UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Integrated Omic Networks Reveal Regulatory Elements of Transcription and Translation

Permalink

https://escholarship.org/uc/item/6mf184kr

Author

Sartor, Ryan

Publication Date

2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Integrated Omic Networks Reveal Regulatory Elements of Transcription and Translation

A dissertation submitted in partial satisfaction of the requirements for the

degree Doctor of Philosophy

in

Biology

by

Ryan Charles Sartor

Committee in charge:

Professor Steven Briggs, Chair
Professor Vineet Bafna
Professor Mark Estelle
Professor Jose Pruneda-Paz
Professor Scott Rifkin

2017

The Dissertation of Ryan Charles Sartor is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

DEDICATION

For the vast majority of human history, we have had little choice but to make decisions based on instinct, gut feeling and subjective opinion.

For the last few centuries, we have seen unimaginable benefits stem from the adoption of scientific, fact-based thinking in our decision-making. However, we have not always had the correct data or analysis to rightfully inform such decisions.

We are now standing on the fringes of a new world. Over the next century, we will achieve the capability to progress into a data-driven society. This will not necessarily make decisions easier but it will allow us to fully understand and accurately predict the effects of our decisions and not remain slaves to our own instincts.

The following dissertation is dedicated to everyone who values objective scientifically founded evidence over subjective opinions. That is, anyone who values truth over fantasy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge Professor Steven Briggs. The two most significant attributes that impacted me during my time in his lab are Steve's pure and genuine love of thinking and his ability to constantly try to grasp the whole picture. These two attributes make for an excellent scientist and are essential to solving complex problems but are also beneficial in many aspects of life. As a scientist, it's too easy to get caught up in details of a project, or in details of the job itself. It is never easy, seldom glamorous and usually frustrating plus the pay is terrible. But if you look at the big picture, we are getting paid by the rest of society to think and solve problems. Like philosophers of the renaissance sitting around in marble-pillared rooms discussing the biggest questions of our time. Steve has a way of bringing about this kind of view. Steve is a risk taker. Two of his favorite pass times are skydiving and scuba diving. Two sports where you are always one equipment failure away from never having to beg the government for money again. This love of risk comes through to the projects that Steve likes to attempt. He has shown me that sometimes it's hard to tell a crazy idea from an ingenious one. Despite a lot of skepticism on my side, most of the major discoveries in this dissertation were seeded or wholly cultivated by Steve's ideas. Lastly, Steve has an unending torrent of tasks every day but always has time to talk. At no point have I doubted that his top priority is to train the next generation of scientists and I was very fortunate to be under his guidance.

I would like to acknowledge Dr. Zhouxin Shen. Zhouxin has been a second advisor, mentor and great friend. It is no exaggeration to say that none of this work would have been possible without him. Zhouxin is responsible for much the mass spectrometry sample preparation and all of the data acquisition and raw data processing. For the small amount that I did myself, I was trained by Zhouxin. He has taught me everything I know about mass spectrometry, given invaluable input on much of my analyses, provided technical support throughout my PhD and answered an unimaginable number of my questions. I was extremely fortunate to have him.

I would like to acknowledge my wife, Colleen for her unwavering support through this long and painful process of earning a PhD. Thanks for always listening. Also my daughter, Isis, who just turned 4, for always asking “Help you?” whenever she sees that I’ve gotten frustrated.

I would like to acknowledge my family. My parents, siblings, aunts, uncles and cousins. I am who I am today because of all of them.

I would like to acknowledge the other two major members of the Briggs lab: Dr. Justin Walley, and Laura Gates for all their help, companionship and support throughout my PhD.

I would like to acknowledge my dissertation committee, Dr. Vineet Bafna, Dr. Mark Estelle, Dr. Jose Pruneda-Paz and Dr. Scott Rifkin for all their

help, thoughtful advice and suggestions. And for sitting through all of my grueling, way-to-long committee meetings.

Chapter 1, in full, is a reprint of the material as it appears in: Walley, J. W.; Sartor, R. C.; Shen, Z.; Schmitz, R. J.; Wu, K. J.; Urich, M. A.; Nery, J. R.; Smith, L. G.; Schnable, J. C.; Ecker, J. R.; Briggs, S. P. “Integration of omic networks in a developmental atlas of maize”. *Science* **353**, 814–818 (2016). The dissertation author was primary investigator and co-first author of this work.

Chapter 2, in full is currently being prepared for submission for publication of the material. Sartor, R. C., Noshay, J., Springer, N. M., Briggs, S. P. The dissertation author was primary investigator and first author of this work.

Chapter 3, in full is currently being prepared for submission for publication of the material. Sartor, R. C., Walley J. W., Shen Z., Briggs, S. P. The dissertation author was primary investigator and first author of this work.

VITA

Education

2007 Bachelor of Science, Biochemistry and Molecular Biology
Michigan State University, East Lansing, MI

2007 Bachelor of Science, Computer Science
Michigan State University, East Lansing, MI

2017 Doctor of Philosophy, Biology
University of California, San Diego

Publications

Chapman, E. J.; Greenham, K.; Castillejo, C.; **Sartor, R**.; Bialy, A.; Sun, T. ping; Estelle, M. Hypocotyl transcriptome reveals auxin regulation of growth-promoting genes through GA-dependent and -independent pathways. *PLoS ONE* **7**, (2012).

Huffaker, A.; Pearce, G.; Veyrat, N.; Erb, M.; Turlings, T. C. J.; **Sartor, R**.; Shen, Z.; Briggs, S. P.; Vaughan, M. M.; Alborn, H. T.; Teal, P. E. A.; Schmelz, E. A. Plant elicitor peptides are conserved signals regulating direct and indirect antiherbivore defense. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5707–5712 (2013).

Walley, J. W.; Shen, Z.; **Sartor, R**.; Wu, K. J.; Osborn, J.; Smith, L. G.; Briggs, S. P. Reconstruction of protein networks from an atlas of maize seed proteotypes. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E4808-17 (2013).

Christensen, S. A.; Huffaker, A.; Kaplan, F.; Sims, J.; Ziemann, S.; Doehlemann, G.; Ji, L.; Schmitz, R. J.; Kolomiets, M. V; Alborn, H. T.; Mori, N.; Jander, G.; Ni, X.; **Sartor, R. C**.; Byers, S.; Abdo, Z.; Schmelz, E. A. cyclopente (a) nones , display activity as cytotoxic phytoalexins and transcriptional mediators. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 11407–11412 (2015).

Walley, J. W.; **Sartor, R. C**.; Shen, Z.; Schmitz, R. J.; Wu, K. J.; Urich, M. A.; Nery, J. R.; Smith, L. G.; Schnable, J. C.; Ecker, J. R.; Briggs, S. P. Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).

ABSTRACT OF THE DISSERTATION

Integrated Omic Networks Reveal Regulatory Elements of Transcription and Translation

By

Ryan Charles Sartor

Doctor of Philosophy in Biology

University of California, San Diego, 2017

Professor Steven P. Briggs, Chair

Our world is becoming quantifiable. The IBM Corporation estimates that our society is collecting data at a rate of 2.5 quintillion bytes per day. To give some perspective, 90% of the data that humans now have access to has been collected in the last two years. Biological science is no exception. This

information holds enormous potential but the biggest challenge now lies in data analysis and interpretation.

In biology, this data revolution has been led by sequencing technology. Therefore most data is either genomic or transcriptomic in nature. This dissertation focuses on protein mass spectrometry. We find that by integrating multiple data sets, we achieve the most powerful systems-level descriptions of biological systems. In the following dissertation we show how proteomic data can be integrated with both transcriptomic and epigenomic data sets to provide critical insight into biological systems.

In the first chapter, we show that proteomic and transcriptomic measurements have fundamental differences and lead to different specific results but similar “big-picture” conclusions. We use both to re-construct gene regulatory networks and find that the most accurate network results from integrating both data types.

In the second chapter, we expand on observations made in chapter one and incorporate DNA methylation data. We discover that, using random forest machine learning models and genic DNA methylation data, we are able to classify the subset of expressed mRNAs with high accuracy. Most interestingly, after incorporating proteomic data, we achieve near perfect classification accuracy and go on to discover a surprising association between genic DNA methylation and translations. Such models can be used to annotate the functional subset of maize genes with equal or better accuracy than

current manual annotations. These models show excellent accuracy in a diverse set of maize inbreds, leading to speculation that DNA methylation is playing a large role in crop domestication.

In the final chapter we use a novel method to integrates protein and transcript data to discover quantitative trait loci that are specifically controlling protein abundance in a mRNA independent manor in arabidopsis. We then demonstrate how transcript data can be used to prioritize causative genes.

CHAPTER 1

Integration of omic networks in a developmental atlas of maize

Reprint of “Integration of omic networks in a developmental atlas of maize”.

Published August 2016 in *Science*.

years, respectively (Fig. 4B) (Monte-Carlo simulations: first-year control males: $P = 0.024$, $n = 22$, and females: $P = 0.075$, $n = 15$; second-year treatment males: $P = 0.046$, $n = 10$; all others: $P > 0.05$) (*24*).

Overall, we have demonstrated experimentally that by acoustically signaling high ambient temperatures to their embryos before hatching, zebra finch parents can program the developmental trajectories of their offspring in response to this key environmental variable. Our findings therefore provide both an adaptive function for prenatal communication and a type of maternal effect where parental control over signal production can be unambiguously tested. By uncovering a mechanism for a transgenerational effect of temperature on development in endotherms, our study also advances our understanding of the acclimatization capacities of organisms to rising temperatures.

REFERENCES AND NOTES

1. B. Dantzer *et al.*, *Science* **340**, 1215–1217 (2013).
2. R. A. Duckworth, V. Belloni, S. R. Anderson, *Science* **347**, 875–877 (2015).
3. T. A. Mousseau, C. W. Fox, *Trends Ecol. Evol.* **13**, 403–407 (1998).
4. C. Teplitsky, J. A. Mills, J. S. Alho, J. W. Yarrall, J. Merilä, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13492–13496 (2008).
5. O. Vedder, S. Bouwhuis, B. C. Sheldon, *PLOS Biol.* **11**, e1001605 (2013).
6. J. M. Donelson, P. L. Munday, M. I. McCormick, C. R. Pitcher, *Nat. Clim. Chg.* **2**, 30–32 (2012).
7. F. R. Groeters, H. Dingle, *J. Evol. Biol.* **1**, 317–333 (1988).
8. S. Salinas, S. B. Munch, *Ecol. Lett.* **15**, 159–163 (2012).
9. D. B. Miller, G. Gottlieb, *Anim. Behav.* **26**, 1178–1194 (1978).
10. D. Colombelli-Négrel *et al.*, *Curr. Biol.* **22**, 2155–2160 (2012).
11. D. Colombelli-Négrel, M. S. Webster, J. L. Dowling, M. E. Hauber, S. Kleindorfer, *Auk* **133**, 273–285 (2016).

GENE REGULATION

Integration of omic networks in a developmental atlas of maize

Justin W. Walley,[1,2]* Ryan C. Sartor,[1]* Zhouxin Shen,[1] Robert J. Schmitz,[3,4]† Kevin J. Wu,[1] Mark A. Urich,[3,4] Joseph R. Nery,[4] Laurie G. Smith,[1] James C. Schnable,[5] Joseph R. Ecker,[3,4,6] Steven P. Briggs[1]‡

Coexpression networks and gene regulatory networks (GRNs) are emerging as important tools for predicting functional roles of individual genes at a system-wide scale. To enable network reconstructions, we built a large-scale gene expression atlas composed of 62,547 messenger RNAs (mRNAs), 17,862 nonmodified proteins, and 6227 phosphoproteins harboring 31,595 phosphorylation sites quantified across maize development. Networks in which nodes are genes connected on the basis of highly correlated expression patterns of mRNAs were very different from networks that were based on coexpression of proteins. Roughly 85% of highly interconnected hubs were not conserved in expression between RNA and protein networks. However, networks from either data type were enriched in similar ontological categories and were effective in predicting known regulatory relationships. Integration of mRNA, protein, and phosphoprotein data sets greatly improved the predictive power of GRNs.

Predicting the functional roles of individual genes at a system-wide scale is a complex challenge in biology. Transcriptome data have been used to generate genome-wide gene regulatory networks (GRNs) (*1–4*) and coexpression networks (*5–7*), the design of which was based on the presumption that mRNA measurements are a proxy for protein abundance measurements. However, genome-wide correlations between the levels of proteins and mRNAs are weakly positive (*8–15*), which indicates that cellular networks built solely on transcriptome data may be enhanced by

[1]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA. [2]Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA 50011, USA. [3]Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [4]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [5]Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68583, USA. [6]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA.

*These authors contributed equally to this work. †Present address: Department of Genetics, Davison Life Sciences, 120 East Green Street, Athens, GA 30602, USA. **‡Corresponding author. Email: sbriggs@ucsd.edu**

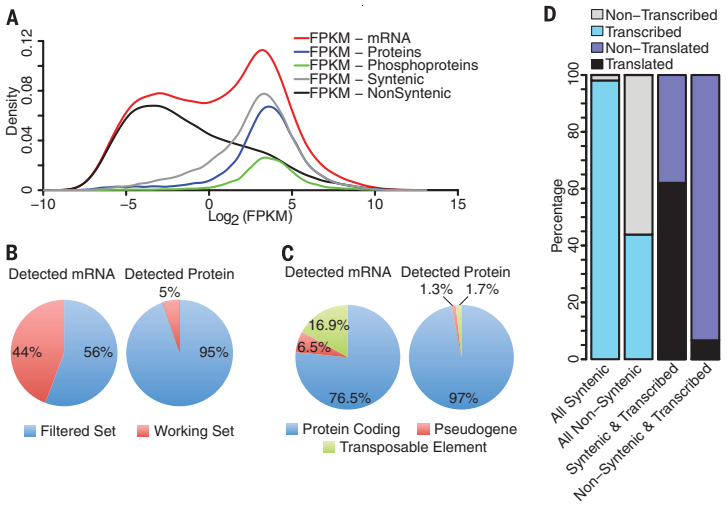


Fig. 1. Comparison of transcriptome and proteome data sets. (**A**) FPKM distribution of mRNA abundance (red). FPKM values of transcripts corresponding to quantified proteins (blue), phosphopeptides (green), syntenic genes conserved between maize and sorghum (gray), and nonsyntenic genes (black) are shown. Data are the average expression from the 23 tissues profiled. (**B**) Percentage of quantified mRNA and proteins in the annotated filtered (high-confidence gene models) and working (all gene models) gene sets. (**C**) Breakdown of detected mRNA and proteins, based on annotations. (**D**) Percentages of all annotated genes that are transcribed and percentages of all transcribed genes that are translated, for both the syntenic and nonsyntenic gene sets.

integration with proteomics data. We generated an integrated developmental atlas of the transcriptome, proteome, and phosphoproteome of the model organism *Zea mays* (maize) and then used these three different cellular descriptions to generate transcriptome- and proteome-based networks.

We profiled 23 tissues spanning vegetative and reproductive stages of maize development to generate a comprehensive and integrated gene expression atlas. Specifically, transcriptome profiling by mRNA sequencing (mRNA-seq) (three biological replicates, 23 tissues) was carried out on a subset of the samples used for proteome profiling (three to seven biological replicates, 33 tissues) by electrospray ionization tandem mass spectrometry (*14*, *16*–*19*) (tables S1 to S3). We assessed reproducibility of the biological replicates by calculating Pearson correlations and found an average of 0.9, 0.84, and 0.7 for the transcriptome, proteome, and phosphoproteome data sets, respectively (table S4). Transcripts were observed from 62,547 genes. Proteins and phosphoproteins were observed from 16,946 and 5587 genes, respectively. The RNA-seq data were bimodal, as reported for mouse and human (*20*, *21*), with nearly all proteins and phosphoproteins arising from the 34,455 transcripts in the high-abundance population (right peak), with an average FPKM (fragments per kilobase of exon per million fragments mapped) greater than 1 (Fig. 1A). Proteins were observed from 46% of these transcripts (right peak). To determine whether coverage of the transcriptome was constrained by the diversity of tissues sampled, we generated proteomics data from an additional 10 tissue types yielding proteins from a total of 18,522 genes (proteins from 17,862 genes and phosphoproteins from 6185 genes), but this only increased coverage of the high-abundance transcriptome to 48%.

There are a variety of possible technical and biological explanations for why we detect proteins from less than half of the high-abundance transcript-producing genes and why we do not observe corresponding mRNA for 245 quantified proteins. Previously, we found evidence for multiple mechanisms that may explain the detection of proteins but not mRNA. These mechanisms include (i) differential stability of mRNA and proteins; (ii) transport of proteins between tissues; and (iii) diurnal, out-of-phase accumulation of mRNAs and cognate proteins (*14*). The heightened sensitivity of transcriptomics relative to proteomics likely provides a partial explanation for why we detect proteins corresponding to less than half of the transcript-producing genes. Additionally, we observed a greater percentage of proteins arising from the annotated filtered gene set, which consists of 39,656 high-confidence gene models that exclude transposons, pseudogenes, and other low-confidence members present in the working gene set (Fig. 1B). Furthermore, a higher proportion of proteins than transcripts arise from genes annotated as protein coding (Fig. 1C), which suggests that transcripts from many genes may not produce proteins. Genes conserved at syntenic orthologous locations between maize and sorghum exhibited a unimodal, high-expression pattern, in contrast to genes in nonsyntenic locales (Fig. 1A). Considering all genes that expressed mRNAs, syntenic genes were nine times more likely than nonsyntenic genes to express proteins (Fig. 1D). To show that this observation is not due to the higher average transcript expression level of syntenic genes, we examined a range of transcript abundance cutoffs and obtained similar results, even when looking at the highest-abundance syntenic and nonsyntenic transcripts (fig. S1). A greater frequency of protein expression is a possible mechanistic explanation for the eightfold enrichment of genes responsible for visible mutant phenotypes among syntenically conserved genes in maize (*22*).

We next examined how genes and biological processes change throughout development. Initially, we focused on transcription factors (TFs), as they are key regulators of development, growth, and cell fate. Of the 2732 annotated TFs and transcriptional co-regulators, we detected 2627 as mRNA (23 tissues), 1026 as protein (33 tissues), and 559 as phosphoprotein (33 tissues). We used hierarchical clustering to identify 712 (mRNA), 469 (protein), and 419 (phosphoprotein) TFs that exhibited tissue-specific enrichment (figs. S2 to S4 and table S5). We also examined expression trends at the TF family level. First, we used traditional overrepresentation analysis to identify TF families whose members are detected in a given tissue at a greater frequency than chance (figs. S5A, S6A, and S7A). To augment the overrepresentation analysis, we also examined TF family–level expression profiles by quantifying the total amount of each TF family's mRNA, protein, and/or phosphoprotein present in given tissue (figs. S5B, S6B, and S7B). Taken together, these data describe the spatiotemporal expression pattern of individual TFs and TF families across development.

We expanded our analyses to examine the patterns of all gene types across maize development. We used the weighted gene coexpression network analysis (WGCNA) R package (*23*) to group similarly expressed genes—detected as mRNA (23 tissues), protein (33 tissues), or phosphoprotein (33 tissues) in at least four tissues—into modules (clusters). This approach enabled us to group 31,447 mRNAs, 13,175 proteins, and 4267 phosphoproteins into coexpression modules (fig. S8 and table S6). We next plotted the eigengene profile for each module in order to assign the tissue(s) in which each module is highly expressed (figs. S9 to S12). We observed that 36 well-characterized genes required for maize development—including the homeobox TFs Knotted1 [KN1, Maize Genetics and Genomics Database (MGGD) accession number GRMZM2G017087] (*24*) and Rough Sheath 1 (RS1,

MGGD accession number GRMZM2G028041) (*25*), as well as the transcriptional co-repressor Ramosa1 Enhancer Locus2 (REL2, MGGD accession number GRMZM2G042992) (*26*) (table S6)—are present in mRNA, protein, and phosphoprotein modules that correspond to dividing and meristematic tissues. The phosphorylation pattern of these proteins is similar to their mRNA profile and occurs in tissues known to have altered developmental phenotypes in mutant plants, which suggests that phosphorylation of these proteins might positively regulate their function. Finally, we determined overrepresentation of MapMan functional categories in each module (table S6). As expected, we found that genes involved in photosynthetic light reactions have mRNA and protein that are enriched predominantly in the mature leaf. We did not detect an enrichment of light-reaction phosphoproteins in the mature leaf module, which suggests that phosphorylation is not a major regulator of the light reactions (fig. S11 and table S6).

Biological networks can be constructed based on many different types of data and serve to elucidate the structure underlying complex systems. Typically, transcript profiling data are used to generate various types of gene expression networks. However, we observed a weakly positive correlation between mRNA and protein levels in our data set (supplementary text, figs. S13 to S17, and table S7), in agreement with research done in a range of organisms (*8*–*15*). Although the modest correlation between mRNA and protein levels is well documented, a major outstanding question is whether transcriptome-based networks predict the same relationships as proteome-based networks. Given our extensive developmental gene expression atlas, we addressed this question by generating two different types of networks: coexpression networks and GRNs. We first generated coexpression networks (table S8), which are undirected networks where nodes are genes connected on the basis of highly correlated expression patterns (Fig. 2A) (*5*–*7*). For these network reconstructions, we used 10,979 genes that were detected as both transcripts and proteins in at least 5 of the 23 developmental gene expression atlas tissues in which we profiled both mRNA and protein. Pairwise mRNA-to-mRNA and protein-to-protein coexpression networks were built with Spearman correlations using WGCNA (fig. S18 and table S8). The biweight midcorrelation yielded similar results (figs. S19 and S20). To directly compare the mRNA- and protein-based coexpression networks and compile a high-confidence coexpression data set, each network was constrained to include only edges with a correlation score >0.75 (top 1 million edges), which is a frequently used correlation threshold for coexpression networks (table S8). As a measure of similarity, we calculated edge conservation by dividing the set intersect by the union (known as the Jaccard index) and reported this as a percentage. We found that 122,029 of the combined 2 million edges (6.1%) were conserved in both networks (Fig. 2B). Though this edge overlap is greater than the 0.8% expected by chance (P value = 0), the majority of relationships between genes were specific to each network, even when we expanded the network size to 10 million edges (fig. S20).

To examine whether the lack of edge overlap was due to experimental noise, we used single biological replicates (three mRNA and three protein networks) to create six new coexpression networks. Pairwise comparisons revealed a similar low level of edge conservation (5%) between the mRNA and protein coexpression networks. However, 46% of mRNA-to-mRNA edges and 36% of protein-to-protein edges were conserved between replicate coexpression networks (fig. S21). These data suggest that biological phenomena underpin the observed lack of edge conservation between transcriptome- and proteome-derived coexpression networks.

A key feature of scale-free networks is a small number of highly interconnected hubs. Because hubs are more likely than nonhubs to be required for network integrity and organism survival, the identification of so-called "hub genes" is of interest (*23*, *27*–*30*). We therefore determined the highly interconnected hub genes in each coexpression network, which we categorized as nodes in the top 10th percentile for most edges (Fig. 2C and fig. S22A). When we compared the hub genes from each network, we found that the majority (85%) were not shared between the mRNA and protein coexpression networks (Fig. 2C and fig. S22).

Groups of coexpressed genes (modules) were derived from the two networks. Each module was examined for over- or underrepresentation of MapMan categories (table S9). The majority of modules from each network (mRNA: 17 of 19; protein: 18 of 25) showed significant enrichment for one or more categories (adjusted P value < 0.05). Overall, we observed similar enrichment of categories between the two coexpression networks (fig. S23). Whereas the overall degree of enrichment was very similar for most categories in both coexpression networks, the actual genes that accounted for the significantly enriched categories were mostly specific to one network (35% protein-specific, 27% mRNA-specific, and 38% shared) (Fig. 3). Taken together, these results demonstrate that transcript- and protein-based coexpression networks yield differing predictions of gene relatedness and function. Presumably, the discrepancy between transcriptome and proteome coexpression networks arises from the limited correlation between mRNA and protein abundance, which has been attributed to a range of factors that include differing stabilities of mRNA and protein, translational control, and protein movement from the tissue of synthesis (*8*, *14*, *31*).

To further explore the regulatory patterns of gene expression across maize development, we generated GRNs, which are directed networks of TFs and their target genes (Fig. 4A) (*1*). Unsupervised GRNs were created using GENIE3, which takes advantage of the random forest machine learning algorithm and was the top-performing method in the DREAM4 and -5 GRN reconstruction challenges (*32*, *33*). Three independent GRNs were generated from the 23 tissues in which we profiled both mRNA and protein. To construct these networks, we varied whether the TFs (termed "regulators") were quantified as mRNAs (2200 TFs), proteins (545 TFs), or phosphopeptides (441 TFs) and used a common set of 41,021 quantified mRNAs (termed "target genes") (table S10). We evaluated the GRNs by using published data for two classical maize TFs, the homeobox TF KN1 and the bZIP TF Opaque2 (O2). These TFs were chosen as benchmarks because they have been the subject of high-quality RNA-seq and chromatin immunoprecipitation (ChIP)–seq studies in both wild-type and null mutant backgrounds, and they represent two distinct types of TFs with key developmental roles (*24*, *34*). Target genes are bound by their TF in a ChIP-seq assay, and their mRNA levels change when their TF is knocked out. Using the published direct targets of KN1 and O2, we generated

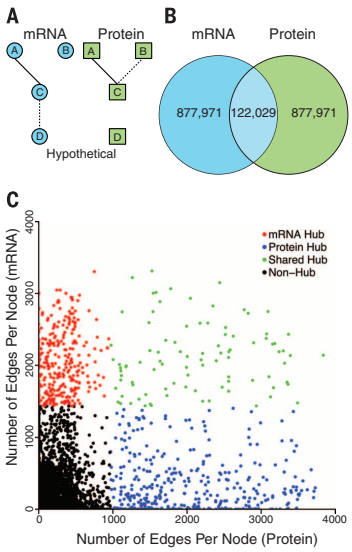


Fig. 2. Coexpression network analyses. (**A**) Hypothetical undirected coexpression subnetwork showing conserved (solid lines) and nonconserved (dotted lines) coexpression edges between mRNA and protein networks. (**B**) Venn diagram depicting edge conservation (solid lines in Fig. 2A) between the two coexpression networks. (**C**) Number of edges a given gene (node) has in the protein (*x* axis) and mRNA (*y* axis) coexpression networks. Nodes above the 90th percentile for the number of edges are considered hubs and are colored according to whether they are hubs in the protein (blue) or mRNA (red) network or both (green). Black dots represent nonhub nodes.

receiver operating characteristic (ROC) and precision-versus-recall curves, which are two methods commonly used to evaluate the power of a predictive model (*35*). These curves showed that the overall qualities of all three GRNs were similar (fig. S24). However, when we looked at the top 500 scoring GENIE3 predictions for KN1 and O2 in each GRN, we observed a performance advantage for the two protein-based GRNs in accurately predicting target genes (Fig. 4B and fig. S25A). Specifically, the KN1 subnetworks accurately predicted 108 (mRNA), 129 (protein), and 125 (phosphopeptide) targets, with the O2 subnetworks performing similarly. Additionally, 44% (KN1) and 31% (O2) of

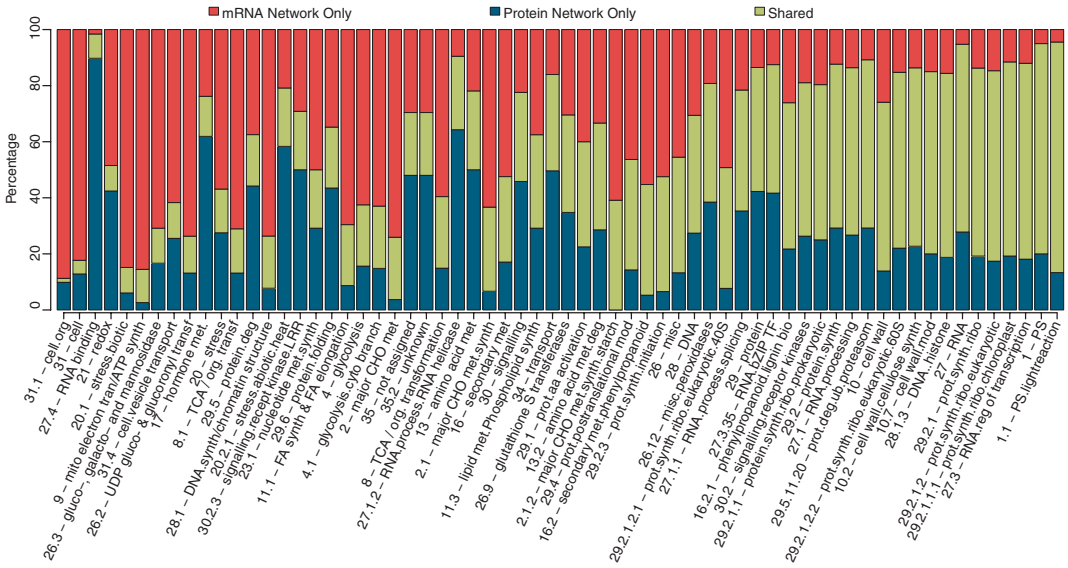


Fig. 3. Categorical enrichment analysis of coexpression modules. Coexpression modules were determined by WGCNA and functionally annotated using MapMan categories. Categories enriched (Benjamini-Hochberg adjusted *P* value ≤ 0.05) in one or more modules are represented by vertical bars and labeled with the bin number and name. For each category, the genes accounting for the enrichment were extracted separately from mRNA and protein modules. Only functional categories with at least 20 genes are shown. Colored bars represent the proportion of genes in each enriched category that are specific to one network (mRNA, red; protein, blue) or shared between the networks (green).

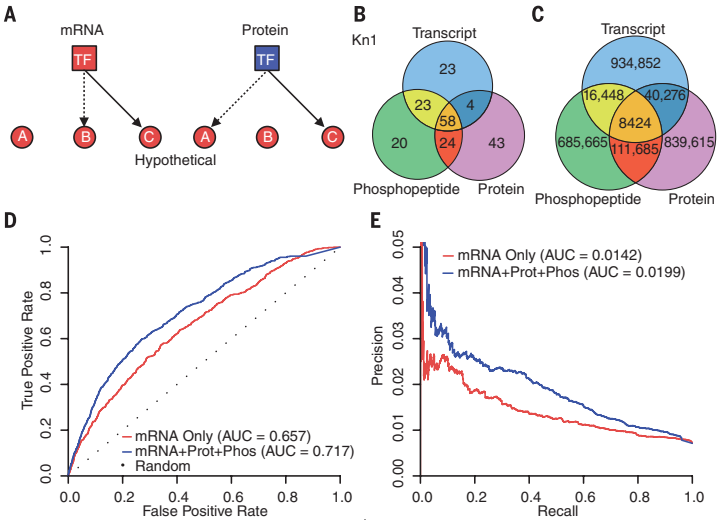


Fig. 4. Unsupervised GRN analyses. (**A**) Hypothetical GRN subnetwork depicting a TF regulator (square) and potential target genes (circle) quantified as mRNA (red) or protein (blue). GRN-specific and -conserved predictions are depicted by dotted and solid lines, respectively. (**B**) Overlap of the true-positive predictions from the top 500 true GRN predictions for KN1 quantified as mRNA, protein, or phosphopeptide. True KN1 targets were identified by Bolduc *et al.* (*24*). (**C**) Overlap of the top 1 million TF target predictions between the GRNs reconstructed using TF abundance quantified at the mRNA, protein, or phosphopeptide level. (**D**) ROC curves and (**E**) precision-recall curves generated using known Kn1 and O2 target genes for a mRNA-only GRN (red) and a fully integrated GRN built by combining mRNA, protein, and phosphoprotein data into a single GRN (blue).

all correctly predicted targets were specific to a single type of GRN (Fig. 4B and fig. S25A). These results indicated that predictions made by all three GRNs were largely complementary to each other.

We expanded our analyses to examine all TFs in the three GRNs. Again, we found that there was low edge conservation between the GRNs, with the vast majority of edges being present in a single GRN (fig. S26). Specifically, when considering one million edges, 93% were present in a single GRN (Fig. 4C). This amount increased to 96% for the 200,000 highest-confidence predictions, which we determined using KN1 precision data as the cutoff (fig. S25, B and C). This finding illustrates that the different accumulation patterns of mRNA, protein, and phosphorylation for a given TF (fig. S27) result in disparate GRN predictions.

The three preceding GRNs were constructed using different-sized sets of TF regulators, which complicated direct comparisons of networks constructed using TF abundance measurements at the mRNA or protein level. Therefore, we used 539 TFs quantified as both mRNAs and proteins to reconstruct GRNs. Evaluation of these GRNs using the KN1 and O2 data indicated quality and accuracy similar to those of the full-sized networks (fig. S28). We still observed a performance advantage for the protein GRN, as well as limited edge conservation between the mRNA- and protein-based GRNs, with only 6% of the top 200,000 edges being shared (figs. S28 and S29). We examined several possible features of the TF regulators to help further our understanding of the limited overlap in TF target predictions. The TFs connected by edges that were present only in the transcript GRN had lower and more variable protein abundance than the TFs connected by edges that were shared with or specific to the protein GRN (fig. S30, A to D). As expected, the mRNA-to-protein correlations were higher for targets of edges present in both GRNs (fig. S30E).

To further validate GRN predictions and test whether network relationships were consistent between different maize varieties, we took advantage of natural variation in regulator abundance arising from the natural genetic variation present in another inbred line, Mo17. Specifically, we compared mRNA and protein abundance in primary roots of Mo17 to B73. Whereas most TFs and target genes were expressed at similar levels in B73 and Mo17, we identified 149 (mRNA), 26 (protein), and 16 (phosphopeptide) regulatory TFs that were expressed at significantly different levels. We found, with high confidence, that for many of these differentially expressed TFs, their GRN predicted target groups were also significantly enriched for differentially expressed transcripts (figs. S31 to S33). Thus, elements of the GRN structure were preserved, and quantitative changes in regulator abundance levels are associated with altered network output and gene expression patterns. Additionally, these findings validated the GRN approaches used in this study and demonstrated the utility of applying this method to examine dynamics of gene regulation.

After analyzing separate mRNA- and protein-based GRNs, we considered integrating the data sets to determine whether the resulting single GRN would have improved inference over the individual GRNs. Specifically, we constructed four additional GRNs, each consisting of combinations of TF regulators quantified as mRNA, protein, and/or phosphopeptides (table S10). Details of how the combined mRNA, protein, and phosphopeptide GRNs were made are described in the supplementary materials. We examined the performance of the resulting networks using the validation set of KN1 and O2 published targets (Fig. 4 and fig. S24). All GRNs reconstructed with combinations of TF regulators performed better than single-input GRNs. This finding demonstrates that integrating readouts of gene expression quantified at different levels results in improved GRN inference. Our use of TF mRNA levels to infer TF activity had provided good GRN predictive power. The area under the ROC curve (AUC) was 0.657, compared with 0.500 for random predictions. When the mRNA measurements were combined with protein abundance and phosphorylation levels to infer TF activity, the AUC increased to 0.717. Thus, if an investigator wished to use network predictions with a false-positive rate of 20%, the mRNA-only network would predict 40% of the true positives, compared with 50% for the combined network (Fig. 4D and fig. S24A). Likewise, examination of Fig. 4E and fig. S24B reveals that if an investigator wished to use network predictions with a precision of 0.021 (which is three times higher than expected at random), then 16% of the true positives would be recalled from the mRNA-only network versus 41% for the combined network.

By quantitatively measuring mRNAs, proteins, and phosphoproteins in parallel in a tissue-specific manner, we discovered unexpected relationships among these cellular readouts across maize development. In particular, our comparison of transcriptome- to proteome-based dendrograms and coexpression networks showed little overlap at the gene level, even though the samples were classified similarly and had similar ontological enrichments. The discovery that most protein-expressing genes are conserved and syntenic also was unexpected. The coexpression networks and GRNs provide a conceptual framework for future detailed studies in a model organism that is central to food security and bioenergy. Our findings highlight the importance of studying gene regulation at multiple levels.

REFERENCES AND NOTES

1. G. Krouk, J. Lingeman, A. M. Colon, G. Coruzzi, D. Shasha, *Genome Biol.* **14**, 123 (2013).
2. T. S. Gardner, J. J. Faith, *Phys. Life Rev.* **2**, 65–88 (2005).
3. Z. Bar-Joseph *et al.*, *Nat. Biotechnol.* **21**, 1337–1342 (2003).
4. R. De Smet, K. Marchal, *Nat. Rev. Microbiol.* **8**, 717–729 (2010).
5. V. van Noort, B. Snel, M. A. Huynen, *Trends Genet.* **19**, 238–242 (2003).
6. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, *Science* **302**, 249–255 (2003).
7. S. Horvath, J. Dong, *PLOS Comput. Biol.* **4**, e1000117 (2008).
8. B. Schwanhäusser *et al.*, *Nature* **473**, 337–342 (2011).
9. C. Vogel *et al.*, *Mol. Syst. Biol.* **6**, 400 (2010).
10. S. Ghaemmaghami *et al.*, *Nature* **425**, 737–741 (2003).
11. K. Baerenfaller *et al.*, *Science* **320**, 938–941 (2008).
12. A. Ghazalpour *et al.*, *PLOS Genet.* **7**, e1001393 (2011).
13. L. Ponnala, Y. Wang, Q. Sun, K. J. van Wijk, *Plant J.* **78**, 424–440 (2014).
14. J. W. Walley *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, E4808–E4817 (2013).
15. M. P. Washburn *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3107–3112 (2003).
16. H. Qiao *et al.*, *Science* **338**, 390–393 (2012).
17. K. N. Chang *et al.*, *eLife* **2**, e00675 (2013).
18. M. R. Facette, Z. Shen, F. R. Björnsdóttir, S. P. Briggs, L. G. Smith, *Plant Cell* **25**, 2798–2812 (2013).
19. C. Marcon *et al.*, *Plant Physiol.* **168**, 233–246 (2015).
20. D. Hebenstreit *et al.*, *Mol. Syst. Biol.* **7**, 497 (2011).
21. N. Nagaraj *et al.*, *Mol. Syst. Biol.* **7**, 548 (2011).
22. J. C. Schnable, M. Freeling, *PLOS ONE* **6**, e17855 (2011).
23. P. Langfelder, S. Horvath, *BMC Bioinformatics* **9**, 559 (2008).
24. N. Bolduc *et al.*, *Genes Dev.* **26**, 1685–1690 (2012).
25. R. G. Schneeberger, P. W. Becraft, S. Hake, M. Freeling, *Genes Dev.* **9**, 2292–2304 (1995).
26. A. Gallavotti *et al.*, *Development* **137**, 2849–2856 (2010).
27. M. S. Mukhtar *et al.*, *Science* **333**, 596–601 (2011).
28. R. Albert, H. Jeong, A. L. Barabasi, *Nature* **406**, 378–382 (2000).
29. M. A. Calderwood *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7606–7611 (2007).
30. H. Jeong, S. P. Mason, A. L. Barabási, Z. N. Oltvai, *Nature* **411**, 41–42 (2001).
31. K. Baerenfaller *et al.*, *Mol. Syst. Biol.* **8**, 606 (2012).
32. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, *PLOS ONE* **5**, e12776 (2010).
33. D. Marbach *et al.*, *Nat. Methods* **9**, 796–804 (2012).
34. C. Li *et al.*, *Plant Cell* 10.1105/tpc.114.134858 (2015).
35. M. Schrynemackers, R. Küffner, P. Geurts, *Front. Genet.* **4**, 262 (2013).

ACKNOWLEDGMENTS

We thank V. Walbot for tassel and anther tissues, F. Hochholdinger for root tissues, and J. Fowler for pollen tissues. This work was supported by the NSF (grant 0924023 to S.P.B. and grants MCB-0929402 and MCB1122246 to J.R.E.), an NIH National Research Service Award Postdoctoral Fellowship (F32GM096707 to J.W.W.), and the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (grant GBMF3034 to J.R.E.). Sequence data can be downloaded from the National Center for Biotechnology Information's Sequence Read Archive (accession number GSE50191). Raw mass spectra have been deposited at the Mass Spectrometry Interactive Virtual Environment (MassIVE) repository (accession numbers MSV000079290 and MSV000079303). Normalized expression data have been integrated into the genome browser and individual gene model pages at the central Maize Genetics and Genomics Database (www.maizegdb.org/).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6301/814/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S33
Tables S1 to S10
References (*36*–*45*)

10 May 2016; accepted 25 July 2016
10.1126/science.aag1125

Supplemental Materials and Methods

Plant material

For paired RNA read and peptide profiling we collected 23 tissues from the B73 inbred including vegetative meristem (16-19 days), 6-7th internode and 7-8th internode (28-30 days), 3 zones from leaf 8 (30 day old), mature leaf 8 (43 days old), intact primary and secondary roots (5 days old), dissected primary roots (meristem, cortex, and elongation zones), mature pollen, female spikelet, ear primordium (2-4 mm), ear primordium (6-8 mm), unpollinated silks, endosperm (12 days after pollination; DAP), endosperm crown (27 DAP), pericarp/aleurone (27 DAP), embryo (20 and 38 DAP), and germinated embryo (2 days after imbibition; DAI). Additionally, we sampled 5-day-old primary roots from the Mo17 inbred. For proteome analyses only we also sampled tissue from B73 including primary root stele, juvenile leaf 3 (19 day), germinated pollen, ear primordium (1 mm), and endosperm (8 and 10 DAP) as well as the W23 inbred, which included 2 cm tassels, 1 mm anthers, 2 mm anthers, and mature pollen.

Mass spectrometry

Approximately 1-2 grams of frozen tissue were ground in liquid nitrogen using a mortar/pestle for 15 minutes to a fine powder and then transferred to a 50ml conical tube. Proteins were precipitated and washed with 50ml -20^{o}C methanol with 0.2mM Na_3VO_4 three times, then 50ml -20^{o}C acetone three times. Protein pellets were aliquoted into four 2ml Eppendorf tubes and dried in a SpeedVac at 4 ^{o}C.

Protein pellets were suspended in 1 ml extraction buffer (0.1% SDS, 1mM EDTA, 50mM HEPES buffer, pH 7). Cysteines were reduced and alkylated using 1 mM Tris (2-carboxyethyl)phosphine (Fisher, AC36383) at 95 °C for 5 minutes then 2.5 mM iodoacetamide (Fisher, AC12227) at 37°C in dark for 15 minutes, respectively. Protein amount was quantified using a Bradford assay (Pierce). Proteins were digested with trypsin (Roche, 03 708 969 001, enzyme:substrate w:w ratio = 1:100) overnight. A second digestion (enzyme:substrate w:w ratio = 1:100) was perform the next day for 4 hours. Digested peptides were purified on a Waters Oasis MCX cartridge to remove SDS. Peptides were eluted from the MCX column with 1ml 50% isopropyl alcohol and 400mM NH_4HCO_3 (pH 9.5) and then dried in a vacuum concentrator at 4^{o}C. Peptide amount was quantified following MCX using the Pierce BCA Protein assay kit. For non-modified proteome profiling peptides were re-suspended in 1% formic acid to a final pH of 3 and used for mass spectrometry analysis. For phospho-proteome profiling peptides were re-suspended in 3% TFA to a final pH of 1 and then used for phosphopeptide enrichment.

Phosphopeptide enrichment was performed using CeO_2 affinity capture. 1% colloidal CeO_2 (Sigma, 289744) was added to the acidified peptide solution (CeO_2:peptide w:w ratio = 1:10). After brief vortexing, CeO_2 with captured phosphopeptides was spun down at 1,000g for 1 minute. Supernatant was

removed and the CeO_2 pellet was washed with 1ml of 1% TFA. Phosphopeptides were eluted by adding eluting buffer (200mM $(NH_4)_2HPO_4$, 2M $NH_3.H_2O$, 10mM EDTA, pH 9.5; same volume as the added 1% colloidal CeO_2) and vortexing briefly. CeO_2 was precipitated by adding 10% formic acid with 100mM citric acid (same volume as the added 1% colloidal CeO_2) to a final pH of 3. Sample was centrifuged at 16,100 g for 1 minute. The supernatant containing phosphopeptides was removed and used for mass spec analysis.

An Agilent 1100 HPLC system (Agilent Technologies) delivered a flow rate of 600 nL min^{-1} to a 3-phase capillary chromatography column through a splitter. Using a custom pressure cell, 5 µm Zorbax SB-C18 (Agilent) was packed into fused silica capillary tubing (250 µm ID; 360 µm OD; 30 cm long non-modified and 20 cm long phospho) to form the first dimension reverse phase section of the column (RP1). A 5 cm long strong cation exchange (SCX) section of the column packed with 5 µm PolySulfoethyl (PolyLC) was connected to RP1 using a zero dead volume 1 µm filter (Upchurch, M548) attached to the exit of the RP1 column. A fused silica capillary (200 µm ID, 360 µm OD, 20 cm long) packed with 5 µm Zorbax SB-C18 (Agilent) was connected to SCX as the analytical section of the column (RP2). The electrospray tip of the fused silica tubing was pulled to a sharp tip with the inner diameter smaller than 1 µm using a laser puller (Sutter P-2000). The peptide mixtures were loaded onto the RP1 column section using the custom pressure cell. Then the 3 sections were joined and mounted on a custom electrospray adapter for on-line nested elutions. A new set of columns was used for each LC-MS/MS analysis. Peptides were first eluted from the RP1 column section to the SCX column section using a 0 to 80% acetonitrile gradient for 150 minutes. The peptides were then fractionated by the SCX column section using a series of 19 step salt gradients for phosphoproteome (0mM, 5mM, 6mM, 7mM, 8mM, 9mM, 10mM, 12mM, 15mM, 20mM, 30mM, 40mM, 50mM, 60mM, 70mM, 80mM, 90mM, 100mM, and 1M ammonium acetate for 20 minutes) and 29 step salt gradients for the non-modified proteome (0mM, 5 mM, 10 mM, 15 mM, 20 mM, 22.5 mM, 25 mM, 27.5 mM, 30 mM, 32.5 mM, 35 mM, 37.5 mM, 40 mM, 42.5 mM, 45 mM, 47.5 mM, 50 mM, 52.5 mM, 55 mM, 57.5 mM, 60 mM, 65 mM, 70 mM, 75 mM, 80 mM, 85 mM, 90 mM, 150 mM, 1M ammonium acetate for 20 minutes), followed by high-resolution reverse phase separation on the RP2 section of the column using an acetonitrile gradient of 0 to 80% for 120 minutes.

Spectra were acquired on a LTQ Velos linear ion trap tandem mass spectrometer (Thermo Electron Corporation, San Jose, CA) employing automated, data dependent acquisition. The mass spectrometer was operated in positive ion mode with a source temperature of 250 °C. As a final fractionation step, gas phase separation in the ion trap was employed to separate the peptides into 3 mass classes prior to scanning; the full MS scan range was divided into 3 smaller scan ranges (300–800, 800–

1,100, and 1,100–2,000 Da) to improve dynamic range. Each MS scan was followed by 5 MS/MS scans of the most intense ions from the parent MS scan. A dynamic exclusion of 1 minute was used to improve the duty cycle.

The raw data were extracted and searched using Spectrum Mill v3.03 (Agilent Technologies). MS/MS spectra with a sequence tag length of 1 or less were considered to be poor spectra and were discarded. The remaining MS/MS spectra were searched against maize B73 RefGen_v2 5a Working Gene Set downloaded from www.maizesequence.org. The enzyme parameter was limited to full tryptic peptides with a maximum mis-cleavage of 1. All other search parameters were set to Spectrum Mill's default settings (carbamidomethylation of cysteines, +/- 2.5 Da for precursor ions, +/- 0.7 Da for fragment ions, and a minimum matched peak intensity of 50%). Ox-Met, n-term pyro-Gln, and phosphorylation on Serine, Threonine, or Tyrosine were defined as variable modifications for phosphoproteome data. A maximum of 2 modifications per peptide was used. A 1:1 concatenated forward-reverse database was constructed to calculate the false discovery rate (FDR). The tryptic peptides in the reverse database were compared to the forward database, and were shuffled if they matched to any tryptic peptides from the forward database. Cutoff scores were dynamically assigned to each dataset to obtain the false discovery rates (FDR) of 0.02% for spectra, 0.14% for peptides, and 1.004% for proteins in the non-modified proteome (Table S2). FDRs for the phosphoproteome (Table S3) were 0.13% for spectra and 0.63% for phosphopeptides. Phosphorylation sites were localized to a particular amino acid within a phosphopeptide using the variable modification localization (VML) score in Agilent's Spectrum Mill software (*36*). Proteins that share common peptides were grouped using principles of parsimony to address protein database redundancy. Thus, proteins within the same group share the same set or subset of peptides. Phosphorylation levels were quantified by spectral counting. Spectral counts for each protein represent the total number of peptide spectral matches to that protein (*14*, *18*, *19*, *37*, *38*). Non-modified abundance was quantified using the distributed normalized spectral abundance factor (dNSAF) method (*39*), which distributes the spectral counts from shared peptides between the matching protein isoforms. Mass spectrometry technical replicate runs, when present, were summed to yield biological replicates. All biological replicates were normalized so that the total number of spectral counts was equal for each run. The proteomics quantification methodology was previously validated for the seed and leaf samples and shown to match known protein accumulation patterns (*14*, *18*).

RNA-seq library preparation, sequencing and analysis

Total RNA was isolated by TRIzol extraction (Life Technologies) from the same tissue samples used for proteomics. TRIzol isolated RNA was further purified using Qiagen RNeasy kit with on-column DNase treatment (Qiagen Inc.). Approximately four micrograms of total RNA was used as input. Each sequencing library was constructed using the TruSeq RNA Sample Prep Kit V2 (Illumina, San Diego, CA) according to manufacturer's instructions with the following modifications to confer strand-specificity. The polyA-selected RNA was used for first strand synthesis and the resulting cDNA:RNA hybrid was purified using RNAClean XP beads (Beckman, Brea, CA). The second strand synthesis was performed using a dNTP mix containing dUTPs (10mM dATPs, 10mM dGTPs, 10mM dCTPs, and 20mM dUTPs) and DNA Polymerase I (New England Biolabs, Ipswich, MA). The purified ligation products were incubated with Uracil DNA Glycosylase (Fermentas) before PCR amplification. The completed libraries were pooled in sets of 24 and each pool was sequenced on 4 lanes. RNA-seq libraries were sequenced using the Illumina HiSeq 2500 (Illumina) instrument as per manufacturer's instructions. Sequencing of libraries was performed up to 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline**.** RNA-seq libraries were prepared as biological triplicates for each tissue except for the vegetative meristem samples. Illumina HiSeq2500 output files in the FASTQ format were aligned to the *Zea mays* reference genome version AGPv2.0 (http://ftp.maizesequence.org/current/assembly/) using Bowtie version 2.1.0 (*40*) and Tophat version 2.0.8b (*41*) (flags = –F 0 –i 30 -M). Gene expression values were calculated using Cufflinks version 2.1.1 (flags = -u --library type fr-firststrand -b) (*42*). The B73 RefGen_v2 5a annotation file was used for quantitation of gene expression (http://ftp.maizesequence.org/current/working-set/).

Syntenic and non-syntenic gene annotations

As described previously *(42)*, 24,249 genes conserved at syntenic orthologous locations between maize and sorghum as well as 82,974 non-syntenic genes were identified from the maize B73 RefGen_v2 5a Working Gene Set.

mRNA abundance distributions

For all 62547 detected mRNAs, the abundance in all non-zero samples was averaged to get average FPKM. The average FPKM was then log2 transformed. A kernel density estimation of this log2 FPKM distribution was calculated using the R function density() with default settings. Other kernel density estimates were calculated in the same way for the subsets of mRNAs where proteins (17,582) and phosphoproteins (6,105) were observed as well as the set of syntenic (23,783) and non-syntenic (36,394) genes. The distributions were plotted using R (plot(), lines() and legend()).

Principal Component Analysis

Data from averaged biological replicates were used for both protein and mRNA. The full protein set in the 23 tissues with RNA-seq data was used (16,946 proteins) as well as all RNA-seq data from genes that were detected with > 1 FPKM in at least one tissue sample (43,994 genes). All values were incremented by 1 in order to eliminate zero values and the data was log transformed using base 2. Principal component analysis (PCA) was carried out separately on each data set using the R function prcomp() from the stats package with default parameters. The summary() function was used on the prcomp objects to extract the proportion of variance explained by each principal component (PC). The values of the first 4 PCs for each tissue was extracted by indexing the "x" matrix of each prcomp object (prcomp$x[,1:4]) and used to generate 2D PCA plots using the R plot() function. For both protein and mRNA data, using these 4 PC values for each tissue, the Euclidean distance was calculated between each tissue type using the R function dist(method="euclidean"). These distances were used for hierarchical clustering using average linkage with the R function hclust(method="average"). The two resulting dendrograms were manually reordered using the R reorder() function to make for easier visualization. For both the protein and mRNA data sets, the contributions of each gene to each of the first 4 principal components (the PCA loadings) were extracted from the prcomp objects by indexing the "rotation" matrix (prcomp$rotation[,1:4]). The absolute values were taken from each matrix. The union of genes represented in either the protein or mRNA data was used and zero values were filled in for genes that were not represented in either set. For all 4 principal components, the protein contributions were plotted against the corresponding mRNA contributions for each gene using the R plot() function. The Pearson correlation coefficients of each of these comparisons was calculated using the cor() function.

Tissue-wise mRNA-to-protein correlations

All detected mRNAs were filtered by retaining only genes that have average FPKM >= 1 in at least one of the 23 samples to give a set of 34,455 genes. All proteins detected in these 23 samples were filtered by retaining only those with at least 1 uniquely mapped peptide to give a set of 15,743 genes. The intersection of these two lists is a set of 15,364 genes. All available mRNA biological replicates were used for this set of genes along with the matching set of protein biological replicates. For each gene, the biological replicates were averaged to yield protein and mRNA abundance. For each sample, the genes with detected protein and mRNA were used to calculate the Spearman correlation coefficient using the R function cor().

Spearman correction

Using the separate biological data from the set described above for tissue-wise mRNA-to-protein correlations, the corrected Spearman correlation coefficient was calculated for each sample using the method described in (*43*).

mRNA-to-protein correlations within MapMan Bins

The set of genes described above in tissue-wise mRNA-to-protein correlations was used. The average of all biological replicates were used for both mRNA and protein abundance measurements. For each MapMan Bin, in each sample, calculations were done on the set of genes that had non-zero abundance for both mRNA and protein in that tissue. For sets with less than 3 genes, NA was returned. The R function cor() was used to calculate both Spearman and Pearson correlation coefficients for each set.

Gene-wise mRNA-to-protein correlations

For these calculations the average of all biological replicates was used as abundance measurements. A set of 16089 genes was found that have mRNA abundance >= 1 FPKM in at least one tissue and detectable protein in one of the 23 tissues profiled by RNA-seq. For each gene, the Spearman correlation coefficient was calculated for the abundance data across all 23 samples using the R function cor().

mRNA and protein CV binning and correlation coefficient distributions

The coefficient of variation was calculated for each protein abundance measurement using R (sd(Biological Replicates) / mean(Biological Replicates)). Five separate gene bins were constructed using CV ranges: [0,0.37), [0.37,0.65), [0.65,1.02), [1.02,1.73) and [1.73,2.45). The criteria for each bin were that the gene must have >= 10 samples (out of 23) where the protein CV is in that bin or lower. This results in overlapping bins where a gene will also be in every bin with higher CVs. This redundancy was eliminated by retaining genes only in the lowest CV bin that they could be put into. The correlation scores were then calculated using only the samples that met the given cutoff. The kernel density estimates of these four distributions along with the full gene set distribution from above (Gene-wise Protein Vs. mRNA Correlations) were calculated using the R function density() with default parameters and plotted using the R functions plot(), lines() and legend().

Co-expression networks

The WGCNA R package (*23*) was used to build both co-expression networks. All available biological replicates were averaged. A set of 3 networks was constructed in order to capture all available information and using all available genes that were detected in at least 4 tissues. These networks consist of 31,447 mRNA, 13,175 proteins, and 4,267 phosphoproteins. Another set of two networks was created for comparison between mRNA and protein. For this set, the 16,089 genes from above (Gene-wise mRNA-to-protein Correlations) was filtered down to 10,979 by removing genes that had fewer than 5 non-zero expression values in the 23 samples. The protein and mRNA data were used separately to create two networks. The parameters used for each network were identical. A soft power threshold of 12 was found to be satisfactory across all networks and both correlation methods. Adjacency matrices were built using the adjacency() function with type=”signed” using either Spearman correlation or biweight

midcorrelation (bicor). Next, topographical overlap matrices (TOMs) were constructed using the TOMsimilarity() function with default parameters. The TOM scores were used as edge weights in the analysis. Co-expression modules were constructed through hierarchical clustering of the TOM distance (1-TOM) using hclust() function with (method="average"). Modules were derived using the cutreeDynamic() function with parameters (deepSplit=2, pamRespectDendro=F, minClusterSize=30). Finally, similar modules were merged using the mergeCloseModules() function with (cutHeight=0.15).

Categorical enrichment

Enrichment analysis was carried out on the modules constructed from the co-expression networks. Each module represents a list of co-expressed genes. A custom R script was written to carry out the analysis separately on each module using the MapMan categories. For each module, every MapMan bin that was represented was examined for enrichment. A hypergeometric test was performed using the R function phyper() from the stats package. The total set of genes (Number of black and white balls) used was the intersection of genes with MapMan annotations and genes in the co-expression analysis. For this test, the number of white balls was the total number of genes annotated with the MapMan category. The number draws was the number of genes in the module and the number of white balls drawn was the number of genes in the module with the MapMan annotation – 1 (1 must be subtracted for right-tail calculations due to the implementation of this function). Within each module, p-values were corrected for multiple testing using the p.adjust() function with (method="fdr") which performs the Benjamini & Hochberg correction.

Comparison of categorical enrichment

All categories and their adjusted p-values for over-representation were extracted from every module in both the protein and mRNA networks. The p-values were converted to –log10(p-value). For each category, these scores were summed. The results for the protein and mRNA networks were plotted against each other using the R functions plot(). To compare the genes that are accounting for the category enrichments, all over-represented categories with adj. p-values < 0.05 were extracted from each network. All genes that were responsible for enrichment were extracted from each category. The categories with significant enrichment in both networks that contained at least 20 genes between the networks (union)

were retained. The genes representing each category fell into three groups: 1) genes accounting for enrichment in the protein network only, 2) genes accounting for enrichment in the mRNA network only and 3) genes accounting for enrichment in both networks. The fraction of the total number of genes represented from each category was calculated for the above 3 groups. These fractions were plotted as a percentage of the total genes using the R functions barplot() and axis().

Gene Regulatory Networks

For all Gene Regulatory Networks (GRNs), the biological replicates for each sample were averaged. A common set of 41,021 mRNAs were used as potential targets in the analyses. These mRNAs were observed with FPKM >= 1 in at least 1 tissue and also have non-zero values in at least 3 tissues. Transcription Factors (TFs) were defined using the GRASSIUS transcription factor list (*44*). For the three full GRNs, every available Transcription Factor (TF) was used that was detected in at least 3 samples resulting in 2,200 transcript, 545 protein, and 441 phosphopeptide quantified TFs being used as potential regulators. For the two normalized GRNs, the 539 TFs quantified as both mRNAs and proteins were used. GRNs were constructed using the GENIE3 algorithm (*32*). The GENIE3 R code was downloaded from the author's web page, http://www.montefiore.ulg.ac.be/~huynh-thu/software.html. This code was modified to take in separate regulator and target data.

Combining Multiple GRNs

To consolidate two or three networks, a new network was generated using the union of all TF expression data from the single networks as regulator inputs into the network and the same set of 41,021 target transcripts. This results in a network with redundancy at the gene level for TFs regulators that were quantified with multiple data types. To alleviate this redundancy, and obtain a combined score for each TF-Target edge, the product of all redundant edges was taken. When only a single edge existed (i.e. the TF was only quantified in one data type) when combining two data types, the square of the edge score was taken. For the final combined network consisting of all three data types, if only one edge was present, the edge score was cubed. If two edges were present, the product of the two edges was multiplied by the average of the two edges.

For the phosphorylation data, the networks were constructed using phosphopeptide quantification but when combined, all phosphopeptides from a given protein were averaged in order to get phosphoprotein level information.

ROC and PR curves

ChIP-Seq and RNA-seq studies in both wild-type and null mutant backgrounds have been previously performed for the maize transcription factors KNOTTED1 (KN1) (*24*) and Opaque2 (O2) (*34*). In these studies, gene sets were defined for TF bound genes (ChIP-Seq) as well as genes whose transcript abundance was modulated as a result of a null mutation in each TF. For our study, we used the intersection of these predefined lists as our standard set of targets for each TF. For each GRN, the GENIE3 score for the total set of potential mRNA targets was used for either KN1 or O2. This score was used to rank the predictions. The predictions for KN1 and O2 were then combined into one ordered list. Using the ROCR R package, these ranked lists were compared to the standard target sets using the prediction() function. Using the outputs of the prediction function, the performance() function was used to generate curve objects. For ROC curves, the parameters were (measure="tpr", x.measure="fpr") and for PR curves, the parameters were (measure="prec", x.measure="rec"). To calculate the area under the ROC curves, the performance() function was used with (measure="auc"). For the area under the PR curves, the PR curve objects were integrated by taking the mean of the y-values. The curves were plotted using the R functions plot(), lines(), abline() and legend().

Feature analysis of transcription factors that were preferential to each GRN

We examined 5 features of TFs weighted by their frequency in each network. In this way, distributions of abundance and CV were generated for both protein and mRNA data as well as distributions of protein vs. mRNA correlation scores. Using the protein and mRNA networks with the common 539 TF regulators, the top 200,000 edges were examined and each edge was distributed into 3 categories [1] unique to protein GRN, [2] shared between protein and mRNA GRNs and [3] unique to mRNA GRN. Next, the TF of each of these edges was used to create a redundant list for each category. The redundancy results from the same TF being predicted to regulate several targets and therefore having multiple edges in the

network. This redundancy was retained and acts as a weight to represent each TFs preference in the three categories. . Using these weighted lists, for all three categories, the distributions of the 5 features mentioned above were examined. In order to determine if the median of each distribution was high or low, a permutation test was conducted independently for each of the 3 categories for all 5 features. The permutation test was carried out by randomly sampling the 539 TFs and calculating the median value of the feature in question. For the randomly sampled lists, the weights (number of times a TF was repeated in the list) were eliminated and the number of unique TFs was randomly sampled. The original weights were then reapplied to the random list. This process was repeated 10,000 times for each distribution. The average median from the 10,000 permutations is plotted as a grey line **(Fig S30)**. P-values were calculated by dividing the number of permuted medians that were either greater than or less than the true distribution median by 10,000.

Functional annotations

The MapMan functional annotation file Zm_B73_5b_FGS_cds_2012 was downloaded from http://mapman.gabipd.org (*45*). GRASSIUS transcription factor annotations were downloaded from http://grassius.org on 12-3-2013 (*44*). CoGe "classical maize genes" were downloaded from http://genomevolution.org/wiki/index.php/Classical_Maize_Genes on 3-27-2012 (*22*).

Clustering Transcription Factors

Hierarchical clustering of the 393 transcription factors that were detected as mRNA, protein, and phosphoprotein was performed using MultiExperiment Viewer (MeV v4.8; http://www.tm4.org/mev/) software. Hierarchical clustering results were visualized as heat maps in MeV following row normalization using the "Normalize Gene/Row Vectors" row adjustment.

Supplementary Text

Recent studies, ranging in scope and complexity, have examined mRNA-to-protein correlation values and found a weakly positive relationship (*8–15*). Our integrated transcriptome and proteome datasets enabled correlation analyses across development encompassing 4,277 to 9,334 genes per tissue (**Table S7**). Both Spearman and Pearson correlations were calculated but we focused on Spearman correlations because they are less susceptible to outliers and do not rely on a linear dependence between variables. Within a given tissue the Spearman correlations ranged from 0.31 to 0.65 (**Fig. S13A** and **Table S7**), which is consistent with mRNA-to-protein correlations observed in a range of eukaryotic organisms (*8–15*). Recently, Csárdi et al. (*43*) reported that using a Spearman correction to account for experimental noise increases correlation values in yeast. However, applying their method to our data provided only a slight improvement for a corrected range of 0.39 to 0.74 (**Fig. S13A and Table S7**). Thus, other factors besides experimental noise underlie the modest correlations. Principal component analysis of the transcriptome and proteome uncovered similar relationships between tissue types (**Fig. S14-16**). However, the principal components of the transcriptome were different from those of the proteome (**Fig. S17**).

Similar to previous findings, we observed a wide range of mRNA-protein correlations dependent upon MapMan functional category (**Table S7**) (*12*, *14*). Intriguingly, the correlations for many MapMan categories varied widely depending on the tissue (**Table S7**). For example, the correlation for genes involved in Cell Wall Degradation varied more than ten-fold from 0.06 in the Mature Leaf to 0.8 in the Secondary Root. Thus, the outcome of a specific gene regulatory pathway within a given tissue is subject to complex regulatory relationships, which underlie the unique developmental biology of each tissue type.

As an alternative to looking at correlations between genes within a tissue type we calculated correlation values across all samples for each gene. These data validated the inference of protein levels from mRNA measurements for 1,468 genes (9.1%) whose Spearman correlations of mRNA to protein abundance were ≥ 0.75 (**Table S7**). Conversely, 1,247 genes (7.6%) exhibited negative correlations. We also examined the impact of experimental noise from protein data on gene-level correlations by calculating the average coefficient of variation (CV) of the biological replicates for protein abundance of each gene. We then grouped genes into discrete bins based upon maximum CV to examine general patterns among genes with low versus high CV values (**Fig. S13B)**. We observe that the bins with larger CV have lower correlation scores. However, even the highest confidence measurements (Lowest CVs) show a modest median correlation score of 0.45, which is a small shift from the median of 0.41 that we see in the background of all observed genes. These data highlight the complex mechanisms of regulation that act on protein-coding genes and illustrate the need for examining both mRNA and protein outputs of gene expression.

References and Notes

1. G. Krouk, J. Lingeman, A. Colon, G. Coruzzi, D. Shasha, *Genome Biol.* **14**, 123 (2013).
2. T. S. Gardner, J. J. Faith, *Phys. Life Rev.* **2**, 65–88 (2005).
3. Z. Bar-Joseph *et al.*, *Nat. Biotechnol.* **21**, 1337–42 (2003).
4. R. De Smet, K. Marchal, *Nat. Rev. Microbiol.* **8**, 717–29 (2010).
5. V. van Noort, B. Snel, M. A. Huynen, *Trends Genet.* **19**, 238–42 (2003).
6. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, *Science*. **302**, 249–55 (2003).
7. S. Horvath, J. Dong, *PLoS Comput. Biol.* **4**, e1000117 (2008).
8. B. Schwanhausser *et al.*, *Nature*. **473**, 337–342 (2011).
9. C. Vogel *et al.*, *Mol. Syst. Biol.* **6** (2010).
10. S. Ghaemmaghami *et al.*, *Nature*. **425**, 737–741 (2003).
11. K. Baerenfaller *et al.*, *Science*. **320**, 938–41 (2008).
12. A. Ghazalpour *et al.*, *PLoS Genet*. **7**, e1001393 (2011).
13. L. Ponnala, Y. Wang, Q. Sun, K. J. van Wijk, *Plant J.* **78**, 424–40 (2014).
14. J. W. Walley *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4808–4817 (2013).
15. M. P. Washburn *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3107–12 (2003).
16. H. Qiao *et al.*, *Science (80-.).* **338**, 390–393 (2012).
17. K. N. Chang *et al.*, *Elife*. **2**, e00675 (2013).
18. M. R. Facette, Z. Shen, S. P. Briggs, L. G. Smith, *Plant Cell.* ***25,*** 2798-2812 (2013).
19. C. Marcon *et al.*, *Plant Physiol.* **168**, 233–46 (2015).
20. D. Hebenstreit *et al.*, *Mol. Syst. Biol.* **7**, 497 (2011).
21. N. Nagaraj *et al.*, *Mol. Syst. Biol.* **7**, 548 (2011).
22. J. C. Schnable, M. Freeling, *PLoS One*. **6**, e17855 (2011).
23. P. Langfelder, S. Horvath, *BMC Bioinformatics*. **9**, 559 (2008).
24. N. Bolduc *et al.*, *Genes Dev.* **26**, 1685–90 (2012).
25. R. G. Schneeberger, P. W. Becraft, S. Hake, M. Freeling, *Genes Dev.* **9**, 2292–2304 (1995).
26. A. Gallavotti *et al.*, *Development*. **137**, 2849–56 (2010).
27. M. S. Mukhtar *et al.*, *Science*. **333**, 596–601 (2011).
28. R. Albert, H. Jeong, A. Barabasi, *Nature*. **406**, 378–82 (2000).
29. M. A. Calderwood *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7606–11 (2007).
30. H. Jeong, S. P. Mason, A. L. Barabási, Z. N. Oltvai, *Nature*. **411**, 41–2 (2001).
31. K. Baerenfaller *et al.*, *Mol. Syst. Biol.* **8** (2012).
32. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, *PLoS One*. **5**, 10 (2010).
33. D. Marbach *et al.*, *Nat. Methods*. **9**, 796–804 (2012).
34. C. Li *et al.*, *Plant Cell Online*, tpc.114.134858 (2015).
35. M. Schrynemackers, R. Küffner, P. Geurts, *Front. Genet.* **4**, 262 (2013).
36. R. J. Chalkley, K. R. Clauser, *Mol. Cell. Proteomics*. **11**, 3–14 (2012).
37. E. L. Huttlin *et al.*, *Cell*. **143**, 1174–1189 (2010).
38. H. Liu, R. G. Sadygov, J. R. Yates, *Anal. Chem.* **76**, 4193–4201 (2004).
39. Y. Zhang, Z. Wen, M. P. Washburn, L. Florens, *Anal. Chem.* **82**, 2272–2281 (2010).
40. B. Langmead, C. Trapnell, M. Pop, S. Salzberg, *Genome Biol*. **10**, R25 (2009).
41. D. Kim *et al.*, *Genome Biol.* **14**, R36 (2013).
42. C. Trapnell *et al.*, *Nat Biotech*. **31**, 46–53 (2013).
43. G. Csárdi, A. Franks, D. S. Choi, E. M. Airoldi, D. A. Drummond, *PLOS Genet.* **11**, e1005206 (2015).
44. A. Yilmaz *et al.*, *Plant Physiol.* **149**, 171–180 (2009).
45. O. Thimm *et al.*, *Plant J.* **37**, 914–939 (2004).

Supplemental Figures

Figure S1. Non-syntenic vs syntenic expression. For transcribed genes, syntenic genes are several times more likely to be detected as protein than non-syntenic genes. The "ratio" represents the ratio of blue bars (Translated Syntenic/ Translated Non-syntenic). An array of 6 mRNA expression level cutoffs was used to show that this observation is independent of transcript expression level.

Figure S2. Hierarchical clustering of transcription factor mRNA expression levels. Transcription factors enriched in specific tissues are listed in Table S5.

Figure S3. Hierarchical clustering of transcription factor protein expression levels. Transcription factors enriched in specific tissues are listed in Table S5.

Figure S4. Hierarchical clustering of transcription factor phosphoprotein expression levels. Transcription factors enriched in specific tissues are listed in Table S5.

Figure S5. Family-wise analysis of transcription factors at the mRNA level. (A) For each TF family, over-enrichment of family members was determined for each tissue. (B) The mRNA abundance of each TF for a given family is summed for each tissue, then hierarchically clustered and row-normalized.

Figure S6. Family-wise analysis of transcription factors at the protein level. (A) For each TF family, over-enrichment of family members was determined for each tissue. (B) The protein abundance of each TF for a given family is summed for each tissue, then hierarchically clustered and row-normalized.

Figure S7. Family-wise analysis of transcription factors at the phosphoprotein level. (A) For each TF family, over-enrichment of family members was determined for each tissue. (B) The phosphoprotein abundance of each TF for a given family is summed for each tissue, then hierarchically clustered and row-normalized.

Figure S8. Co-expression Network Clusters. WGCNA derived co-expression dendrograms and corresponding modules (colored boxes) for (A) mRNA network, (B) protein network and (C) phosphoprotein network. Colors correspond to co-expression modules and are manually annotated based on the tissue(s) of highest expression. Module expression and membership is described in Figs S9-12 and TableS6.

Figures S9-S12. Expression of co-expression clusters. The eigengenes derived from all WGCNA clusters (modules) depict a summarized expression vector of the whole module. Modules for mRNA (red), protein (blue) and/or phosphoprotein (green) are plotted. Each panel represents a different module. Modules with similar expression from different data types are depicted in the same panel. The titles of each plot along with the colored boarder are manual annotations based on the tissue(s) of highest expression and correspond to the module color in Figure S8. Each module has a separate tab in Table S6 with the same name as the plot in this figure. This tab lists all the genes in this module as well as the enriched MapMan categories.

Figure S13. Correlation of protein vs. mRNA expression. (A) The sample-wise spearman correlation was calculated for each tissue. A correction for correlation scores based on replicate reproducibility was also calculated (35). (B) The gene-wise spearman correlation was calculated by examining the protein and

mRNA levels for each gene individually across all tissues and plotting the distribution (black dashed line). The genes were then binned based on the CV of protein expression and the distribution for each bin is shown. The upper left legend displays the CV range for each bin, the number of genes in each bin ("n") and the mean normalized protein abundance for each bin ("dNSAF")

Figure S14. Comparison of principle component analysis (PCA) between transcript and protein data sets. (A) The proportion of total variance explained by each principal component from mRNA data. (B) The proportion of total variance explained by each principal component from the protein data.

Figure S15. PCA results for transcript and protein data sets. Two-dimensional principal component plots of the first 4 principal components plotted pairwise against each other for both protein (upper right triangle) and mRNA (bottom left triangle). Colors represent developmentally similar tissues. See figure S16 for color-to-tissue assignments

Figure S16. Comparison of PCA between transcript and protein data sets. Hierarchical clustering dendrograms constructed using values of the first 4 principal components for each sample type. Colors are the same as in Fig. S15.

Figure S17. Comparison of PCA between transcript and protein data sets. For the first 4 principal components (PC), the contributions (PCA loadings) of each gene from the protein data (x-axis) and mRNA data (y-axis) are plotted against each other. The Pearson correlation coefficient (PCC) is indicated for each PC.

Figure S18. Assessment of soft thresholds used in WGCNA to generate the Spearman based co-expression networks. A range of soft thresholds was evaluated by looking at (A) Scale independence for the mRNA co-expression network. (B) Mean connectivity in the mRNA co-expression network. (C) Scale independence for the protein co-expression network. (D) Mean connectivity in the protein co-expression network.

Figure S19. Assessment of soft thresholds used in WGCNA to generate the Bicor based co-expression networks. A range of soft thresholds was evaluated by looking at (A) Scale independence for the mRNA co-expression network. (B) Mean connectivity in the mRNA co-expression network. (C) Scale independence for the protein co-expression network. (D) Mean connectivity in the protein co-expression network.

Figure S20. Edge overlap of mRNA and Protein co-expression networks. Using co-expression networks generated using either spearman correlation or biweight midcorrelation from a common set of detected proteins and mRNAs, the edge overlap is shown vs. network size (blue) compared to random (black dashed). For reference, the minimum edge score (correlation) is also shown vs. network size (red).

Figure S21. Heatmap showing the jaccard index (intersect/union) of co-expression networks built using single biological replicates of protein or mRNA abundance measurements.

Figure S22. mRNA to protein co-expression hub overlap (A) For the Co-expression networks built using the biweight midcorrelation and a common set of detected mRNA and protein, the number of edges a given gene (node) has in the protein (x-axis) and mRNA (y-axis) is shown. Nodes above the 90^{th} percentile for number of edges (degree) are considered hubs and colored based on whether they are a hub

in the protein (blue), mRNA (red), or both (green) networks. Black dots represent non-hub nodes. (B) For the Co-expression network built using spearman correlation (Fig 2), the percentage of all hub genes that are unique to the RNA network, unique to the protein network or shared between networks is plotted as a function of network size.

Figure S23. MapMan functional category enrichment in mRNA vs. protein co-expression network modules. In total 1,089 MapMan categories were enriched. Data are the sum of the $\log^{-10}$ P-value for every module in which a given MapMan category was present.

Figure S24. Quality of the full gene regulatory networks. Both single networks and combined networks are shown. KN1 and O2 target genes were obtained from previously published ChIP-seq datasets (*24*, *34*). (A) Receiver operating characteristic (ROC) curve. (B) Precision-recall curve. Phosphorylation modifications were predominantly localized to a specific site on both KN1 (GRMZM2G017087, peptide – NILSSGSSEEDQEG**s**GGETELPEVDAHGVDQELK, site-S225) and O2 (GRMZM2G015534, peptide – DP**s**PSDEDMDGEVEILGFK, site-S225).

Figure S25. Comparison of the full-sized gene regulatory networks. GRNs were reconstructed using TF regulators quantified as mRNAs (2,200 TFs), proteins (545 TFs), or phosphopeptides (441 TFs), and 41,021 shared potential target genes were quantified as mRNAs. (A) Overlap of the true positive predictions from the top 500 true GRN predictions for O2 quantified as mRNA, protein, or phosphopeptide. (B) GRN precision as a function of network size was calculated using KN1 by comparing the number of true positive (TP) vs false positive (FP) predictions. As the number of edges increased the prediction score decreased. A cutoff of 200,000 edges (vertical dashed line) was used to select the set of high-confidence predictions for all three GRNs. (C) Overlap of the TF-target predictions for the top 200,000 scoring predictions in each GRN.

Figure S26. Edge Overlap of full-sized gene regulatory networks. Using the GRNs that were constructed using all available TF information for each data type, the edge overlap was evaluated for an array of network sizes ranging from 100,000 to 10 million. For each size, all edges were categorized as being specific to one network or shared between two or all networks. The sizes of each category are represented by stacked colored bars as a percentage of all edges represented.

Figure S27. Differential expression of 393 TF genes measured as mRNAs, proteins, and phosphoproteins. (A) Heat maps ordered by hierarchical clustering of mRNA abundance. (B) Heat maps ordered by hierarchical clustering of protein abundance. (C) Heat maps ordered by hierarchical clustering of phosphoprotein abundance.

Figure S28. Quality of the GRNs reconstructed using 539 TFs quantified by their mRNA or protein abundance. (A) Receiver operating characteristic (ROC) curve. (B) Precision-recall curve. Standard sets based on KN1 and O2 target genes, obtained from previously published ChIP-seq datasets (*24*, *34*). Overlap of the true positive predictions from the top 500 true GRN predictions for (C) Kn1 and (D) O2.

Figure S29. Comparison of predictions in GRNs made using only 539 TFs. (A) Percentage of predictions (edges) that were conserved between GRNs made using the mRNA or protein to measure TF abundance. (B) Overlap of the TF-target predictions for the top 200,000 scoring predictions in each GRN.

Figure S30. TF regulator expression. TF regulators were weighted based on the number of edges they have in each of 3 categories; [1] edges unique to the protein network, [2] edges shared between the two networks and [3] edges unique to the transcript network. Features of the TFs were then examined for each category after applying the category weights. (A) The distributions of weighted protein abundance of TF regulators in the 3 categories. (B) The distributions of weighted coefficient of variation (CV) for protein abundance of TF regulators in each category. (C) The distributions of weighted transcript abundance of TF regulators in each category. (D) The distributions of weighted CVs for transcript abundance of TF regulators in each category. (E) The distributions of Spearman correlations of mRNA-to-protein abundance for TF regulators in each category. P-values were determined using a permutation test with 10,000 repetitions. Grey lines represent the average median of all permutation tests.

Figure S31. Conservation of mRNA GRN predictions in Mo17. TF regulators from the mRNA GRN who's mRNA is differentially expressed between B73 and Mo17 are shown on the left. The "N=" number next to the gene accession is the number of predicted targets for that regulator. On the right, bars represent the percentage of a given regulators target genes that are also differentially expressed between B73 and Mo17. Colored bars indicate a significant overrepresentation in differentially expressed target genes. P-values for this overrepresentations are printed to the right of each bar.

Figure S32. Conservation of protein GRN predictions in Mo17. TF regulators from the protein GRN whose protein is differentially expressed between B73 and Mo17 are shown below the x-axis. The "N=" number next to the gene accession is the number of predicted targets for that regulator. Bars above represent the percentage of a given regulators target genes who are also differentially expressed between B73 and Mo17. Colored bars indicate a significant overrepresentation in differentially expressed target genes. P-values for this overrepresentations are printed on top of each bar.

Figure S33. Conservation of phosphopeptide GRN predictions in Mo17. TF regulators from the phosphopeptide GRN that are differentially expressed between B73 and Mo17 are shown below the x-axis. The "N=" number next to the gene accession is the number of predicted targets for that regulator. Bars above represent the percentage of a given regulators target genes who are also differentially expressed between B73 and Mo17. Colored bars indicate a significant overrepresentation in differentially expressed target genes. P-values for this overrepresentations are printed on top of each bar.

Supplemental Table Legends

Table S1. Transcriptome. Abundance values are FPKM.

Table S2. Non-modified proteome. Abundance values are dNSAF.

Table S3. Phosphoproteome. Abundance values are spectral counts of both phosphopeptides and phosphoproteins.

Table S4. Pearson correlations of biological replicates within each tissue.

Table S5. Transcription factors and the tissues they are enriched in. Tissue enrichment was determined using hierarchical clustering (Figs. S3-S5).

Table S6. Overrepresentation of MapMan functional categories and gene lists for co-expression modules generated using WGCNA. For this analysis all genes quantified as mRNA, protein, or phosphoproteins were used. The tissue(s) corresponding in each module were determined from Figures S9-S12.

Table S7. Correlation of mRNA-to-protein abundance. Spearman and Pearson correlations at the tissue, MapMan bin, and gene levels.

Table S8. Pairwise transcript and protein co-expression networks. For these network reconstructions we used only the 10,979 genes that were detected both as transcripts and proteins in at least 5 of the 23 developmental gene expression atlas tissues in which we profiled both mRNA and protein.

Table S9. Transcript and proteins present in co-expression modules derived from Table S8. Lists all genes that group into co-expression modules in either the transcript or protein co-expression networks as well as the MapMan functional categories that were enriched within each module.

Table S10. Full-sized gene regulatory networks. The top 1 million edges (predictions) for GRNs reconstructed using the mRNA, protein, and/or phosphopeptide abundance of quantified transcription factors (regulators) to predict genes that they regulate (quantified as mRNA).

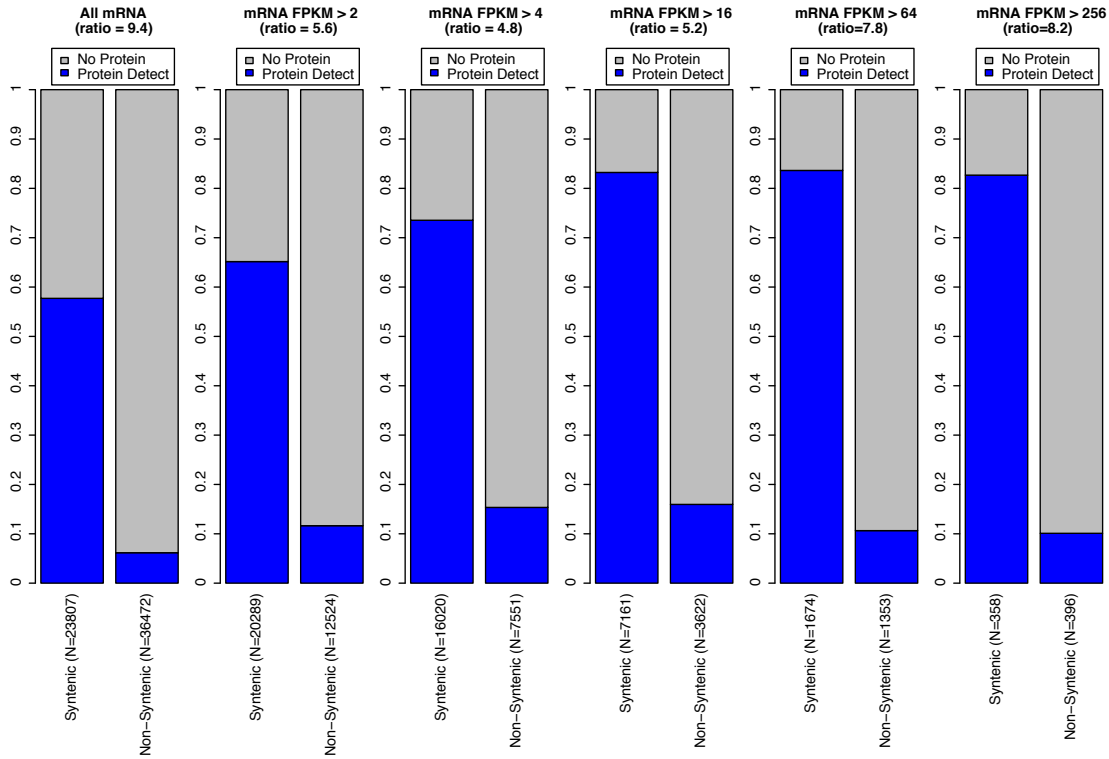


Figure S1. Non-syntenic vs syntenic expression. For transcribed genes, syntenic genes are several times more likely to be detected as protein than non-syntenic genes. The "ratio" represents the ratio of blue bars (Translated Syntenic/ Translated Non-syntenic). An array of 6 mRNA expression level cutoffs was used to show that this observation is independent of transcript expression level.

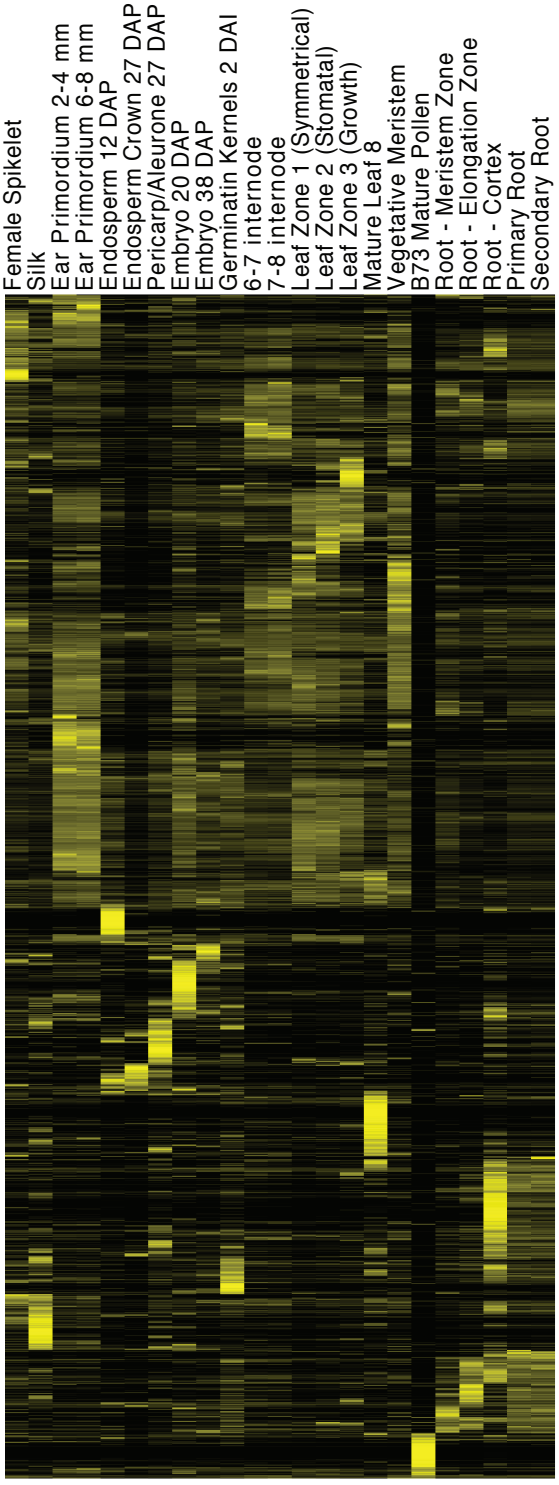


Figure S2. Hierarchical clustering of transcription factor mRNA expression levels. Transcription factors enriched in specific tissues are listed in Table S5.

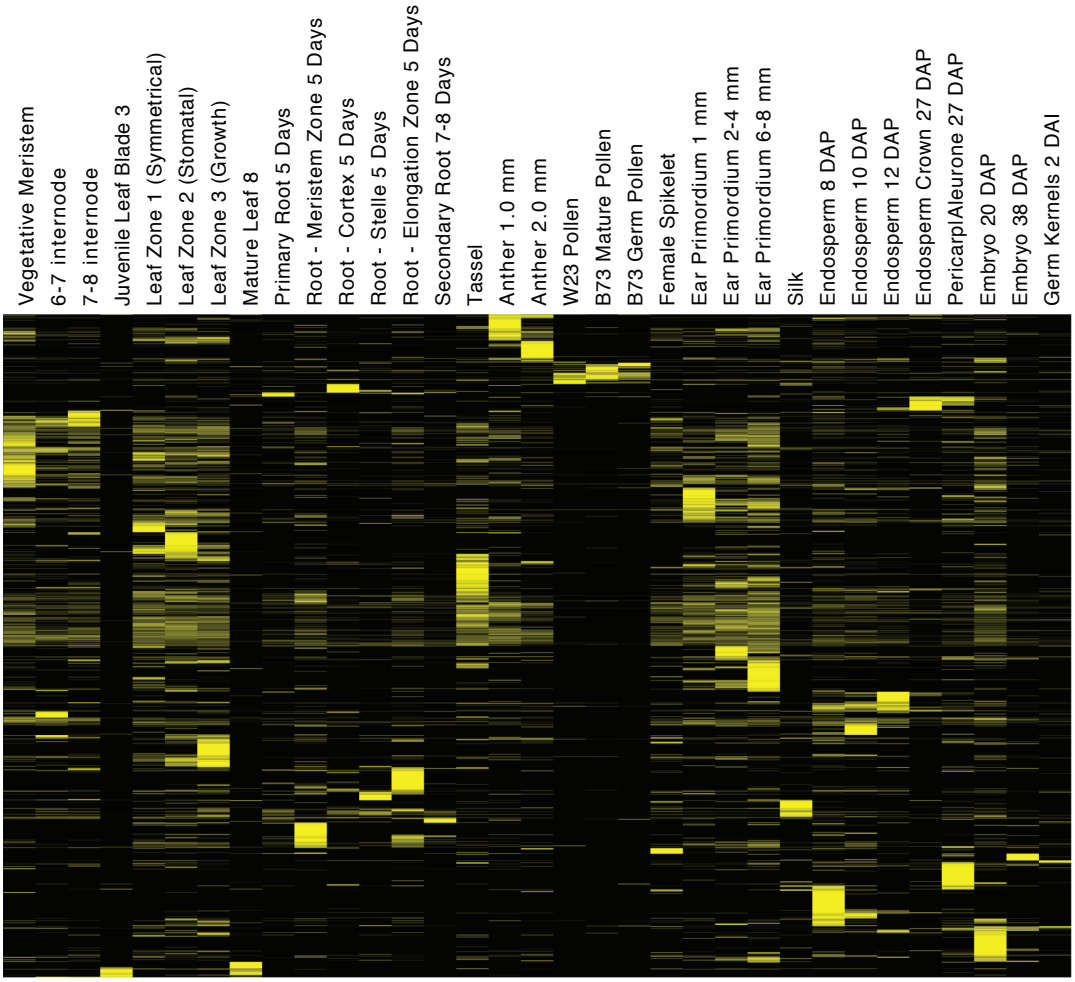


Figure S3. Hierarchical clustering of transcription factor protein expression levels. Transcription factors enriched in specific tissues are listed in Table S5.

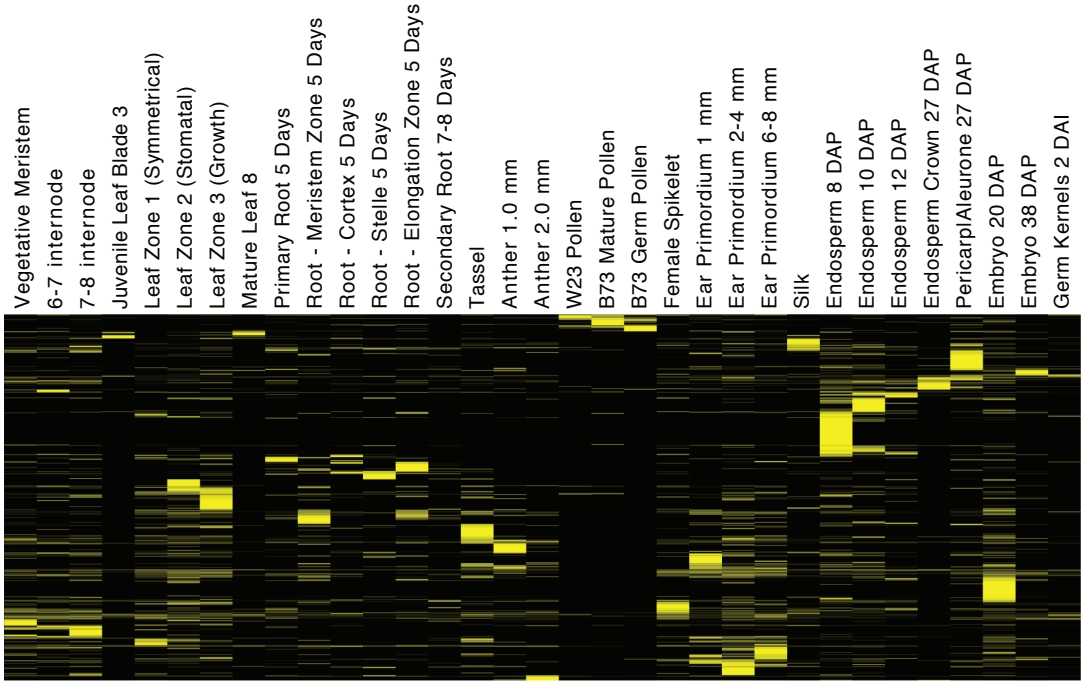


Figure S4. Hierarchical clustering of transcription factor phosphoprotein expression levels. Transcription factors enriched in specific tissues are listed in Table S5.

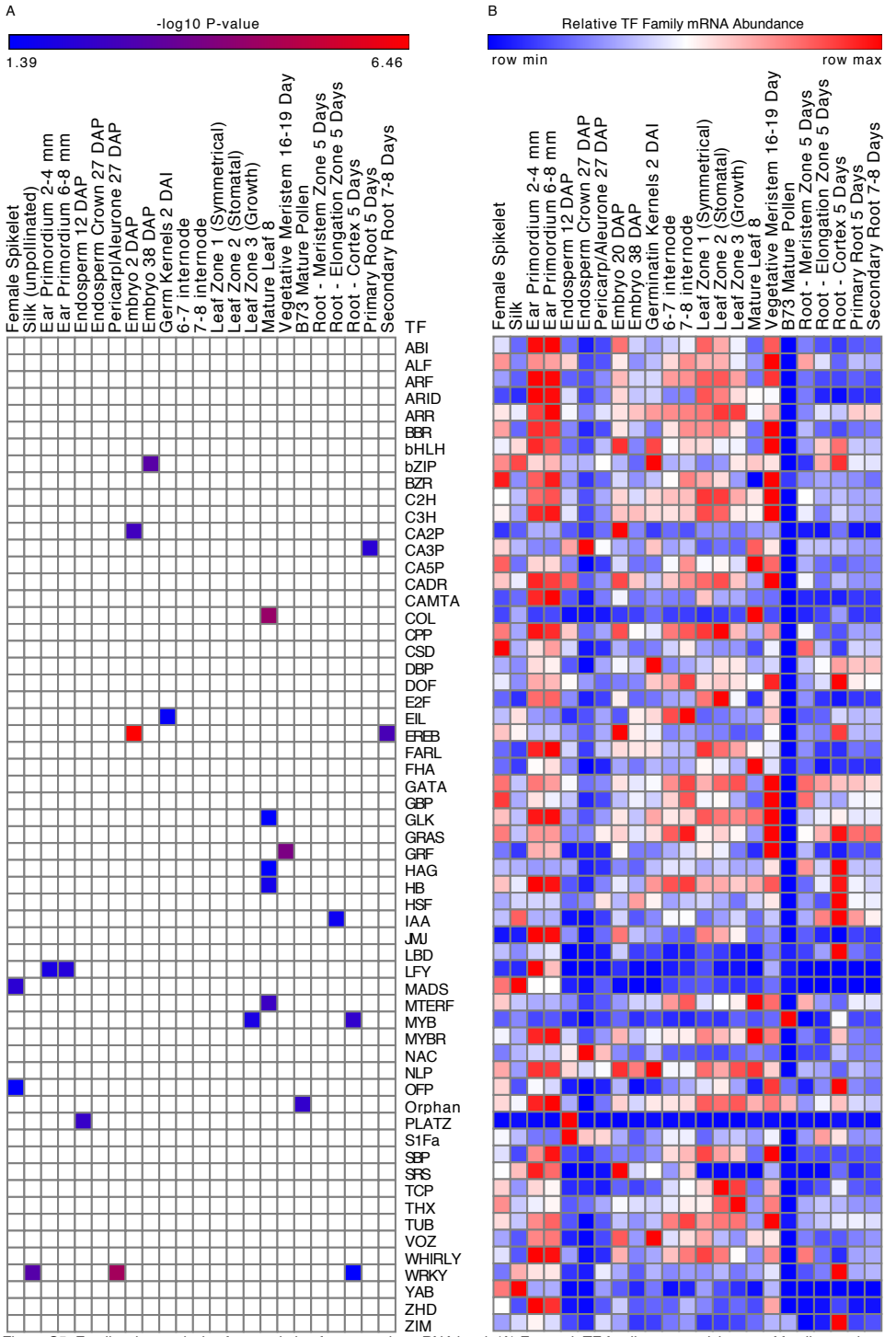


Figure S5. Family-wise analysis of transcription factors at the mRNA level. (A) For each TF family, over-enrichment of family members was determined for each tissue. (B) The mRNA abundance of each TF for a given family is summed for each tissue, then hierarchically clustered and row-normalized.

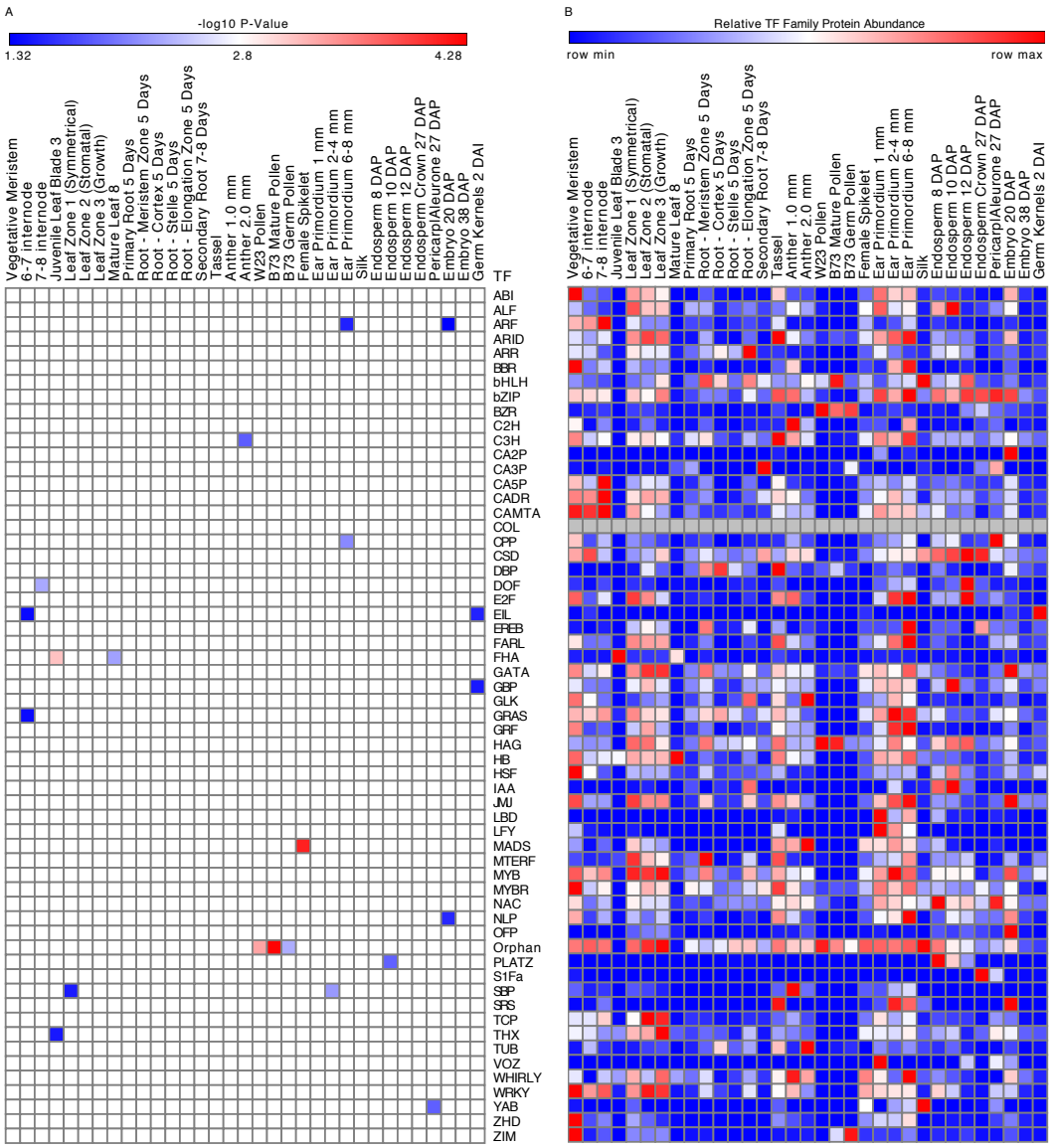


Figure S6. Family-wise analysis of transcription factors at the protein level. (A) For each TF family, over-enrichment of family members was determined for each tissue. (B) The protein abundance of each TF for a given family is summed for each tissue, then hierarchically clustered and row-normalized.

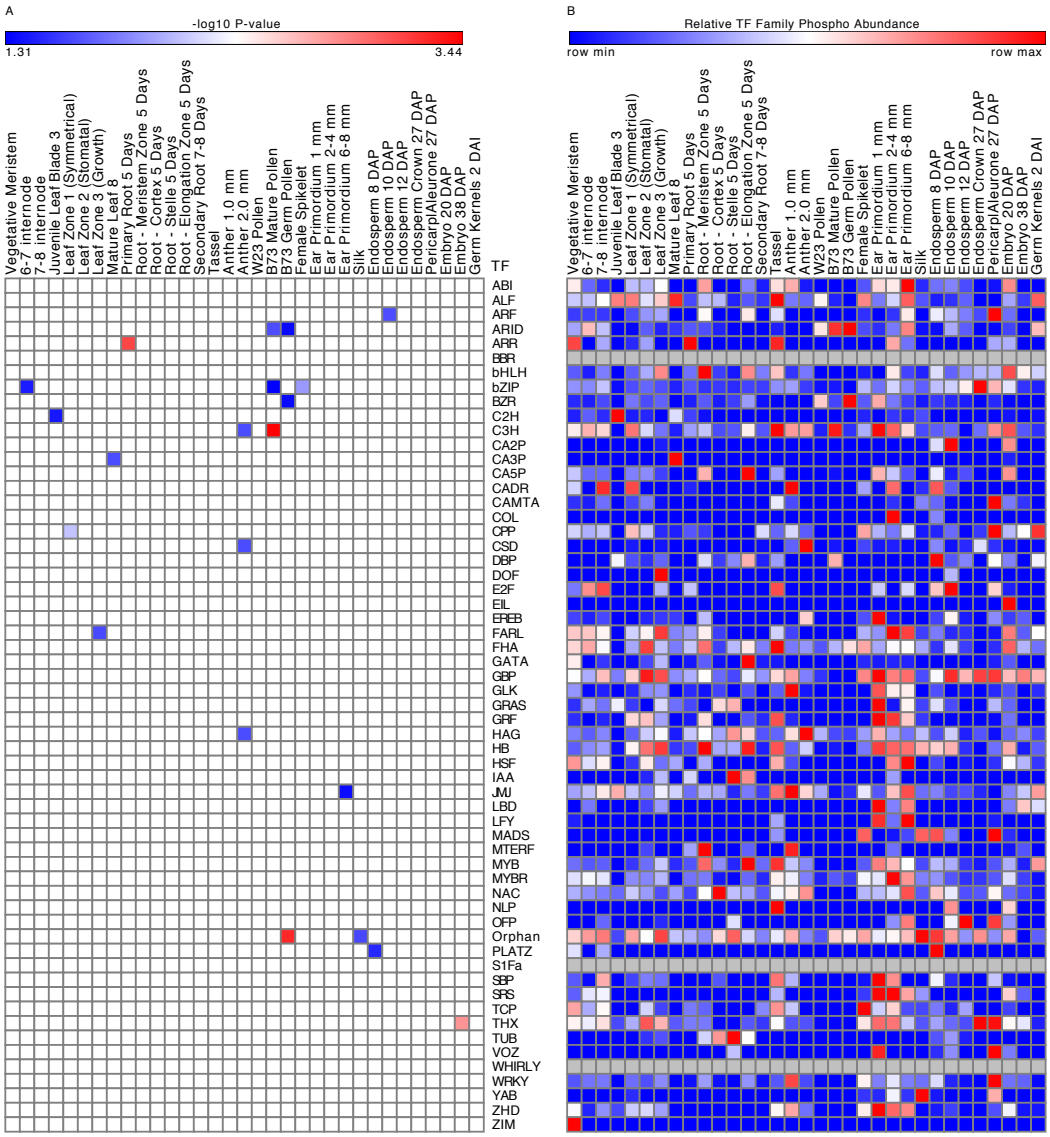


Figure S7. Family-wise analysis of transcription factors at the phosphoprotein level. (A) For each TF family, over-enrichment of family members was determined for each tissue. (B) The phosphoprotein abundance of each TF for a given family is summed for each tissue, then hierarchically clustered and row-normalized.

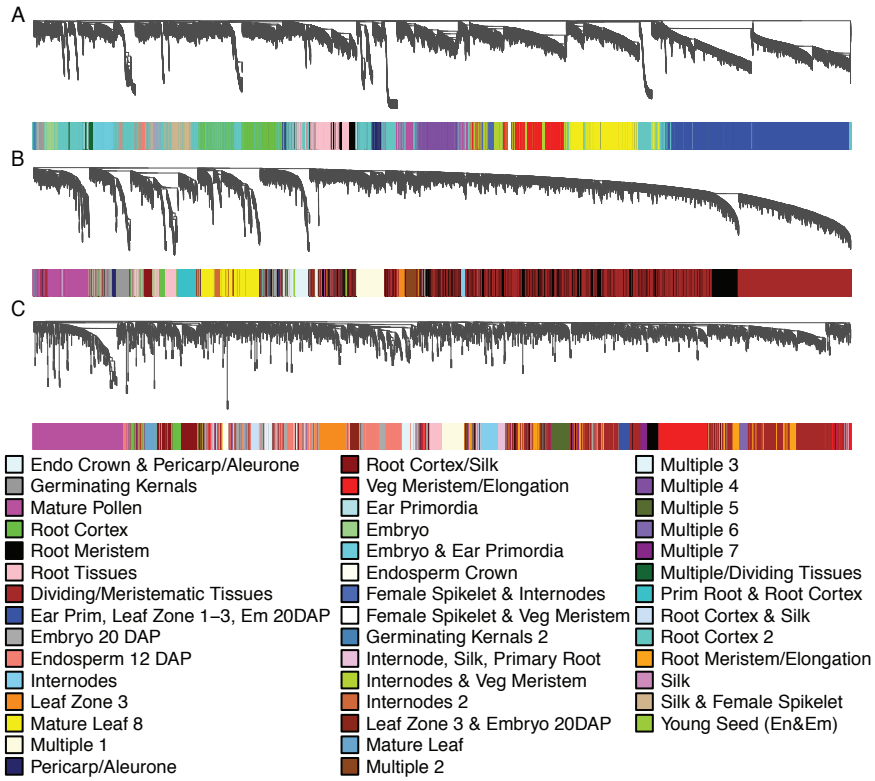
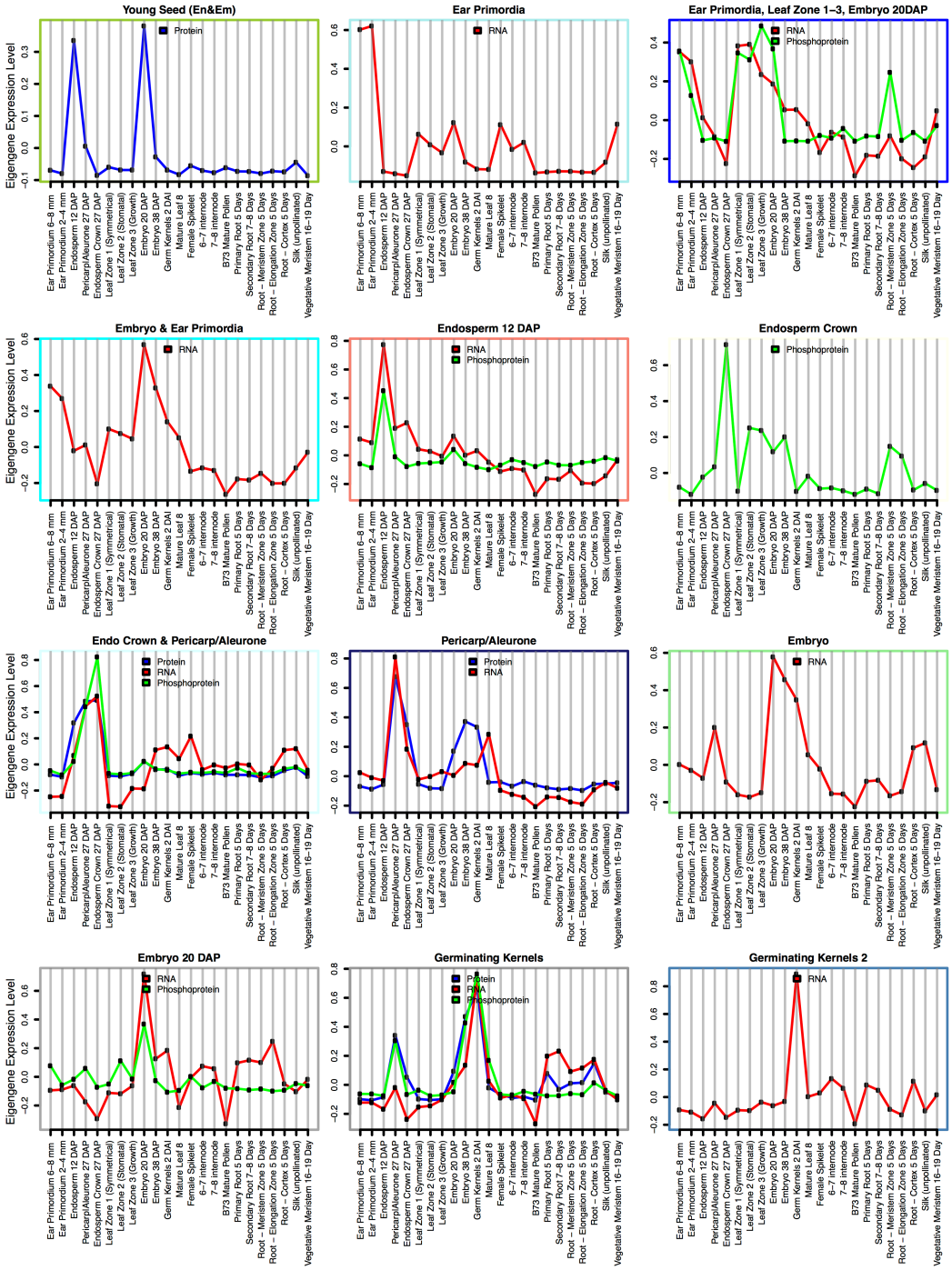
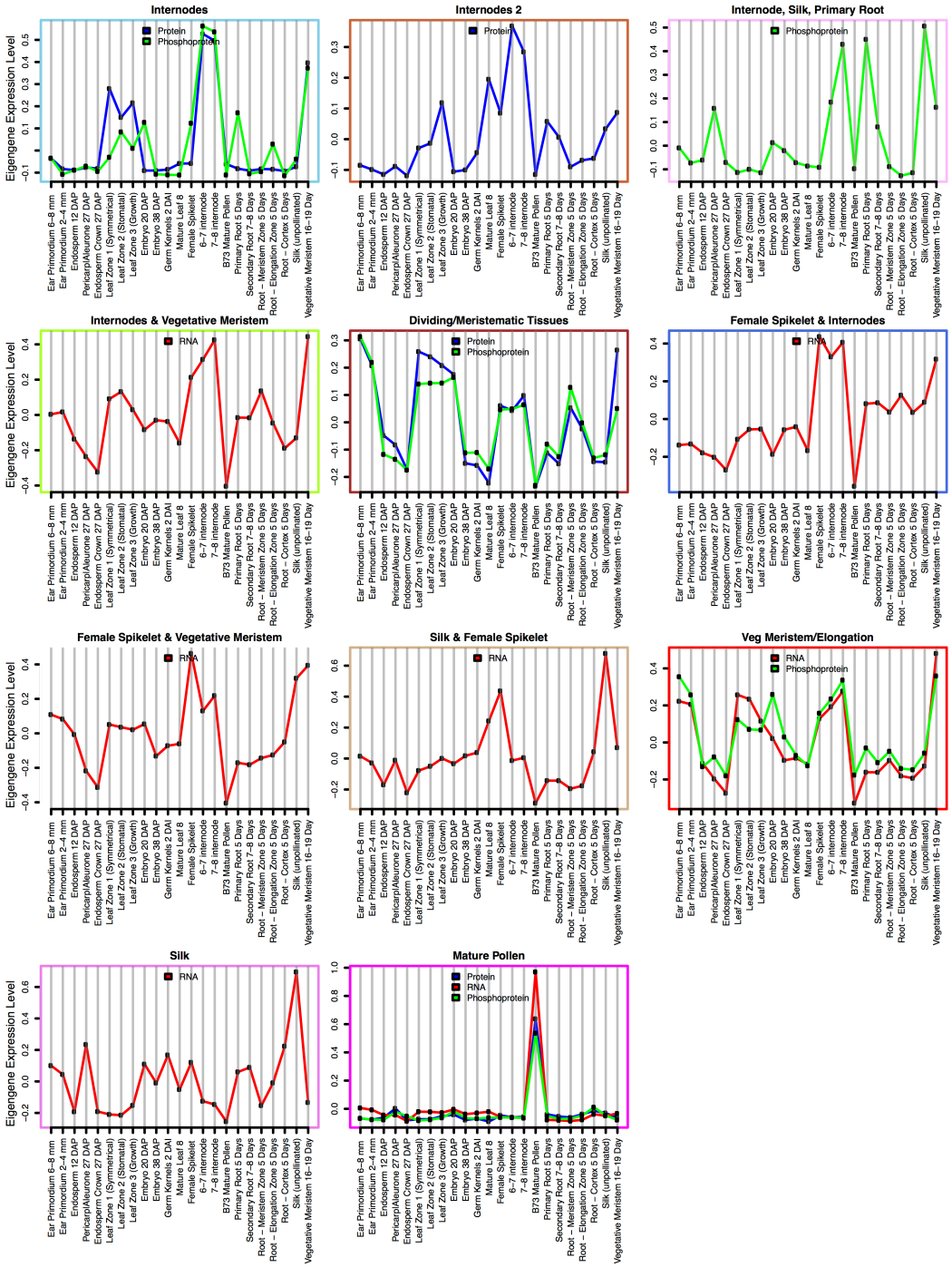


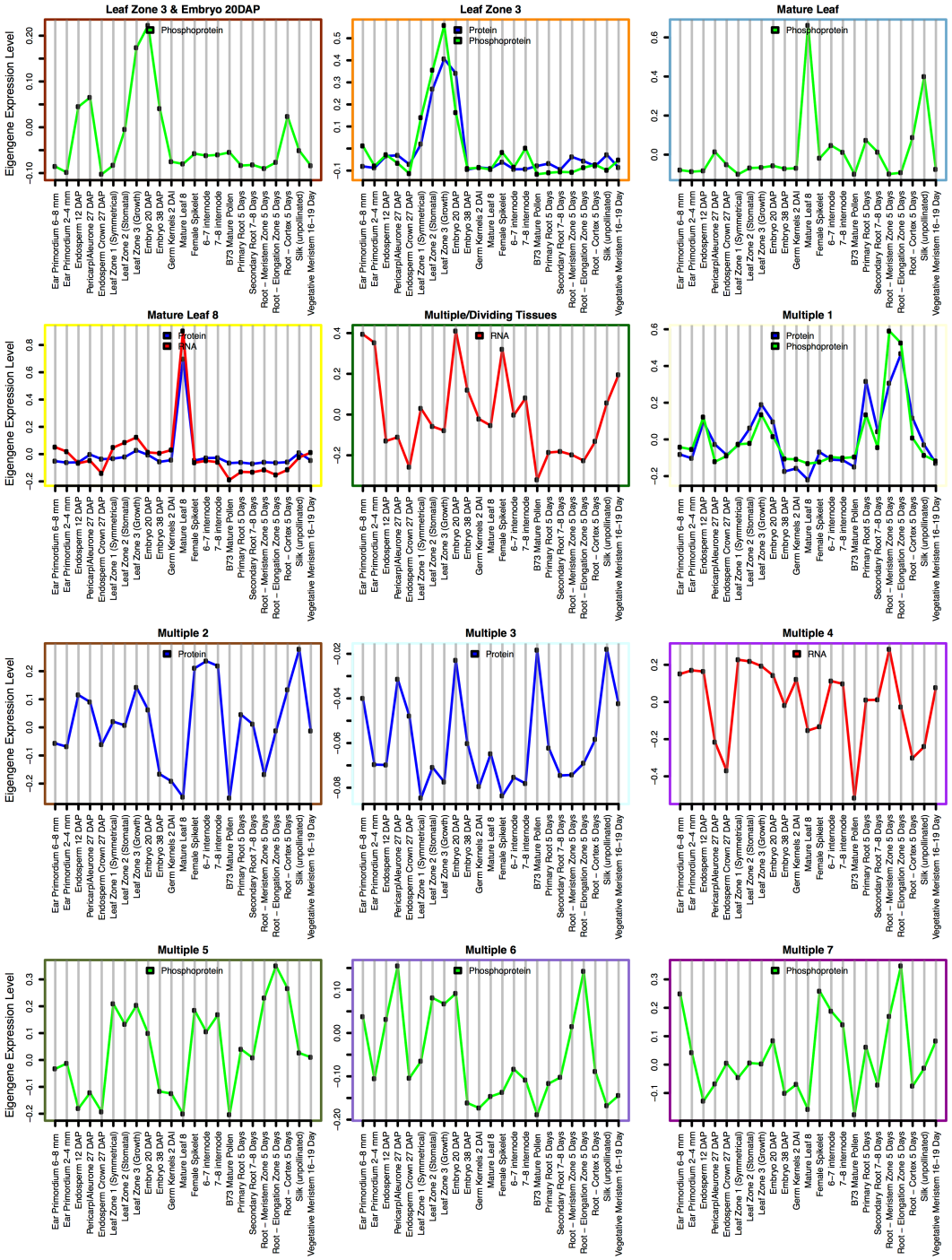
Figure S8. Co-expression Network Clusters. WGCNA derived co-expression dendrograms and corresponding modules (colored boxes) for (A) mRNA network, (B) protein network and (C) phosphoprotein network. Colors correspond to co-expression modules and are manually annotated based on the tissue(s) of highest expression. Module expression and membership is described in Figs S9-12 and TableS6.



Figures S9..Expression of co-expression clusters. The eigengenes derived from all WGCNA clusters (modules) depict a summarized expression vector of the whole module. Modules for mRNA (red), protein (blue) and/or phosphoprotein (green) are plotted. Each panel represents a different module. Modules with similar expression from different data types are depicted in the same panel. The titles of each plot along with the colored boarder are manual annotations based on the tissue(s) of highest expression and correspond to the module color in Figure S8. Each module has a separate tab in Table S6 with the same name as the plot in this figure. This tab lists all the genes in this module as well as the enriched MapMan categories.



Figures S10. Expression of co-expression clusters. The eigengenes derived from all WGCNA clusters (modules) depict a summarized expression vector of the whole module. Modules for mRNA (red), protein (blue) and/or phosphoprotein (green) are plotted. Each panel represents a different module. Modules with similar expression from different data types are depicted in the same panel. The titles of each plot along with the colored boarder are manual annotations based on the tissue(s) of highest expression and correspond to the module color in Figure S8. Each module has a separate tab in Table S6 with the same name as the plot in this figure. This tab lists all the genes in this module as well as the enriched MapMan categories.



Figures S11. Expression of co-expression clusters. The eigengenes derived from all WGCNA clusters (modules) depict a summarized expression vector of the whole module. Modules for mRNA (red), protein (blue) and/or phosphoprotein (green) are plotted. Each panel represents a different module. Modules with similar expression from different data types are depicted in the same panel. The titles of each plot along with the colored boarder are manual annotations based on the tissue(s) of highest expression and correspond to the module color in Figure S8. Each module has a separate tab in Table S6 with the same name as the plot in this figure. This tab lists all the genes in this module as well as the enriched MapMan categories.

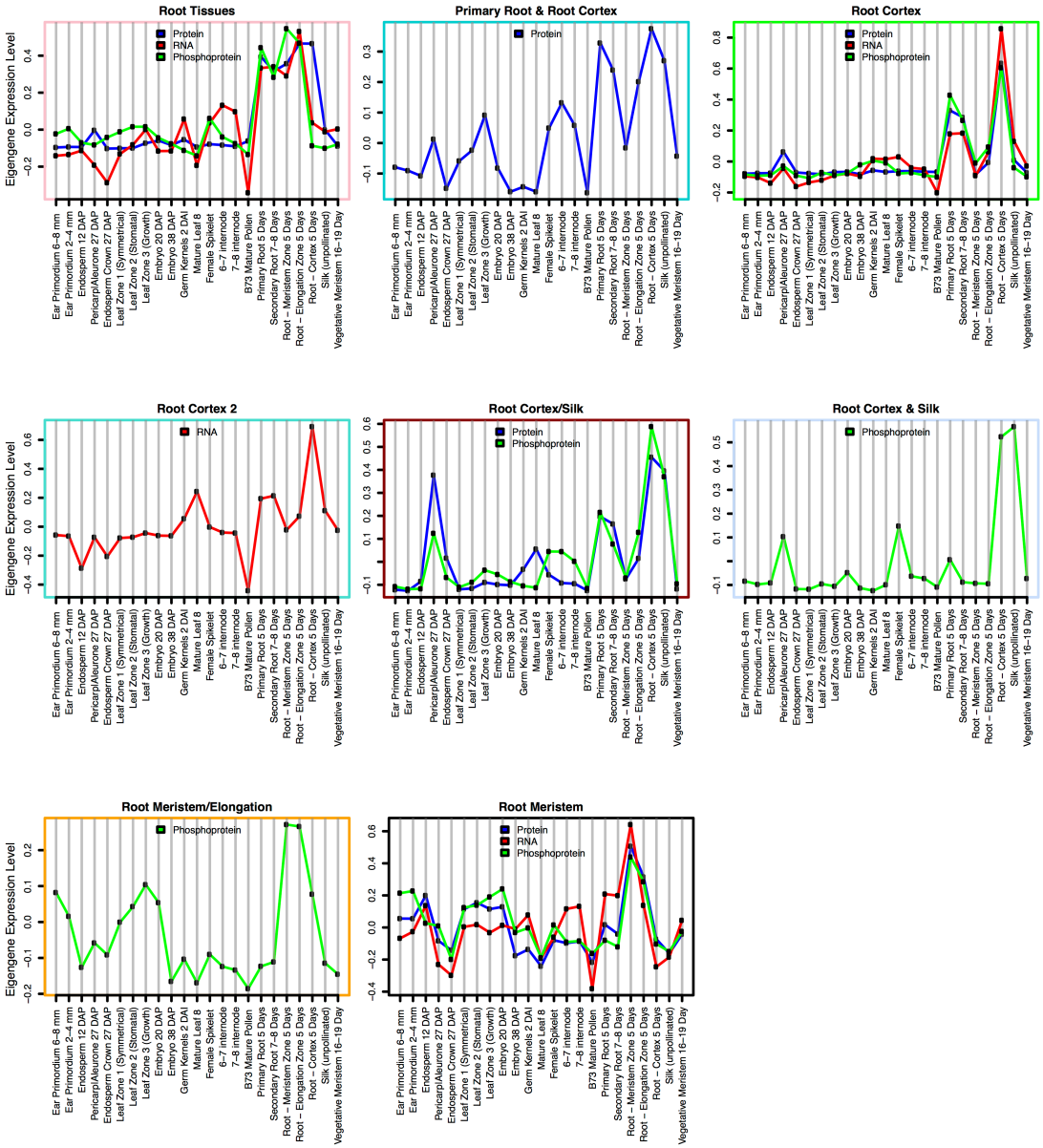


Figure S12. Expression of co-expression clusters. The eigengenes derived from all WGCNA clusters (modules) depict a summarized expression vector of the whole module. Modules for mRNA (red), protein (blue) and/or phosphoprotein (green) are plotted. Each panel represents a different module. Modules with similar expression from different data types are depicted in the same panel. The titles of each plot along with the colored boarder are manual annotations based on the tissue(s) of highest expression and correspond to the module color in Figure S8. Each module has a separate tab in Table S6 with the same name as the plot in this figure. This tab lists all the genes in this module as well as the enriched MapMan categories.

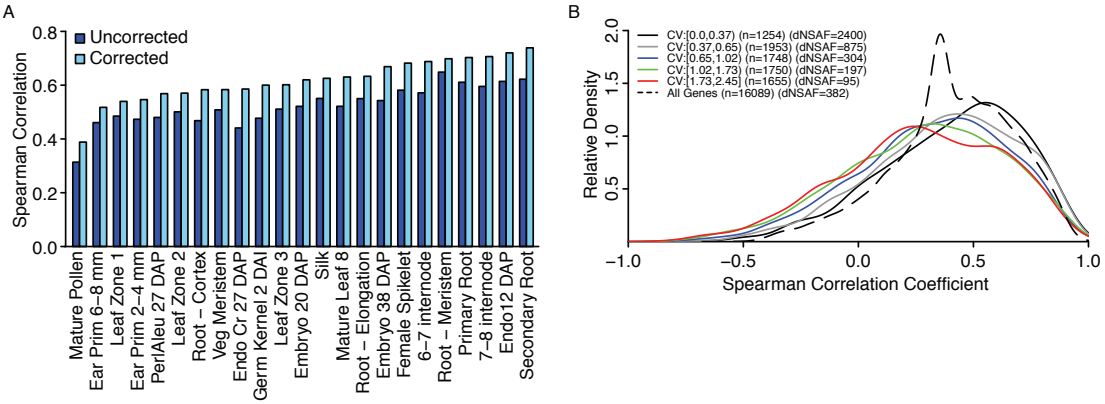


Figure S13. Correlation of protein vs. mRNA expression. (A) The sample-wise spearman correlation was calculated for each tissue. A correction for correlation scores based on replicate reproducibility was also calculated (35). (B) The gene-wise spearman correlation was calculated by examining the protein and mRNA levels for each gene individually across all tissues and plotting the distribution (black dashed line). The genes were then binned based on the CV of protein expression and the distribution for each bin is shown. The upper left legend displays the CV range for each bin, the number of genes in each bin (“n”) and the mean normalized protein abundance for each bin (“dNSAF”)

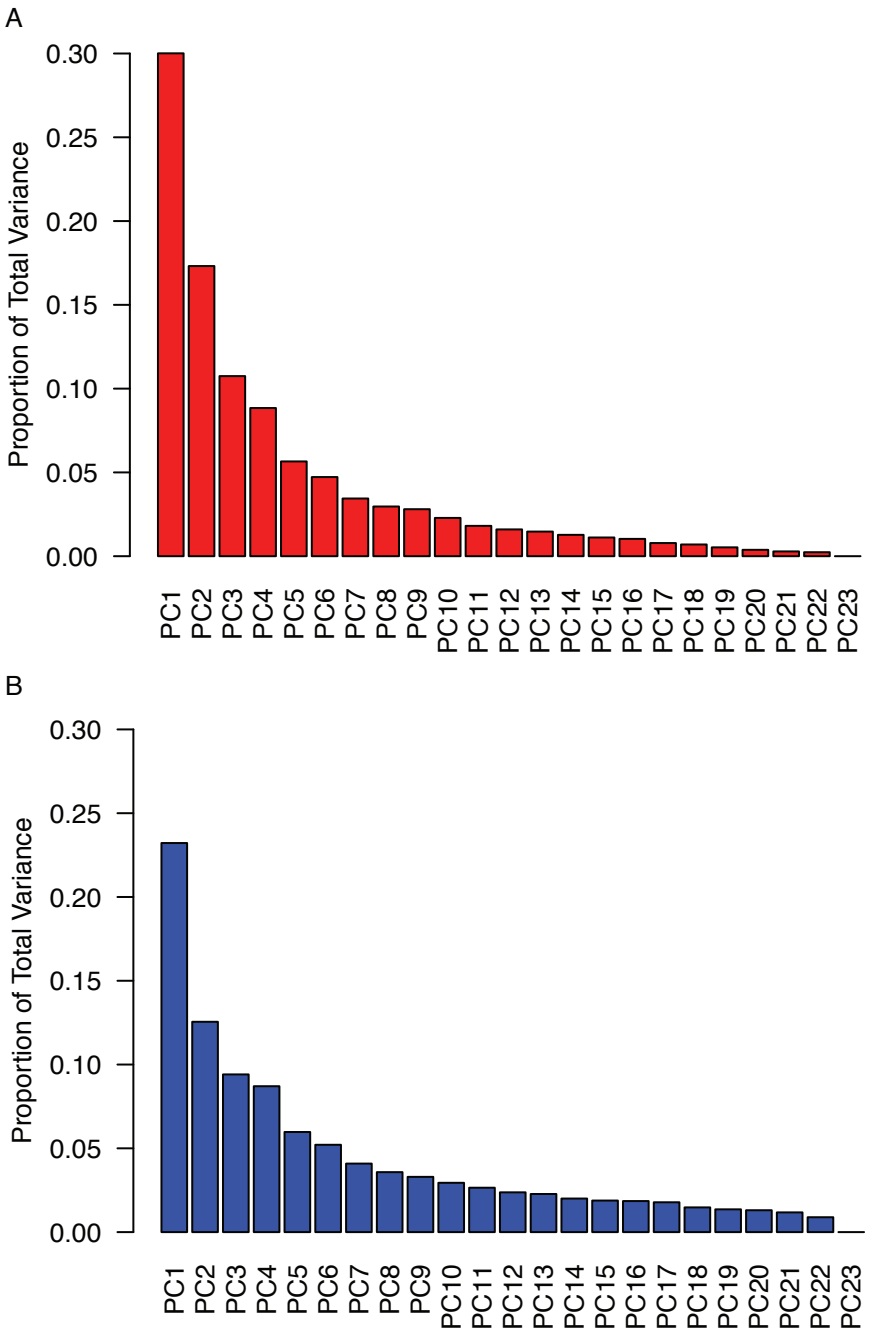


Figure S14. Comparison of principle component analysis (PCA) between transcript and protein data sets. (A) The proportion of total variance explained by each principal component from mRNA data. (B) The proportion of total variance explained by each principal component from the protein data.

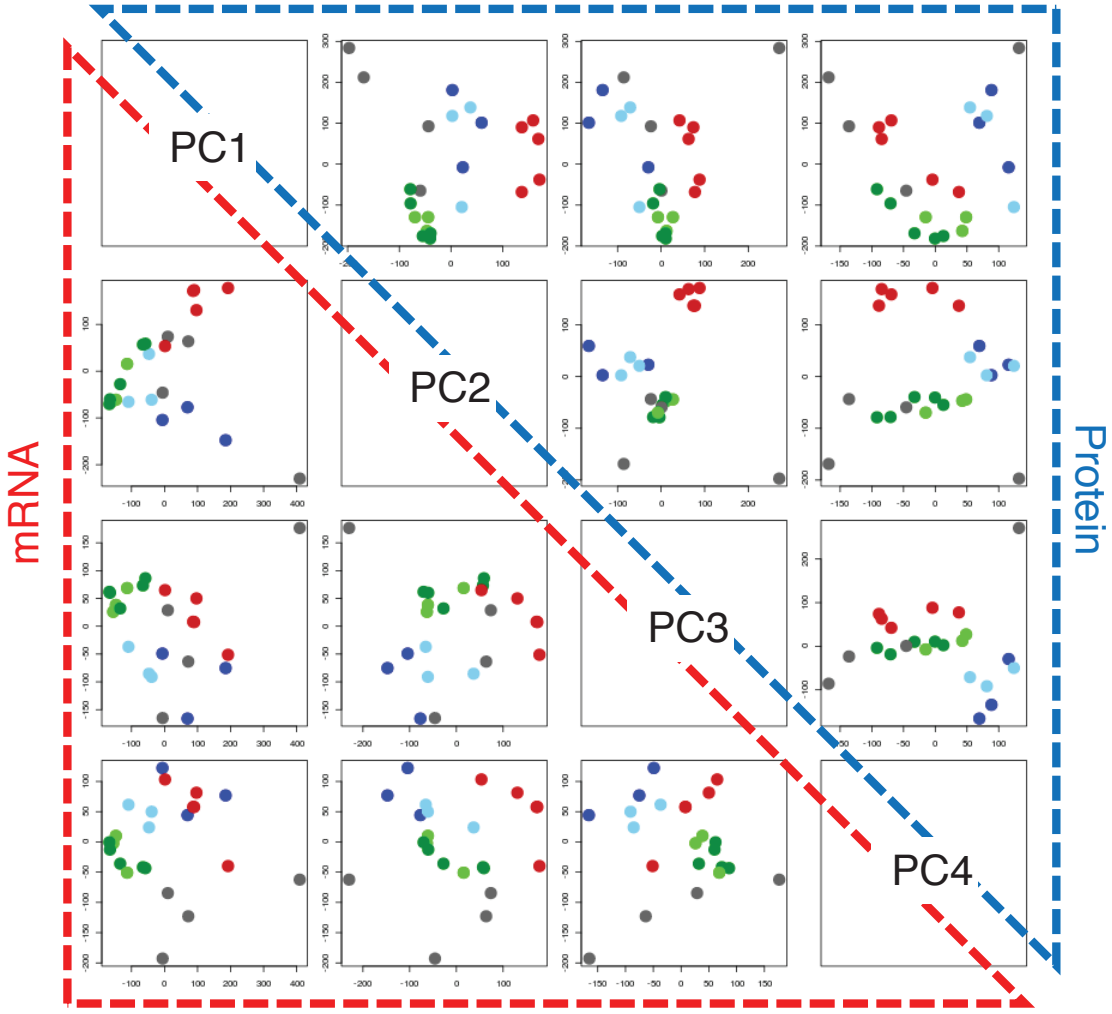


Figure S15. PCA results for transcript and protein data sets. Two-dimensional principal component plots of the first 4 principal components plotted pairwise against each other for both protein (upper right triangle) and mRNA (bottom left triangle). Colors represent developmentally similar tissues. See figure S16 for color-to-tissue assignments

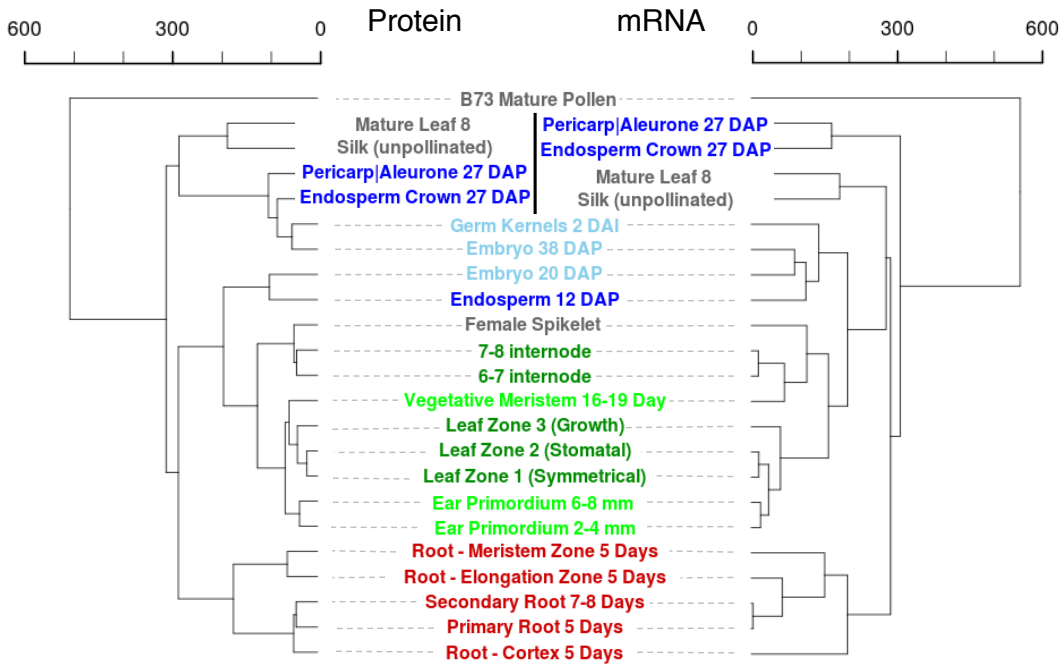


Figure S16. Comparison of PCA between transcript and protein data sets. Hierarchical clustering dendrograms constructed using values of the first 4 principal components for each sample type. Colors are the same as in Fig. S15.

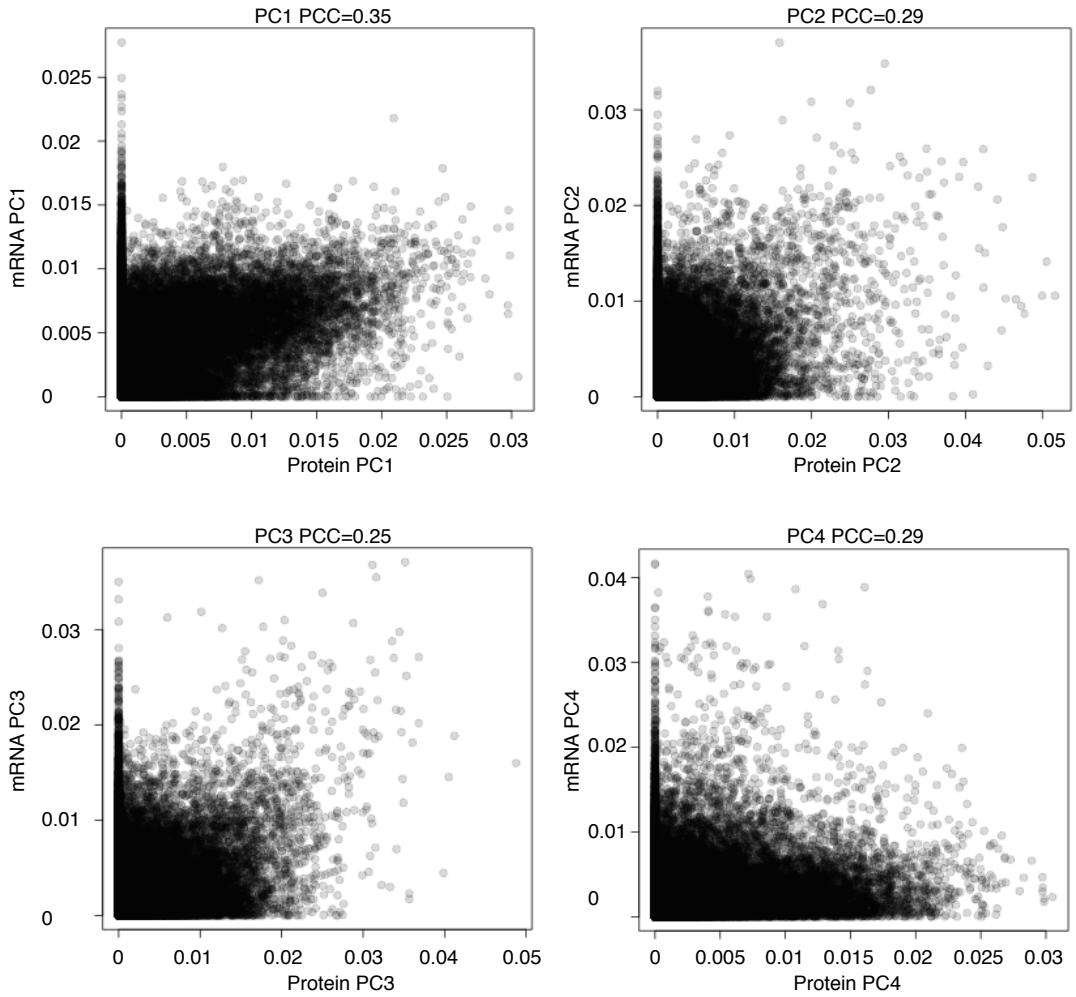


Figure S17. Comparison of PCA between transcript and protein data sets. For the first 4 principal components (PC), the contributions (PCA loadings) of each gene from the protein data (x-axis) and mRNA data (y-axis) are plotted against each other. The Pearson correlation coefficient (PCC) is indicated for each PC.

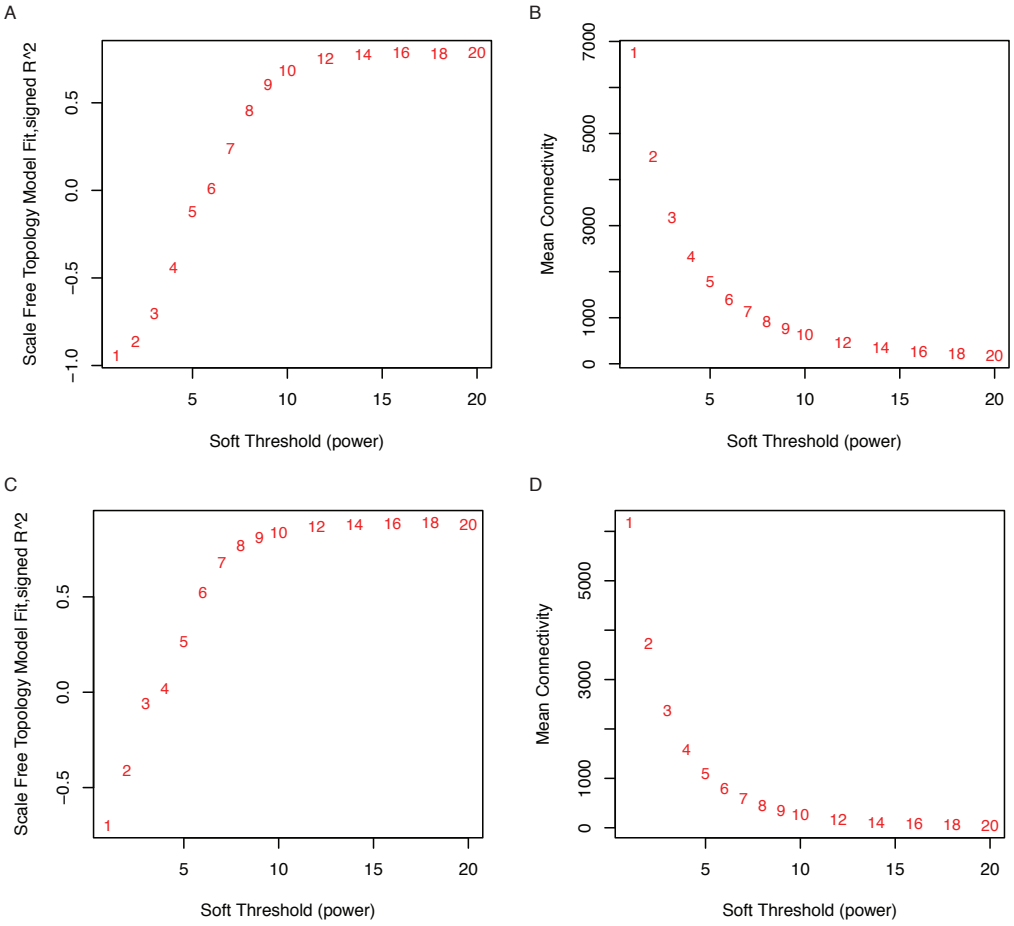


Figure S18. Assessment of soft thresholds used in WGCNA to generate the Spearman based co-expression networks. A range of soft thresholds was evaluated by looking at (A) Scale independence for the mRNA co-expression network. (B) Mean connectivity in the mRNA co-expression network. (C) Scale independence for the protein co-expression network. (D) Mean connectivity in the protein co-expression network.

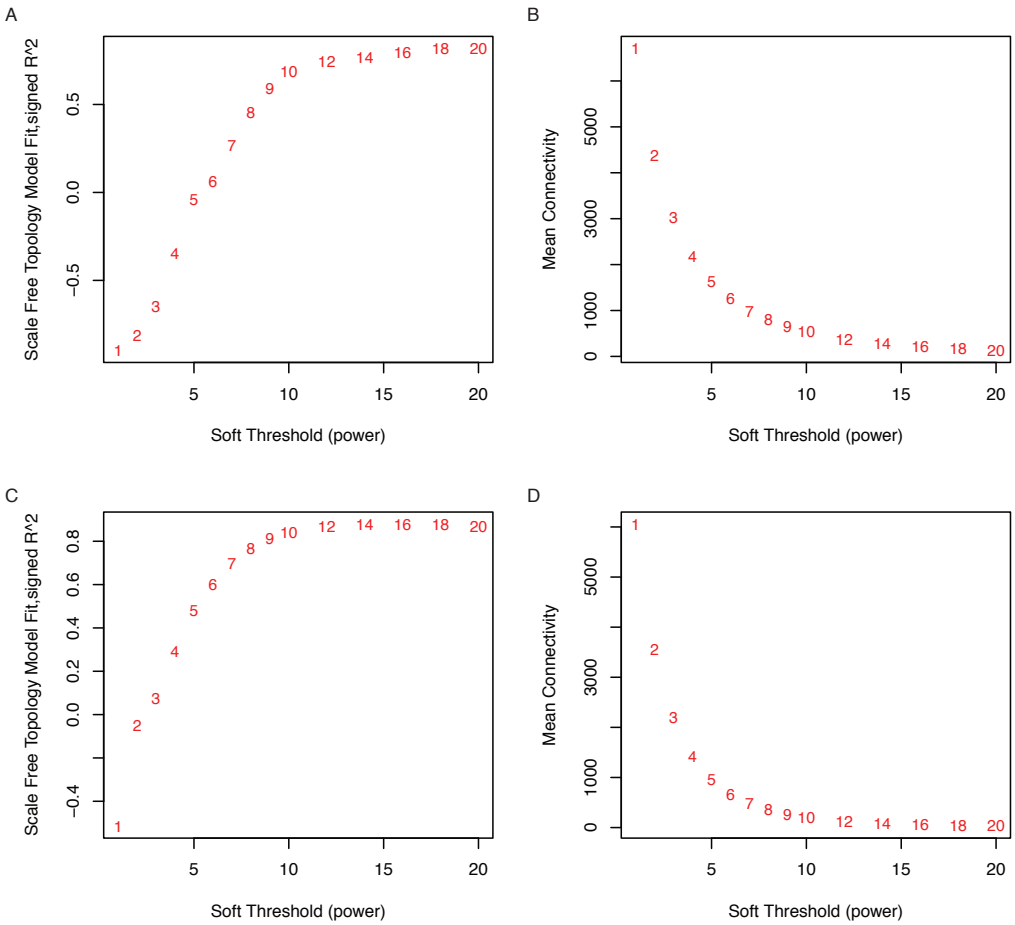


Figure S19. Assessment of soft thresholds used in WGCNA to generate the Bicor based co-expression networks. A range of soft thresholds was evaluated by looking at (A) Scale independence for the mRNA co-expression network. (B) Mean connectivity in the mRNA co-expression network. (C) Scale independence for the protein co-expression network. (D) Mean connectivity in the protein co-expression network.

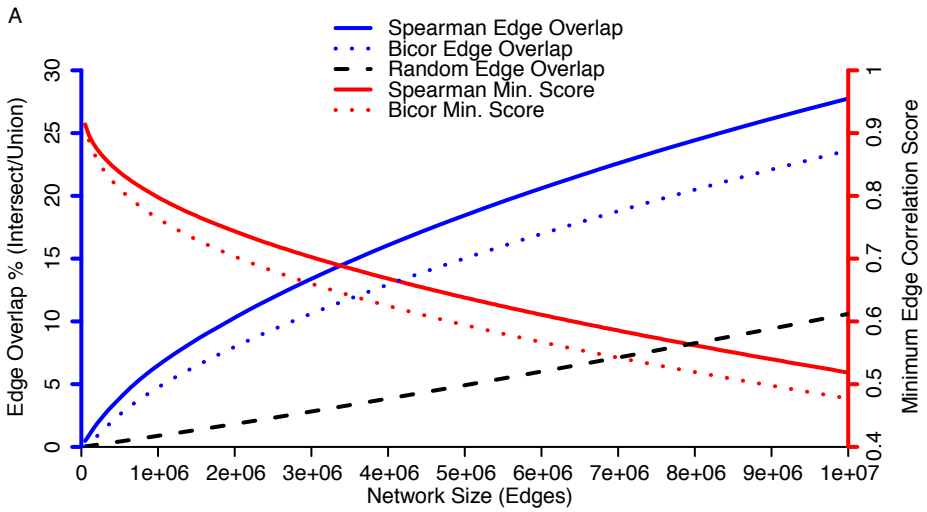


Figure S20. Edge overlap of mRNA and Protein co-expression networks. Using co-expression networks generated using either spearman correlation or biweight midcorrelation from a common set of detected proteins and mRNAs, the edge overlap is shown vs. network size (blue) compared to random (black dashed). For reference, the minimum edge score (correlation) is also shown vs. network size (red).

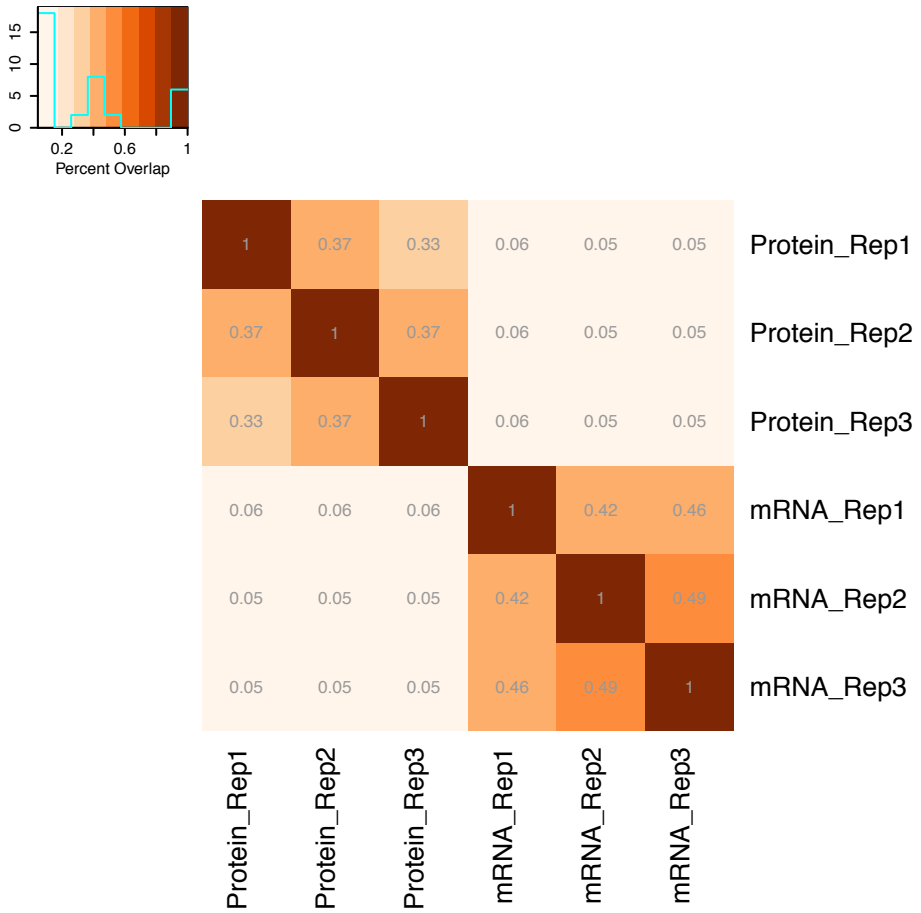


Figure S21. Heatmap showing the jaccard index (intersect/union) of co-expression networks built using single biological replicates of protein or mRNA abundance measurements.

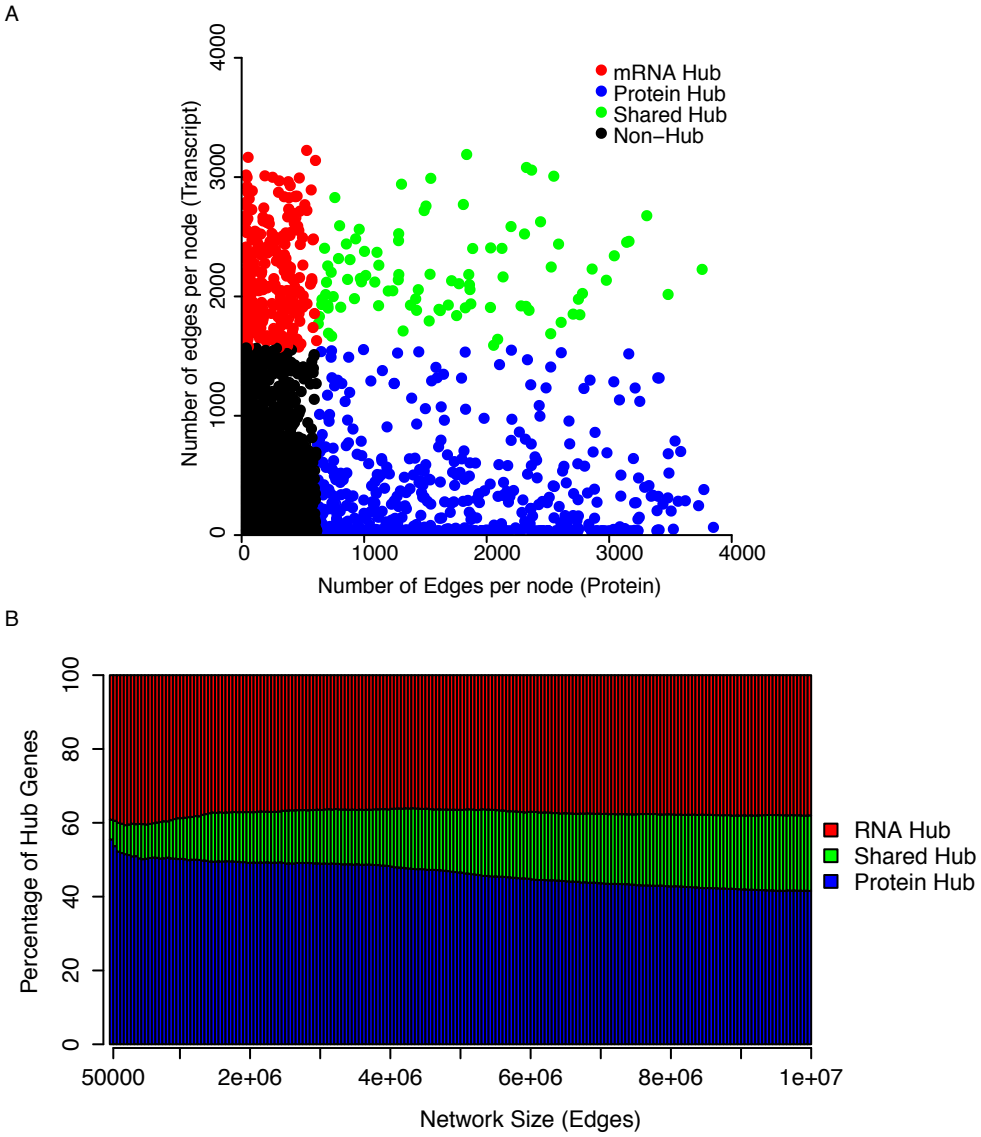


Figure S22. mRNA to protein co-expression hub overlap (A) For the Co-expression networks built using the biweight midcorrelation and a common set of detected mRNA and protein, the number of edges a given gene (node) has in the protein (x-axis) and mRNA (y-axis) is shown. Nodes above the 90th percentile for number of edges (degree) are considered hubs and colored based on whether they are a hub in the protein (blue), mRNA (red), or both (green) networks. Black dots represent non-hub nodes. (B) For the Co-expression network built using spearman correlation (Fig 2), the percentage of all hub genes that are unique to the RNA network, unique to the protein network or shared between networks is plotted as a function of network size.

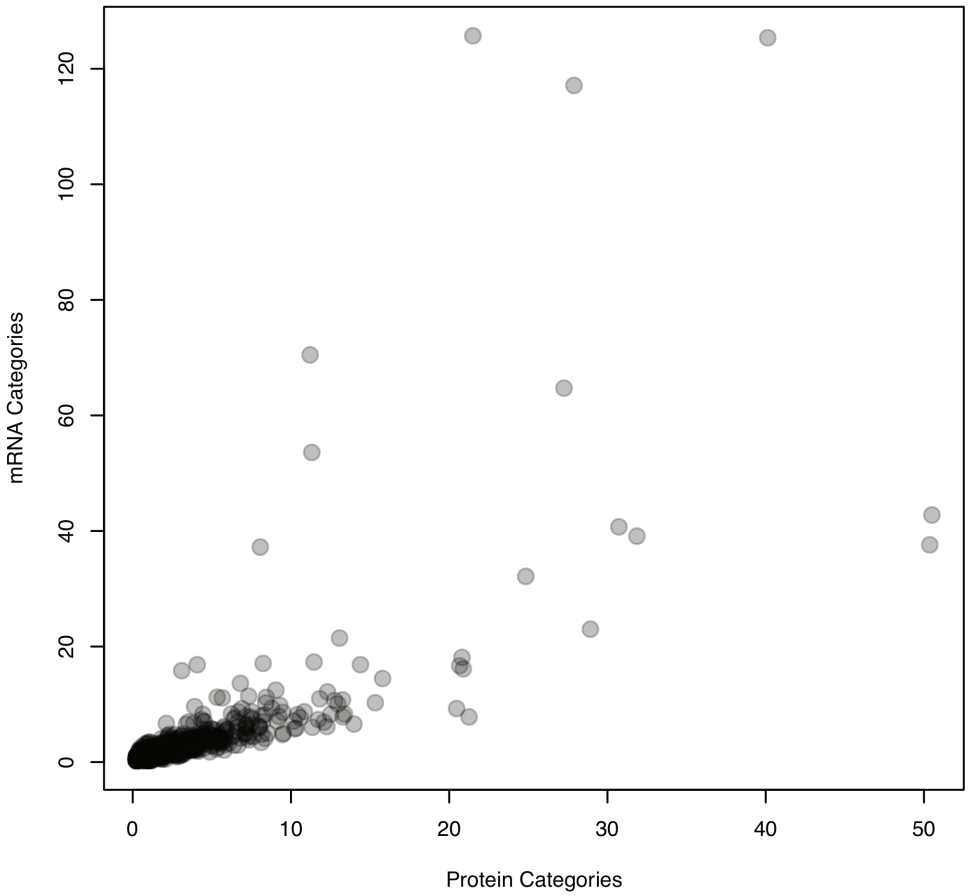


Figure S23. MapMan functional category enrichment in mRNA vs. protein co-expression network modules. In total 1,089 MapMan categories were enriched. Data are the sum of the log-10 P-value for every module in which a given MapMan category was present.

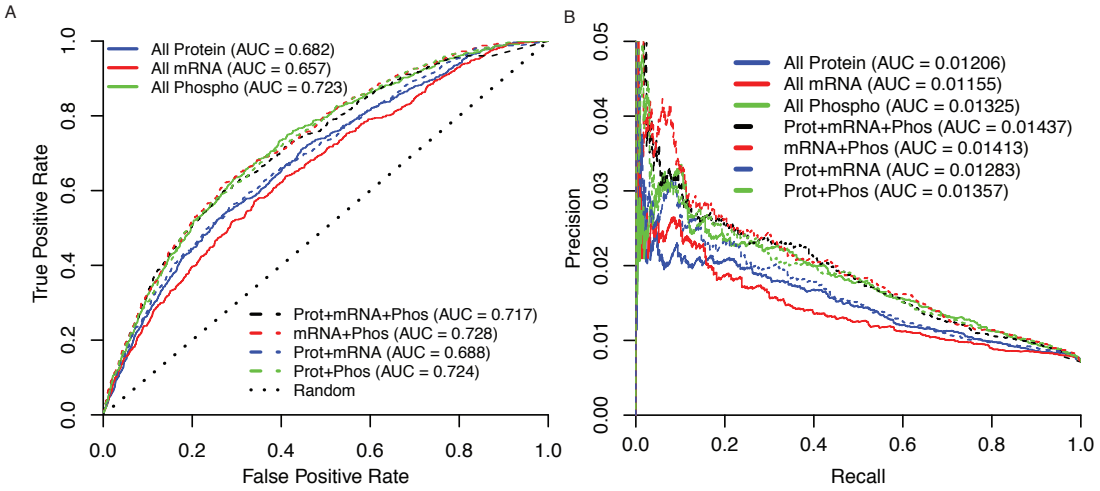


Figure S24. Quality of the full gene regulatory networks. Both single networks and combined networks are shown. KN1 and O2 target genes were obtained from previously published ChIP-seq datasets (24, 34). (A) Receiver operating characteristic (ROC) curve. (B) Precision-recall curve. Phosphorylation modifications were predominantly localized to a specific site on both KN1 (GRMZM2G017087, peptide – NILSSGSSEEDQEGsGGETELPEVDAHGVDQELK, site-S225) and O2 (GRMZM2G015534, peptide – DPsPSDEDMDGEVEILGFK, site-S225).

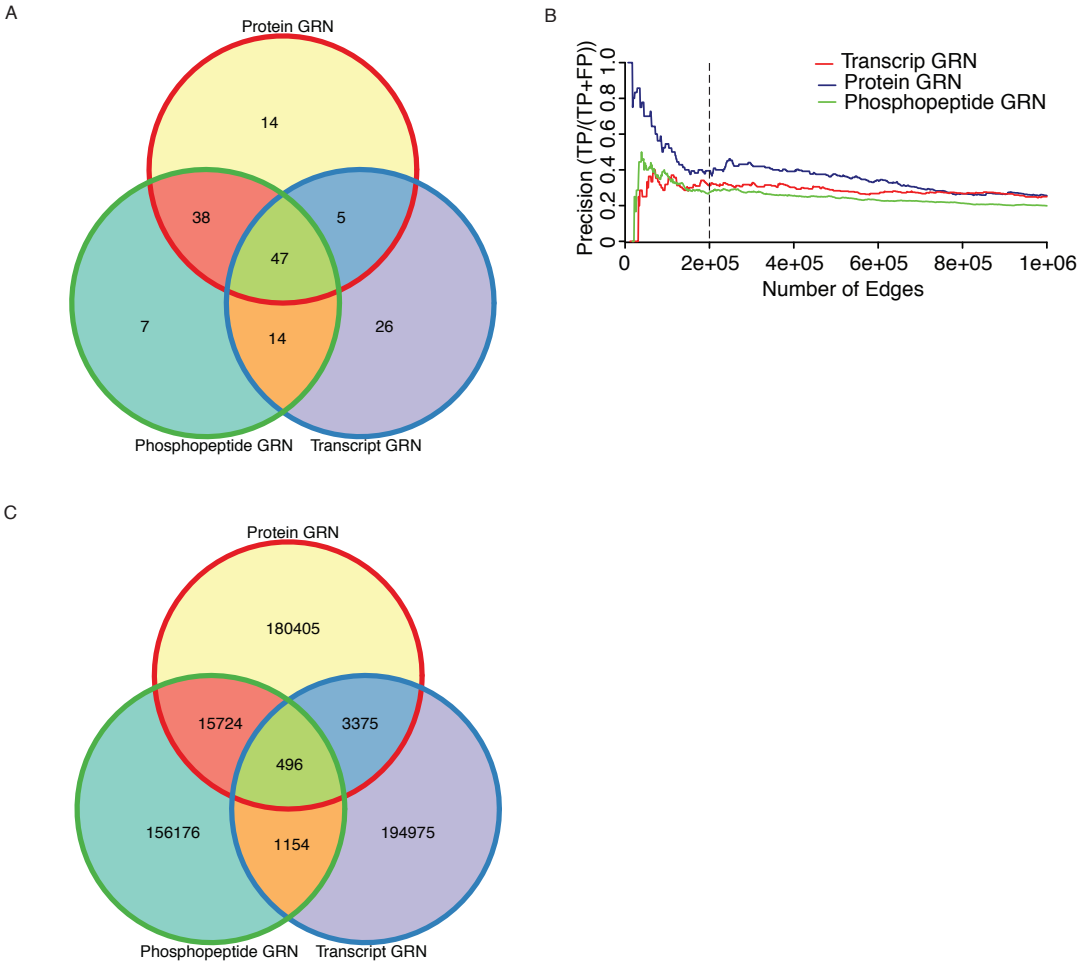


Figure S25. Comparison of the full-sized gene regulatory networks. GRNs were reconstructed using TF regulators quantified as mRNAs (2,200 TFs), proteins (545 TFs), or phosphopeptides (441 TFs), and 41,021 shared potential target genes were quantified as mRNAs. (A) Overlap of the true positive predictions from the top 500 true GRN predictions for O2 quantified as mRNA, protein, or phosphopeptide. (B) GRN precision as a function of network size was calculated using KN1 by comparing the number of true positive (TP) vs false positive (FP) predictions. As the number of edges increased the prediction score decreased. A cutoff of 200,000 edges (vertical dashed line) was used to select the set of high-confidence predictions for all three GRNs. (C) Overlap of the TF-target predictions for the top 200,000 scoring predictions in each GRN.

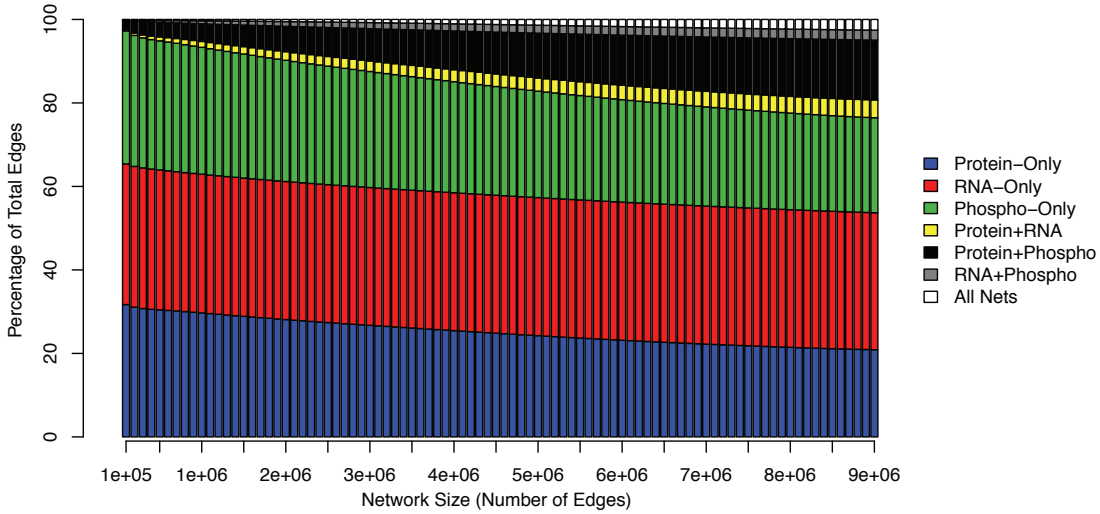


Figure S26. Edge Overlap of full-sized gene regulatory networks. Using the GRNs that were constructed using all available TF information for each data type, the edge overlap was evaluated for an array of network sizes ranging from 100,000 to 10 million. For each size, all edges were categorized as being specific to one network or shared between two or all networks. The sizes of each category are represented by stacked colored bars as a percentage of all edges represented.

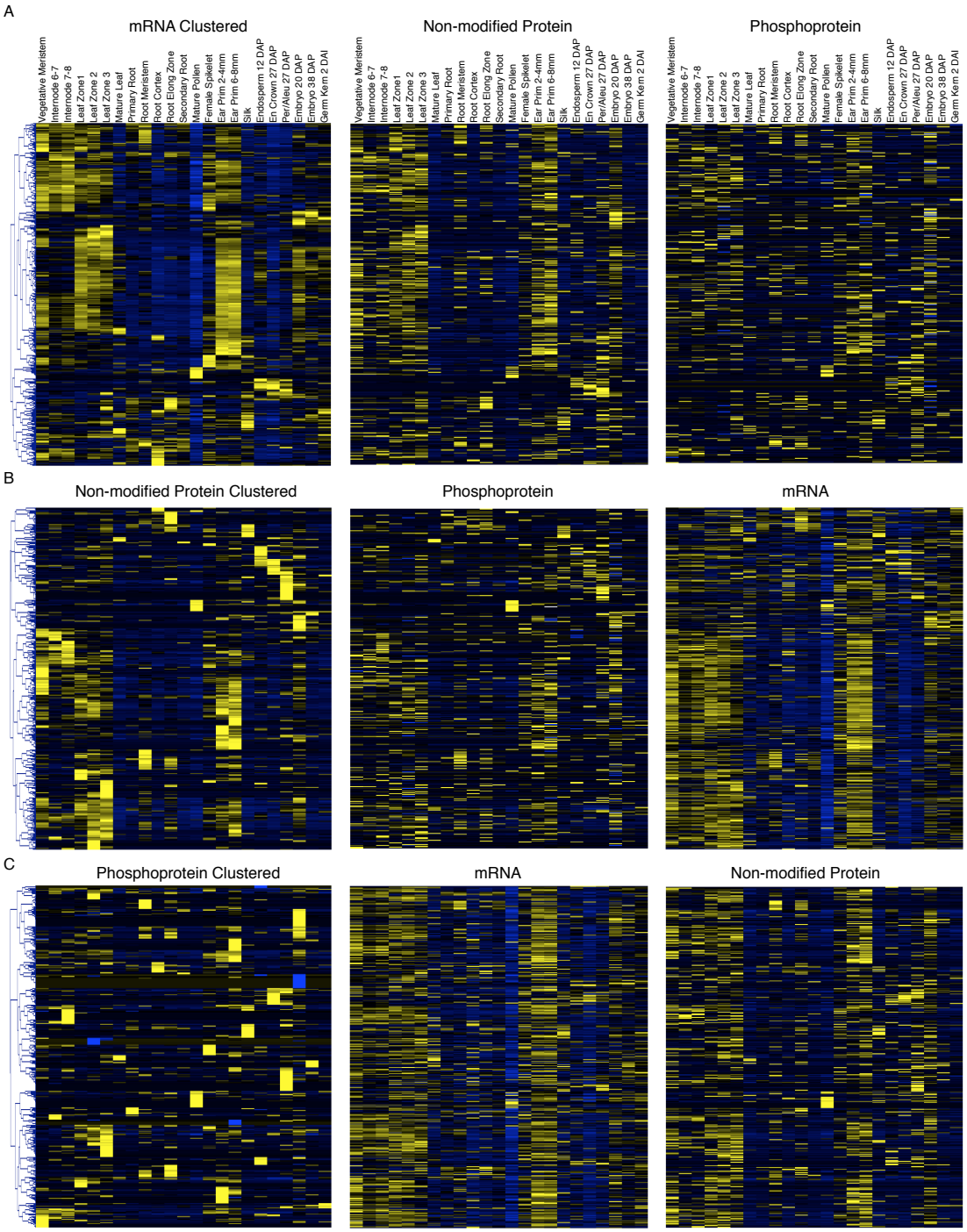


Figure S27. Differential expression of 393 TF genes measured as mRNAs, proteins, and phosphoproteins. (A) Heat maps ordered by hierarchical clustering of mRNA abundance. (B) Heat maps ordered by hierarchical clustering of protein abundance. (C) Heat maps ordered by hierarchical clustering of phosphoprotein abundance.

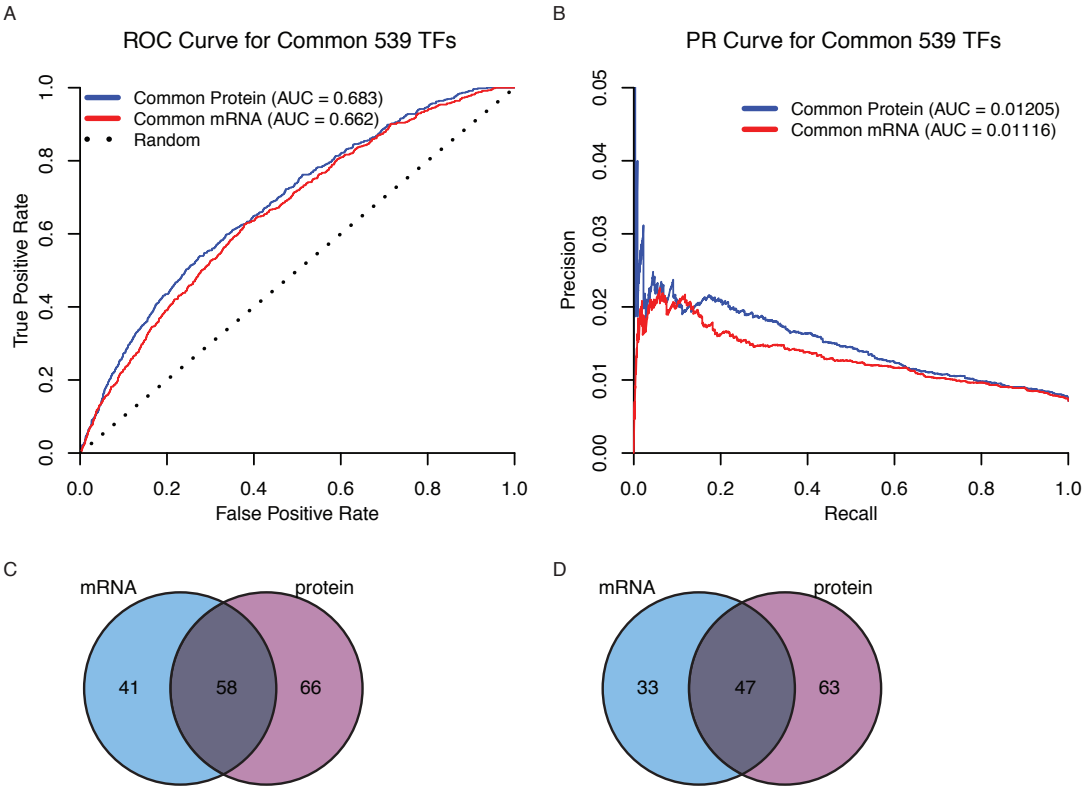


Figure S28. Quality of the GRNs reconstructed using 539 TFs quantified by their mRNA or protein abundance. (A) Receiver operating characteristic (ROC) curve. (B) Precision-recall curve. Standard sets based on KN1 and O2 target genes, obtained from previously published ChIP-seq datasets (24, 34). Overlap of the true positive predictions from the top 500 true GRN predictions for (C) Kn1 and (D) O2.

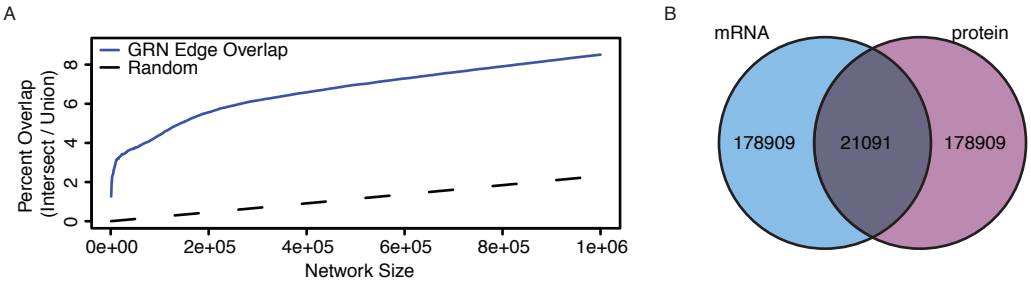


Figure S29. Comparison of predictions in GRNs made using only 539 TFs. (A) Percentage of predictions (edges) that were conserved between GRNs made using the mRNA or protein to measure TF abundance. (B) Overlap of the TF-target predictions for the top 200,000 scoring predictions in each GRN.

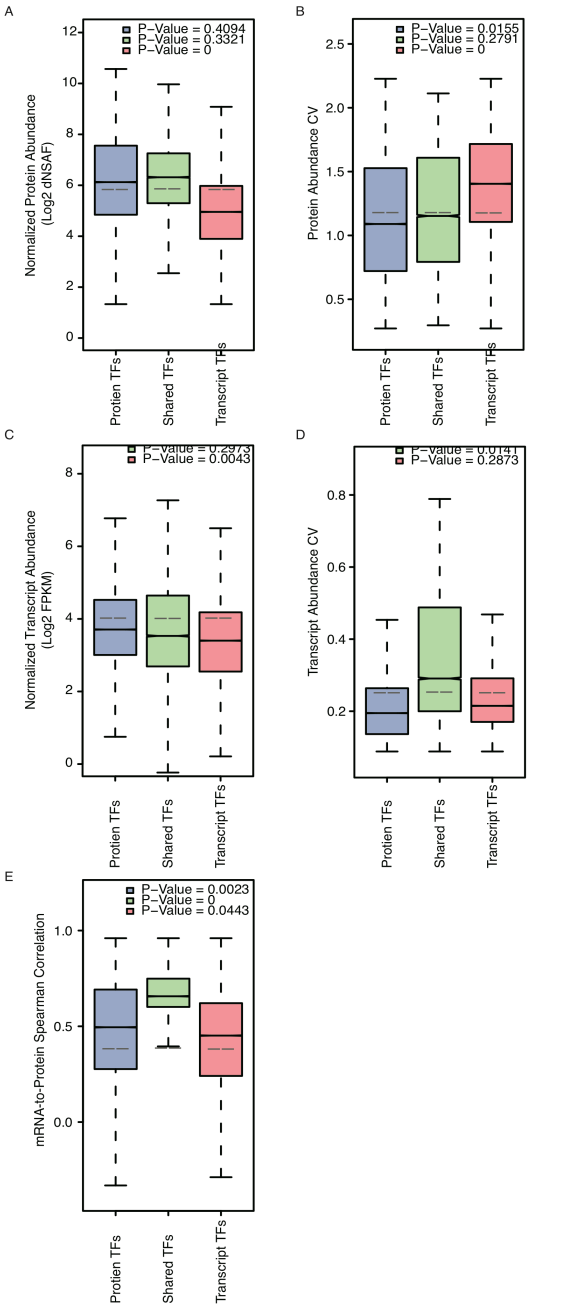


Figure S30. TF regulator expression. TF regulators were weighted based on the number of edges they have in each of 3 categories; [1] edges unique to the protein network, [2] edges shared between the two networks and [3] edges unique to the transcript network. Features of the TFs were then examined for each category after applying the category weights. (A) The distributions of weighted protein abundance of TF regulators in the 3 categories. (B) The distributions of weighted coefficient of variation (CV) for protein abundance of TF regulators in each category. (C) The distributions of weighted transcript abundance of TF regulators in each category. (D) The distributions of weighted CVs for transcript abundance of TF regulators in each category. (E) The distributions of Spearman correlations of mRNA-to-protein abundance for TF regulators in each category. P-values were determined using a permutation test with 10,000 repetitions. Grey lines represent the average median of all permutation tests.

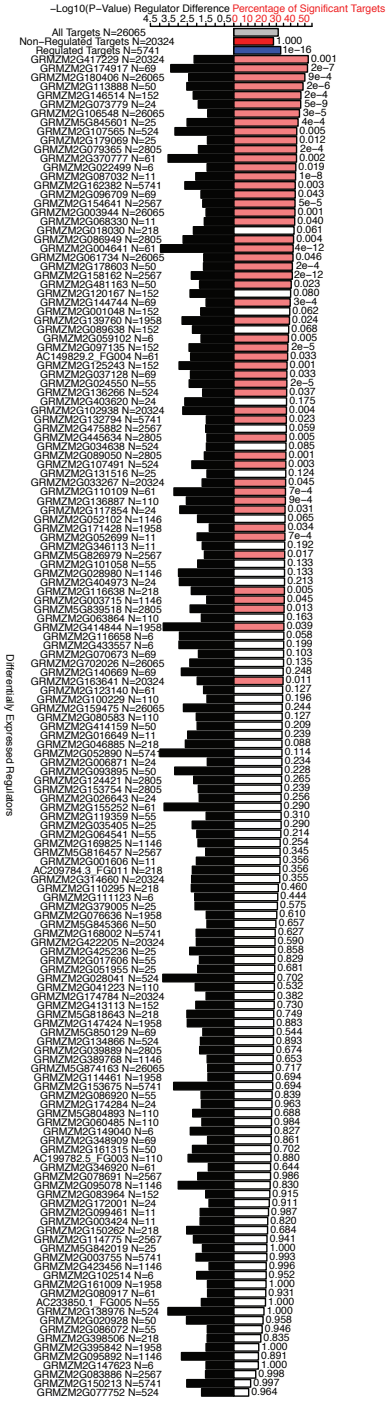


Figure S31. Conservation of mRNA GRN predictions in Mo17. TF regulators from the mRNA GRN who's mRNA is differentially expressed between B73 and Mo17 are shown on the left. The "N=" number next to the gene accession is the number of predicted targets for that regulator. On the right, bars represent the percentage of a given regulators target genes that are also differentially expressed between B73 and Mo17. Colored bars indicate a significant overrepresentation in differentially expressed target genes. P-values for this overrepresentations are printed to the right of each bar.

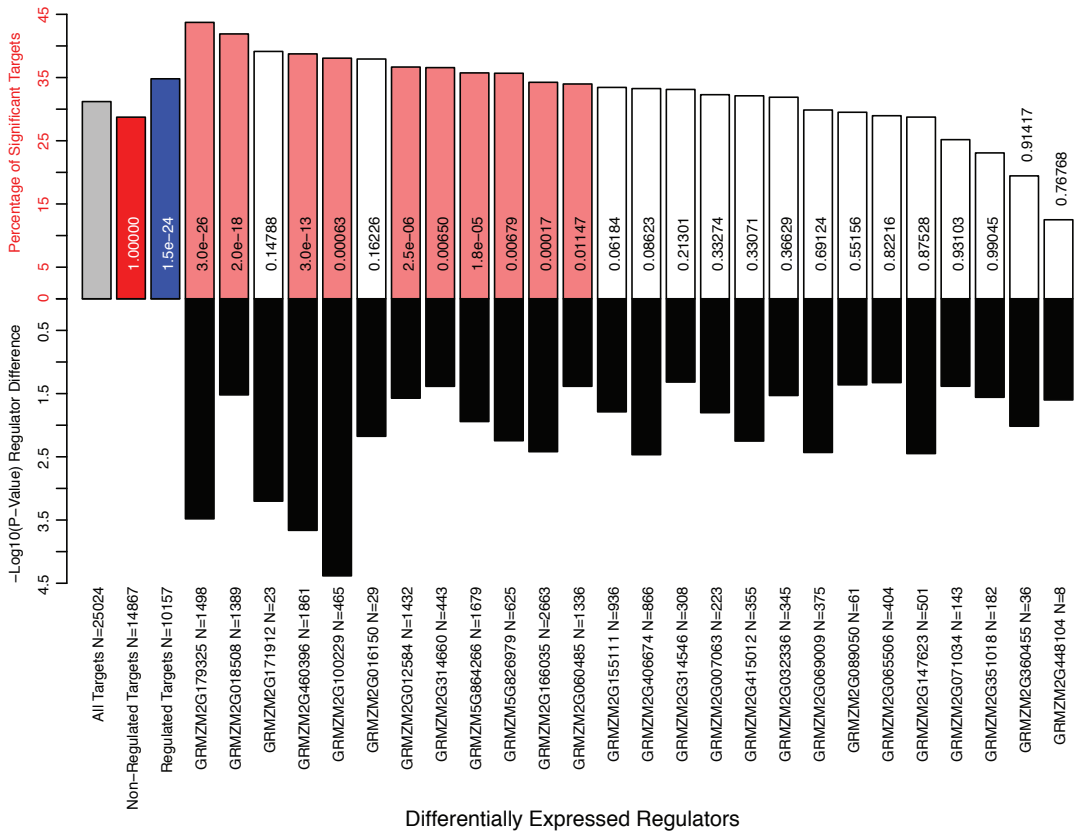


Figure S32. Conservation of protein GRN predictions in Mo17. TF regulators from the protein GRN whose protein is differentially expressed between B73 and Mo17 are shown below the x-axis. The “N=” number next to the gene accession is the number of predicted targets for that regulator. Bars above represent the percentage of a given regulators target genes who are also differentially expressed between B73 and Mo17. Colored bars indicate a significant overrepresentation in differentially expressed target genes. P-values for this overrepresentations are printed on top of each bar.

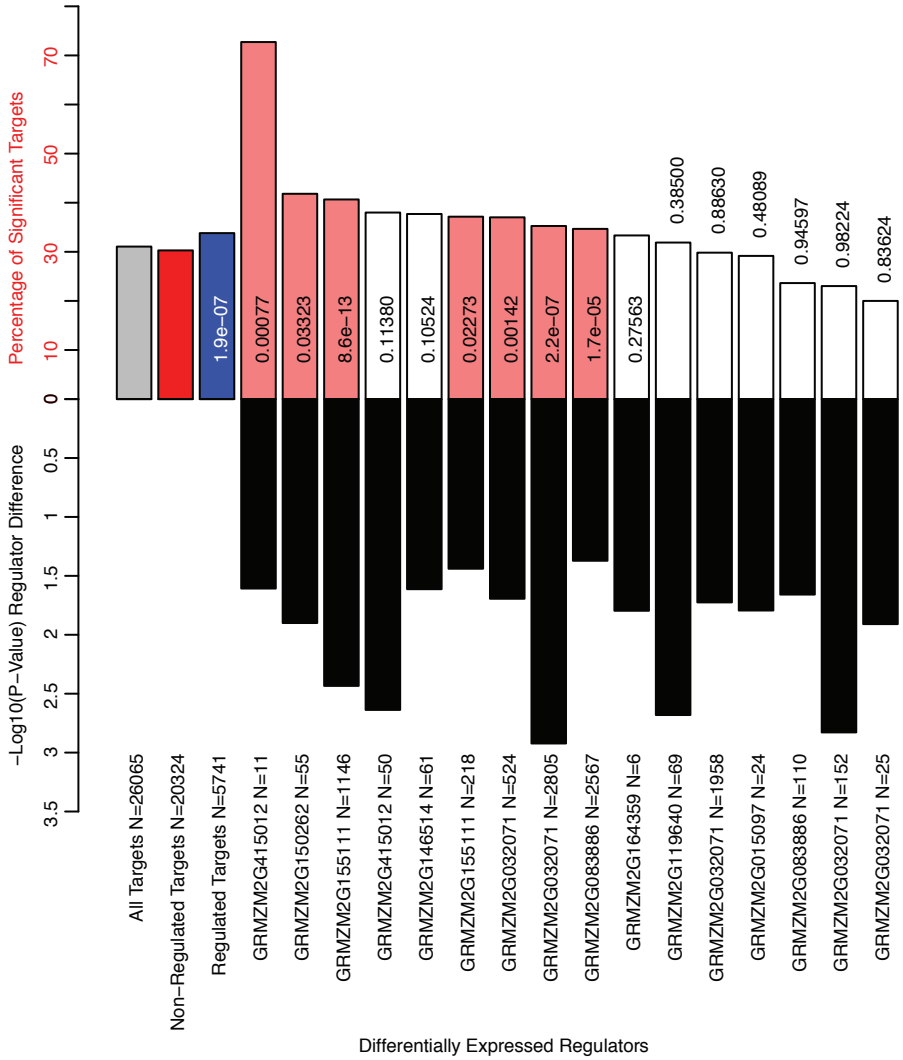


Figure S33. Conservation of phosphopeptide GRN predictions in Mo17. TF regulators from the phosphopeptide GRN that are differentially expressed between B73 and Mo17 are shown below the x-axis. The “N=” number next to the gene accession is the number of predicted targets for that regulator. Bars above represent the percentage of a given regulators target genes who are also differentially expressed between B73 and Mo17. Colored bars indicate a significant overrepresentation in differentially expressed target genes. P-values for this overrepresentations are printed on top of each bar.

References and Notes

1. G. Krouk, J. Lingeman, A. M. Colon, G. Coruzzi, D. Shasha, Gene regulatory networks in plants: Learning causality from time and perturbation. *Genome Biol*. **14**, 123 (2013). Medline doi:10.1186/gb-2013-14-6-123

2. T. S. Gardner, J. J. Faith, Reverse-engineering transcription control networks. *Phys*. *Life Rev*. **2**, 65–88 (2005). Medline doi:10.1016/j.plrev.2005.01.001

3. Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, D. K. Gifford, Computational discovery of gene modules and regulatory networks. *Nat*. *Biotechnol*. **21**, 1337–1342 (2003). Medline doi:10.1038/nbt890

4. R. De Smet, K. Marchal, Advantages and limitations of current network inference methods. *Nat*. *Rev*. *Microbiol*. **8**, 717–729 (2010). Medline

5. V. van Noort, B. Snel, M. A. Huynen, Predicting gene function by conserved co-expression. *Trends Genet*. **19**, 238–242 (2003). Medline doi:10.1016/S0168-9525(03)00056-8

6. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003). Medline doi:10.1126/science.1087447

7. S. Horvath, J. Dong, Geometric interpretation of gene coexpression network analysis. *PLOS Comput*. *Biol*. **4**, e1000117 (2008). Medline doi:10.1371/journal.pcbi.1000117

8. B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach, Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011). Medline doi:10.1038/nature10098

9. C. Vogel, R. de Sousa Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, L. O. Penalva, Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol*. *Syst*. *Biol*. **6**, 400 (2010). doi:10.1038/msb.2010.59

10. S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, J. S. Weissman, Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003). Medline doi:10.1038/nature02046

11. K. Baerenfaller, J. Grossmann, M. A. Grobei, R. Hull, M. Hirsch-Hoffmann, S. Yalovsky, P. Zimmermann, U. Grossniklaus, W. Gruissem, S. Baginsky, Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320**, 938–941 (2008). Medline doi:10.1126/science.1157956

12. A. Ghazalpour, B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian, I. N. Mungrue, C. R. Farber, J. Sinsheimer, H. M. Kang, N. Furlotte, C. C. Park, P. Z. Wen, H. Brewer, K. Weitz, D. G. Camp 2nd, C. Pan, R. Yordanova, I. Neuhaus, C. Tilford, N. Siemers, P. Gargalovic, E. Eskin, T. Kirchgessner, D. J. Smith, R. D. Smith, A. J. Lusis, Comparative analysis of proteome and transcriptome variation in mouse. *PLOS Genet*. **7**, e1001393 (2011). Medline doi:10.1371/journal.pgen.1001393

13. L. Ponnala, Y. Wang, Q. Sun, K. J. van Wijk, Correlation of mRNA and protein abundance in the developing maize leaf. *Plant J*. **78**, 424–440 (2014). Medline doi:10.1111/tpj.12482

14. J. W. Walley, Z. Shen, R. Sartor, K. J. Wu, J. Osborn, L. G. Smith, S. P. Briggs, Reconstruction of protein networks from an atlas of maize seed proteotypes. *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*. **110**, E4808–E4817 (2013). Medline doi:10.1073/pnas.1319113110

15. M. P. Washburn, A. Koller, G. Oshiro, R. R. Ulaszek, D. Plouffe, C. Deciu, E. Winzeler, J. R. Yates III, Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*. **100**, 3107–3112 (2003). Medline doi:10.1073/pnas.0634629100

16. H. Qiao, Z. Shen, S. S. Huang, R. J. Schmitz, M. A. Urich, S. P. Briggs, J. R. Ecker, Processing and subcellular trafficking of ER-tethered EIN2 control response to ethylene gas. *Science* **338**, 390–393 (2012). Medline doi:10.1126/science.1225974

17. K. N. Chang, S. Zhong, M. T. Weirauch, G. Hon, M. Pelizzola, H. Li, S. S. Huang, R. J. Schmitz, M. A. Urich, D. Kuo, J. R. Nery, H. Qiao, A. Yang, A. Jamali, H. Chen, T. Ideker, B. Ren, Z. Bar-Joseph, T. R. Hughes, J. R. Ecker, Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in *Arabidopsis*. *eLife* **2**, e00675 (2013). Medline doi:10.7554/eLife.00675

18. M. R. Facette, Z. Shen, F. R. Björnsdóttir, S. P. Briggs, L. G. Smith, Parallel proteomic and phosphoproteomic analyses of successive stages of maize leaf development. *Plant Cell* **25**, 2798–2812 (2013).

19. C. Marcon, W. A. Malik, J. W. Walley, Z. Shen, A. Paschold, L. G. Smith, H. P. Piepho, S. P. Briggs, F. Hochholdinger, A high-resolution tissue-specific proteome and phosphoproteome atlas of maize primary roots reveals functional gradients along the root axes. *Plant Physiol*. **168**, 233–246 (2015). Medline doi:10.1104/pp.15.00138

20. D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, S. A. Teichmann, RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol*. *Syst*. *Biol*. **7**, 497 (2011). Medline doi:10.1038/msb.2011.28

21. N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo, M. Mann, Deep proteome and transcriptome mapping of a human cancer cell line. *Mol*. *Syst*. *Biol*. **7**, 548 (2011). Medline doi:10.1038/msb.2011.81

22. J. C. Schnable, M. Freeling, Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLOS ONE* **6**, e17855 (2011). Medline doi:10.1371/journal.pone.0017855

23. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). Medline doi:10.1186/1471-2105-9-559

24. N. Bolduc, A. Yilmaz, M. K. Mejia-Guerra, K. Morohashi, D. O'Connor, E. Grotewold, S. Hake, Unraveling the KNOTTED1 regulatory network in maize meristems. *Genes Dev*. **26**, 1685–1690 (2012). Medline doi:10.1101/gad.193433.112

25. R. G. Schneeberger, P. W. Becraft, S. Hake, M. Freeling, Ectopic expression of the knox homeo box gene rough sheath1 alters cell fate in the maize leaf. *Genes Dev*. **9**, 2292–2304 (1995). Medline doi:10.1101/gad.9.18.2292

26. A. Gallavotti, J. A. Long, S. Stanfield, X. Yang, D. Jackson, E. Vollbrecht, R. J. Schmidt, The control of axillary meristem fate in the maize ramosa pathway. *Development* **137**, 2849–2856 (2010). Medline doi:10.1242/dev.051748

27. M. S. Mukhtar, A. R. Carvunis, M. Dreze, P. Epple, J. Steinbrenner, J. Moore, M. Tasan, M. Galli, T. Hao, M. T. Nishimura, S. J. Pevzner, S. E. Donovan, L. Ghamsari, B. Santhanam, V. Romero, M. M. Poulin, F. Gebreab, B. J. Gutierrez, S. Tam, D. Monachello, M. Boxem, C. J. Harbort, N. McDonald, L. Gai, H. Chen, Y. He, J. Vandenhaute, F. P. Roth, D. E. Hill, J. R. Ecker, M. Vidal, J. Beynon, P. Braun, J. L. Dangl; European Union Effectoromics Consortium, Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* **333**, 596–601 (2011). Medline doi:10.1126/science.1203659

28. R. Albert, H. Jeong, A. L. Barabasi, Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000). Medline doi:10.1038/35019019

29. M. A. Calderwood, K. Venkatesan, L. Xing, M. R. Chase, A. Vazquez, A. M. Holthaus, A. E. Ewence, N. Li, T. Hirozane-Kishikawa, D. E. Hill, M. Vidal, E. Kieff, E. Johannsen, Epstein-Barr virus and virus human protein interaction maps. *Proc*. *Natl*. *Acad*. *Sci*. *U*.*S*.*A*. **104**, 7606–7611 (2007). Medline doi:10.1073/pnas.0702332104

30. H. Jeong, S. P. Mason, A. L. Barabási, Z. N. Oltvai, Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001). Medline doi:10.1038/35075138

31. K. Baerenfaller, C. Massonnet, S. Walsh, S. Baginsky, P. Bühlmann, L. Hennig, M. Hirsch-Hoffmann, K. A. Howell, S. Kahlau, A. Radziejwoski, D. Russenberger, D. Rutishauser, I. Small, D. Stekhoven, R. Sulpice, J. Svozil, N. Wuyts, M. Stitt, P. Hilson, C. Granier, W. Gruissem, Systems-based analysis of *Arabidopsis* leaf growth reveals adaptation to water deficit. *Mol*. *Syst*. *Biol*. **8**, 606 (2012). doi:10.1038/msb.2012.39

32. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE* **5**, e12776 (2010). doi:10.1371/journal.pone.0012776

33. D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky; DREAM5 Consortium, Wisdom of crowds for robust gene network inference. *Nat*. *Methods* **9**, 796–804 (2012). Medline doi:10.1038/nmeth.2016

34. C. Li, Z. Qiao, W. Qi, Q. Wang, Y. Yuan, X. Yang, Y. Tang, B. Mei, Y. Lv, H. Zhao, H. Xiao, R. Song, Genome-wide characterization of *cis*-acting DNA targets reveals the transcriptional regulatory framework of *Opaque2* in maize. *Plant Cell* 10.1105/tpc.114.134858 (2015). Medline doi:10.1105/tpc.114.134858

35. M. Schrynemackers, R. Küffner, P. Geurts, On protocols and measures for the validation of supervised methods for the inference of biological networks. *Front*. *Genet*. **4**, 262 (2013). Medline doi:10.3389/fgene.2013.00262

36. R. J. Chalkley, K. R. Clauser, Modification site localization scoring: Strategies and performance. *Mol*. *Cell*. *Proteomics* **11**, 3–14 (2012). Medline doi:10.1074/mcp.R111.015305

37. E. L. Huttlin, M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausoleil, J. Villén, W. Haas, M. E. Sowa, S. P. Gygi, A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010). Medline doi:10.1016/j.cell.2010.12.001

38. H. Liu, R. G. Sadygov, J. R. Yates III A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal*. *Chem*. **76**, 4193–4201 (2004). Medline doi:10.1021/ac0498563

39. Y. Zhang, Z. Wen, M. P. Washburn, L. Florens, Refinements to label free proteome quantitation: How to deal with peptides shared by multiple proteins. *Anal*. *Chem*. **82**, 2272–2281 (2010). Medline doi:10.1021/ac9023999

40. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10**, R25 (2009). Medline doi:10.1186/gb-2009-10-3-r25

41. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. **14**, R36 (2013). Medline doi:10.1186/gb-2013-14-4-r36

42. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat*. *Biotechnol*. **31**, 46–53 (2013). Medline doi:10.1038/nbt.2450

43. G. Csárdi, A. Franks, D. S. Choi, E. M. Airoldi, D. A. Drummond, Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLOS Genet*. **11**, e1005206 (2015). Medline doi:10.1371/journal.pgen.1005206

44. A. Yilmaz, M. Y. Nishiyama Jr., B. G. Fuentes, G. M. Souza, D. Janies, J. Gray, E. Grotewold, GRASSIUS: A platform for comparative regulatory genomics across the grasses. *Plant Physiol*. **149**, 171–180 (2009). Medline doi:10.1104/pp.108.128579

45. O. Thimm, O. Bläsing, Y. Gibon, A. Nagel, S. Meyer, P. Krüger, J. Selbig, L. A. Müller, S. Y. Rhee, M. Stitt, MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*. **37**, 914–939 (2004). Medline doi:10.1111/j.1365-313X.2004.02016.x

ACKNOWLEDGMENTS

Chapter 1, in full, is a reprint of the material as it appears in: Walley, J. W.; Sartor, R. C.; Shen, Z.; Schmitz, R. J.; Wu, K. J.; Urich, M. A.; Nery, J. R.; Smith, L. G.; Schnable, J. C.; Ecker, J. R.; Briggs, S. P. “Integration of omic

networks in a developmental atlas of maize". *Science* **353**, 814–818 (2016). The dissertation author was primary investigator and co-first author of this work.

CHAPTER 2

Genic DNA Methylation Plays a Key Role in Establishing the Maize Expressome: A Machine Learning-Based Approach to Systems-level Discovery.

Sartor, R. C.[1], Noshay, J.[2], Springer, N. M.[2], Briggs, S. P.[1]
[1]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA.
[2]Department of Plant Biology, University of Minnesota, Saint Paul, Minnesota, United States of America

ABSTRACT

We used a machine learning approach to classify genes according to their expression potentials. The algorithmic models use only genic DNA methylation, operate on a genome-wide scale, and perform as good or better than manually curated classifications. Methylation patterns distinguish between gene classes with high accuracy: 8,065 can express mRNA only; 32,995 can express both mRNA and protein; 57,236 are constitutively silent. Naturally occurring differential methylation of genes between genetically diverse inbreds is associated with differences in expression predicted by the models. Therefore, while more than half of the protein-coding genes are silent in a given inbred, many can be expressed in the diverse germplasm of the species, providing a reservoir of adaptive potential that may play a role in plant breeding and evolution.

INTRODUCTION

Methods for genome-wide annotation of protein-coding genes are based on homology to genes in related species and on mRNA expression data. Except for a relatively small number of studied genes, these two criteria provide the primary evidence for gene functions. While useful, these methods of annotation are incomplete because in most species a significant fraction of genes are not known to be expressed or they have no known homologs. Furthermore, a large proportion of transcribed genes are not known to express proteins. Genes naturally form two groups according to their mRNA expression levels of high vs. low/no expression (Walley & Sartor, 2016; Hebenstreit, 2011; Nagaraj, 2011). Nearly all genes shown to have a function are in the high mRNA expression group. Correspondingly, nearly all observed protein-expressing genes are in this group and 90% of these observed protein-expressing genes are syntenic orthologs. DNA methylation within the gene body is lower in syntenic genes (Eichten, 2011). DNA methylation can repress the expression and transposition of transposable elements (TEs) (Slotkin & Martienssen, 2007; Zemach, 2010; Regulski, 2013; West, 2014). Spreading of methylation from TEs to nearby genes is thought to influence their expression (Weil and Martienssen, 2008; Eichten, 2012). To explore the possibility of a genome-wide silencing system that extends to all protein-coding genes, we developed a method for algorithm-based annotation that describes the

expression potentials for all protein-coding genes. This revealed that mRNA-expressing genes have no/little methylation at their 5' and 3' ends and protein-expressing genes additionally have methylation in their centers.

RESULTS

Algorithmic Models of Expression Potential

Genes can be grouped into categories based on the abundance of their transcripts. We used three categories: i) High abundance mRNA (HR), Low abundance mRNA (LR) and Non-Observed mRNA (NR) (Fig2.1_B). Each category can be split into two based on whether a gene has: Observed Protein (OP) or Non-Observed Protein (NP) resulting in 6 categories total (Fig2.1_B). We examined three sets of genes for which there is independent, published evidence of expression or function and saw that these genes reside in the high abundance mRNA group (S2.1). We trained the random forest machine learning algorithm (Breiman, 2001) to classify genes of the inbred B73 by using genic methylation as predictive features to explain gene expression: genes were classified as able to express both mRNAs and proteins; as able to express only mRNAs; or as silent (Fig2.1). Genic methylation was summarized into multiple features. All three methylation sequence contexts (CHG, CpG and CHH) were quantified separately and summarized within gene regions (Fig2.1_A); gene regions included the transcription start site (TSS), 5' UTR, 3'

UTR, introns, exons, and summed across the whole gene model. These 6 regions combined with 3 methylation contexts provided 18 features that were used for training the algorithmic models (S2.2). The LR and HR were separated using an mRNA abundance value of 1 Fragment Per Kilobase per Million reads (FPKM) (Fig2.1_B). The NR and LR were generally grouped together because it is likely that the NR is an extension of the lower tail of the LR that is below the detection limit.

The methylation data were used as training features (independent variables) in every model. Several different class variables (dependent variables) were tested, each being used to train a different model. These class variables were determined using transcript abundance or a combination of transcript and protein abundance (Walley & Sartor, 2016). Two classifiers were built. The first used a combination of protein and transcript data. The silent class consists of genes with no observed mRNA or protein (NR/NP). The express-able class consists of genes with high mRNA and observed proteins (HR/OP). This classifier is therefore defining two populations of genes based on mRNA abundance but also conditioned upon protein observation and attempting to differentiate between the two populations using genic DNA methylation. We refer to this as the Express-able Protein Classifier (EPC). The Express-able mRNA Classifier (ERC) does not use protein data and the class variable is defined using all HR genes vs. all NR genes (Fig2.1_C).

Random forest feature importance is shown in Figure S2.2. Based on these scores, multiple features were deleted from the model. Most of the classification power was from CHG and CpG methylation in exons, introns, or the aggregation of all features (labeled “Gene”). Therefore, these elements were retained in the models. In addition, methylation levels were seen to be variable across the gene. Patterns of methylation were observed (S2.3 & S2.4) especially with CpG methylation and to a lesser extent CHG methylation. To characterize these patterns we divided each gene into 5 equivalent bins. Bin 1 covers the 5’ end of the gene through the first 20% of the gene model and bin 5 covers the 3’ end and the last 20% of the gene model (Fig2.1-A). Use of the aggregated features; exonic; or intronic regions; the 5 bins; and CpG or CHG methylation provided 30 methylation features (Fig2.2-C).

Classification accuracies were determined using the random forest out-of-bag cross-validation (Fig2.2_A). Both the EPC and ERC models had Receiver Operator Characteristic (ROC) and Precision vs. Recall (PR) curves with areas under the curve (AUC) of 0.94 or higher. The EPC achieved a near-perfect area under the ROC curve of 0.99 indicating that nearly all of the genes (true positives and false positives) were classified correctly.

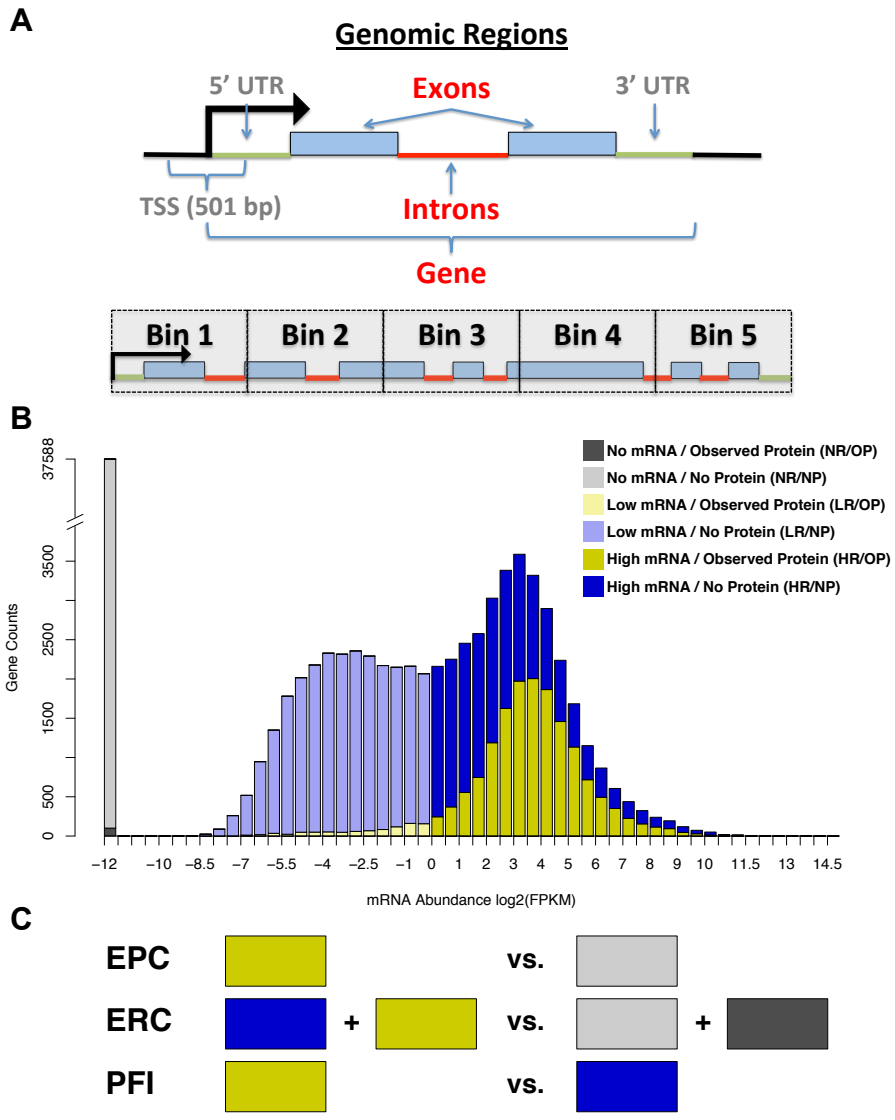


Figure 2.1 Overview of model features and training set definitions. (A) The various genomic regions where DNA methylation levels were quantified and used as features for classification. Grey features were discarded after initial testing and only the whole gene, exons and introns were used as features in the final models. Because methylation levels vary across the gene body, each gene was also split into five equivalent regions, called bins and separate features were quantified in each bin. (B) The distribution of detected mRNA abundance is bimodal with nearly all detected proteins existing in the High mRNA population. The two mRNA populations can be roughly separated using an FPKM of 1. Here the non-detected mRNA (No mRNA) is represented as a separate population and given an artificial value of -12. Red dashed lines separate all populations. Each population can be further refined into observed vs. non-observed protein (No Protein) to yield 6 different groups of genes to use as training sets, indicated by the different colors. (C) Three separate random forest models were built using classes defined in B.

The random forest feature importance was determined using the mean decrease in accuracy upon random permutation of each individual variable (Breiman, 2001). This value is unsigned, so to determine whether silent genes are caused by high or low methylation levels, a t-statistic was calculated for the values of each feature and the sign of the t-statistic was assigned to feature importance (Fig2.2_C).

The structures of the EPC and ERC models were very similar based on feature importance (Fig2.2_C). In each case, the beginnings and ends of genes (bins 1 & 5) were the most important. The exons were more important than introns and CHG methylation contributed more than CpG. All of the relationships between methylation and class variables were negative. That is, in both the ERC and EPC we observed that high methylation levels had a silencing effect on expression and genes with high methylation across the entire gene body (particularly at both ends) were classified as silent (Fig2.2_C, S2.6_A&B).

We tested whether genic methylation can be used to predict quantities of mRNAs and proteins. Random forest models were run using the same methylation training data but replacing the binary class vector with quantitative mRNA or protein abundance for the ERC and EPC, respectively. We call these models the Protein Expression-level Predictor (PEP) and the mRNA Expression-level Predictor (REP). We found that methylation levels did not predict expression levels. When examining the full set of predictions

(S2.5_C&D), we see good R^2 values, but this correlation is entirely due to the ability of the model to separate observed from non-observed proteins or mRNAs, which is analogous to the classifiers described above. When looking only at the genes with detectable expression (Fig2.2_B and S2.5_B), we see very low R^2 values.

Genome-Wide Re-Annotation of Protein-Coding Genes

The ERC and EPC were used to classify the expression potentials for 98,296 protein-coding genes for which we have whole genome bisulfite sequencing (WGBS) coverage. This identified 41,060 genes with potential to express mRNAs and 32,995 genes with potential to express proteins; only three genes in the latter set were missing from the former. For breakdowns of training vs. test data, see S2.11. The final classifications for the genes in the training set were determined using the random forest out-of-bag cross-validation.

The maize genome contains over 110,000 annotated genes. This set is called the working gene set (WGS). The MaizeGDB group (Andorf, 2016) curates the genome. They have annotated 45,354 genes

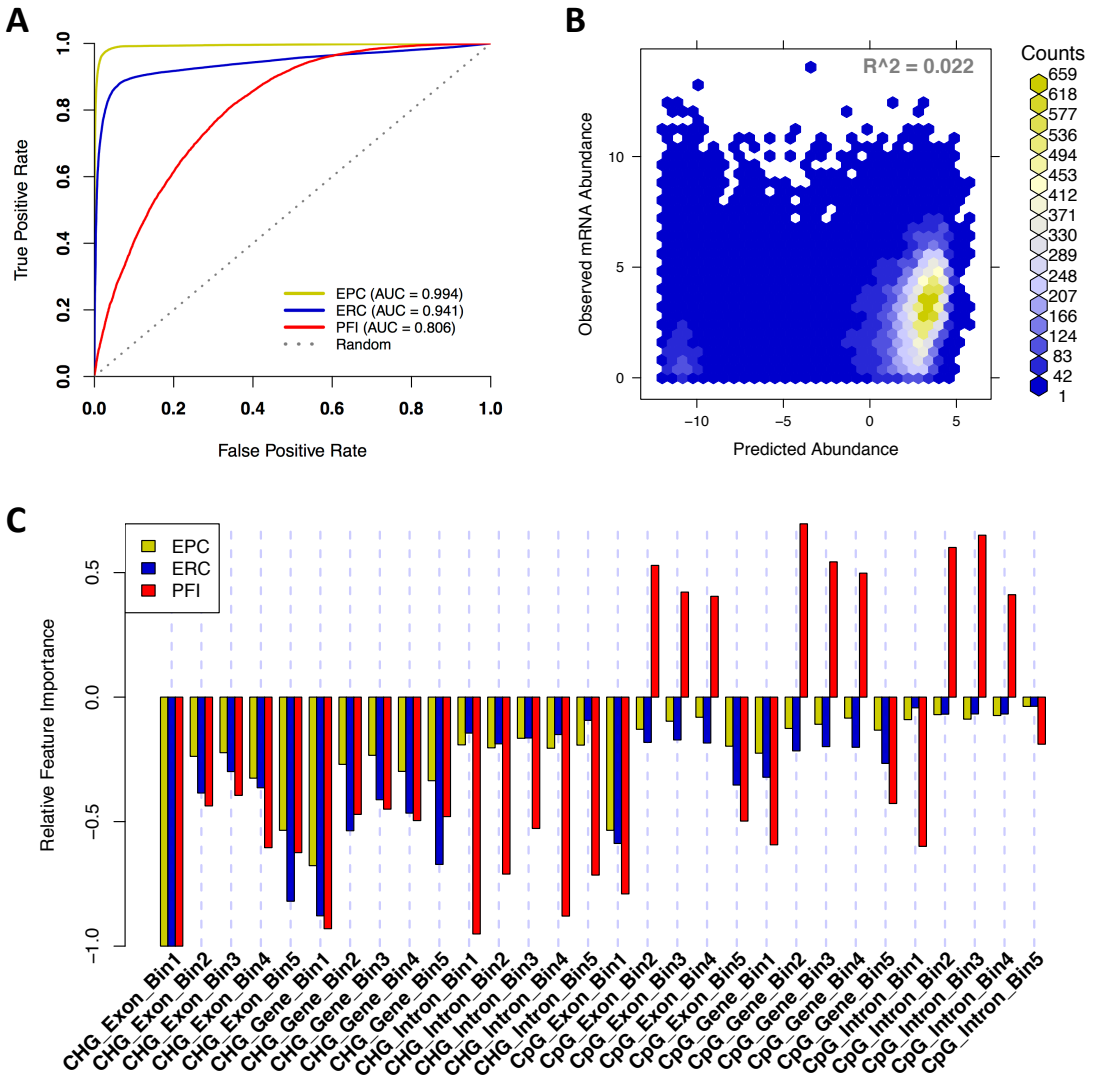


Figure 2.2 Results for random forest models. (A) Receiver operating characteristic (ROC) curves showing classification accuracy of the EPC, ERC, and PFI models. The random forest "votes" from the out-of-bag cross-validated classifications were used for all models. (B) Prediction accuracy for quantitative mRNA abundance model. (C) Signed feature importance measures. The values reflect the random forest "mean decrease in accuracy" measure of feature importance. The sign is based on the relationship of the feature values to the training class assignments. Positive values indicate a positive correlation between the feature and either protein observation (EPC and PFI) or high mRNA (ERC).

(ZmB73_5b.60_FGS) as the most likely to be functional and named them the Filtered Gene Set (FGS). The FGS is derived from the WGS using multiple avenues of evidence including TE and pseudo-gene exclusion, orthology to rice/sorghum, protein length, CDS completeness, synteny to close species, and RNA-seq expression. We compared the 43,700 FGS genes that have WGBS data to our DNA methylation-based gene sets. The ERC set (41,068 genes) excluded 10,176 genes of the FGS and contained 7,544 genes absent from the FGS (Fig2.3_A). The EPC set (32,986 genes) excluded 14,303 genes of the FGS and contained 3,592 genes absent from the FGS.

We evaluated the FGS, ERC, and EPC gene sets by comparison to four high-confidence operative gene sets: i) a set of 434 “classical” maize genes (Schnable, 2011) (Fig2.3_B); ii) 24,092 syntenic orthologs between maize and sorghum (Walley & Sartor et al., 2016) (Fig2.3_C); iii) 9,940 genes with observed full length cDNA (FLcDNA) (Soderlund et al., 2009) (Fig2.3_D); and iv) 4,329 genes that have been curated by MaizeGDB (S2.11_C). The methylation-based classifications out-performed the FGS for fold-enrichment, precision, and false-positive rates. The FGS had higher recall rates as a result of curation. The EPC model not only performed best; it enabled reconstruction of the mRNA bimodal distributions to distinguish each gene in the LR and HR populations (S2.10_A).

We found that genic methylation-associated gene silencing is widespread amongst protein-coding genes. The ERC/EPC classified 23/33%

of FGS genes as silent; these proportions rose to 58/66% for silencing of WGS genes. The MaizeGDB project (Andorf, 2016) annotates a biotype to each gene model, which can be used to filter out all likely TEs and pseudo-genes, to yield 63,331 probable protein-coding genes. Of these, 60,295 have coverage in the WGBS data. Using the higher confidence EPC classifier, 48% (28,779) were classified as silent (S2.12) and nearly all of the TEs (97%) and pseudo-genes (94%) in the WGS were classified as silent.

High levels of CHG and CpG methylation repress expression of transposable elements (TEs) and repetitive elements (REs) (Slotkin & Martienssen, 2007 ; Zemach, 2010 ; Regulski, 2013 ; West, 2014). Due to the abundance of these elements in the genome, many gene models in the WGS are TEs that have escaped sequence masking. Of the 110,028 gene models, 29,082 have been categorized as likely TEs. In addition, we identified 7,612 gene models that have a high blast hit to one or more reference TE sequences (Wessler et al., 2015). Many of these are likely protein-coding genes with TEs inserted into the gene body. To determine the extent to which these previously characterized, highly methylated elements are affecting our classifiers and conclusions, we re-built all the classification models after filtering out all 36,694 TEs and TE-containing gene models. The new classifier is nearly identical to the original both in classification accuracy (S2.9_A&B) and in feature importance (S2.9_C). We have left the TEs in the final models because our goal is to examine the relationship between genic methylation and

expression potential. A subset of these 36,694 TEs and TE-containing genes were observed as proteins (2423) or highly expressed mRNAs (5,065).

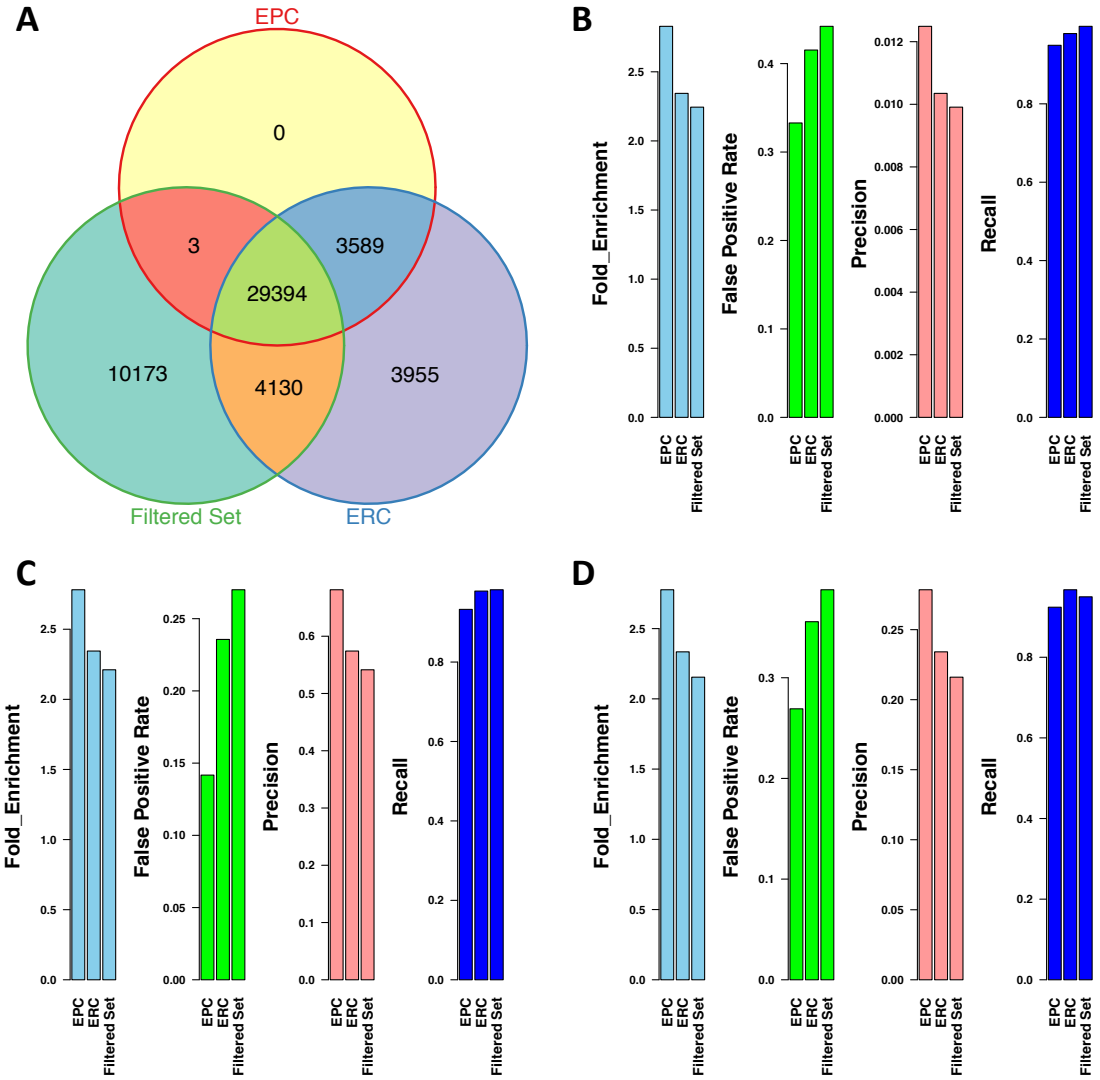


Figure 2.3 Express-able gene set annotations. (A) Overlap between the EPC express-able class genes, the ERC express-able class genes and the pre-defined maize filtered gene set (FGS). (B) Comparison between EPC, ERC and FGS using the pre-defined set of maize "classical" genes as a gold standard. (C) Comparison between EPC, ERC and FGS using the pre-defined set of syntenic orthologs between maize and sorghum as a gold standard. (D) Comparison between EPC, ERC and FGS using the pre-defined set of maize full-length cDNAs as a gold standard.

Expression Potential Associated with Genic Methylation Varies between Inbreds and between Tissues

To determine whether inbred-specific variation in genic methylation is associated with inbred-specific gene expression potentials, the EPC, which had been based on data from inbred B73, was extended to genetically diverse inbreds Mo17, CML322, Oh43 and Tx303. The new methylation data for each inbred were processed by quantifying weighted methylation levels for consecutive 100bp tiles along the genome. A new classifier, EPC-2, was trained on B73 14 day-old seedling data using the same class variable and methylation features as the original EPC but it was quantified using the tiled WGBS data; RNA-seq data from the other inbreds also came from 14 day-old whole seedlings. The new inbred WGBS samples were then used as test data and classified using EPC-2.

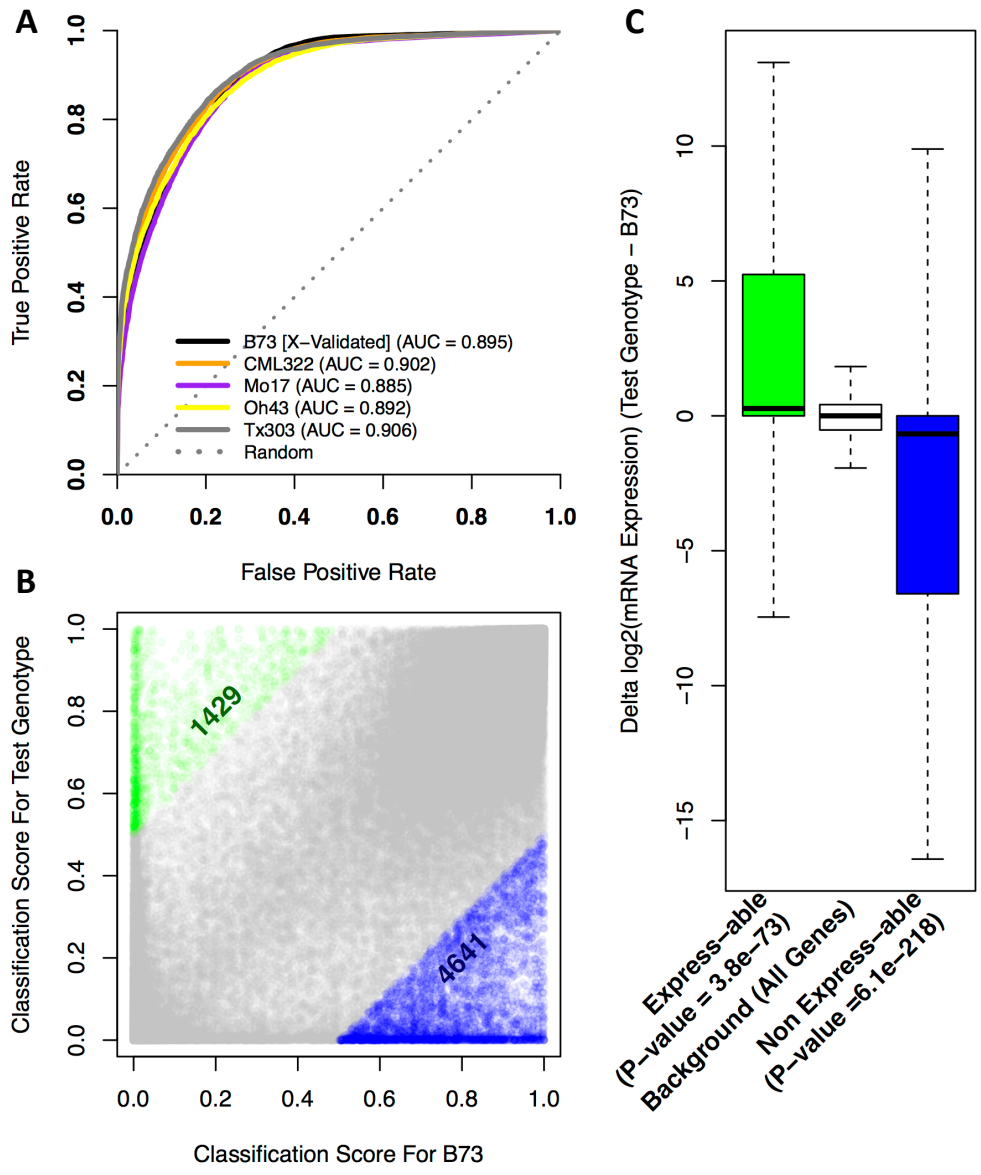


Figure 2.4 Testing random forest classifiers on separate maize inbred lines. (A) Receiver operating characteristic (ROC) curve showing prediction accuracies achieved when the EPC model is tested with new methylation data from different maize inbreds. (B) Scatter plot of cross-validated classification scores from the whole seedling (training set) vs. the other 3 test sets (tissues). Each point represents one gene for one seedling–to-tissue comparison. Upper left (green) and lower right (blue) sections represent genes that are classified differently in one tissue compared to the whole seedling. (C) Boxplot showing the difference in mRNA expression between whole seedling and individual tissues for all genes (white) and the differentially classified genes from B (green and blue). P-values were determined with a two-sided t-test for each subset (blue or green) against the background of all genes (white).

EPC-2 was accurate for all inbred lines (Fig2.4_A), with area under ROC curves of 0.88 or higher. As expected, the EPC-2 classification scores between B73 and other inbreds were highly correlated for most genes. However, some genes had reversed classification compared to B73 because of differential DNA methylation (Fig2.4_B blue and green dots). Upon examination of the corresponding changes in mRNA expression for these uncorrelated genes, we saw the expected differences in mRNA abundance (Fig2.4_C). That is, compared to B73, when genes in another inbred were re-classified as silent (blue), we observed lower/no expression; when genes that are silent in B73 were re-classified as expressed, a gain in expression was often observed (green).

EPC-2 was additionally used to determine whether developmentally regulated differences in genic methylation are associated with tissue-specific gene expression potentials. Three B73 tissues (developing ear, flag leaf and shoot apical meristem or SAM) were compared to 14 day-old whole seedlings. Each sample was characterized using WGBS and RNA-seq. We observed that EPC-2 accurately predicted loss of gene expression from genes that were re-classified as silent according to tissue-specific differences in genic methylation (blue in Fig2.5). However, silent genes that were re-classified (green) did not show a gain of expression.

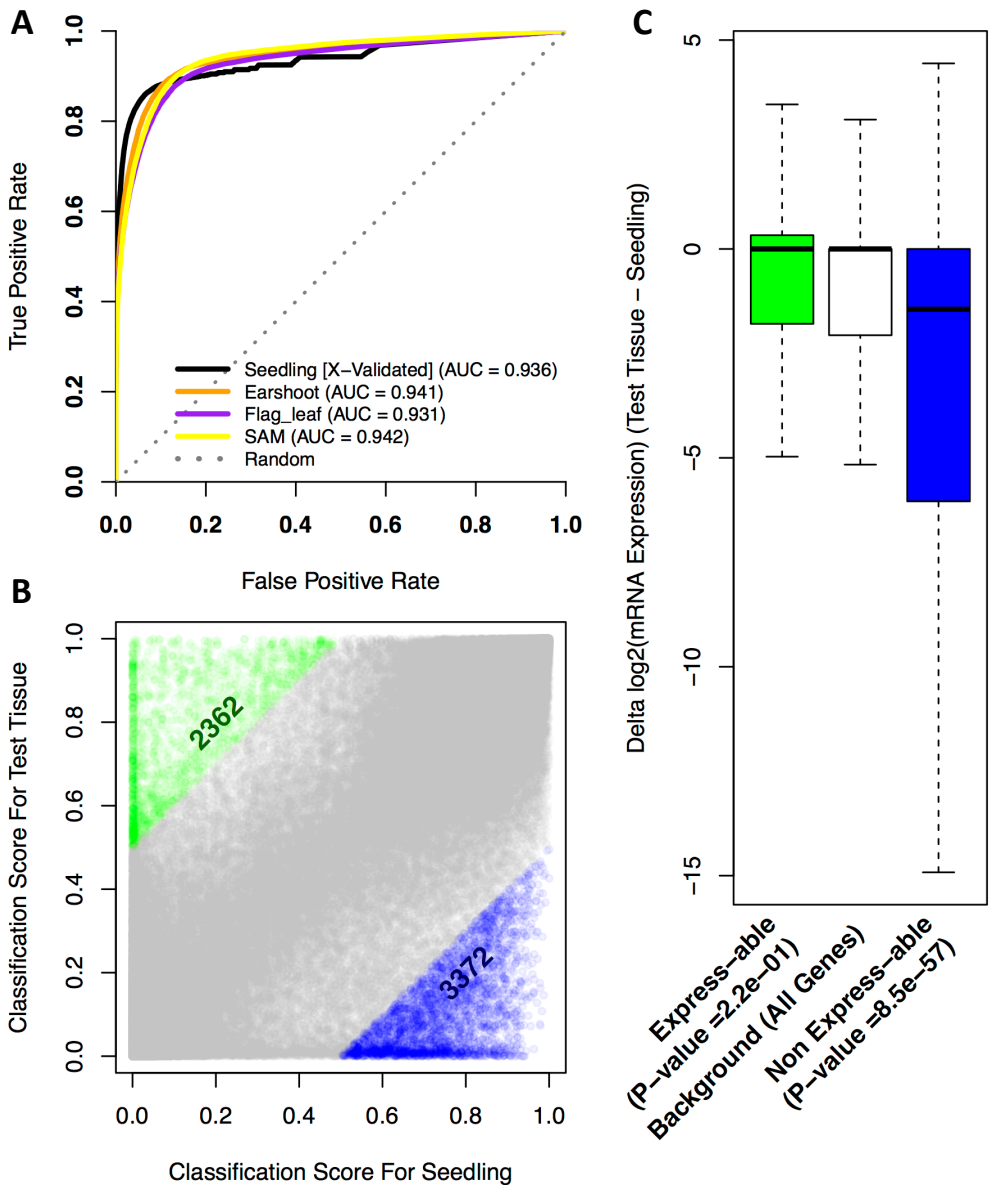


Figure 2.5 Testing random forest classifiers on separate B73 tissues. (A) Receiver operating characteristic (ROC) curve showing prediction accuracies achieved when the EPC model is tested with new methylation data from different maize tissues. (B) Scatter plot of cross-validated classification scores from the whole seedling (training set) vs. the other 3 test sets (tissues). Each point represents one gene for one seedling–to-tissue comparison. Upper left (green) and lower right (blue) sections represent genes that are classified differently in one tissue compared to the whole seedling. (C) Boxplot showing the difference in mRNA expression between whole seedling and individual tissues for all genes (white) and the differentially classified genes from B (green and blue). P-values were determined with a two-sided t-test for each subset (blue or green) against the background of all genes (white).

Genic Methylation Patterns Associated with Protein Expression

Of the 33,696 genes with observed mRNAs in the HR, less than half (15,421) had observed proteins. A third random forest model, the Protein-specific Feature Illuminator (PFI), was built to find genic methylation patterns which distinguish between genes that express high mRNA levels but no proteins and genes that have high mRNA levels plus observed proteins (HR/NP vs. HR/OP).. The PFI was able to differentiate between the HR/NP and HR/OP with good accuracy (Fig2.2_A & S2.5_A), achieving an area under the ROC curve of 0.8. Comparison of the feature importance between PFI and the EPC/ERC showed that most of the important features were shared including an inverse association between high CHG methylation in exons and protein expression. However, there were key differences. We observed a change in sign for mid-gene CpG methylation (bins 2-4), indicating an association between protein expression and high CpG methylation in the middle of genes (Fig2.2_C), a pattern previously described as gene body methylation (gbM) (Zhang, 2006 ; Zilberman, 2007). In particular, CpG methylation of introns and exons in the middle region of genes marked genes that expressed proteins compared with genes that only expressed high mRNA levels (S2.5_E).

We examined the association between gbM and protein expression directly. Genes with < 50% methylation in bins 1 and 5 plus > 50% methylation in at least one of bins 2-4 were defined as having gbM. Of these 9,071 genes,

59% had observed proteins, which is 3.5 times more than expected by chance (p-value = 0, based on hypergeometric test using the upper tail) (S2.8_A). High mRNA/ no protein genes showed a lesser enrichment of 1.4 fold (p-value = 1e-62) while low mRNA and non-expressed genes were under-enriched at 0.4 and 0.1 fold respectively (p-values = 0 for both, based on hypergeometric test using the lower tail). Of the 9,071 gbM genes, 88% were classified by the EPC as able to express proteins (S2.8_B).

Intronic regions were very important in the PFI model. Of the 110,028 genes in the WGS, only 55,558 (50%) contain introns; 70% of the FGS contain introns. Of genes with observed proteins, 88% contain 1 or more introns (S2.7_A). A link between introns and protein expression has been reported previously (Hir, 2003). We observed that 93% of genes with gbM contained introns; this is consistent with the hypothesis that gbM plays a role in RNA splicing (Wang, 2016 ; Maor, 2015; Regulski, 2013). We observed that the presence of introns distinguishes protein expressing genes from the other high mRNA expressing genes (S2.7_B).

DISCUSSION

Silencing and Protein Expression both Require Genic Methylation

Evidence supporting the hypothesis that genic DNA methylation affects gene expression has been developing for decades (Razin & Riggs, 1980). Genic cytosine methylation occurs from the transcription start site (TSS) to the

transcription termination site (TTS) in the context of CHH, CHG, and CpG. Advances in whole-genome bisulfite sequencing (WGBS) and mRNA profiling (RNA-seq) have made it possible to examine impacts of methylation in detail. It is known that, "Retrotransposons, viruses, transgenes, or repetitive genes are subject to silencing by RdDM" (Pikaard review CSHL). Genic methylation can silence protein-coding genes (Cubas, 1999; Melquist, 1999; Silveira, 2013) or small groups of genes with hyper-methylation at their transcription start sites (West, 2014). We found that silencing of protein-coding genes by genic methylation occurs on a genome-wide scale, affecting tens of thousands of genes.

Using random forest (RF) algorithmic classifiers, we identified patterns of DNA methylation in the gene body that robustly identified which genes were: able to express mRNA only; able to express mRNA plus protein; or were unable to express either mRNA or protein (S2.10). Examination of the most important features of these algorithmic classifiers revealed genic methylation patterns that corresponded to the mRNA and protein expression potentials of all protein-coding genes. In this study we demonstrated a strong association between levels of DNA methylation within gene models and their classification into populations based on observed proteins and/or mRNA abundance. More specifically, we found that high levels of DNA methylation from both symmetrical sequence contexts (CpG and CHG) at the beginning and ends of gene models is associated with silent genes. This does not prove causation

but the most parsimonious explanation is that genic methylation plays a key role in silencing a large proportion of protein-coding genes, which we refer to as Symmetrical DNA methylation-based Gene Silencing (SDGS).

A different pattern of methylation, known as gene-body methylation (gbM) has been observed in both plants and animals. This pattern is specifically defined as CpG hyper-methylation that occurs in the middle of the gene while both the 5' and 3' end of the gene body remain hypo-methylated (Zilberman, 2007). The occurrence of gbM in plants appears to be specific to angiosperms (Niederhuth, *BBA-Gene Regulatory Mechanisms,* 2016). However, within angiosperms, there are several reports of species that do not have detectable gbM (Bewick, 2016 ; Niederhuth, *Genome Bio., 2016*). gbM has been shown to be associated with constitutive mRNA expression and these mRNAs also tend have relatively high abundance (Zhang, 2006 ; Zilberman, 2007). The purpose of gbM remains unknown. One interesting hypotheses is that it acts to block TE insertion (Regulski, 2013) and therefore prevents mutagenesis of critical genes. Our finding that gbM is specifically associated with protein expression within the context of SDGS is unexpected and it suggest that gbM confers necessary properties for expression onto mRNAs. However, the level of expression is unrelated to the level of genic methylation.

Genome-Wide Annotations of Gene Expression Potentials

Both the EPC and ERC performed better than the FGS; the EPC was best and its application to the 98,296 members of the WGS for which we had WGBS data identified 32,995 genes with potential to express proteins (Fig2.3). The main advantage of our models appears to arise from exclusion of many thousand false-positive genes in the FGS. It should be noted that the classical and curated gene lists used for validation may not be completely independent from the FGS. It is possible that these lists themselves could have been used to inform the FGS. It is difficult to determine how much influence the former had on the later. Also, the FGS is not independent of the syntenic gene list since it incorporates synteny as one piece of evidence for inclusion. These dependencies may or may not contribute to the near prefect recall for the FGS on each of these validation sets.

We demonstrated that a single model, trained on B73 whole-seedling data, had the power to accurately classify diverse maize inbred lines (Fig2.4). These results show that each inbred has a unique “expressome” arising from its specific genic methylation patterns. We were able to accurately predict the “pan-expressome” of diverse inbreds based on features of their methylomes. This suggests that SDGS signatures can be used in genome selection-like models for breeding.

Our findings may shed light on the mechanisms of evolution and domestication. Plant genomes are very large. During the course of evolution, plant genomes were subjected to whole genome duplications (for recent

review, see [Panchy, 2017]). SDGS may be a mechanism for temporarily silencing genes after genome duplication. This nascent genetic material can then exist as a reservoir for potential variation that is tapped during the course of evolution/domestication as certain genes become un-methylated and express-able.

MATERIALS AND METHODS

Data Processing and Manipulation

The R Statistical Programing Language (R Core Team, 2016 ; Liaw and Wiener, 2002) along with the Rstudio integrated development environment (Rstudio, 2015) was used for all data processing and manipulation unless otherwise specified.

Figure Plotting

All figure were generated using R software. All details can be found in the supplemental source code file (Supplemental_Text01.R: Lines 170:905)

Quantitation of Whole Genome Bisulfite Sequencing Data

Raw sequencing reads for Whole Genome Bisulfite Sequencing (WGBS) of maize B73 14-day old seedlings were downloaded form the NCBI Sequence Read Archive (**SRA**) (SRA# SRR850328). Files were converted to fastq format using fastq-dump from the SRA toolkit. The Trim Galore program

(Krueger, 2007) was used with the "--paired" argument to remove paired-end adapter sequences. The Bismark program (Krueger, 2011) was used to identify methylated cytosines against the maize B73 RefGen_v2 genome, allowing for 1 miss-match "-N 1". Individual methylated cytosines were summarized separately for the 3 methylation contexts (CpG, CHG and CHH) using "bismark_methylation_extractor" with arguments "--paired-end" and "--no-overlap" which prevents duplicated methyl-C counts from read pairs. Gene model chromosomal coordinates were obtained from the B73 RefGen_v2 working gene set (known as 5a_WGS). These loci were extended 250 bp upstream in order to include sequences around the transcription start sight (TTS). All three methylation contexts were then mapped to these loci using the R statistical computing language. Several genomic regions were identified for each gene model using the 5a WGS annotations. These regions included 500 base pairs around the TTS (250 bp up and down stream), the 5'UTR, All Exons, All Introns, the 3' UTR and the whole gene model (from TSS to transcription stop sight). Methylation of each of these regions was quantified separately. For each context, the fraction of methylated cytosines relative to the total detected cytosines was used as a measure of methylation level, known as the weighted methylation level (Schultz, 2012).

Gene-wise Binning of DNA Methylation Levels

All gene models were divided into 5 equal sized, consecutive sections called bins (Fig2.1_A). Quantitation of DNA methylation levels was carried out separately for each bin.

Handling of Missing DNA Methylation Data

Of the 110028 gene models in the maize B73 5a Working Gene Set, 11732 did not have coverage in the WGBS data. These genes were discarded from further analysis. For genes with coverage along the gene model, not all genomic features have WGBS coverage. In this case, missing data was simply substituted with a place-holder value of 0.5 which represents an un-informative number as this feature is neither hypo or hyper-methylated.

Observed Protein List and Protein Abundance Data

All protein data is from (Walley & Sartor, et al., 2016).

mRNA Abundance Data

mRNA abundance used for the Express-able mRNA Classified (ERC) and the Protein-specific Feature Illuminator (PFI) was take from (Walley & Sartor, et al., 2016). **RNA-sequencing of Multiple B73 Tissues.**

For comparison of mRNA abundance to DNA methylation in various tissues, the closest approximation of one or more samples from (Walley & Sartor, et al.) was used to quantify mRNA in 3 different B74 tissues. For the

shoot apical meristem (SAM), this same tissue was sampled. For the flag leaf, the mRNA data for mature leaf 8 was used and for the ear shoot, two mRNA samples form 2-4 mm ear primordial and 6-8 mm ear primordial were averaged.

Filtered Gene Set Annotation

The list of 45348 gene models (parent genes) was taken from the Maize GDB annotation file "ZmB73_5b.60_FGS.translations.fasta" found on the Maize GDB ftp site (http://ftp.maizegdb.org/). This list is constructed and maintained by the Maize GDB project. Its construction is described on the site as:

"The Filtered Gene Set (FGS) is a subset of the Working Gene Set intended to exclude transposons, pseudogenes, contaminants, and other low-confidence annotations. [Maizesequence.org] used essentially the same method as described in (Schnable et al. 2009. 326:1112) but with modifications. First, the inclusion criterion of synteny (relative rice, sorghum, and Brachypodium distachyon) was given higher precedence than the exclusion criteria of pseudogene and transposon. This measure was taken to avoid exclusion of possibly legitimate genes that may have been miss-assembled or miss-annotated. Second, selection of the FGS was additionally informed by evidence of expression, taking advantage of RNA-seq data displayed on this site (Li et al 2010. Nature Genetics 42:1060). In addition, a

total of 786 genes in the 4a filtered set (Schnable et al) were rejected this time around primarily because they physically overlap with other genes in the FGS and have inferior evidence/confidence."

Gene Biotype Annotation

Gene Biotypes were extracted for the gff annotation file "ZmB73_5a_WGS.gff" that was downloaded from the Maize GDB ftp site (http://ftp.maizegdb.org/). Biotype information is listed under "gene" features in the information field.

Annotation of the Number of Introns

For each gene model in the working gene set, the number of introns was determined by counting all unique introns listed in the gff annotation file "ZmB73_5a_WGS.gff" that was downloaded from the Maize GDB ftp site (http://ftp.maizegdb.org/).

Validation Lists of Known Functional Genes

Four separate gene lists were used to validate classification results of this study: i) The Classical maize genes are a list of 464 manually curated gene models where at least 3 publications are listed on maize GDB that are associated with that locus (Schnable, 2011). ii) The Syntenic genes are a list of 24,092 gene models conserved at syntenic orthologous locations between maize and sorghum, take from (Walley & Sartor, et al., 2016). iii) The Full

Length cDNA genes are a list of 9,940 genes. These are gene models taken from the Maize Full Length CDNA Project (Soderlund, 2009) that could be cross-referenced with accessions used for the RefGen_v2 annotations. iv) The Curated genes are a list of 4,329 gene models that have been manually assigned a name by Maize GDB (Andorf, 2016).

Construction of Classification Models (ERC and EPC)

(Supplemental_Text01.R: Lines 57 – 118)

All models used for classification were constructed in the same way, with differences only in the class variable on which they were trained. Random Forest classification models were used (Breiman, 2001) as implemented in the R statistical programing language. For training sets, a matrix of methylation data described above was used as features (independent variables) for every model and a unique classification variables (dependent variables) was used to train the Express-able Protein Classified (EPC) and the Express-able mRNA Classifier (ERC) separately. These class variables were defined using transcript abundance data (ERC) or a combination of transcript and protein abundance data (EPC). Therefore two classifiers were built. For the ERC classification training vector, the “Non Express-able” class was defined as genes with no detectable mRNA abundance (No RNA: 37588 genes) and the “Express-able” Class was defined as high abundance mRNA genes with FPKM >1 (High RNA: 33696 genes). For the EPC, this classification vector

was further refined by also requiring no detectable Protein in the “Non Express-able” class (No RNA / No Protein: 37487 genes) and detected protein in the “Express-able” class (High RNA / Observed Protein: 15421 genes). The random forest models were built with the classification vector as a factor type using 1000 trees and importance = T which returns the “Mean Decrease in Accuracy” measure of importance. The DNA-methylation data used when looking across different genotypes and different B73 tissues (Fig2.4 and 2.5) was summarized independently and 100bp tiling was done for quantitation (see below). Therefore a new EPC model was re-trained on this tiled data. The same training classification vector for the EPC was used (Supplemental_Text01.R: Line 627-686)

Classification of Test Data

(Supplemental_Text01.R: Lines 150 – 157)

The Random forest classifiers (ERC and EPC) were used to classify the remaining genes that were not represented in the training set. The Random Forest “predict()” function was used along with the DNA methylation data for the test genes. This same procedure was used to classify genes based on methylation data from the different maize inbred seedlings (Supplemental_Text01.R: Lines 831 – 833) and different B73 tissues (Supplemental_Text01.R: Lines 690-692). The classifications for genes in the

training set were obtained form the Out-of-Bag cross-validated predictions of each model ("$predicted" slot from the Random Forest object).

Construction of Regression Model (PFI)

(Supplemental_Text01.R: Lines 57 - 118)

The PFI regression model was built in the same manor as the classification model described above. The only difference is the classification vector used for training. For the PFI, the first class was defined as genes with mRNA expression > 1 FPKM and non-observed protein (High RNA / No Protein: 18275 genes). The other class is composed of genes with mRNA expression >1 and observed protein (High RNA / Observed Protein: 15421 genes).

Construction of Quantitative Expression Predictors

(Supplemental_Text01.R: Lines 101 - 125)

Two predictive models were constructed in an attempt to predict quantitative mRNA and Protein expression levels using DNA methylation data. The mRNA Expression-level Predictor (REP) and Protein Expression-level Predictor (PEP) were built using random forest models in much the same way as the ERC and EPC (described above), with one difference. The dependent variable training vector was a continuous numeric variable representing either observed mRNA or observed Protein log2(abundance). Non observed values

are represented as the integer less than the lowest observed value for the data set (since log2(0) = -infinity). This is -12 for the mRNA data and -1 for the protein data. The Out-Of-Bag cross-validated results were extracted from the "$predicted" slot in the Random Forest objects. These values were used for further analysis and plotting (Fig2.2_B & S2.5_B, C & D).

Feature Importance (Regression) Measure

(Supplemental_Text01.R: Line 129)

The feature importance was determined by the random forest algorithm, using the mean decrease in accuracy upon random permutation of each individual variable (Breiman, 2001). This requires the parameter "importance = TRUE" in the call to the "randomForest()" function.

Determination of the Mathematical Sign (Direction) of Relationships Between Methylation Features and Classification

(Supplemental_Text01.R: Lines 133 - 146)

The sign of the relationship between DNA-methylation features and classifications was determined as follows. Genes in the training set were split between the two training classes (Non Express-able and Express-able) to yield two populations (S2.06). For each feature, the corresponding feature values were assigned to each of the class populations and a student's t-test was carried out between the populations. The sign of the resulting t-statistic

was assigned to each feature while the magnitude of each was taken from the RF variable importance (Fig2.2_C).

Receiver Operating Characteristic (ROC) Curves and Precision Vs. Recall Curves

(Supplemental_Text02.R: PlotROCPR(): Lines 88 - 220)

The “ROCR” R package was used to generate all ROC curves. For each gene, the number of votes in the random forest model was used as a quantitative classification score and this was evaluated against the classification vector. For the ERC and EPC, the votes were taken from the out-of-bag cross-validated predictions (Supplemental_Text01.R: Line 313). When evaluating model accuracies of independent test samples (different B73 tissues and different maize inbreds) the random forest model trained on B73 seedling data was used to classify these independent test sets. Again, the number of votes was used as a quantitative classification score. These scores were evaluated against a response vector that was generated from separate RNA-seq data sets corresponding to each test sample. In these response vectors, “positive” genes were ones with mRNA abundance > 1 FPKM and “negative” genes were ones with undetected mRNA abundance (Supplemental_Text01.R: Lines 703-714 & 846 – 857).

For all curves, the “prediction()” function from the ROCR package was used to evaluate the model. For the ROC curves, the “performance()” function was used with arguments “tpr” and “fpr” to generate curves and the argument “auc” was used to calculate the area under the curves (Supplemental_Text02.R: Lines 118-119). For PR curves, the “performance()” function was used with arguments “prec” and “rec” to generate curves (Supplemental_Text02.R: Line 120). The area was calculated by estimating a function for the curve with “approxfun()” and integrating across its entire length (Supplemental_Text02.R: Lines 226-240)

CpG Gene Body Methylation Genes

(Supplemental_Text02.R: RetrurnGBMGenes(): Lines 359 - 371)

All genes with DNA-methylation data were tested for CpG Gene body methylation (gbM). A gene was said to have gbM if the average methylation of the end bins (1 and 5) was less than 0.5 and the average methylation of the center bins (2-4) was greater than 0.5.

Validation of Express-able Gene Lists Vs. Known Functional Gene Sets

(Supplemental_Text01.R: Lines 561 – 576)

All validation lists (listed above) were compared to the Express-able protein and Express-able mRNA classes and the set of all genes with methylation data was used as a background. A number of statistics were

computed for each (Supplemental_Text02.R: GetOverLapStats: Lines 295 – 317). These include the Fold Enrichment (Observed Frequency / Expected Frequency), False Positive Rate, Precision and Recall.

SUPPELEMENTAL FIGURES

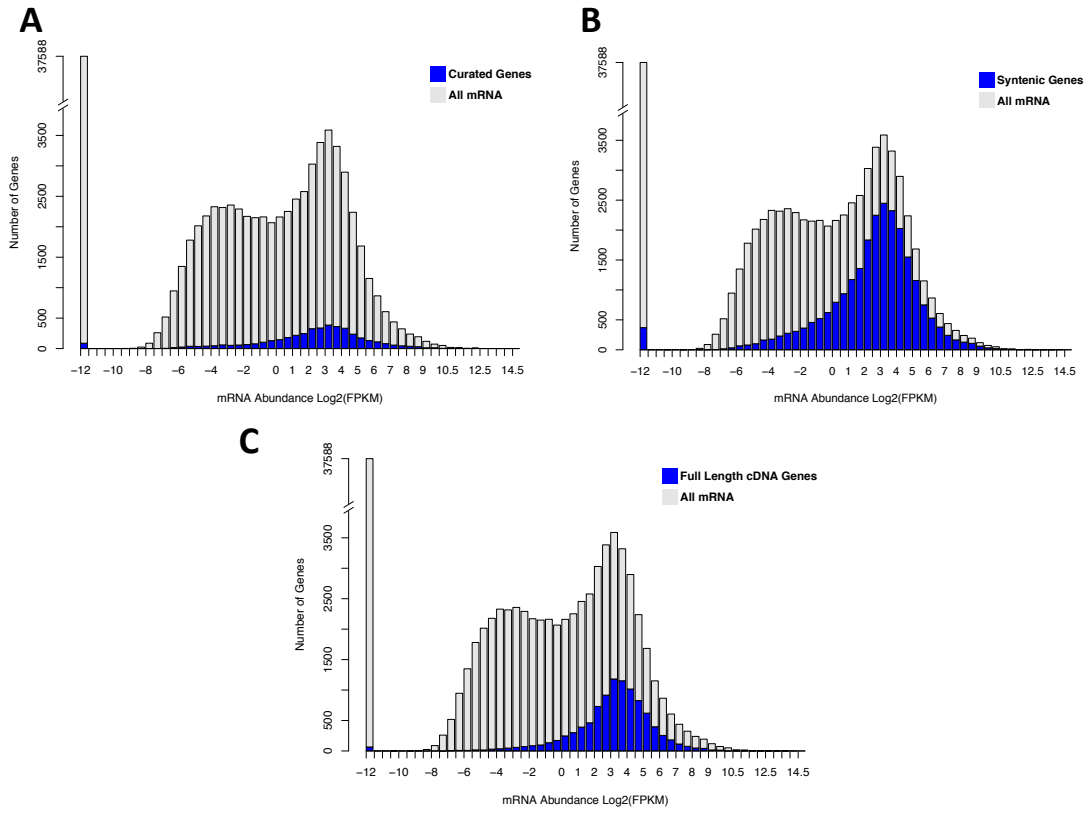


Figure S2.1 mRNA abundance distributions with various sub-distributions of annotated gene sets highlighted (blue). (A) Shows the set of curated genes from the Maize GDB project. (B) Shows the set of maize syntenic orthologs against sorghum. (C) Shows the set of maize full-length cDNAs.

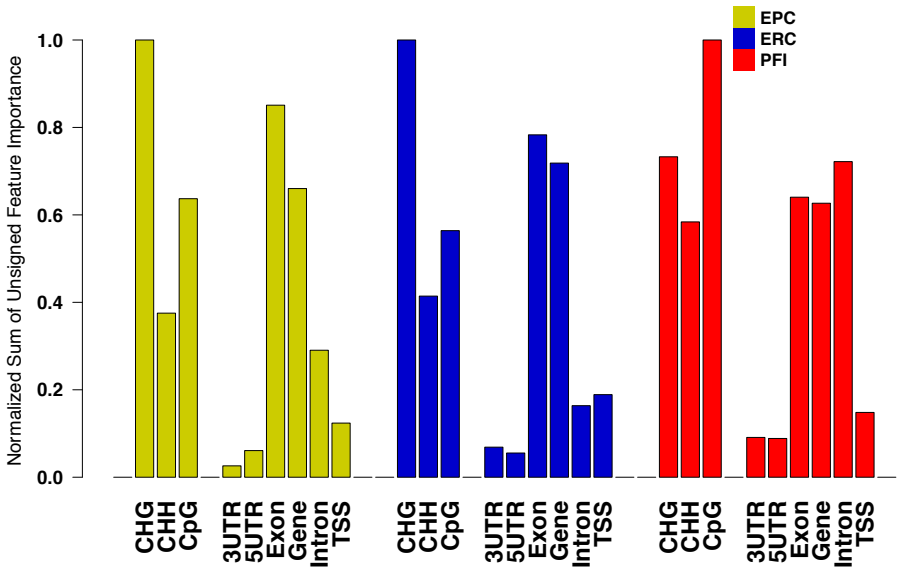


Figure S2.2 Summarized measures of feature importance summed over various methylation features.

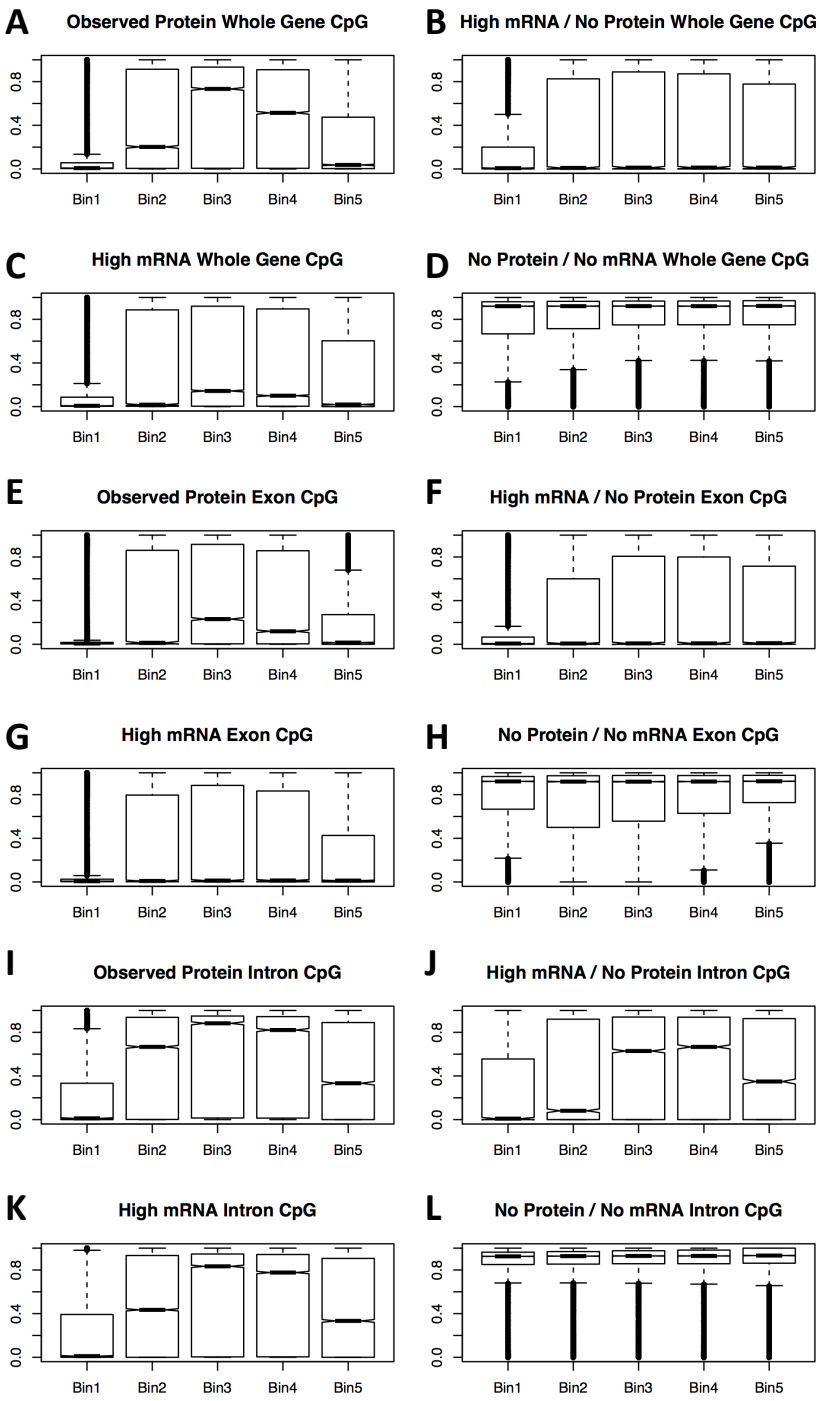


Figure S2.3 Binned DNA methylation levels in the CpG context. Boxes show distributions of the proportion of (methylated cytosines / All cytosines) for all 5 bins for various gene sets determined based on mRNA abundance and the presence of observed or non-observed proteins. (A-D) Summarized over the whole gene models. (E-H) Summarized over exons only. (I-L) Summarized over introns only.

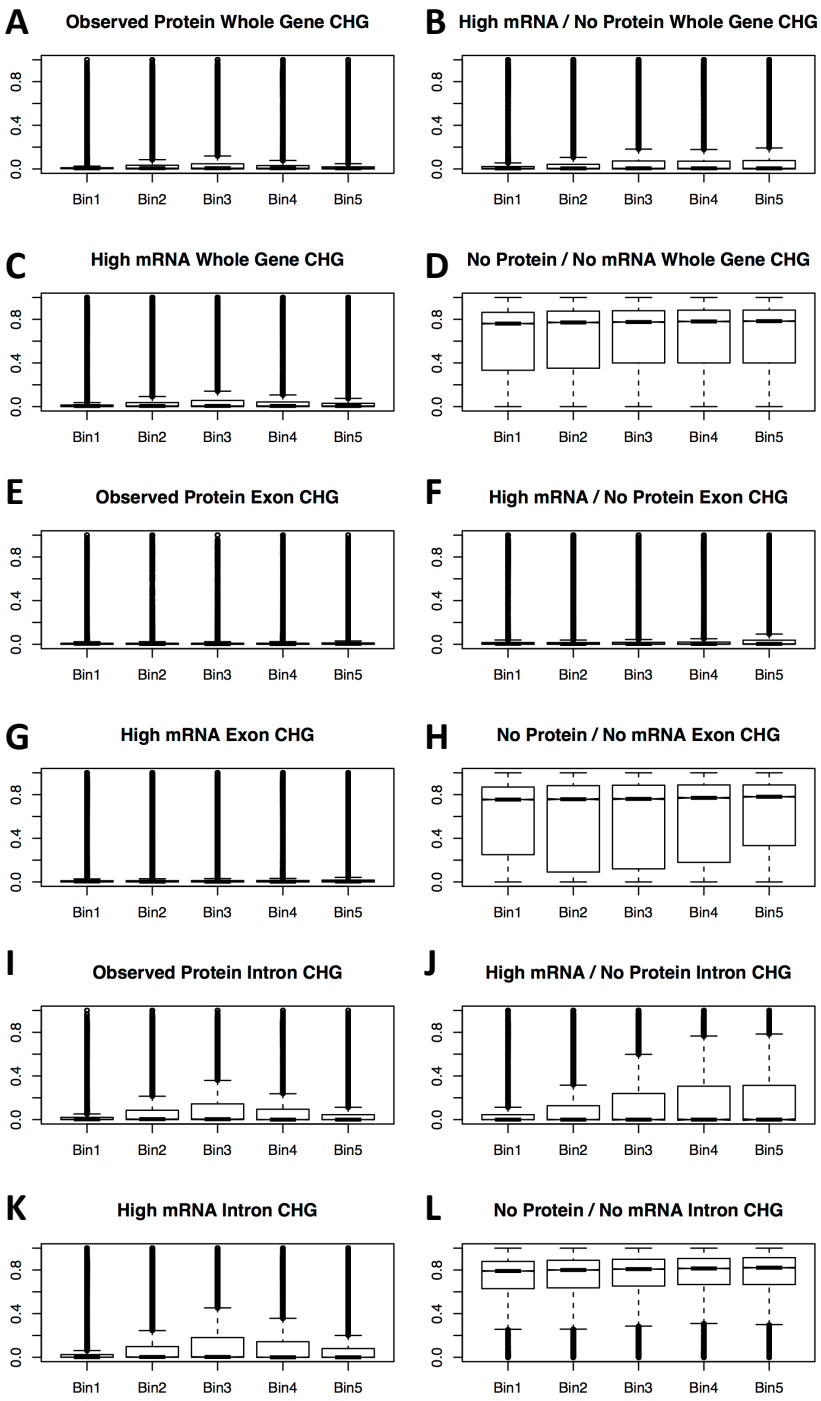


Figure S2.4 Binned DNA methylation levels in the CHG context. Boxes show distributions of the proportion of (methylated cytosines / All cytosines) for all 5 bins for various gene sets determined based on mRNA abundance and the presence of observed or non-observed proteins. (A-D) Summarized over the whole gene models. (E-H) Summarized over exons only. (I-L) Summarized over introns only.

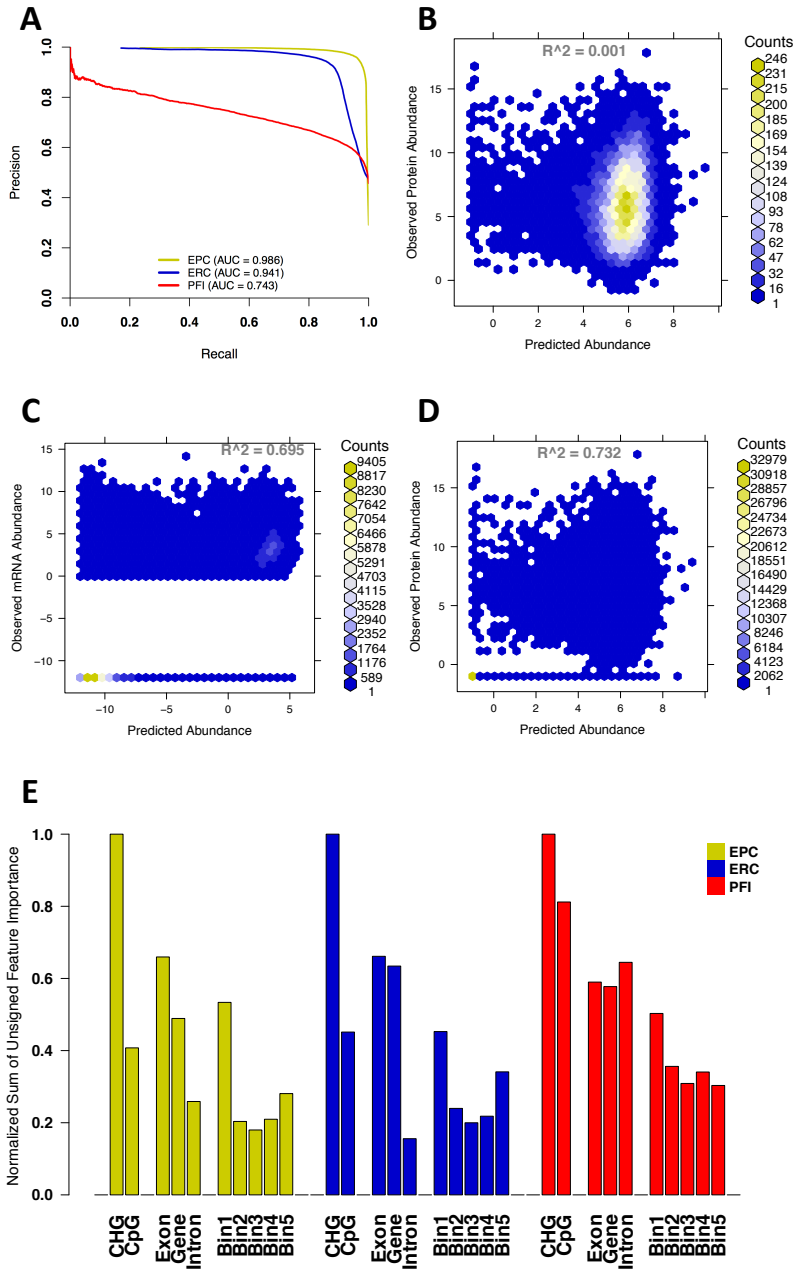


Figure S2.5 Results for random forest models. (A) Precision vs. Recall (PR) curves showing classification accuracy of the EPC, ERC, and PFI models. The random forest "votes" from the out-of-bag cross-validated classifications were used for all models. (B) Prediction accuracy for quantitative protein abundance model, only looking at genes with observed proteins. (C) Prediction accuracy for quantitative mRNA abundance model, looking at all genes with methylation data. (D) Prediction accuracy for quantitative protein abundance model, looking at all genes with methylation data.

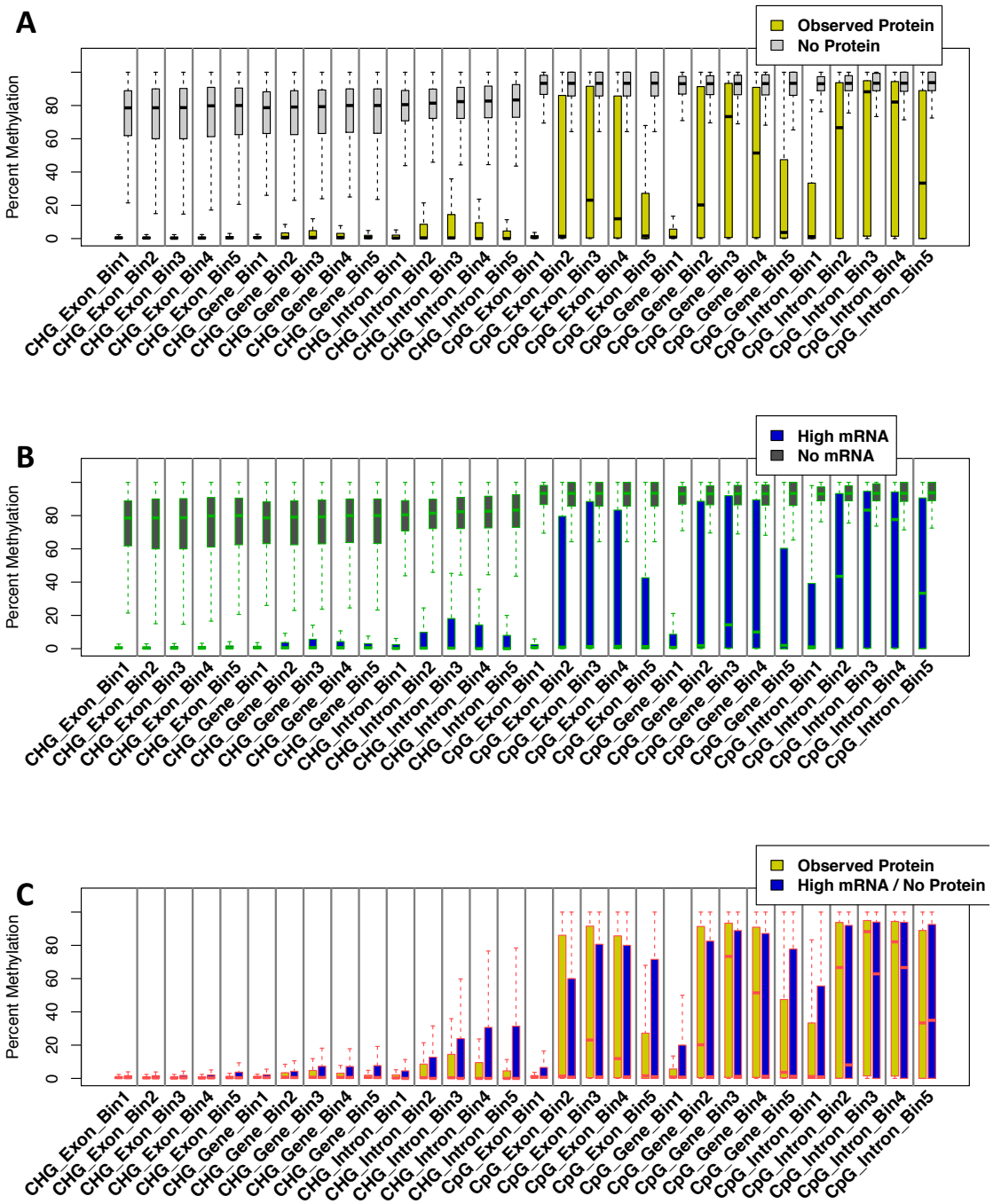


Figure S2.6 Boxplots showing all methylation feature levels for the two different training classes of each random forest model. (A) Results for the EPC model. (b) Results for the EPC model and (C), results for the PFI regression model.

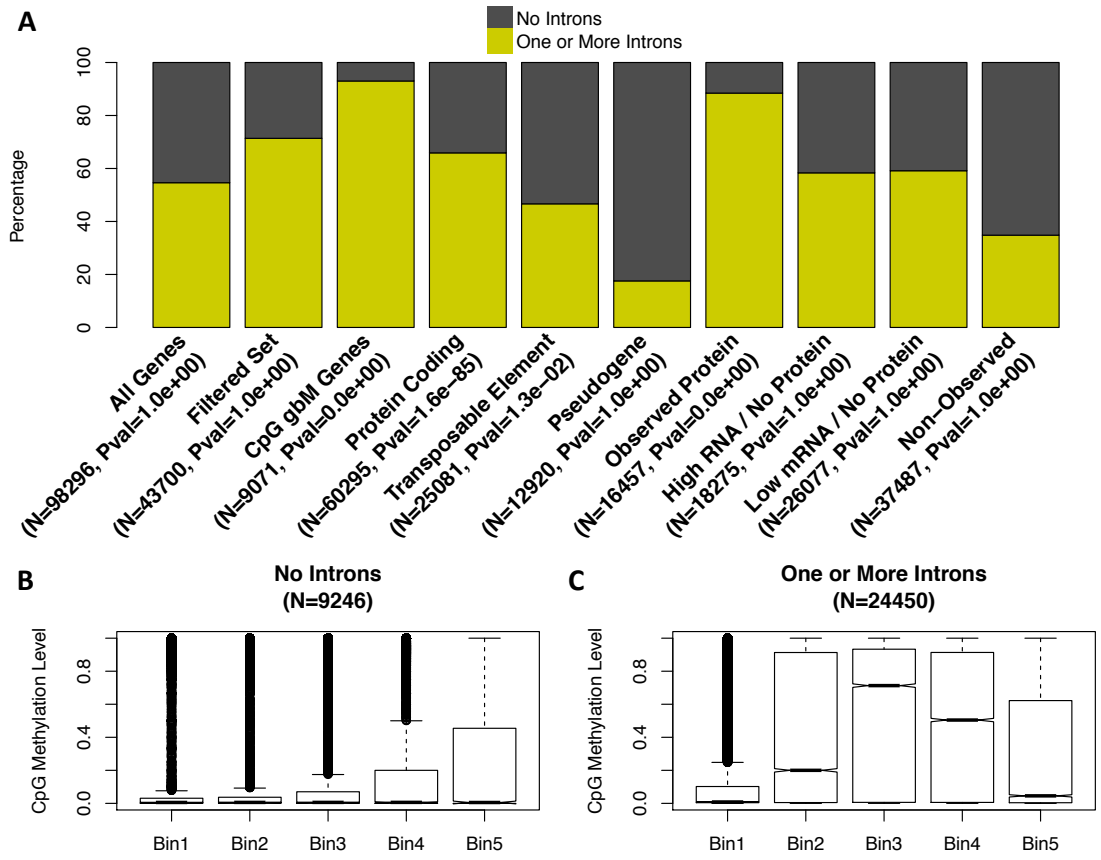


Figure S2.7 Analysis of genes with one or more introns. (A) The relative proportions of intron-containing and non-intron genes in various pre-defined gene sets. “N” specifies the total number in that set. “Pval” specifies the p-values calculated for the significance of enrichment of intron-containing genes for each category relative to the filtered set, using a hypergeometric test and the upper tail. It should be noted that this calculation, by default, only includes genes present in the filtered set. Therefore, the ratios used in the p-value calculation will differ slightly from those shown in the plot that are relative to all genes. (B) The distribution of CpG methylation levels across all 5 bins for genes in the HR with no introns. (C) The distribution of CpG methylation levels across all 5 bins for genes in the HR with one or more introns.

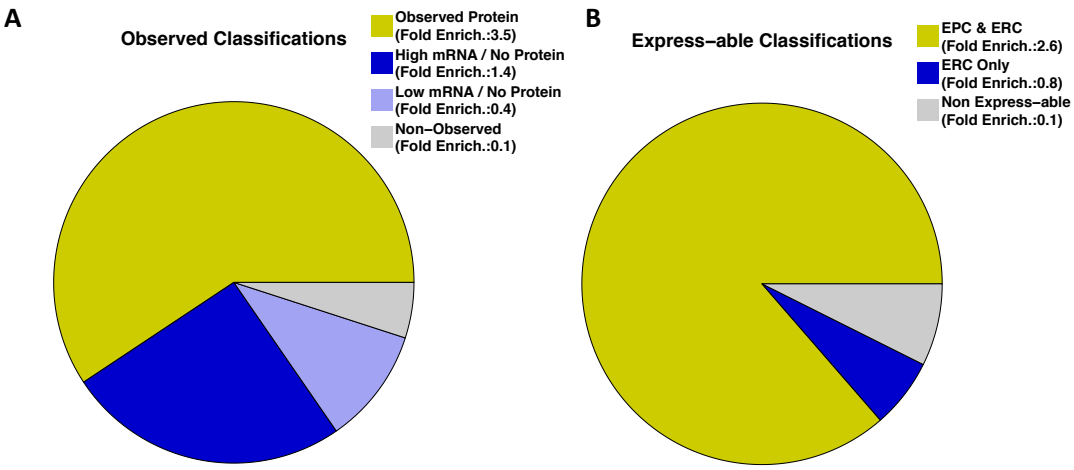


Figure S2.8 Classes of genes with CpG gene body methylation. (A) Pie chart displaying the observed distribution of 9071 genes with CpG gene body methylation patterns into various expression populations along with the fold enrichment ratio (Observed / Expected) for each set. (B) Pie chart displaying the distribution of Express-able Protein Classifier and Express-able RNA Classifier results for the 9071 genes with CpG gene body methylation patterns. Note that no “EPC only” is shown because only 1 gene exists in this category and it is not gbM.

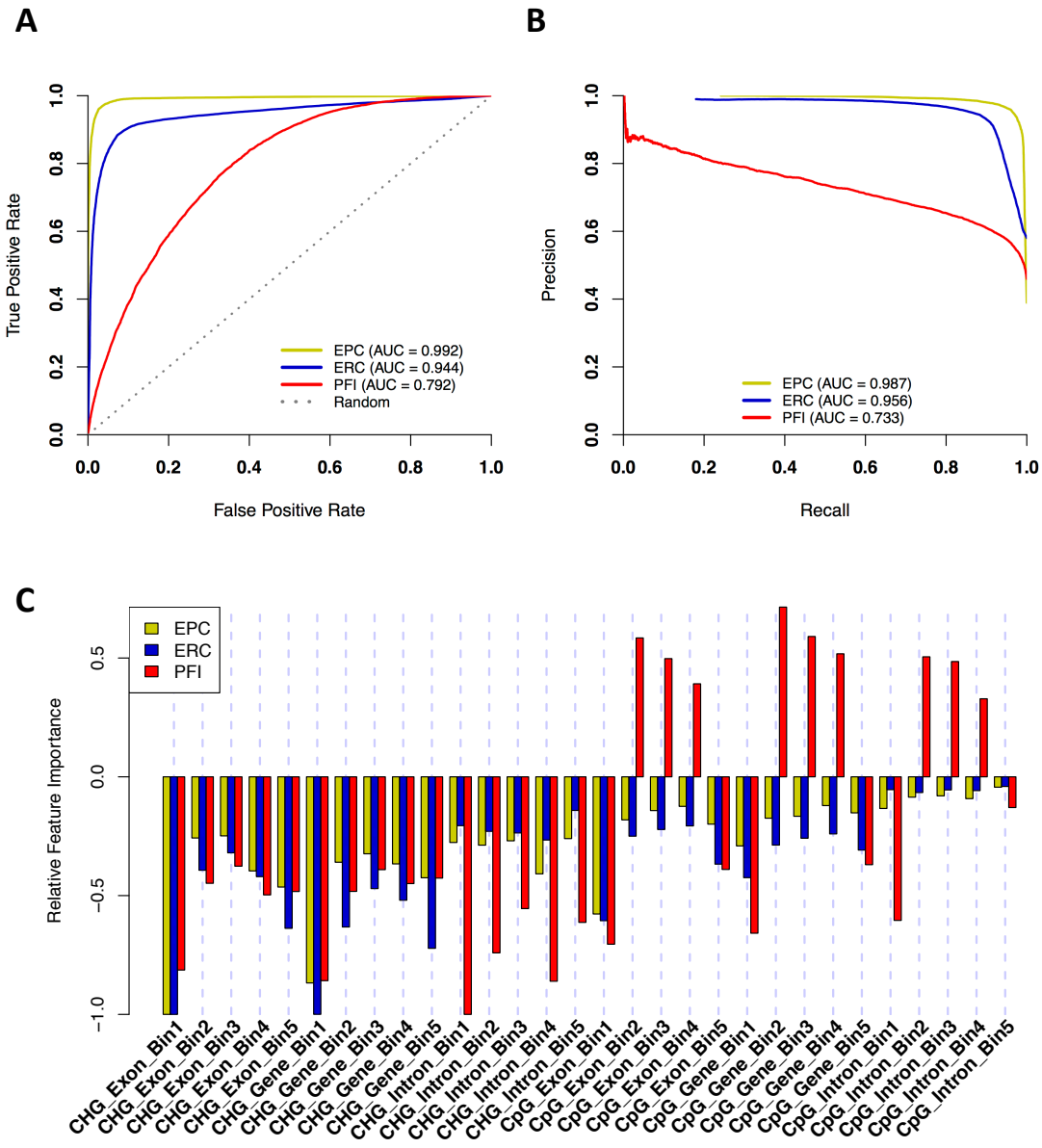


Figure S2.9 Classifiers built with no transposable elements. Out of the 98,296 genes with methylation data, 32,400 were identified as likely transposable elements (TEs). After filtering out these TEs, classifiers were re-built. (A) Receiver Operating Characteristic (ROC) curve and (B) Precision vs. Recall (PR) showing predictive accuracy of out-of-bag cross validated classification results against observed mRNA and Protein abundance-based classifications. (C) Feature importance measures for each methylation feature used in the classification models. The sign indicates the sign of the relationship between the quantity of the feature vs. the positive (Observed Protein and/or High mRNA) class.

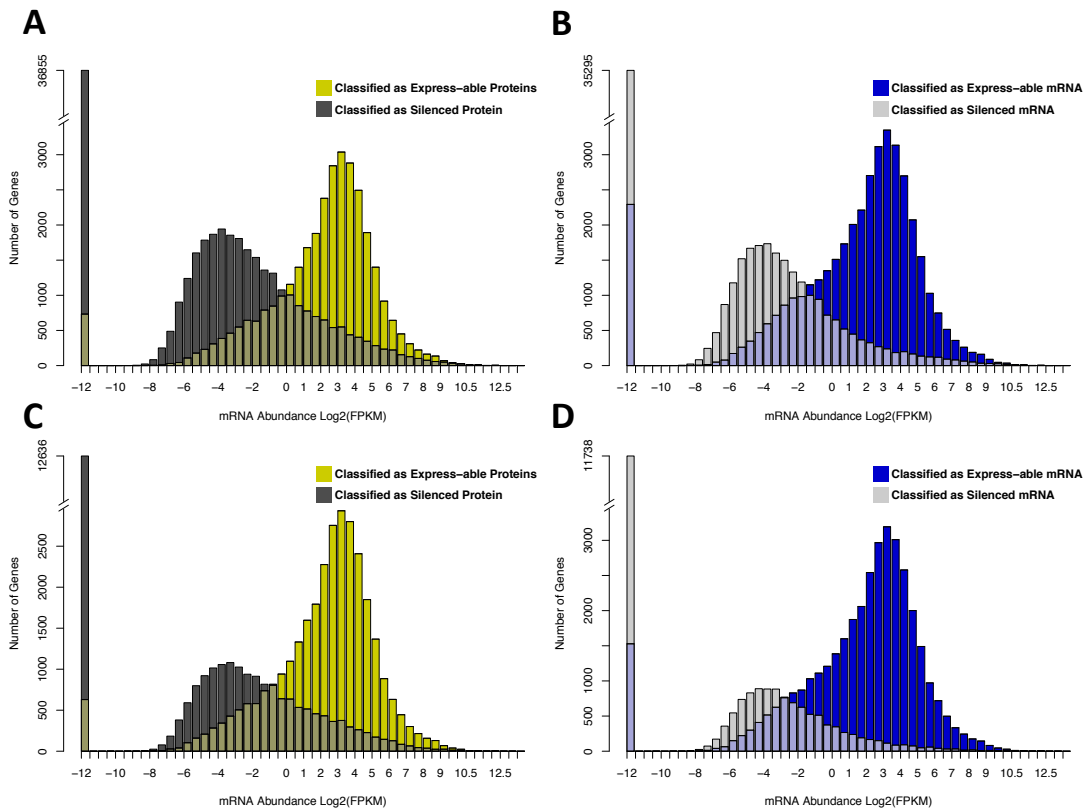


Figure S2.10 mRNA abundance distributions of classifier results. (A) Results from the EPC of all genes with methylation data. (B) Results from the ERC of all genes with methylation data. (C) Results from the EPC of all genes with "protein coding" biotype. (D) Results from the ERC of all genes with "protein coding" biotype.

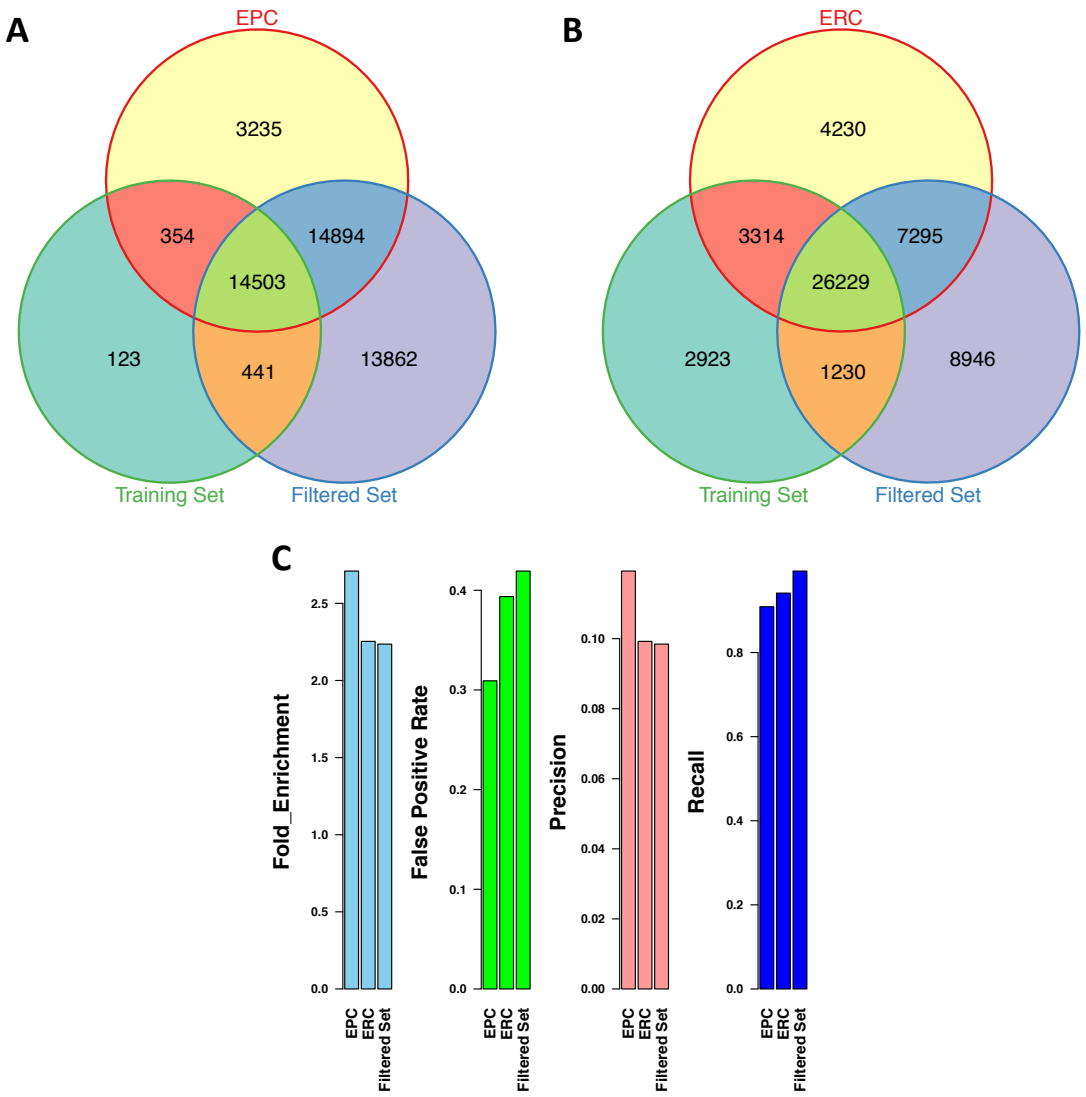


Figure S2.11 Results from EPC and ERC classification. (A) Venn diagram showing the EPC results for genes classified as express-able proteins (top circle), all genes with observed proteins in the training set (bottom left) and the maize filtered set (bottom right). (B) Venn diagram showing the EPC results for genes classified as express-able mRNA (top circle), all genes with high mRNA in the training set (bottom left) and the maize filtered set (bottom right). (C) Comparison between EPC, ERC and the filtered gene set using the pre-defined set of maize GDB named genes as a gold standard.

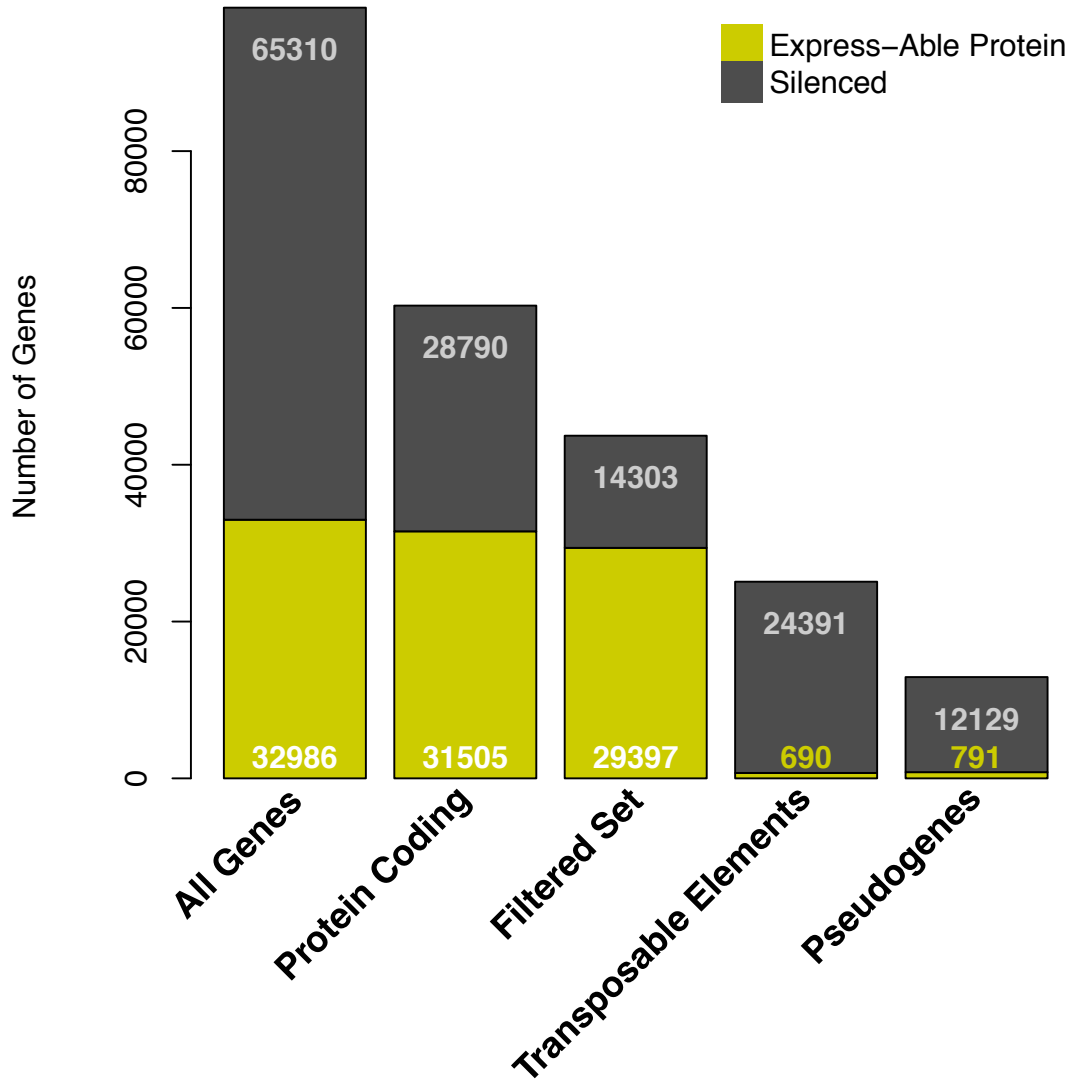


Figure S2.12 Classification results from the EPC classifier. Stacked bar plot showing multiple pre-defined gene sets and the relative proportions that are classified as express-able or silenced by the EPC model.

ACKNOWLEDGMENTS

Chapter 2, in full is currently being prepared for submission for publication of the material. Sartor, R. C., Noshay, J., Springer, N. M., Briggs, S. P. The dissertation author was primary investigator and first author of this work.

REFERENCES

Andorf, C. M.; Cannon, E. K.; Portwood, J. L.; Gardiner, J. M.; Harper, L. C.; Schaeffer, M. L.; Braun, B. L.; Campbell, D. A.; Vinnakota, G.; Sribalusu, V. V; Huerta, M.; Cho, K. T.; Wimalanathan, K.; Richter, J. D.; Mauch, E. D.; Rao, B. S.; Birkett, S. M.; Sen, T. Z.; Lawrence-dill, C. J. MaizeGDB update : new tools , data and interface for the maize model organism database. *Nucleic Acids Research* **44**, 1195–1201 (2016).

Bewick, A. J.; Ji, L.; Niederhuth, C. E.; Willing, E.-M.; Hofmeister, B. T.; Shi, X.; Wang, L.; Lu, Z.; Rohr, N. A.; Hartwig, B.; Kiefer, C.; Deal, R. B.; Schmutz, J.; Grimwood, J.; Stroud, H.; Jacobsen, S. E.; Schneeberger, K.; Zhang, X.; Schmitz, R. J. On the origin and evolutionary consequences of gene body DNA methylation. *Proceedings of the National Academy of Scien* **113**, 9111–9116 (2016). [PMID:27457936]

Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001). [PMID:21816105]

Cubas, P.; Vincent, C.; Coen, E. An epigenetic mutation responsible for natural variation in ¯ oral symmetry. *Nature* **401**, 157–161 (1999).

Eichten, S. R.; Swanson-Wagner, R. A.; Schnable, J. C.; Waters, A. J.; Hermanson, P. J.; Liu, S.; Yeh, C. T.; Jia, Y.; Gendler, K.; Freeling, M.; Schnable, P. S.; Vaughn, M. W.; Springer, N. M. Heritable epigenetic variation among maize inbreds. *PLoS Genetics* **7**, (2011). [PMID:22125494]

Eichten, S. R.; Ellis, N. A.; Makarevitch, I.; Yeh, C.; Gent, J. I.; Guo, L.; Mcginnis, K. M.; Zhang, X.; Schnable, P. S.; Vaughn, M. W.; Dawe, R. K.; Springer, N. M. Spreading of Heterochromatin Is Limited to Specific Families of Maize Retrotransposons. *PLoS Genetics* **8**, (2012).

Hebenstreit, D.; Fang, M.; Gu, M.; Charoensawan, V.; Van Oudenaarden, A.; Teichmann, S. a. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology* **7**, 497 (2011). [PMID:21654674]

Hir, H. L.; Nott, A.; Moore, M. J. How introns influence and enhance eukaryotic gene expression. *TRENDS in Biochemical Sciences* **28**, 215–220 (2003).

Krueger, F. , Trim Galore!. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (2007)

Krueger, F.; Andrews, S. R. Bismark : a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

Maor, G. L.; Yearim, A.; Ast, G. The alternative role of DNA methylation in splicing regulation. *Trends in Genetics* **31**, 274–280 (2015).

Melquist, S.; Luff, B.; Bender, J. Arabidopsis PAI Gene Arrangements , Cytosine Methylation and Expression. *Genetics* **153**, (1999).

Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology* **7**, 548 (2011). [PMID:22068331]

Niederhuth, C. E.; Bewick, A. J.; Ji, L.; Alabady, M.; Kim, K. Do; Page, J. T.; Li, Q.; Rohr, N. A.; Rambani, A.; Burke, J. M.; Udall, J. A.; Egesi, C.; Schmutz, J.; Grimwood, J.; Jackson, S. A.; Springer, N. M.; Schmitz, R. J. Widespread natural variation of DNA methylation within angiosperms. *Genome Biology* **17**, 194 (2016). [PMID:25246403]

Niederhuth, C. E.; Schmitz, R. J. Putting DNA methylation in context: From genomes to gene expression in plants. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (2016).

Panchy, N.; Lehti-shiu, M.; Shiu, S. Evolution of Gene Duplication in Plants. *Plant Physiology* **171**, 2294–2316 (2016).

Pikaard, C. S.; Scheid, O. M. Epigenetic Regulation in Plants. *Cold Spring Harbor Perspect Biol* **6**, (2014).

Razin, A.; Riggs, A. DNA Methylation and Gene Function. *Science* **210**, 604–610 (1980).

Regulski, M.; Lu, Z.; Kendall, J.; Donoghue, M. T. A.; Reinders, J.; Llaca, V.; Deschamps, S.; Smith, A.; Levy, D.; McCombie, W. R.; Tingey, S.; Rafalski, A.; Hicks, J.; Ware, D.; Martienssen, R. A. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Research* **23**, 1651–1662 (2013). [PMID:23739895]

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Schnable, J. C.; Freeling, M. Genes Identified by Visible Mutant Phenotypes Show Increased Bias toward One of Two Subgenomes of Maize. *PLoS ONE* **6**, e17855 (2011).

Schultz, M. D.; Schmitz, R. J.; Ecker, J. R. “Leveling” the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics* **28**, 583–585 (2012). [PMID:23131467]

Silveira, A. B.; Trontin, C.; Cortijo, S.; Barau, J.; Eduardo, L.; Del, V.; Loudet, O.; Colot, V.; Vincentz, M. Extensive Natural Epigenetic Variation at a De Novo Originated Gene. *PLoS Genetics* **9**, e1003437 (2013).

Slotkin, R. K.; Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* **8**, 272–285 (2007).

Soderlund, C.; Descour, A.; Kudrna, D.; Bomhoff, M.; Boyd, L.; Currie, J.; Angelova, A.; Collura, K.; Wissotski, M.; Ashley, E.; Morrow, D.; Walbot, V.; Yu, Y. Sequencing , Mapping , and Analysis of 27 , 455 Maize Full-Length cDNAs. *PLoS Genetics* **5**, e1000740 (2009).

Walley, J. W.; Sartor, R. C.; Shen, Z.; Schmitz, R. J.; Wu, K. J.; Urich, M. A.; Nery, J. R.; Smith, L. G.; Schnable, J. C.; Ecker, J. R.; Briggs, S. P. Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).

Wang, X.; Hu, L.; Wang, X.; Li, N.; Xu, C.; Gong, L.; Liu, B. DNA Methylation Affects Gene Alternative Splicing in Plants : An Example from Rice. *Molecular Plant* **9**, 305–307 (2016).

Weil, C.; Martienssen, R. Epigenetic interactions between transposons and genes : lessons from plants. *Current Opinion in Genetics and Development* **18**, 188–192 (2008).

Wessler, S. R., Bennetzen, J. L., Dawe, K. R., Jiang, N., SanMiguel, P., Freeman, B. Maize Transposable Element Data Base. Retrieved from: http://maizetedb.org/~maize/. (2015, Feb 12).

West, P. T.; Li, Q.; Ji, L.; Eichten, S. R.; Song, J.; Vaughn, M. W.; Schmitz, R. J.; Springer, N. M. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. *PLoS ONE* **9**, e105267 (2014). [PMID:25122127]

Zemach, A.; McDaniel, I. E.; Silva, P.; Zilberman, D. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science* **328**, 916–919 (2010).

Zhang, X.; Yazaki, J.; Sundaresan, A.; Cokus, S.; Chan, S. W. L.; Chen, H.; Henderson, I. R.; Shinn, P.; Pellegrini, M.; Jacobsen, S. E.; Ecker, J. R. Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell* **126**, 1189–1201 (2006). [PMID:16949657]

Zilberman, D.; Gehring, M.; Tran, R. K.; Ballinger, T.; Henikoff, S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics* **39**, 61–69 (2007). [PMID:17128275]

CHAPTER 3

A Novel Protein vs. mRNA Correlation QTL Identifies Arabidopsis Genes Involved in Translational Control

Sartor, R. C.[1], Walley, J. W.[1,2], Shen, Z.[1], Briggs, S.P.[1]
[1]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA.
[2]Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA 50011, USA. 3Plant

INTRODUCTION

The *Arabidopsis* Bay-0 x Sha Recombinant Inbred Line (**RIL**) population (Loudet, 2002) is a structured population of arabidopsis lines that was developed for the purpose of mapping phenotypic traits to genetic loci. Most phenotypic traits of interest are quantitative in nature and manifest on a continuous distribution. The genomic regions that control such traits are known as quantitative trait loci (**QTL**). The idea behind any QTL experiment is to associate regions of the genome with a certain trait that they are likely to influence. Traditionally, the trait would be something of agricultural interest such as grain yield (Veldboom, 1994) or pathogen resistance (Kump, 2011). However, more recently, molecular traits such as mRNA levels are being investigated in eQTL studies to map loci that control transcript expression (West, 2007). This study employs a similar method but uses protein abundance as the trait of interest, known as a **pQTL** analysis. We have attempted to map QTL that regulate protein abundance with a focus on the FLS2 signaling pathway. FLS2 is a well-studied Arabidopsis Leucine-Rich Repeat

Receptor-Like Kinase (**LRR-RLK**). It perceives bacterial flagella and initiates signaling that activates the plant defense response (Chinchilla et al., 2006)

We have successfully developed a multiple reaction monitoring (**MRM**) assays, also known as selective reaction monitoring (**SRM**) to quantify the abundance of each protein. An MRM assay is run on a type of mass spectrometer called a triple quadrupole **(QQQ)** (see Fig3.1). A QQQ is joined to an HPLC system. Whole proteins are extracted from a sample and enzymatically digested into peptides. The digested sample is separated in the LC and then runs into the mass spectrometer. First, the peptides are ionized. This simply adds 1 or more positive charges to each peptide turning them into ions. Once the peptides are ionized, they are able to get passed into the mass spectrometer. A QQQ is three quadrupole mass analyzers set up in a straight line with a detector at the end. After a peptide of interest is ionized, it is know as the precursor ion. The first quadrupole is used to filter out every ion except ones with the specific mass/charge ratio (m/z) of our precursor ion. Unfortunately, our samples are so complex that this one filter is not enough to isolate a particular peptide and an unacceptable amount of noise will make it through this first step. The second quadrupole is set up as a collision cell. The chamber is filled with nitrogen and an electrical voltage is applied across the cell, which excites the nitrogen causing it to collide with the peptide ions and break each one into two fragments. These resulting fragment ions are known as product ions (Fig3.2). Each peptide has a distinct and reproducible product ion abundance distributions and by monitoring

multiple product ions, the peptide of interest can be both identified and quantified. Finally the third quadrupole is used as another mass analyzer and filters out everything except a specific product ion. These multiple levels of filtering enable us to quantify specific peptides with high specificity and sensitivity.

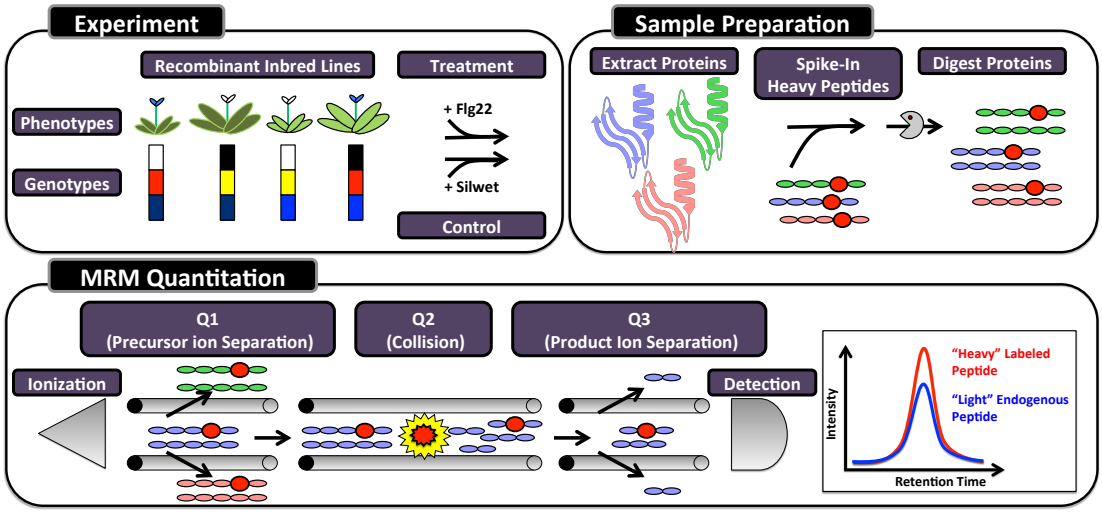


Figure 3.1 Experimental Setup. The Bay-0 x Sha RIL population was treated with the defense elicitor peptide Flg22 or a non-elicitor control. Proteins were extracted and enzymatically digested. Synthetic versions of selected peptides were produced with heavy isotopes and spiked in at known quantities. Endogenous and heavy peptides are quantified in tandem on a triple quadrupole mass spectrometer.

RESULTS

Initial Protein QTL

We have developed a quantitative mass spectrometry assay for a panel of 33 proteins (Table 3.1). 22 of these are known to be involved in plant defense. The remaining proteins are either proteins used for controls or proteins for which we attempted, unsuccessfully, to quantify modifications. Of these 33, 30 were observed in a published eQTL analysis (West et al., 2007). The remainder of the

analysis was conducted on this set of 30. By measuring the abundance of these proteins in the RIL population, we have identified significant QTLs (P-value <= 0.01) that are controlling the abundance of 8 of these proteins (Fig3.3_A).

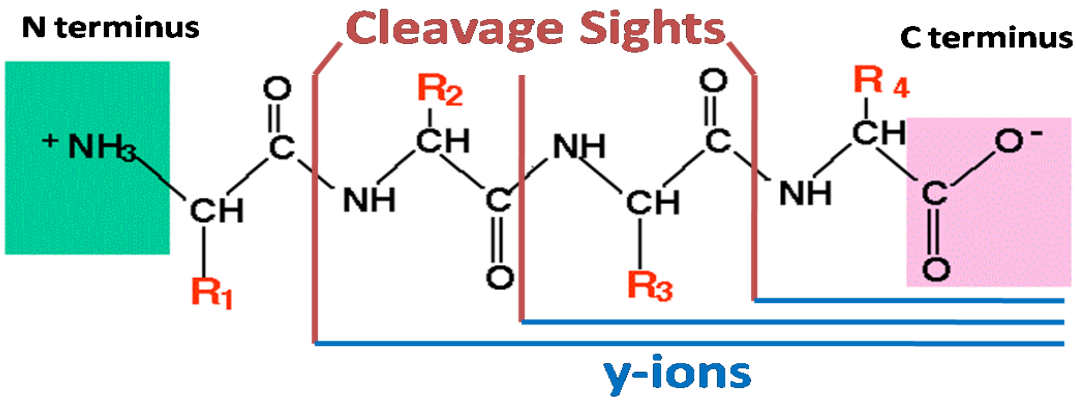


Figure 3.2 Diagram of peptide fragmentation. When a peptide is fragmented by cleavage of a single peptide bond (red lines), the C terminal side of the cleavage site is known as the y ion. This peptide has 3 possible cleavage sites, resulting in 3 possible y-ions. The ion from the N-terminal half is known as a b-ion. However, any bond along the backbone may be cleaved, resulting in 6 possible ions (a,b,c,x,y,z) for every amino acid.

Protein QTL Hotspots Overlap with eQTL Hotspots

A single locus on the top of chromosome 2 accounts for a large number of these significant associations (Fig3.3_A). Such an event is known as a QTL hotspot. This is where a single locus has a large effect on the expression of many genes. QTL Hotspots are also routinely observed in eQTL analyses where they represent one or more genes that have a large effect on the expression of many different transcripts.

Table 3.1 Protein panel for MRM assay. Grey rows indicate genes that were not observed in eQTL data.

Accession	Name	Description	Type
AT1G42970	GAPB	glyceraldehyde-3-phosphate dehydrogenase B subunit	Control
AT1G67090	RBCS1A	ribulose bisphosphate carboxylase small chain 1A	Control
AT3G18780	ACT2	actin 2	Control
ATCG00490	RBCL	ribulose-bisphosphate carboxylases	Control
AT1G20440	COR47	cold-regulated 47	Defense
AT1G32060	PRK	phosphoribulokinase	Defense
AT1G59870	PEN3	ABC-2 and Plant PDR ABC-type transporter family protein	Defense
AT1G75040	PR5	pathogenesis-related gene 5	Defense
AT2G06050	OPR3	oxophytodienoate-reductase 3	Defense
AT2G13790	SERK4	somatic embryogenesis receptor-like kinase 4	Defense
AT2G18960	AHA1	H(+)-ATPase 1	Defense
AT2G30770	CYP71A13	cytochrome P450, family 71, subfamily A, polypeptide 13	Defense
AT3G02260	BIG	auxin transport protein (BIG)	Defense
AT3G12780	PGK1	phosphoglycerate kinase 1	Defense
AT3G21220	MKK5	MAP kinase kinase 5	Defense
AT3G25070	RIN4	RPM1 interacting protein 4	Defense
AT3G26830	PAD3	Cytochrome P450 superfamily protein	Defense
AT3G45140	LOX2	lipoxygenase 2	Defense
AT3G48090	EDS1	alpha/beta-Hydrolases superfamily protein	Defense
AT3G52430	PAD4	alpha/beta-Hydrolases superfamily protein	Defense
AT4G01370	MPK4	MAP kinase 4	Defense
AT4G33430	BAK1	BRI1-associated receptor kinase	Defense
AT5G20480	EFR	EF-TU receptor	Defense
AT5G24780	VSP1	vegetative storage protein 1	Defense
AT5G42650	AOS	allene oxide synthase	Defense
AT5G47910	RBOHD	respiratory burst oxidase homologue D	Defense
AT1G48030	mtLPD1	mitochondrial lipoamide dehydrogenase 1	Modified
AT2G43750	OASB	O-acetylserine (thiol) lyase B	Modified
AT3G09630		Ribosomal protein L4/L1 family	Modified
AT4G35230	BSK1	BR-signaling kinase 1	Modified
AT4G40040		Histone superfamily protein	Modified
AT5G17310	UGP2	UDP-glucose pyrophosphorylase 2	Modified
AT5G22880	HTB2	histone B2	Modified

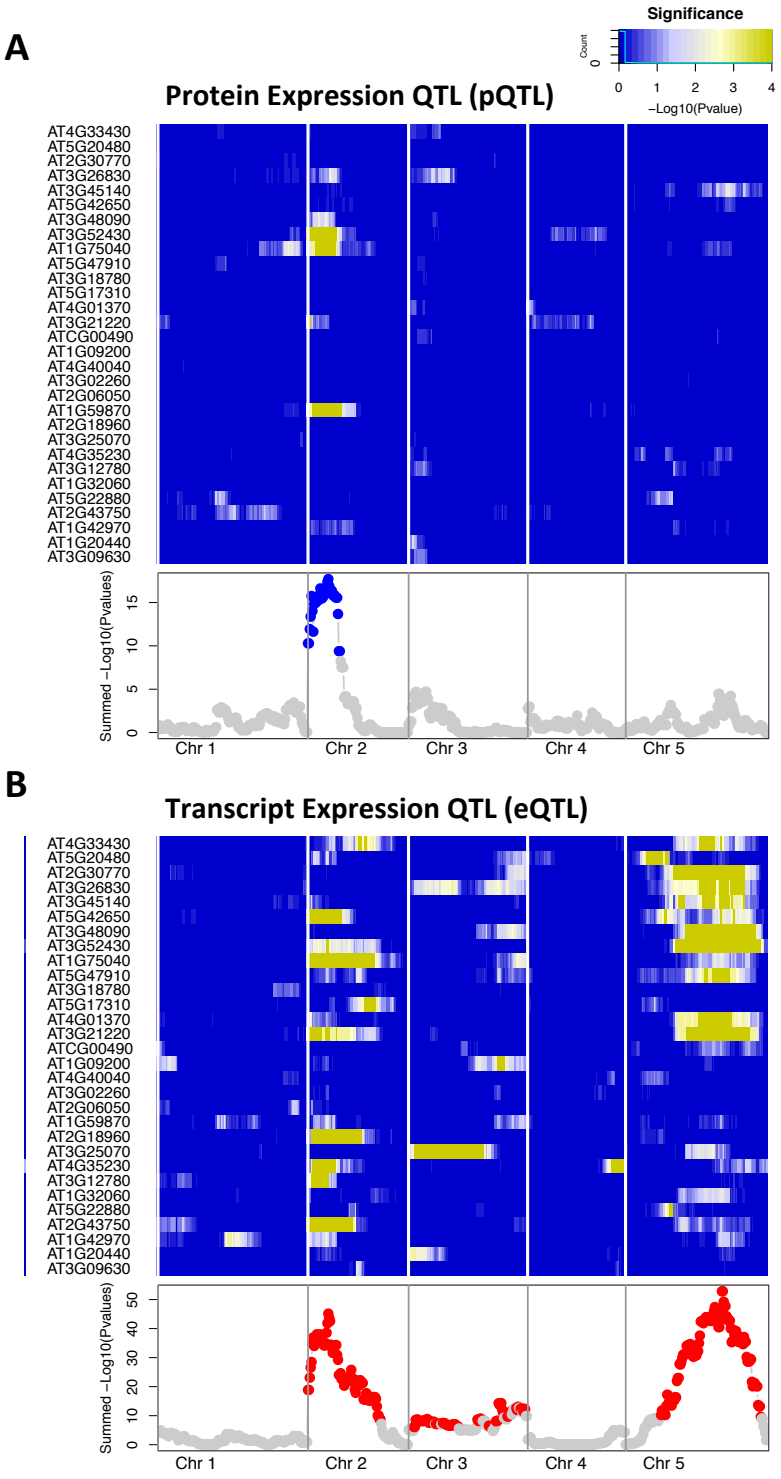


Figure 3.3 pQTL and eQTL Hotspots Overlap. (A) A novel protein QTL hotspot was identified on the top of chromosome 2. (B) However, an analogous transcript expression QTL has been identified (Data from West et al. 2007) in the same region and likely represents the underlying cause of the protein abundance variation.

This protein QTL hotspot is of interest because it likely harbors one or more regulators that act on multiple members of the FLS2 response pathway. However, after comparing to a similar eQTL study using the same population (West, 2007). We found that a strong eQTL hotspot underlies our pQTL hotspot (Fig3.3_B). This indicates that the locus is effecting transcriptional regulation and may not be specific to protein regulation.

A Novel QTL Method Based on Protein-mRNA Correlation Uncovers Candidate Genes Involved in the Regulation of Defense-Related Proteins.

By making use of the published eQTL dataset (West, 2007), the protein vs. mRNA abundance correlation can be examined across the RIL population for the 30 proteins that were represented in both sets. In addition, the population can be split into two sets at each marker to give Bay-O and a Sha subpopulations. The protein vs. mRNA spearman correlation was calculated for each sub-population at each marker and the difference in correlation scores between the sub-populations was taken (Fig3.4). This measurement represents the change in protein vs. mRNA correlation between the Bay-0 and Sha versions of any given locus. When a large change is observed, this would indicate that something at that locus is causing protein and mRNA to become more or less correlated. Another view is that it's a way to examine protein

expression while factoring out mRNA expression and any causative gene we identify is likely to have an effect down-stream of transcription.

This procedure was carried out looking specifically at the list of 20 defense related proteins and 14 significant (P-value <= 0.01) protein-transcript correlation QTL (ptcQTL) were identified in 13 different proteins (Fig3.5_A). One small hotspot was identified at the top of chromosome 5 where 2 proteins (BAK1:AT4G33430 and PAD3:AT3G26830) had the same significant ptcQTL (Fig3.5_B). One additional protein, PEN3:AT1G59870, shares this locus with P-value < 0.05. This hotspot represents the locus that likely has the strongest effect on the overall proteome and therefore is the focus of the remainder of the investigation. It should be noted that this is a different locus than the eQTL hotspot that is on the other end of chromosome 5 (Fig3.3_B).

A combined score from BAK1 and PAD3 was calculated at each marker by taking the sum of the delta spearman correlation scores for both proteins. Another permutation test was carried out on the random data using this same summed score. This results in a single combined ptcQTL trait for our two proteins and allows for a finer QTL interval to be determined. The 95% Bayes credible interval was estimated (Broman, 2003) (Fig3.5_C). This interval was found to contain 542 genes.

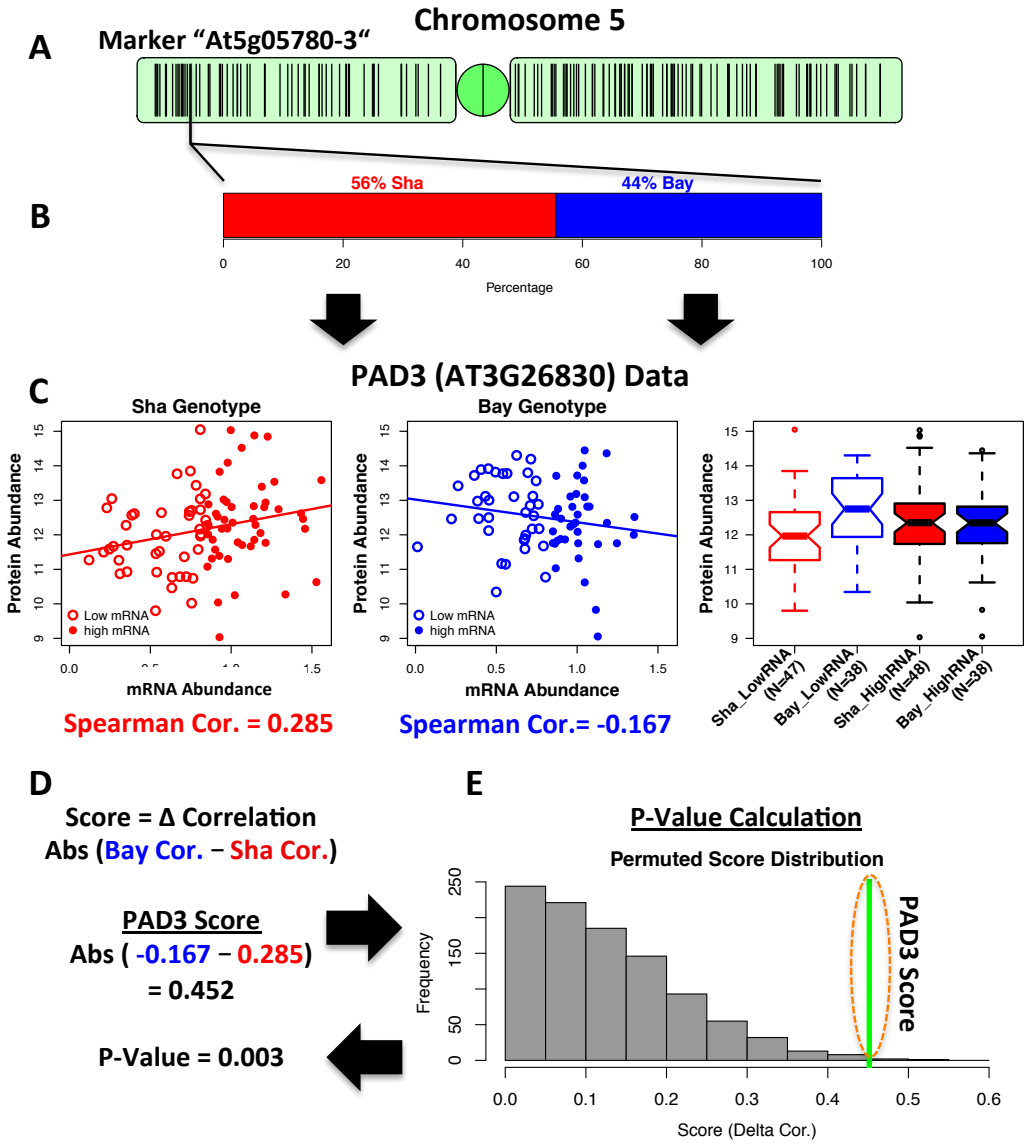


Figure 3.4 Procedure and example of protein vs. transcript QTL analysis. (A) Chromosomes are scanned and the following procedure is carried out for each marker. This example depicts a marker at the top of chromosome 5. (B) At the At5g05780-3 marker, the RIL population has 56% of the lines with the Sha version and 44% with the Bay-0 version. (C) The PAD3 gene is shown at this locus. A positive protein vs. mRNA correlation is shown for the Sha sub-population and a negative correlation for the Bay-0 sub-population. Boxplots are summarized data splitting each point at the median mRNA abundance. (D) Calculation for score statistic. (E) Example null distribution generated by repeated random permutation of Bay-0/Sha genotypes at the specific marker.

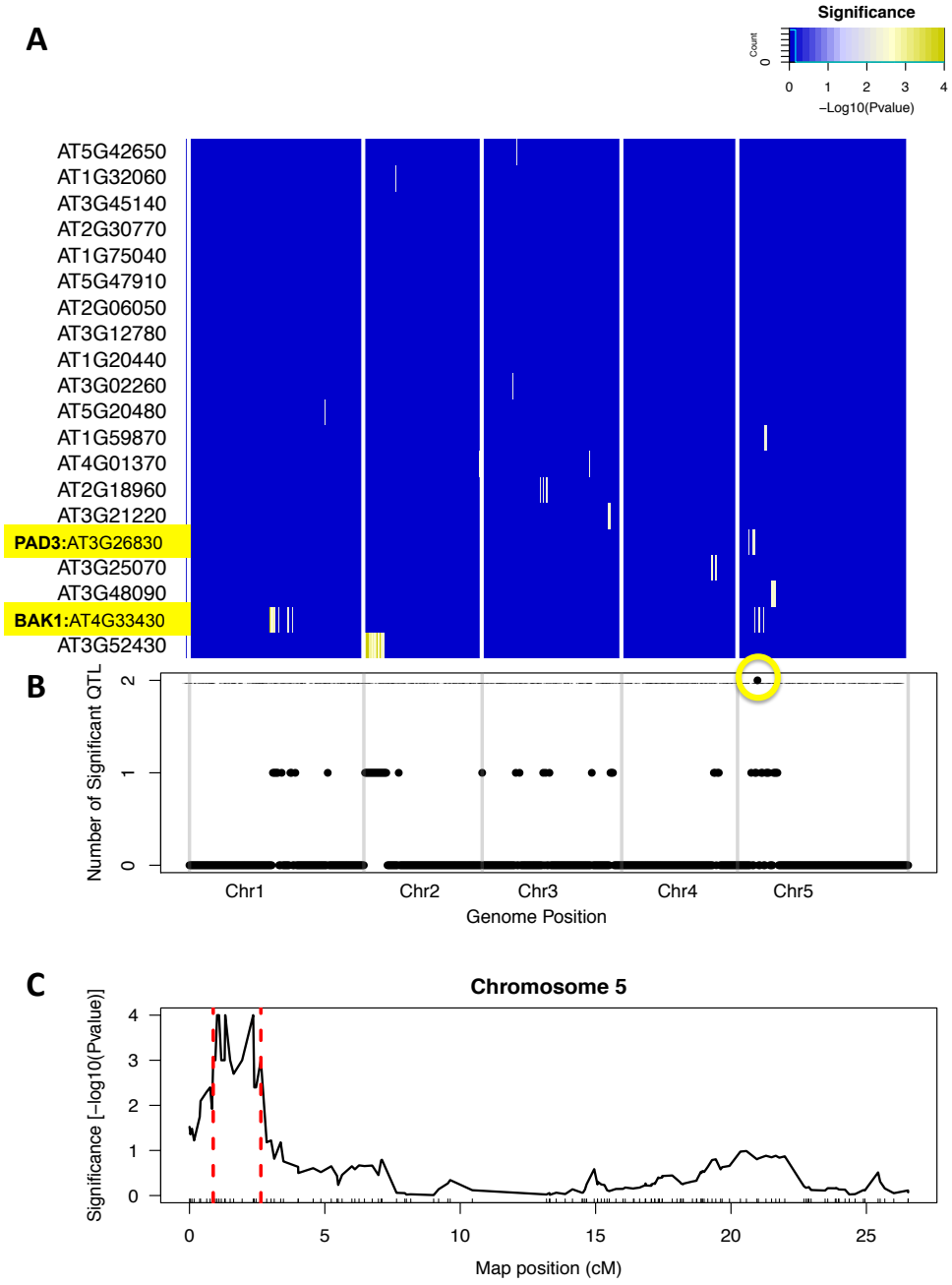


Figure 3.5 Significant protein-transcript correlation QTL (ptcQTL). (A) A heatmap displaying only significant (P-value <= 0.01) ptcQTL. Each column is a marker arrayed along 5 chromosomes and each row is a genes whose protein-transcript correlation trait was examined. (B) The number of significant ptcQTL was summed for each marker, showing one with a ptcQTL for multiple genes, indicating a ptcQTL hotspot (yellow circle). (C) The chromosome 5 hotspot is shown. The y-axis is a significance score calculated from combining the two significant ptcQTL hits at the hotspot locus (BAK1 and PAD3). The 95% confidence interval of this locus is shown with red vertical lines.

Utilization of eQTL Transcript Abundance Data To Prioritize Potential Causative Genes

Within any QTL that was mapped using an RIL population, the number of potential causative genes in the interval is very large sometimes up to several thousand genes. In most cases, the most difficult challenge comes in identifying causative genes. To do this, we turned back to the eQTL transcript expression data. There are two general probable mechanisms by which natural genomic variation can have an effect on a given trait; i) One of the alleles causes a change in protein sequence of the causative gene which results in a protein with decreased or increased effectiveness related to the trait. ii) One of the alleles causes a change in transcript expression of the causative gene that negatively or positively impacts the trait of interest. Both of these mechanisms can be queried given genomic sequences of the RIL parent lines or transcript expression from the RIL population respectively. The Bay-0 and Sha re-sequencing data was used to determine which genes have potential non-synonymous single nucleotide polymorphisms (SNPs). 253 genes in the interval are predicted to have amino acid polymorphisms. While this is a significant decrease, it still leaves us with too many to attempt to validate. Next we attempted to look for extreme transcript variation between the Bay-0 and Sha alleles for each gene in the interval. This is analogous to mapping cis-eQTL within the interval. These are expression QTL that have an

effect on a transcript whose gene is within the QTL interval. Mechanistically, this is simply where a genetic polymorphism is effecting the transcript abundance for a gene that is nearby. A common example would be a polymorphism in the promoter of a gene that effects transcription of that same gene. A standard eQTL was carried out and all cis-eQTL were identified. A LOD score cutoff of 45 was used to discover 11 cis eQTL with extreme effect within the interval (Fig3.6_A). Each of these genes has an extreme expression difference between RILs with the Bay-0 and Sha alleles. These 11 were further refined by manually examining the known or predicted gene function for each. 3 were chosen that have a potential impact on protein life-cycle (Fig3.6_B). AT5G05230 is a RING/U-box domain containing protein that likely functions as an E3 ubiquitin ligase and is presumably involved in ubiquitin mediated protein degradation. This protein will be referred to as the hot spot RING protein (HsRING). AT5G05750 is a DNAJ domain containing heat shock protein likely involved in promoting correct protein folding and stability. This protein will be referred to as the hot spot DNAJ protein (HsDNAJ). AT5G05760 (SYP31) is a syntaxin involved in ER to Golgi vesicle trafficking (Chatre et al., 2005 ; Bubeck et al., 2008).

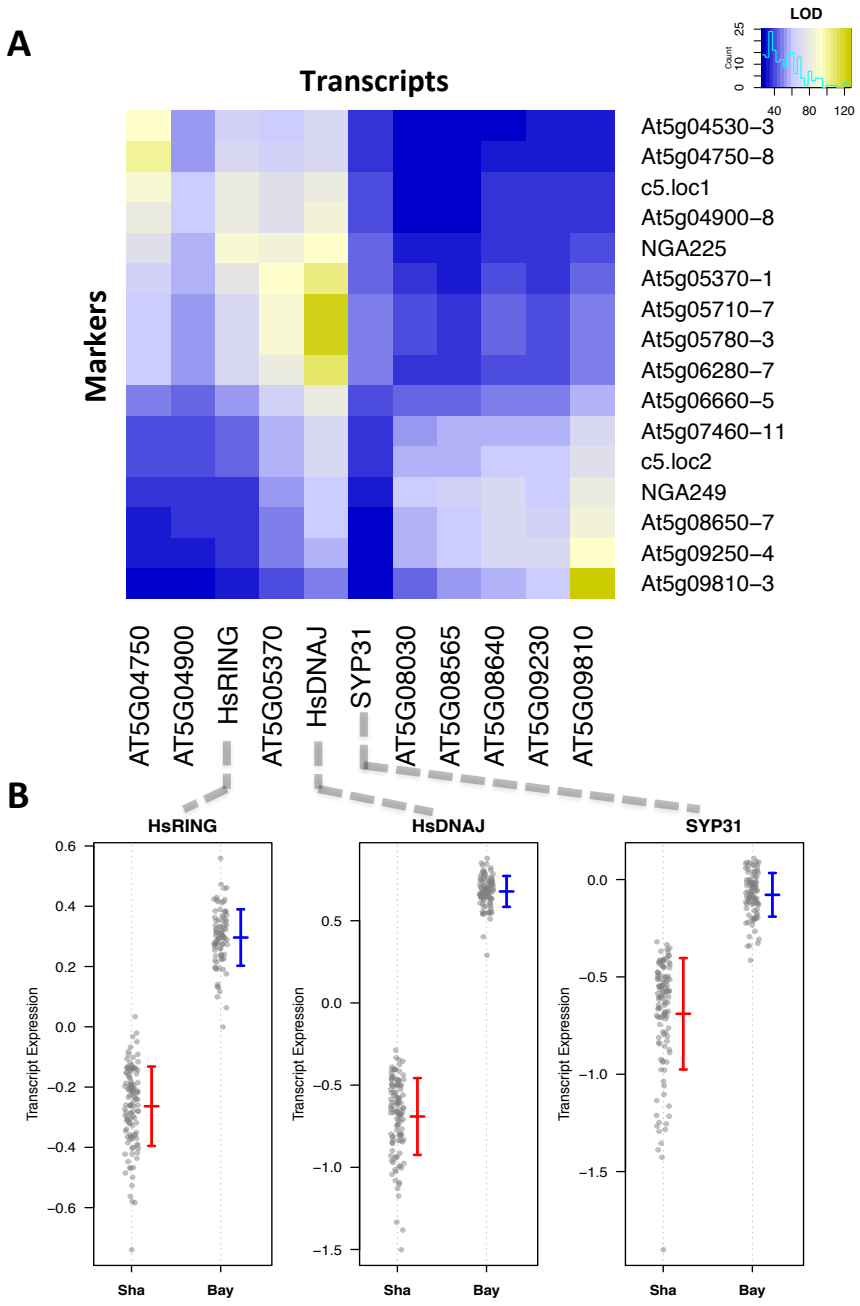


Figure 3.6 RIL Transcript expression is used to select potential causative genes. (A) Heatmap of eQTL LOD scores for 11 genes in the ptcQTL chromosome 5 hotspot interval with extreme cis-eQTL (LOD > 45). These genes represent cases where a very large difference exists in transcript level between all RILs containing the Bay-0 and Sha versions of a given locus. (B) Three of these genes have a predicted function with high likelihood of effecting protein life-cycle. These three were chosen for validation. Scatter plots show relative expression levels of each gene in all RILs with the Sha and Bay versions of the hotspot locus. Bars are showing standard deviation.

Validation of Candidate Genes

For validation we chose to obtain T-DNA insertion lines (O'Malley, 2015) for each of our candidate genes and query the proteomes of these lines. For each of these candidate genes, we see a significant change in transcript expression between the Bay-0 and Sha genotypes. In each case, we find that the Col-0 genotype expresses the transcript at approximately the same level as Bay-0. Therefore we refer to the Bay-0 expression as "wildtype" and say that in the Sha genotype, all three of our candidate genes are significantly knocked down or possibly knocked out. Since the Arabidopsis T-DNA collection uses the Col-0 background, a knock out or significantly knocked down version of any or our candidate genes would mimic Sha. Based on our analysis, we would expect the mutant line of our causative gene to have decreased protein expression for PAD3 and BAK1. However, this specific change in protein expression is only expected for lines expressing relatively low mRNA. We chose this locus because we observe 2 out of 20 defense genes modulated specifically at the protein level. We expect that many more genes may be influenced by this locus and therefore we would also consider a candidate gene successful if we observe a significantly large modulation of the proteome while the transcriptome remains relatively constant.

We are analyzing two separate T-DNA inserts for both HsRING and SYP31, and 3 inserts for HsDNAJ (Fig3.7). Currently, we have no results for HsRING. We have obtains homozygous lines for all 3 hsdnaj inserts and both

syp31 inserts. Quantitative real-time PCR (**qPCR**) was carried out on each line to determine if the respective transcript was knocked out or knocked down. None of the hsdnaj insertions show decreased HsDNAJ expression, but hsdnaj_3 shows significant over-expression (Fig3.7_B). This can results from a constitutive promoter that is present in the T-DNA insert. It is unknown to us if this transcript results in a viable protein as HsDNAJ remains undetected at the protein level. Both of the syp31 insertion lines resulted in a significant knock down (Fig3.7_C). We went on to query the proteomes of each of these syp31 lines as well as the hsdnaj_3 overexpressor.

The ptcQTL hotspot locus was discovered by looking specifically at proteins involved in defense and the FLS2 pathway. Two out of twenty proteins share this ptcQTL, with an additional protein at P-value < 0.05. It is possible that this locus has an effect specific to or enriched for defense related proteins. Or it is possible that it is not related to defense and has a general effect on some subset of the proteome. In order to address both possibilities, we grew mutant and wildtype (Col-0) seedlings in both mock treatment (H2O) and flg22 treatment to elicit a defense response. We harvested seedlings 3 hours after treatment and quantified proteins using nano-LC tandem mass spectrometry.

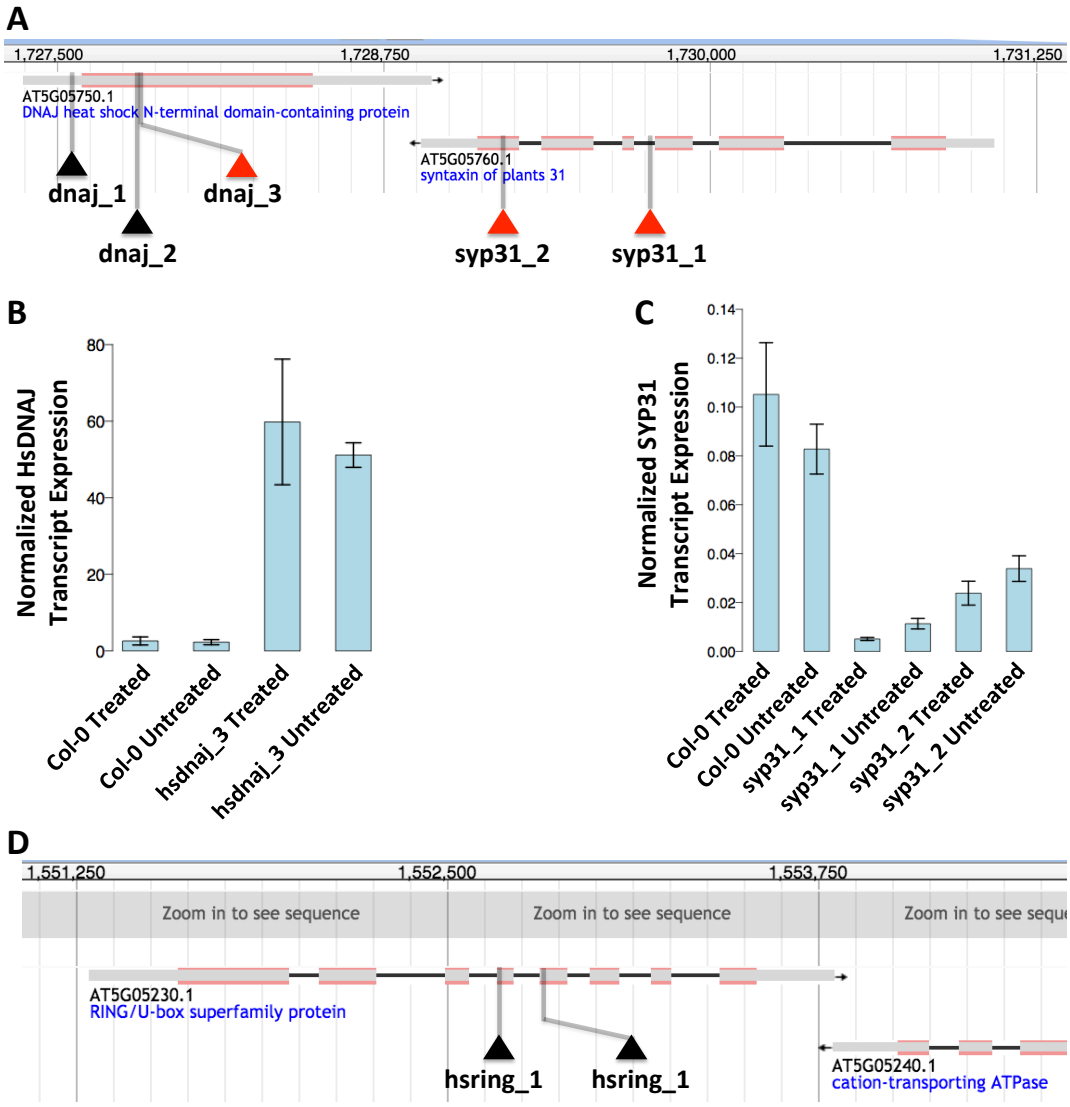


Figure 3.7 Schematic of T-DNA insertion lines. (A) Three separate insertions were analyzed for HsDNAJ and two separate insertions for SYP31. Red triangles indicate lines that were shown to have significant alteration of the target gene transcript levels and were assayed by mass spectrometry. (B) hsdnaj_3 was the only line that showed significant expression difference compared to wild type and it appears to be an overexpressor. (C) Both syp31 lines show significant decreases in SYP31 expression. (D) Two separate insertions are being examined for HsRING.

Of the three mutant lines that were assayed, all were tested for significant protein differential expression (DE) using a linear model based DE

analysis (Ritchie et al., 2016 ; Phipson et al., 2016). Two tests were carried out on each data set for all genes: (i) between mutant and wildtype and (ii) an interaction effect between the mutation and defense elicitor treatment. This interaction effect would indicate genes that show a different response to treatment in mutant compared to wildtype plants. An adjusted P-value <= 0.05 was deemed significant. The HsDNAJ overexpressor (hsdnaj_3) showed no significant differential expression of any proteins. The SYP31 knockdown, syp31_1 showed only one gene differentially expressed between mutant and wildtype. However, syp31_2 shows a massive impact on the overall proteome. 6935 proteins had sufficient data to calculate DE. 1203 of these show differential protein abundance between syp31_2 and Col-0 (Fig3.8_A), with similar numbers of “Up” and “Down” genes. Only one gene (AT1G29700) shows significant interaction for flg22 response between syp31_2 mutant and Col-0. We are currently in the process of running RNA-seq samples for this mutant in order to determine if these expression changes are specific to the proteome (as we hypothesize) or if they are underpinned by the transcriptome.

Next we examined the set of DE proteins in the syp31_2 line. We see very little categorical enrichment for proteins that are down-regulated in the mutant (Fig3.8_A). However, we see that the up-regulated proteins show strong enrichment for eukaryotic ribosomal proteins (P-value = 2.9e-17), where 59 of the 77 (77%) DE ribosomal proteins show significant up-regulation (Fig3.8_B).

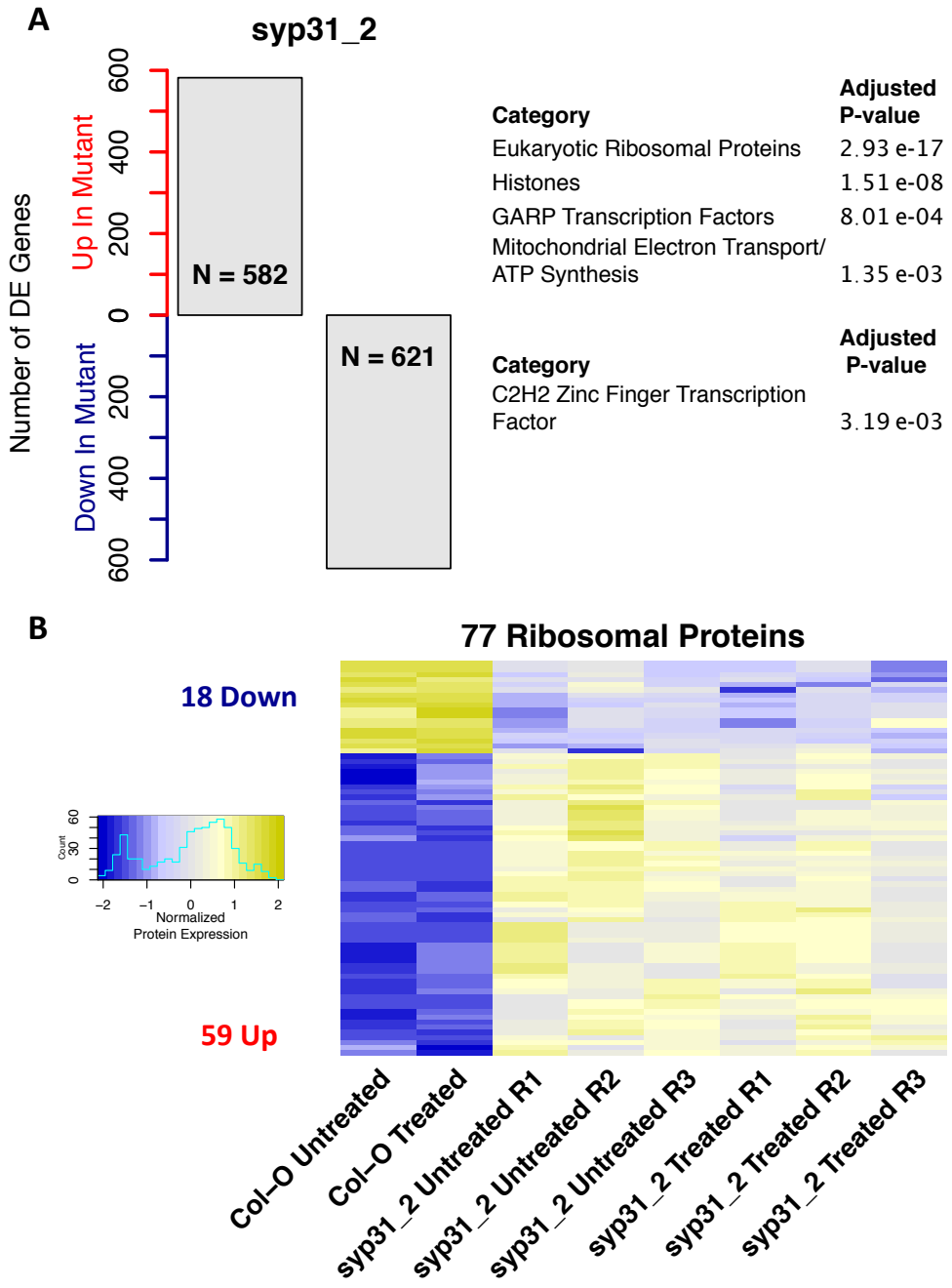


Figure 3.8 Results for SYP31 validation. (A) The syp31_2 T-DNA insertion line shows a large number (1203) of differentially expressed proteins with similar numbers of up and down regulated proteins. Down regulated proteins show very little categorical enrichment. However, Up regulated proteins show strong enrichment for ribosomal proteins. (B) 77 ribosomal proteins were differentially expressed between Col-0 and syp31_2, with a majority (77%) being up regulated.

Conclusion

The above work describes the development and implementation of a novel QTL-based analysis that specifically identifies loci involved in modulating protein abundance while changes in mRNA abundance are normalized out. This method was used to identify 14 ptcQTL that have an impact on the protein abundance of 13 different proteins. This method makes it possible carry out pQTL analyses and discover loci that are affecting post-transcriptional steps. Most large scale data analysis has a very transcription-oriented focus and for good reason. We show here that even at the QTL level, we see transcription playing a major role in determining protein abundance. However, it is not the whole story and much regulation happens post-transcriptionally. This method allows one to examine such processes.

Here, we have demonstrated the use of published eQTL data to aid in the identification of causative loci through a transcript-based approach. We have prioritized 3 candidate genes for validation. At present we have only been able to test one of these genes, SYP31. This prioritization approach has very broad application. If shown to be effective, it could be used on any mapped trait in any population that has existing eQTL data to significantly narrow down the number of potential causative genes. Hundreds of QTL analyses exist in dozens of organisms. The vast majority only report a locus but were unable to identify the causative gene because testing hundreds or thousands is usually far beyond the ability of most projects. In studies done for

populations with published eQTL data, this method would make it possible to prioritize very small numbers of potential causative genes with no extra data or experiments required.

We have shown that knocking down SYP31 via a T-DNA insert has a dramatic effect on the plant proteome with over 1,200 proteins being differentially expressed in the syp31_2 mutant compared to wildtype Col-0. The most strongly effected group of proteins is ribosomal proteins where we see significant enrichment of ribosomal proteins in the group that is up-regulated in the mutant.

SYP31 is a syntaxin protein that is involved in ER to Golgi vesicle trafficking (Chatre et al., 2005 ; Bubeck et al., 2008). Syntaxins are required to form a SNARE complex that initiates membrane fusions. Transient over-expression of arabidopsis SYP31 in tobacco protoplast inhibits ER to Golgi vesicle trafficking (Bubeck et al., 2008). We cannot say what the effect on ER to Golgi trafficking is in the syp31_2 mutant but we speculate that this primary step in cellular protein trafficking is the cause of the proteome modulation that we observe. However, it is also unclear how this relates to ribosomal proteins. Additional proteome modulation may also result from altered ribosomal protein composition in the cell that may cause an imbalance in protein synthesis.

FUTURE WORK

We are in the process of obtaining homozygous lines for two HsRING mutants in order to determine the potential for this gene to be causative for the ptcQTL hotspot.

We are also going to profile transcript abundance (RNA-seq) in the syp31_2 mutant as well as any of the HsRING mutant lines that show proteomic modulation. Based on the ptcQTL results, we hypothesize that a causative mutation will show little or no impact on the transcriptome and that the severe proteomic modulation that we observe is a result of post-transcriptional control that specifically effects proteins levels.

MATERIALS AND METHODS

RIL Plant Growth

The Bay x Sha Recombinant Inbred Line Population (Loudet et al., 2002) consist of 211 lines + the two parent. Seeds were germinated in 2' x 2' pots in soil. ~4 seeds were placed in each pot and subsequently culled after germination to result in 2 plants / pot. Plants were grown under short days until the first RIL lines began to bolt.

RIL Treatment and Harvesting

Plant surfaces were sprayed with an aerosol solution of 10uM flg22 elicitor peptide in 0.01% silwet surfactant (treatment) or just 0.01% silwet

(control). 20 minutes after treatment, all areal parts of the plant was harvested and immediately frozen in liquid nitrogen. Two plants were harvested per biological replicate. One replicate was taken for treatment and one for control.

MRM Assay Development

The process of MRM assay development, involves determining the precursor and product m/z settings that yield the highest amount of detectable product ions for each peptide. The combination of these two numbers is known as a transition. This is followed by a procedure to optimize the energy applied to the collision cell (Q2). For this project, we started with 174 peptides total and so we chose to develop in a more high-throughput way. We computationally predict the 5 best possible transitions for each peptide. This method makes several assumptions, only considering precursors with +2 or +3 charge state and only y-ions (i.e. ions that result from C-terminal side of a fragmentation where the peptide backbone is split along the peptide bond). After examining 5 possible y-ion transitions for each heavy peptide, 34 unmodified peptides were eliminated that had low or no signal, leaving 139. For each remaining peptide, assay methods were created by choosing the 3-4 transitions with the highest signal then finding the optimum collision energy for each transition. Using these methods, we have altered the m/z for each transition to reflect the mass of the naturally occurring isotopes and obtain methods for the endogenous (light) peptides.

MRM Sample Prep and Mass Spectrometry Assay

Frozen samples were course ground manually inside a 2mL vial. ~ 2cc of crushed frozen tissue was transferred to a new 2mL vial. Two stainless steel balls were added to the vial. Tubes were place in cartridge designed specifically for 2mL tubes that fit a Qiagen tissue homogenizer. Keeping everything frozen, samples were homogenized at 25Hz for 1 minute. Cartridges were rotated to expose all tubes to the maximum possible wavelength of the shaker and again homogenized at 25Hz for 1 minute.

Samples were washed 3 times with MeOH containing 0.2mM Na3VO4 at -20 °C and then twice with Acetone at -20 °C then vacuum-evaporated until dry.

Samples were re-suspended in 1 mL Extraction Buffer (0.1% SDS in 50 mM Hepes / 1mM EDTA) plus 20 µL 50x Phosphatase inhibitor (125mM NaF, 12.5 mM NaVO4, 12.5 mM Na4P2O7 [Sodium Pyrophosphate tetrabasic decahydrate] and 12.5 mM C3H7Na2O6P * 5H2O [B-Glycerophosphate, Disodium Salt, Pentahydrate]), plus 10 uL of 100x HDAc Inhibitor (100uM TSA [Tricostatin A], 1M Nicotinomide in 50 mM Hepes) and 10 uL of 100mM TCEP.

Samples were incubated at 94 °C for 20 minutes to denature and extract proteins. The pellet was then centrifuged down at max speed for 2 minutes and supernatant was taken to be used for BCA assay.

A BCA assay was then carried out to determine the precise total protein concentration in each sample. Using Thermo Scientific BCA assay kit, assays were performed on 96-well plates with 6 technical replicates of each sample

and 4 replicates of BCA standards (125, 250, 500, 750, 1000,1200, 1500 and 2000 ug/mL).

Tissue pellets were then re-suspended by pipetting and a volume containing 250 ug total protein (as determined by BCA) was transferred to a new 2 mL vial. This was then brought to 1.0 mL with Extraction buffer.

Heavy peptides were synthesized by Thermo Scientific. Optimum concentrations of heavy peptides were determined by attempting to match the heavy peak intensity of each transition to that of the endogenous peak. 6.5uL of the same standard mix of heavy peptides was then spiked into the sample.

Sample pH should be between 7 and 8. Next, 5 uL of 0.5 mg/mL Trypsin was added and peptide digest was allowed to proceed over night in a 37 °C shaker.

Samples were then treated with 5 uL of 500mM iodoacetamide and incubated at 37 °C in the dark to reduce cysteines.

Samples were centrifuged at 25,000g for 5 minutes at room temperature and 0.9uL of supernatant was transferred to a new tube. Pellet was then washed with 0.6 mL Hepes, centrifuged at 25,000g and supernatant was added to previous supernatant. The palette is then discarded.

An additional trypsin digest is carried out by adding 2.5 uL of 0.5 mg/ml trypsin and incubating at 37 °C for 4 hours.

Samples were then acidified by adding 7.5 uL of 100% formic acid and centrifuged at 25,000 for 10 minutes.

Peptides were cleaned up using Oasis MCX LP Extraction kit (1cc size). Columns are rinsed with pure MeOH followed by 2 rinses with pure H2O. Samples are then added to the column and washed two times with pure H2O then with 3mL solution of 80% Acetonitrile / 0.05% formic acid. Peptides are eluted from column using 1.5 mL of 60%IPA/500mM NH_4HCO_3.

Samples are then vacuum-centrifuged overnight at 4°C to dry and then re-suspended in 40uL 2% formic acid.

Finally, any precipitate is removed by centrifuging through a 2.2uM filter at 25,000g for 10 minutes.

8 uL of sample was injected into a 10-inch SCX C-18 nano Liquid Chromatography column packed in-house. This column was eluted with a 3 hour acetonitrile gradient from 10% - 80% that fed directly into the electro-spray source of an Agilent 6410 triple quadrupole mass spectrometer.

MRM Quantification

3 or 4 transitions were measured for each peptide with 1 or two peptides per protein. For each transition, both a heavy (control) and light (endogenous) transition is measured. Quantification of the data was carried out using the Skyline software package (McLean et al., 2010) all peaks were manually annotated for integration. The first round of quality control was carried out at this step and if the heavy peptide peak was not obvious with all precursor peaks aligned at the same retention time, the measurement for that peptide in that run was discarded and not carried through to the next steps.

Next, the dot product for the ratio of heavy to light transition peak areas was used to filter low quality measurements. Any peak with a dot product below 0.9 was discarded. Each peptide was than quantified as the ratio of the total area for all transitions of the light peptide to the total area of all transitions of the heavy peptide. For proteins with multiple peptides, the mean was used for protein quantification.

pQTL and eQTL analysis

The R package R/qtl (Broman et al., 2003) was used to map protein QTL. Each protein quantity measured in the MRM assay across the RIL population was taken as a separate traits. A high density genotype map of the Bay x Sha RIL population was obtained from: http://elp.ucdavis.edu/data/analysis/211_RILs_SFP_map/211_RILs_SFP_map.html and described in (West et al., 2006). Data was imported from .csv format using the read.cross() function with crosstype = “bc”. In order to model an RIL population, the convert2riself() and then the calc.genoprob() functions were run. Haley-Knott regression was used to map QTL with the scanone() function using method = “hk”. Finally, P-values were calculated by generating a null distribution of 1000 random permutations of the input data set for each trait (protein) using the scanone() function with n.perm = 1000.

ptcQTL analysis

This analysis operates by examining each marker separately. At each marker, the genotype for all the RILs is either Bay or Sha. Therefore the RIL

population is split between two groups (Bay and Sha). The spearman correlation of protein vs. mRNA is then calculated for each group separately and the absolute value of the difference in correlation scores is used as the statistic. A large number would indicate that there is a change in protein vs. mRNA correlation between the groups and that this regulatory alteration is linked to the marker in question.

Significance of the delta correlation statistic is then determined by randomly permuting the genotype assignments at each marker 1000 times. The same statistic is calculated for each of these 1000 permutations to give a null distribution. A P-value is than calculated by taking the proportion of random statistics that are greater than the actual statistic.

T-DNA Plant Growth

T-DNA lines indicated in figure 07 were obtained from the Arabidopsis Biological Resource Center (ABRC): https://abrc.osu.edu/ . Next, 24 or 48 plants from All lines were planted in soil. For SALK “homozygous” lines, 6 plants were randomly selected and genotyped as described by (O’Malley et al., 2015) In order to confirm homozygosity. For all other lines, all plants were genotyped and any plant not homozygous for the T-DNA insertion was removed. All lines were bulked for one generation in order to obtain enough seeds.

T-DNA Flg22 treatment and harvest

All T-DNA and Col-0 lines were surface sterilized by first washing with 70% EtOH, followed by 50% Bleach and 0.02% Triton X-100 for 10 minutes. Seeds were germinated and grown in liquid ½ Murashige and Skoog (MS) media (2.2g MS/ Liter) with 0.5% sucrose. Plants were grown in 50 mL sterile culture tubes with 35 mL media. In a shaker at 80 rpm inside a growth chamber. Plants were stratified at 4°C for 2 days and then alowed to grow for ~10 days. When the second set of true leave was ~ 50% developed, seedlings were treated with flg22 peptide disolved in H2O or with H2O control by adding it directly to the liquid media. Flg22 treatment was a final concentration of 1uM in the culture media. Plants were harvested after 3 hours by draining liquid, briefly wraping in paper towel to dry excess liquid and imediately feezing in liquid nitrogen.

iTRAQ Quantitative Proteomics

For global protein abundance profiling 50 µg of peptides from each sample was treated with one tube of 8-plex iTRAQ reagent for 2 hours at room temperature. Labeled samples were dried down in a vacuum concentrator and re-suspended in H2O. Samples tagged with the 8 different iTRAQ reagents were pooled together. Using the above protocol we obtained higher than 95% iTRAQ labeling efficiency.

An Agilent 1200 HPLC system (Agilent Technologies) delivered a flow rate of 600 nL min^{-1} to a 3-phase capillary chromatography column through a splitter. Using a custom pressure cell, 5 µm Zorbax SB-C18 (Agilent) was

packed into fused silica capillary tubing (250 µm ID, 360 µm OD, 30 cm long) to form the first dimension reverse phase column (RP1). A 5 cm long strong cation exchange (SCX) column packed with 5 µm PolySulfoethyl (PolyLC) was connected to RP1 using a zero dead volume 1 µm filter (Upchurch, M548) attached to the exit of the RP1 column. A fused silica capillary (250 µm ID, 360 µm OD, 20 cm long) packed with 2.5 µM C18 (Waters) was connected to SCX as the analytical column (RP2). The electrospray tip of the fused silica tubing was pulled to a sharp tip with the inner diameter smaller than 1 µm using a laser puller (Sutter P-2000). The peptide mixtures were loaded onto the RP1 column using the custom pressure cell. A new set of columns was used for each LC-MS/MS analysis.

Peptides were first eluted from the RP1 column to the SCX column using a 0 to 80% acetonitrile gradient for 150 minutes. The peptides were then fractionated by the SCX column using a series of 19 step salt gradients for non-modified iTRAQ profiling (0, 30, 40, 50, 55, 60, 65, 70, 75, 80, 85, 90, 100, 110, 120, 130, 150, 200, 1000 mM ammonium acetate) followed by high-resolution reverse phase separation using an acetonitrile gradient of 0 to 80% for 120 minutes.

Spectra were acquired using a Thermo Q-Exactive-HF mass spectrometer (Thermo Electron Corporation, San Jose, CA) employing automated, data-dependent acquisition. The mass spectrometer was operated in positive ion mode with a source temperature of 275^{o}C, spray voltage

3,000V, and S-lens RF level of 70. A full MS - data dependent MS2 scan method was used to acquire the data. Full MS scan parameters are: resolution = 60,000; AGC target = 1e5; max IT = 10ms; scan range = 400-2000. Data dependent MS2 scan parameters are: resolution = 15,000; AGC target = 1e5; max IT = 50ms; TopN = 20; isolation windows = 3.0Da; Fixed first mass = 110Da; NCE = 28; charge exclusion = unassigned, 1, 5-8, >8; dynamic exclusion =10 seconds.

The raw data were extracted and searched using Spectrum Mill vB.06 (Agilent). MS/MS spectra with a sequence tag length of 1 or less were considered to be poor spectra and were discarded. The remaining MS/MS spectra were searched against the TAIR10 protein database. The enzyme parameter was limited to fully tryptic peptides with a maximum miscleavage of 1. All other search parameters were set to default settings of Spectrum Mill (carbamidomethylation of cysteines, iTRAQ modification, or K-Ac). A concatenated forward-reverse database was constructed to calculate the in-situ false discovery rate (FDR). There are 70,800 protein sequences in the database (35,386 TAIR10 proteins, 35,386 decoy sequences, and 28 common contaminant proteins such as trypsin, Lys-C, keratins, etc.) Cutoff scores were dynamically assigned to each dataset to maintain the false discovery rate less than 0.1% at the peptide level. Proteins that share common peptides were grouped to address the database redundancy issue. The proteins within the same group shared the same set or subset of unique peptides.

iTRAQ intensities were calculated by summing the peptide iTRAQ intensities from each protein group. Peptides shared among different protein groups were removed before quantitation. Isotope impurities of iTRAQ reagents were corrected using correction factors provided by the manufacturer (Applied Biosystems). Median normalization was performed to normalize the protein iTRAQ reporter intensities in which the log ratios between different iTRAQ tags (114/113, 115/113 ...) are adjusted globally such that the median log ratio is zero. Quantitative analysis was performed on the normalized protein iTRAQ intensities. Protein ratios between the mock and each treatment were calculated by taking the ratios of the total iTRAQ intensities from the corresponding iTRAQ reporter. Protein ratios were then log2 converted. Proteins that significantly changed in each treatment, relative to mock, were determined using t-tests (two tailed, paired). Proteins with more than 1.5 fold change and P-value less than 0.05 were considered significantly changed in abundance.

Differential Protein Expression

The R Limma package (Ritchie et al., 2015 : Phipson et al., 2016) was used to determine differential protein abundance. A design matrix was built using categories of every Genotype-Treatment combination. For each T-DNA insertion line and therefore each iTRAQ run, a separate analysis was done. To find proteins with differential abundance in mutant vs. wildtype, a contrast of (Mutant_Treated + Mutant_Untreated) – (Col-0_Treated + Col-0_Untreated)

was used. To find proteins where and interaction between genotype and treatment was observed, a contrast of (Mutant_Treated - Mutant_Untreated) – (Col-0_Treated - Col-0_Untreated) was used. Genes were determined significant using an adjusted P-value cutoff of 0.05.

ACKNOWLEDGMENTS

Chapter 3, in full is currently being prepared for submission for publication of the material. Sartor, R. C., Walley J. W., Shen Z., Briggs, S. P. The dissertation author was primary investigator and first author of this work.

REFERENCES

Broman, K. W.; Wu, H.; Churchill, G. A. R / qtl : QTL mapping in experimental crosses. **19**, 889–890 (2003).

Bubeck, J.; Scheuring, D.; Hummel, E.; Langhans, M.; Viotti, C.; Foresti, O.; Jurgen, D.; Banfield, D. K.; Robinson, D. G. The Syntaxins SYP31 and SYP81 Control ER – Golgi Trafficking in the Plant Secretory Pathway. *Traffic* **9**, 1629–1652 (2008).

Chatre, L.; Brandizzi, F.; Hocquellet, A.; Hawes, C.; Moreau, P. Sec22 and Memb11 Are v-SNAREs of the Anterograde Endoplasmic Reticulum-Golgi Pathway in Tobacco Leaf Epidermal Cells 1. *Plant Physiology* **139**, 1244–1254 (2005).

Chinchilla, D.; Bauer, Z.; Regenass, M.; Boller, T.; Felix, G. The Arabidopsis Receptor Kinase FLS2 Binds flg22 and Determines the Specificity of Flagellin Perception. *The Plant Cell* **18**, 465–476 (2006).

Kump, K. L.; Bradbury, P. J.; Wisser, R. J.; Buckler, E. S.; Belcher, A. R.; Oropeza-rosas, M. A.; Zwonitzer, J. C.; Kresovich, S.; Mcmullen, M. D.; Ware, D.; Balint-kurti, P. J.; Holland, J. B. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics* **43**, 163–169 (2011).

Loudet, O.; Chaillou, S.; Camilleri, C.; Bouchez, D.; Daniel-Vedele, F. Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics* **104**, 1173–1184 (2002).

Maclean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; Maccoss, M. J. Skyline : an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

O'Malley, R. C.; Barragan, C. C.; Ecker, J. R. HHMI Author Manuscript A User ' s Guide to the Arabidopsis T-DNA Insertional Mutant Collections. *Methods Mol Biol.* **1284**, 323–342 (2015).

Phipson, B.; Lee, S.; Majewski, I.; Alexander, W. S.; Smyth, G. HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT. *The Annals of Applied Statistics* **10**, 946–963 (2016).

Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, (2015).

Veldboom, L. R.; Lee, M. Molecular-marker-facilitated studies of morphological traits in maize . I1 : Determination of QTLs for grain yield and yield components. *Theor Appl Genet* **89**, 451–458 (1994).

West, M. A. L.; Kim, K.; Kliebenstein, D. J.; Leeuwen, H. Van; Michelmore, R. W.; Doerge, R. W.; Clair, D. A. S. Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis. *Genetics* **175**, 1441–1450 (2007).

West, M. A. L.; Van Leeuwen, H.; Kozik, A.; Kliebenstein, D. J.; Doerge, R. W.; St. Clair, D. A.; Michelmore, R. W. High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Research* **16**, 787–795 (2006).

Zipfel, C.; Kunze, G.; Chinchilla, D.; Caniard, A.; Jones, J. D. G.; Boller, T.; Felix, G. Perception of the Bacterial PAMP EF-Tu by the Receptor EFR Restricts Agrobacterium -Mediated Transformation. *Cell* **125**, 749–760 (2006).