us to identify the direct targets of the TF when acting alone or in combination with other TFs that are regulated transcriptionally or post-translationally. If positive results will be in the transient assay system, these constructs will be transformed into wild type as well as mutant plants to create stable overexpressor lines to investigate the effect of modulating the expression of the selected TF at different developmental stages in whole plants (Fig. 11B).

As described above, these experiments will also help us refine our inferred networks by eliminating incorrect edges.

**Aim 4C. Condition-specific testing of TFs towards the creation of low-N adapted crops.** In this final subaim, we will analyze the growth phenotype of mutants and transgenics of selected TFs under specific NPK conditions. We will prioritize the TFs that will be found to be negatively or positively correlated with high biomass, respectively, and we will test their function on the NPK condition that corresponds to the Low-N:Low biomass from Aim 1 (Fig. 11C). If a TF indeed plays a crucial role in supporting growth in N-limiting conditions, we should observe an increase in biomass production compared to wild type plants that grown in the same NPK conditions (Fig. 11C). For TFs that are negatively correlated with biomass, we would expect an increase in biomass in the T-DNA mutants, while for positively correlated TFs, we would expect to observe the same outcome in their overexpressing lines. [Up to here, this is redundant to 4A] We will then perform genome-wide expression studies on the mutants that exhibit an increase in biomass upon growth on the selected low-N NPK combination(s). This analysis will corroborate our hypothesis that the effect of the TF on growth is due to the regulation of the molecular machines that appear to be play a major role according to the transcriptomics data and GO term analysis in Aim 2.

**Epilogue.** The systems approach described here will enable us to identify genes and PK conditions that will enhance the production of biomass under low nitrogen conditions. Our goal is to move up in the regulatory hierarchy to be able to modify not only single genes, but the regulation of entire pathways to effect the nutrient signaling underlies biomass production. We envision the long-term advantage of this study as the ability to target critical regulatory components to create plants that will thrive on low N, therefore yielding energy, environmental and health benefits.