**Figure legends**

Figure 1: Hierarchical clustering on the reduced Iron and Phosphate differential expression patterns of 240 differentially expressed genes. 8 clusters Cells represent log fold of differential up- or down-regulation (red and green). Asterisks indicate statistical significance (p-value < 0.05) of differential expression or repression.

Figure 2: Biological process specific gene set enrichment analysis for all 240 differentially expressed genes (a) and per cluster (b).

Figure 1: The final predicted gene regulatory network. The 10 annotated transcription factors (orange circles) regulating other transcription factors (orange circles) and other genes (gray circles). Mode of regulation between transcription factors and target genes is indicated as red (activating) and green (repressing) links.

Figure 2: Differential expression patterns of putative regulators and targets across all three experimental time series datasets. Cells represent log fold of differential up- or down-regulation (red and green). Asterisks indicate statistical significance (p-value < 0.05) of differential expression or repression.

**Results**

Iron and Phosphate treatment differential expression dataset and cluster analysis (to be edited Hatem & his colleague)

@hatem: Experimental design, growth conditions and treatments, final dataset.

@colleague: three-way anova followed by Tuk key test analysis

Given the three-way anova analysis results, we retained differential expression events (p-value < 0.05) between genes in each experimental treatment (+Ph/-Fe, -Ph /+Fe, -Ph/-Fe) per time point (3h, 6h, 9h) and the corresponding control treatments (+Ph/+Fe), resulting in a total of 240 (0.85 %) out of the 28501 genes being differentially expressed in at least one experiment and time point. These include 13 transcription factors distributed over 9 transcription factor families to be AP2/ERF (3), AUX/IAA (1), bHLH (3), MYB (2), NAC (1), B3 (1), C2H2 (1), PLATZ (1). Most recent transcription factor family annotations were downloaded from iTAK (27717919).

In addition, we curated three short time course (3h, 6h, 9h), differential expression datasets for each treatment (+Ph/-Fe, -Ph /+Fe, -Ph/-Fe) based on computing the log fold change between each treatment per time point and the corresponding control treatment (+Ph/+Fe). We performed hierarchical clustering using complete linkage (cite) on the resulting reduced differential expression matrix containing only the retained 240 genes, choosing 8 clusters after manual inspection (see figure 1).

Gene set enrichment analysis

We acquired gene model annotations from TAIR10 (Athaliana\_167). The dataset contained gene ontology annotations (biological process, molecular function, and cellular component) for in total 7115 genes. We performed biological process specific enrichment analysis for the entire 240 differential expressed genes (figure 2a), as well as genes of all 8 clusters individually (figure 2b).

Transcription factor promotor binding network

We construct a comprehensive and reliable *A. thaliana* transcription factor promoter binding based network from 3 experimentally or literature curated datasets (27203113, [16524982](https://www.ncbi.nlm.nih.gov/pubmed/16524982), [25750178](https://www.ncbi.nlm.nih.gov/pubmed/25750178)). The curated network contains 2854073 links between 762 transcription factors and 32610 target genes.

Transcription factor promotor binding, co-differential expression network

Our final gene regulatory network consisted of 637 links, involving 10 transcription factors distributed over 6 transcription factor families to be AP2/ERF (3), AUX/IAA (1), bHLH (2), MYB (2), NAC (1), PLATZ (1). In particular, we observe an interconnected hub structure among the network, specifically formed by 2 members of the AP2/ERF (AT3G16280, AT1G77200) as well as 1 member of the MYB (AT5G54230) transcription factor family.

Gene regulatory network inference based on heterogeneous data integration using an ensemble approach

In order to infer an iron and phosphate specific gene regulatory network, we select an unsupervised ensemble model strategy. Therefore, we apply unsupervised integration of several heterogeneous features based networks to derive a score per regulatory link between a transcription factor and a putative target gene. Heterogeneous data integration approaches have proven necessary in the context of gene regulatory network inference in eukaryotes (22456606, Banf and Rhee, 2017).

For our ensemble approach we integrated three regulatory link evidences based on: (i) experimentally observed binding of a transcription factor in the promoter of its putative target, (ii) co-differential expression of both the transcription factor and its putative target in iron and phosphate specific experimental datasets, (iii) link strengths between the transcription factors and its target in a genome wide supportive gene network inferred from a curated large-scale root stress specific gene expression dataset using a random forest regression approach (see methods).

Our final gene regulatory network consisted of 637 links, involving 10 transcription factors distributed over 6 transcription factor families to be AP2/ERF (3), AUX/IAA (1), bHLH (2), MYB (2), NAC (1), PLATZ (1). In particular, we observe an interconnected hub structure among the network, specifically formed by 2 members of the AP2/ERF (AT3G16280, AT1G77200) as well as 1 member of the MYB (AT5G54230) transcription factor family.

F

A gene ontology analysis among all 576 target genes in our predicted gene regulatory network, revealed an array of statistically significant (p-value < 0.05, hyper geometric test) terms with at least 3 associated genes molecular function (96), biological process (33) and cellular component (13). In particular, we observed an enrichment membrane associated genes, reporting transferase, catalytic and/or hydrolase activity, as well as ATP or metal ion binding. Predominant biological processes include phosphorylation, oxygen-reduction, transcriptional regulation, response to auxin, as well as signal transduction.

**Materials & Methods**

***Datasets***

Iron and Phosphate treatment differential expression datasets (to be edited Hatem & his colleague)

@hatem: Experimental design, growth conditions and treatments, final dataset.

@colleague: three-way anova followed by Tuk key test analysis

Given the three-way anova analysis, we retained differential expression events (p-value < 0.05) between genes in each experimental treatment (+Ph/-Fe, -Ph /+Fe, -Ph/-Fe) per time point (3h, 6h, 9h) and the corresponding control treatments (+Ph/+Fe), resulting in a total of 240 out of the 26320 genes being differentially expressed in at least one experiment and time point.

In addition, we curated three short time course (3h, 6h, 9h), differential expression datasets for each treatment (+Ph/-Fe, -Ph /+Fe, -Ph/-Fe) based on computing the log fold change between each treatment per time point and the corresponding control treatment (+Ph/+Fe).

Root specific gene differential expression compendium

We acquired a recently published, comprehensive study on gene expression *for A. thaliana* by He et al. (27015116). The collection had been subjected to consistent data processing and quality control. Given this dataset, we only retained conditions related to various stress and nutrient treatments in root for *A. thaliana* ecotype Columbia.

Our final differential expression dataset covered 82 % (21678 out of 26320 genes) of the *A. thaliana* genome and consisted of 380 expression profiles.

Subsequently, we transformed this condition specific gene expression dataset into differential expression profiles computing the log fold change difference between individual treatments and corresponding control conditions. Our final differential expression dataset covered 82 % (21678 out of 26320 genes) of the *A. thaliana* genome and consisted of 380 expression profiles.

***Gene regulatory network inference using an ensemble approach***

We observed 13 transcription factors (based on iTAK annotations) to be differentially expressed in at least one of the iron and phosphate deficiency experiments. To derive the gene regulatory network based on these Nr = 13 transcription factors, we build an ensemble model of transcriptional regulation, integrating several heterogeneous features to derive a score per regulatory link between a transcription factor and a putative target gene .

Therefore, we integrate: (i) experimentally observed binding of in the promoter of , (ii) co-differential expression of and in the iron and phosphate experimental datasets, (iii) link strengths between and in a genome wide supportive gene regulatory network inferred from a large-scale root stress specific gene expression dataset.

We construct a comprehensive and reliable *A. thaliana* transcription factor promoter binding based network from 3 experimentally or literature curated datasets. The curated network contains 2854073 links between 762 transcription factors and 32610 target genes. We only retain genes, i.e. transcription factors and target genes, differentially expressed within the iron and phosphate experimental dataset and excluded self-binding events, resulting in 5 out of the 13 transcription factors, 134 out of the 240 target genes and 235 links.

For the co-differential expression link evidences we assigned binary weights between a transcription factor and a putative target, if differential expression of the target takes place either at the same time or after a putative regulator. The network contained 1094 links based on the 13 transcription factors and the 240 genes. In addition, we derived a putative mode of action from the directionality differences in differential expression of a regulator and their putative targets, i. e. whether a regulator might act as an activator or repressor on its target.

Finally, to estimate the supportive gene regulatory network, we used the curated large-scale differential expression dataset of *A. thaliana* root stress treatments. The dataset contained 11 of the 13 differentially expressed transcription factors, as well as 175 of the 240 differentially expressed genes. We applied a robust random forest regression based approach (26072483). Regression-based approaches to gene regulatory network inference are based on the assumption that the expression profiles of the transcription factors that directly regulate a target gene are the most informative, among all transcription factors, to predict the expression profile of the target gene. Tree-based regression approaches, such as random forests, have proven successful (22796662) as they can handle complex interaction and apply resampling strategies for repeated subsampling of the data, providing an inherent cross-validation. To implement a genome wide background distribution, we ran random forest regression using Nr = 13 differential expressed regulators for all genes covered in the expression dataset. Regression parameters were set for number of split variables and number of trees . Subsequently, we computed the empirical cumulative distribution function over all predictions assigning probabilities as continuous weight between 0 and 1 per regulatory link.

Finally, we defined rank based on heterogeneous data integration as:

However, we retain a links for which co-differential evidence was provided.

, if , otherwise it was removed. Here denotes a threshold parameter. We set in order to prioritize experimentally observed transcription factor DNA binding and co-differential expression.

Accordingly, if a link is supported by experimentally observed transcription factor DNA binding and co-differential expression, the supportive root stress specific regulatory network becomes less important. In contrast, if a link is supported by either co-differentially expressed or experimentally observed transcription factor DNA binding only, the root stress specific regulatory network is invoked to decide on whether to retain an individual regulatory link (if ).

Hence, our final network consisted of 1083 links – again excluding self-regulations – between 13 regulators and 240 putative target genes.

Transcription factor binding information and family annotations

Most recent transcription factor family annotations were downloaded from iTAK (27717919). Further, we acquired the most comprehensive *A. thaliana* transcription factor binding dataset to date based on in-vitro-expressed transcription factors (27203113), consisting of ~ 2.8 million links for 387 transcription factors.

***Phosphate Iron treatment gene expression hierarchy***

The hierarchy of the iron and phosphate specific differential expression experiments (+Ph/-Fe, -Ph/+Fe, -Ph/-Fe) was estimated based on computing the mean differential expression of each gene in each experiment in order to represent similarity in gene expression patterns of experiments.

***Gene regulatory network inference using an ensemble approach***

We observed 13 transcription factors (based on iTAK annotations) to be differentially expressed in at least one of the iron and phosphate deficiency experiments. To derive the gene regulatory network based on these Nr = 13 transcription factors, we build an ensemble model of transcriptional regulation, integrating several heterogeneous features to derive a score per regulatory link between a transcription factor and a putative target gene .

Therefore, we integrate: (i) experimentally observed binding of in the promoter of , (ii) co-differential expression of and in the iron and phosphate experimental datasets, (iii) link strengths between and in a genome wide supportive gene regulatory network inferred from a large-scale root stress specific gene expression dataset.

For unsupervised data integration, we assigned binary weights to the transcription factor promoter binding as well as co-differential expression link evidences as either present or absent. To estimate the supportive gene regulatory network, we used the curated large-scale differential expression dataset of *A. thaliana* root stress treatments, applying a robust random forest regression based approach (26072483). Regression-based approaches to gene regulatory network inference are based on the assumption that the expression profiles of the transcription factors that directly regulate a target gene are the most informative, among all transcription factors, to predict the expression profile of the target gene. Tree-based regression approaches, such as random forests, have proven successful (22796662) as they can handle complex interaction and apply resampling strategies for repeated subsampling of the data, providing an inherent cross-validation. To implement a genome wide background distribution, we ran random forest regression using Nr = 13 differential expressed regulators for all genes covered in the expression dataset. Regression parameters were set for number of split variables and number of trees . Subsequently, we computed the empirical cumulative distribution function over all predictions assigning probabilities as continuous weight between 0 and 1 per regulatory link.

Finally, we defined rank based on heterogeneous data integration as:

We retain a link , if , otherwise it was removed. Here denotes a threshold parameter. We set in order to prioritize experimentally observed transcription factor DNA binding and co-differential expression. Accordingly, if a link is supported by experimentally observed transcription factor DNA binding and co-differential expression, the supportive root stress specific regulatory network becomes less important. In contrast, if a link is supported by either co-differentially expressed or experimentally observed transcription factor DNA binding only, the root stress specific regulatory network is invoked to decide on whether to retain an individual regulatory link (if ).

For the final gene regulatory network, we used the sign of the Pearson’s correlation coefficient (pcc) based on the curated large-scale differential expression dataset of *A. thaliana* root stress treatments to estimate a putative mode of regulation between a transcription factor and the target gene , as either activation (pcc > 0) or repression (pcc < 0). However, since transcription factors typically act in sets onto a given target, this allows for changes in transcriptional modalities of individual transcription factors, such mode inference should be treated with caution.

**Acknowledgements**

M.B. is a Feodor Lynen Postdoctoral Fellow supported by the Alexander Humboldt Foundation.

**Affiliations**

Department of Plant Biology, Carnegie Institution for Science, 93405 Stanford, USA

Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany

**References**

27015116

Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis.

27717919

iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases.

22456606

Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks.

26072483

Integrative random forest for gene regulatory network inference.

Banf and Rhee, 2017

Enhancing gene regulatory network inference through data integration with markov random fields. Nature Scientific Reports, 2017, in print

22796662

Wisdom of crowds for robust gene network inference.

27203113

Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape

[25750178](https://www.ncbi.nlm.nih.gov/pubmed/25750178)

An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors

[16524982](https://www.ncbi.nlm.nih.gov/pubmed/16524982)

AGRIS and AtRegNet: A Plattform to Link cis-regulatory elements and transcription factors into regulatory networks.