# Title

Integrating Gene Co-Expression Networks with GWAS to Detect Causal Genes Driving Elemental Accumulation in Maize

# Authors

Robert J. Schaefer1, Jean-Michel Michno2, Joseph Jeffers3, Owen Hoekenga4, Brian Dilkes5, Ivan Baxter6,7, Chad L. Myers1,3

# Affiliations

1. Biomedical Informatics and Computational Biology Graduate Program, University of Minnesota, Minneapolis, MN, USA
2. Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN, USA
3. Department of Computer Science, University of Minnesota, Minneapolis, MN, USA
4. Cayuga Genetics Consulting Group LLC, Ithaca, NY, USA
5. Department of Biochemistry, Purdue University, West Lafayette, IN, USA
6. Donald Danforth Plant Science Center, St. Louis, MO, USA
7. USDA-ARS Plant Genetics Research Unit, St. Louis, MO, USA

# Abstract

Genome wide association studies (GWAS) have identified thousands of loci linked to hundreds of traits in many different species. However, the causal genes and the cellular processes they contribute to remain unknown for most loci. This problem is even more pronounced in non-model species where functional annotations are sparse. The vast amounts of data available from high throughput sequencing, such as RNA-Seq, are a tantalizing resource to leverage in identifying potential candidates under GWAS single nucleotide polymorphisms (SNPs), though are often underutilized or difficult to interpret. Here, we developed a framework to integrate functional information derived from RNA-Seq co-expression networks directly with GWAS datasets, establishing significance with a robust null-distribution model. We demonstrate the precision of our method by building three whole genome co-expression networks then simulating GWA studies using Gene Ontology (GO) terms. We then applied our method to functionally relate loci identified in a large scale, GWA study examining elemental accumulation in maize. Our results demonstrate that simply taking the closest genes to significant GWAS SNPs will often lead to spurious results demonstrating the need for proper functional modeling and a reliable null-distribution. Additionally, when deriving functional information from gene co-expression networks, the biological context from which the transcription was measured is essential. Inclusion of gene expression data from tissues not relevant to the elemental phenotypes collected abolished the relationships between the co-expression networks and the GWAS SNPs. Furthermore, in the correct biological context, genes linked to GWAS hits for elemental accumulation were more significantly co-expressed than genes within the same GO terms. Our study illustrates the factors that researchers should consider when taking this approach and demonstrates the utility of gene lists produced from GWAS for annotation of genes for systems biology applications.

# Author Summary

# Introduction

Genome wide association (GWA) studies are a powerful tool for understanding the genetic basis of traits. This approach has been successfully applied for hundreds of important traits in different species, including important yield-relevant traits in crops. Sufficiently powered GWA studies often identify tens to hundreds of loci containing hundreds of single nucleotide polymorphisms (SNPs) associated with a trait of interest (McMullen et al. 2009), many of which fall outside annotated gene boundaries (Wallace et al. 2014). For example, in maize alone, GWA studies have identified nearly 40 genetic loci for flowering time (Buckler et al. 2009), 89 loci for plant height (Peiffer et al. 2014), 36 loci for leaf length (Tian et al. 2011), 32 loci for resistance to southern leaf blight (Kump et al. 2011), and 26 loci for kernel protein (Cook et al. 2012). Despite an understanding of the overall genetic architecture and the ability to statistically associate many loci with a trait of interest, a major challenge has been the identification of causal genes and functional alleles associated with these loci, and more broadly, interpreting the biological pathways and processes that are affected by the identified variants. One major hurdle that limits the identification of causal genes is linkage disequilibrium (LD). Genetic markers identified by a GWA study can be relatively far from the actual causal variant, which can implicate a large number of causal genes at each associated locus. For example, LD in maize can vary between 1kb to over 1 million bases (Gore et al. 2009), and this can be even broader in other crop species (Morrell et al. 2005; Caldwell et al. 2006). A related issue is that there is increasing evidence that gene regulatory regions play a significant role in functional variation (Wray 2007), which means that causal variants can be quite far from the gene whose expression is affected. Several examples of non-coding sequences having major functional effects have been previously reported in maize (Clark et al. 2006; Castelletti et al. 2014; Louwers et al. 2009). The combination of these two challenging factors means that even when presented with a variant is strongly associated with a trait, a large number of candidate genes remain that are plausibly causal. Furthermore, these issues are multiplied when studying complex traits involving many loci. And the problem of pinpointing causal genes is exacerbated in crop species where gene annotation is largely incomplete. For example, in maize, only ~1% of genes have functional annotations based on mutant analysis (Andorf et al. 2015).

Thus, even when a list of potential candidate genes can be identified for a particular trait, there are very few other sources of information that can help identify strong candidate genes influencing a trait. The interpretation and narrowing of large lists of highly associated SNPs with complex traits is now the bottleneck in developing new mechanistic understanding of how genes influence traits. Advanced mapping populations developed in crops species have enabled the rapid identification of hundreds of loci that characterize traits critical to important, global issues such as worldwide food supply and crop nutritional quality, yet we lack the tools to understand the wealth of information linking genotypic variation to phenotype, especially when the trait of interest involves many genes that have interactions that a GWA study does not explicitly model.

One informative and easily measurable source of functional information are gene expression levels. Surveying genes’ expression profiles in different contexts, such as throughout tissue development or within different genetic backgrounds, helps establish how a gene’s expression is linked to its biological function, by establishing which genes are in the correct time and place to contribute to a common product or outcome. Variation in the pattern or intensity of gene expression can underlie variation in phenotypes, even where protein sequences maybe identical. Comparing the similarity of two gene’s expression profiles, or co-expression, quantifies the joint response of the genes to various biological contexts, and highly similar expression profiles can indicate shared regulation and function. Analysis of co-expression or co-expression networks have been used successfully for identifying functionally related genes, including in several crop species (R.J. Schaefer et al. 2014; Mochida et al. 2011; Obayashi et al. 2014; Sarkar et al. 2014; Zheng & Zhao 2013; Ozaki et al. 2010; Swanson-Wagner et al. 2012).

Because co-expression provides a global measure of functional relationships, it can serve as a powerful means of interpreting GWAS candidate loci. Specifically, we expect that variation in several different genes contributing to the same biological process would be associated with a given phenotype (Wolfe et al. 2005). Thus, if genetic variation driving the phenotype captured by GWAS is encoded by co-regulated genes, these data sets will non-randomly overlap. This additional logical filter provides an opportunity to prioritize candidate genes linked to GWAS SNPs based on putative functional information captured by a gene co-expression network. Though not all functional relationships are not captured using co-expression (Kim 2015), the comparison of these data still provide a highly informative, and sometimes exclusive, set of clues about genes that have been otherwise not been studied. This principle has been used successfully with other types of networks, for example protein-protein interactions (Li et al. 2008), and co-expression has been used as a basis for understanding GWA studies in mouse and human (Calabrese et al. 2017; Bunyavanich et al. 2014; Taşan et al. 2014).

We developed a freely available, open source computational framework called Camoco (**C**o-**A**nalysis of **mo**lecular **co**mponents), designed specifically for integrating GWAS candidate lists with gene co-expression networks to prioritize individual candidate genes. Camoco takes as input a list of candidate SNPs and corresponding association statistics derived from a typical GWAS study and identifies a ranked list of high confidence candidate genes that are embedded in densely connected functional modules where multiple members of those modules are associated with the phenotype of interest.

We tested this approach in the context of maize, one of the most important agricultural crops in the world, yielding 13.6 billion bushels of grain alone in the USA in 2015 (USDA 2016). We specifically focused on quantitative phenotypes measuring the accumulation of 17 different elements in maize grain ionome (Al, As, B, Ca, Cd, Fe, K, Mg, Mn, Mo, Na, Ni, Rb, S, Se, Sr and Zn). Plants must take up all elements except carbon and oxygen from the soil making the plant ionome a critical component in understanding plant environmental response(Baxter 2010), grain nutritional quality (Guerinot & Salt 2017) and plant physiology (Baxter et al. 2008).

We evaluated the utility of multiple different types of co-expression networks for supporting the application of Camoco, and demonstrate the efficacy of our approach by simulating GWA studies using Gene Ontology (GO) terms to establish maize specific SNP-to-gene mapping parameters as well as a robust null model for GWAS-network overlap. We show that the approach does indeed confirm overlap between functional modules captured by co-expression networks and GWAS candidate SNPs for the maize grain ionome. We present high-confidence candidate genes identified for a variety of different ionomic traits, and more generally, highlight lessons about the connection between co-expression and GWAS loci from our study that are likely to generalize to other traits and other species.

# Results

## A framework for integrating GWAS results and co-expression networks

We developed a computational framework, called Camoco (**C**o-**A**nalysis of **mo**lecular **co**mponents), that integrates the outputs of GWAS with co-expression networks to prioritize high-confidence causal genes associated with a phenotype of interest. The rationale for our approach is that functionally coherent genes identified by a GWA study will have non-random structure in co-expression networks that capture the same biological function. Our approach takes, as input, a list of SNPs associated with a trait of interest along with a table of gene expression values and produces, as output, a list of high priority candidate genes that are near GWAS peaks having evidence of strong co-expression.

The three major components of the Camoco system are a module for SNP-to-gene mapping (Fig. 1A); tools for construction and analysis of co-expression networks (Fig. 1B); and an "overlap" algorithm that integrates GWAS-derived candidate genes with the co-expression networks and identifies high priority candidate genes with strong co-expression support across multiple GWAS loci (Fig. 1C) (see Materials and Methods for details on each component). The overlap algorithm identifies high-priority genes by identifying the subset of genes that are highly connected to other genes also implicated by GWAS SNPs.

To accomplish this, we implemented and evaluated two different network scoring metrics: network density, and network locality (Eq.1 and Eq.2). Network *density* measures the average interaction strength between all pairwise combinations of genes linked to GWAS peaks. Alternatively, network *locality* measures the specificity of co-expression-derived interactions for a given candidate gene to all genes in the GWAS-derived subnetwork as compared to all other genes in the genome (See Materials and Methods for score details). For a given query GWAS trait, statistical significance for both density as well as locality is determined by generating a null distribution based generating random GWAS traits (n=1000) with the same number of implicated loci and corresponding candidate genes. This null-distribution is used to derive a p-value for the observed network density and locality for all putative causal genes (Fig. 1D). The end result of this process is that for an input GWAS trait, Camoco produces a ranked list of candidate causal genes for both network metrics, paired with an associated false discovery rate that indicates the significance of the observed overlap between each candidate causal gene and the co-expression network.

### Fig. 1

### Schematic of the Camoco integration framework

The Camoco framework integrates genes identified by SNPs associated with complex traits with functional information inferred from co-expression networks. **(A)** A typical GWAS result for a complex trait identifies several SNPs (circled) passing the threshold for significance indicating a multi-genic trait. SNP-to-gene mapping windows identifies a varying number of candidate genes for each SNP. Candidate genes are identified based on user-specified window size and a maximum number of flanking genes surrounding a SNP (e.g. 50kb and 2 flanking genes designated in red). **(B)** Independently, gene co-expression networks identify interactions between genes uncovering an unbiased survey of putative biological co-function. Network interactions are identified by comparing gene expression profiles across a diverse set of accessions (e.g. experimental conditions, tissue, samples). Dense subnetworks indicate sets of genes with strongly correlated gene expression profiles. **(C)** Co-analysis of co-expression interactions among GWAS trait candidate genes identifies a small subset of genes with strong network connections. Blue lines indicate genes that have similar co-expression patterns indicating co-regulation or shared function. Starred genes are potential candidate genes associated with GWAS trait based on SNP-to-gene mapping and co-expression evidence. Red stars indicate genes that are not the closest to the GWAS SNP (non-adjacent) that may have been missed using stringent SNP-to-gene mapping. **(D)** Statistical significance of subnetwork interactions are assessed by comparing co-expression strength among genes identified from empirical GWAS datasets to random networks containing the same number of genes. In the illustrated case, the more interesting subnetwork has both high density as well as locality.

## Generating co-expression networks from diverse transcriptional data

The co-expression network is a key component of our approach as it provides the functional relationships that allow us to identify functionally coherent subsets of GWAS-implicated genes. We and others have previously shown that co-expression networks generated from expression data derived from different contexts capture different functional information (Robert J. Schaefer, Briskine, Springer & Myers 2014). For example, experiments will target changes in gene expression stemming from environmental conditions, developmental and organ based variation, or variation that arises from population and ecological dynamics (see (Schaefer et al. 2016) for review). For many species, already published data contains enough experimental accessions (i.e. samples, tissues, conditions, etc.) to build networks from all of these different types of expression experiments. We reasoned that these different sources of expression profiles likely have a strong impact on the utility of the co-expression network for interpreting genetic variation captured by GWAS. Using this rationale, we constructed several different co-expression networks independently, and assessed the ability of each to produce high-confidence discoveries using our Camoco framework.

All datasets used here were generated from whole-genome RNA-Seq analysis, although Camoco could also be applied to microarray derived expression data. The first dataset (ZmPAN hereafter) targeted expression variation that exists between diverse maize accessions, built from whole-seedling transcriptomes on a panel of 503 diverse maize inbred lines from a previously published dataset (Hirsch et al. 2014). Briefly, Hirsch et al. chose these lines to represent major heterotic groups within the US, sweet corns, pop corns, and exotic maize lines while seedling tissue was chosen to due to the number of tissues which could be represented. The second dataset (ZmSAM hereafter) examined genotypic variation of different tissues and across developmental time points. Whole genome RNA-Seq transcriptome profiles from 76 different diverse tissues and developmental time points in the maize reference accession, B73, were used to build a second gene co-expression network using a publically available dataset (Stelpflug et al. 2015). Finally, a third dataset (ZmRoot hereafter) was generated as part of the ionomics GWAS research program and represents genotypic variation existing within a single organ. Gene expression was measured from mature roots from a collection of 46 genotypically diverse maize inbreds. See Materials and Methods for details on building these networks.

### Table 1

#### Significantly Co-expressed GO Terms

Co-expression was measured among genes within each GO term that had co-expression data in each network using both density (Eq.1) and locality (Eq.2). Significance of co-expression metrics was assessed by comparing values to 1000 random gene sets of the same size.

Co-expression networks for each dataset were constructed from gene expression matrices using Camoco (See Supp. File 1). Once built, several statistics were evaluated from interactions that arise from genes in the network (Supp. Fig. 1-3). Co-expression was measured among genes within the same Gene Ontology (GO) term to establish how well density and locality captured terms with annotated biological function.

Density and locality were measured for subnetworks consisting of the set of genes co-annotated to each GO term and compared to scores from 1000 random sets of genes of the same size (See Table 1; Supp. Table 1 for full data). In total, 818 GO Terms of the 1078 tested (76%) were significantly co-expressed in at least one network using density or locality relative to the randomized gene lists of the same size. Broken down by network as well by co-expression score, there was substantial co-expression among GO Terms for both density as well as locality in each network. Density captures the most GO terms in the ZmRoot network while locality performs best in ZmPAN. Considering terms captured by both scores or by either score, overlap between the two co-expression metrics was comparable. As previously reported (Robert J. Schaefer, Briskine, Springer, Myers, et al. 2014), GO terms were often strongly co-expressed in only a subset of the networks indicating that the networks experimental context influenced which biological processes it captured. Overall, while density and locality recover different GO terms, there are substantially more co-expressed GO terms, for either score, than was found by size-matched randomly generated sets of genes (Supp. Table 1).

### Table 2

#### Gene co-expression network cluster assignments

Gene clusters were calculated by running MCL on the co-expression matrix. Cluster values designate network specific gene clusters and are not compared across networks.

In addition to strong co-expression among previously known functional processes, unsupervised network clustering using the Markov-Clustering algorithm (Dongen 2000) showed distinct modules within each network. A large number of clusters were significantly enriched (hypergeometric p-value < 0.01) for genes that are co-annotated for the same GO term, however many strongly co-expressed clusters lacked any previously annotated function (Table 2; see Supp. Table 2 for full data). Additionally, all networks exhibited a truncated power law distribution in the number of significant interactions (degree) each gene had (Supp. Fig. 1-3), which is typical of biological networks (Ghazalpour et al. 2006).

### Accounting for *cis-* gene interactions

Integration of GWAS candidates with co-expression interactions is modeled in Camoco by directly assessing the density or locality of interaction among candidate genes near GWAS SNPs. However, mapping SNPs to surrounding candidate genes has inherent complications that can have strong influence on subnetwork co-expression calculations. While we assume that the majority of informative interactions among candidate genes are between GWAS loci, cis-regulatory elements and other factors can lead to co-expression between linked genes, which can produce skewed distributions in density and locality calculations. Identifying significant overlap between GWAS loci and co-expression networks requires a distinction between co-expression among genes that are in close proximity to one another on a chromosome (*cis*-) compared to genes which are situated far away from each other (*trans*-).

Network interactions for genes located on different chromosome (*trans-*interactions) were compared to interactions for pairs of genes that were less than 50 kb apart (*cis*-interactions). The distributions of the two groups indicate that *cis* genes are much more likely to have a strong co-expression interaction score than *trans* genes (Fig 2). This bias toward linked genes is especially pronounced at the extreme positive co-expression, where we observed substantially stronger enrichment for linked gene pairs as compared to genes on different chromosomes (e.g. Z-score ≥ 3, see Fig 2 inset).

### Fig 2

#### Cis vs Trans Co-Expression Network Interactions

Comparison of co-expression network interaction score distributions between *cis-* and *trans-* sets of genes. Distribution densities of *trans-* gene pairs (green) show interactions between genes on separate chromosomes. Distribution densities of *cis-* gene pairs (blue) show interactions between genes with less than 50kb intergenic distance.

The enrichment of significant co-expression among *cis*- genes, likely due to shared *cis*-regulatory mechanisms, prompted us to remove *cis-* interactions when examining co-expression relationships among candidate genes identified by GWAS SNPs in Camoco. To account for possible cis-regulation within network metrics described here, only interactions that span different GWAS loci (*trans-*) are included in density and locality calculations (See Materials and Methods).

## Evaluation of the Camoco framework

To explore the limits of our co-expression based approach, we examined factors that influence our proposed method of detecting overlap between co-expression networks and causal genes underlying GWAS-associated loci. In an idealized scenario, GWAS SNPs map directly to true causal genes, all of which exhibit strong co-expression network interactions (Fig 3). In practice, SNPs are likely in linkage disequilibrium (LD) with functional alleles leading to a large proportion of SNPs occurring outside of genic regions (Wallace et al. 2014).

We evaluated two major factors that influence SNP-to-gene mapping. The first of these is simply the total number of functionally related genes in a subnetwork, or the fraction of genes involved in a causal biological process, that are simultaneously identified by GWAS. In cases where there are too few genes representing any one of the underlying causal processes, our proposed approach is not likely to perform well. For example, if a biological process is controlled by 10 genes but due to penetrance, limited allelic variation in the mapping population, or extensive gene by environment interactions, only one locus has a significant association in the dataset. We refer to this as the *missing candidate gene rate (MCR)*, or, in other words, the fraction of genes involved in the causal process that simply weren’t identified by the GWAS in question (Fig 3**B**; Eq. 6).

A second key challenge in using the proposed approach to identify causal genes from a collection of significant GWAS loci is the fact that each associated SNP can be potentially linked with a large number of candidate genes. Thus, in cases where the linked regions are large (i.e. imperfect SNP-to-gene mapping), the ability to confidently identify subnetworks of highly co-expressed causal genes, which is the basis of our whole approach, may be compromised. One would expect to find scenarios where the proposed approach does not work simply because there are too many non-causal genes implicated by each GWAS locus, such that the co-expression signal among the true causal genes is diminished by the false candidates linked to those regions. We refer to this as the *false candidate gene rate (FCR)*, the fraction of all genes linked to GWAS loci that are not the causal genes (Fig 3**C**; Eq. 7).

To explore the limits of our co-expression-based approach with respect to these factors, we simulated scenarios where we could precisely control both the rate of missing candidates (MCR) and false candidates (FCR). Of course, in practice, neither of these quantities can be controlled—the missing candidate rate is a function of the genetic architecture of the phenotype as well as the degree of power within the study population of interest, and the false candidate rate is a function of recombination frequency in the GWA study population.

To evaluate the expected performance of our Camoco system for a range of both of these parameters, we simulated ideal GWAS scenarios using co-expressed GO terms (p < 0.05; Table 1). These ideal cases were then either subjected to subsampled genes (i.e. to simulate missing candidates) or adding functionally unrelated genes (i.e. to simulate false candidates introduced by linkage). In both cases, simulated GWAS (GO terms) were subjected to varying levels of either FCR or MCR while tracking the number of GO terms that remained significantly co-expressed at each level. These simulations enabled us to explore a broad range of settings for these key parameters and establish whether our proposed approach actually had the potential to work in the maize GWAS setting.

### Fig 3

#### Simulating GWAS-network overlap using Gene Ontology Terms

Several GWAS scenarios were simulated to assess the effect of noise on co-expression network overlap. Panel (**A**) shows an ideal GWA study were SNPs (blue points) map directly to candidate genes that are within the same biological process (i.e. a GO Term) and have strong co-expression (green lines). Signal is defined as the co-expression among the genes exclusive to the GO term. Noise in the overlap between GWAS and co-expression networks was introduced by varying two parameters: missing candidate gene rate (MCR) and false candidate gene rate (FCR). (**B)** The effect of a large proportion of missing candidate genes (e.g. MCR=2/5) on network signal. (**C**) The effect of false candidate genes on network overlap, either through false positive GWAS SNPs (orange points) or through imperfect SNP-to-Gene mapping (e.g. FCR=3/8). Here the additional candidate genes introduce co-expression noise (orange lines) that impedes the identification of network structure.

### Simulated GWAS datasets show robust co-expression signal to MCR and FCR

Subnetwork density and locality were measured for significantly co-expressed GO terms containing between 50 and 150 genes in each network at varying levels of missing candidate rate (MCR) (see Supp. Table 3). At each MCR level, density and locality among the remaining genes were compared to 1000 random sets of genes of the same size. The proportion of initial GO terms that remained significantly co-expressed was recorded for each network (see Fig. 4, red-curve; See Supp. Fig. 4**A** for absolute term numbers). GO terms were also split into two starting groups based on strength of initial co-expression: moderate (0.001 < p-val ≤ 0.05; blue curve) and strong (p-val ≤ 0.001; violet curve).

### Fig. 4

#### Strength of co-expression among GO Terms at varying levels of MCR

Subnetwork density and locality were measured for all GO terms with strong initial co-expression (p≤ 0.05) comparing co-expression in GO Terms to 1,000 random networks of the same size. Co-expression density and locality were then compared again (n=1000) with varying missing candidate rate (MCR) where a percentage of genes were removed from the term and replaced with random genes to conserve GO Term size. Curves decline with increased MCR as the proportion of strongly co-expressed GO terms (p-val ≤ 0.05, n=1000) decreases compared to the initial number of strongly co-expressed terms in each network (Red curve). GO terms in each network were also split into two subsets based on initial co-expression strength: ‘Strong’, initial co-expression p ≤ 0.001 (Blue curve), and ‘Moderate’, initial 0.001 < p ≤ 0.05 (Purple curve). In all three networks, GO terms with stronger initial co-expression were more robust to missing candidate genes.

As expected, strength of co-expression among GO terms decreased as MCR increases. Fig 3 shows the decay in the proportion of GO terms that exhibit significant co-expression at increasing levels of MCR (red curve). In general, the decay of signal is similar between density and locality, where signal initially decays slowly until approximately 60% MCR, where signal quickly diminishes.

We observed that the initial strength of co-expression impacted the rate at which co-expression signal decayed. Signal among strongly co-expressed GO terms (p-val ≤ 0.001; violet curve) decayed at a substantially lower rate than moderately co-expressed GO terms indicating that non-simulated GWAS datasets will be robust to a moderate level of missing candidate genes when co-expression among true candidate genes is strong. Co-expression signal in relation to MCR was also compared between GO terms split by gene size (See Supp. Fig. 4**B-C**), which did not differentiate the rate at which co-expression signal decayed.

Likewise, the effect of false positive candidate gene rate (FCR) was simulated by taking genes within significantly co-expressed GO terms between 50 and 150 genes (MCR=0) that had significant co-expression (p<0.05; see Supp. Table 3). Simulated SNPs corresponding to the starting base pair of each true GO term gene were used as input to SNP-to-gene mapping (See Materials and Methods). Subnetwork density and locality were calculated for candidate genes at each SNP-to-gene mapping combination in each network to evaluate the decay of co-expression signal in relation to FCR (Fig. 5).

### Fig. 5

#### Simulated GWAS: SNP-to-gene Mapping Density Signal Robustness

Strongly co-expressed GO terms (density or locality p-value ≤ 0.05) were used to simulate the effect of FCR on GWAS results. False candidates were added to GO terms by including flanking genes near true GO term genes according to SNP-to-gene mapping (windowing) parameters. Boxplots show effective FCR of GO terms at each SNP-to-gene mapping parameter. Signal plots show the proportional number GO terms that remain significant at FCR ≥ x.

Candidate genes were considered for each SNP up to 50kb, 100kb and 500kb upstream and downstream, limited to 1, 2, and 5 maximum flanking genes on each side. Given the number of additional candidate genes introduced at each SNP-to-gene mapping combination, false candidate rate (FCR) was calculated for each GO term at each level (See Fig. 5 box-plots). The least permissive SNP-to-gene mapping combination, 50kb and 1 flanking gene, incurred an average of over 50% false candidate genes. As SNP-to-gene mapping parameters become more permissive, FCR increases quickly (Fig. 5 bottom).

Co-expression signal in relation to FCR was assessed by comparing subnetwork density and locality in each GO term at different SNP-to-gene mapping parameters for each of the three co-expression networks to random subnetworks with the same number of genes (n=1000) (Fig. 5 top). The proportion of GO terms that were significantly co-expressed decays at higher levels of FCR (See Supp. Fig. 5**A** for absolute term numbers). We note that many GO terms could not be evaluated at lower FCR levels (<50%) because, by definition, even the most stringent SNP-to-gene mapping parameter set (50kb/1 flank) doubles the number of genes considered. Two additional scenarios were considered where signal was further split based on strength of initial co-expression: moderate (0.001 < p-val < 0.05; blue curve) and strong (p-val < 0.001; violet curve).

Despite high initial false candidate rates, co-expression signal among GO terms remained significant even at 60-70% FCR. Similar to results with MCR, GO terms with stronger initial co-expression are more likely to remain significantly co-expressed at higher FCR levels. Co-expression signal in relation to FCR was also compared between GO terms split by gene size (See Supp. Fig. 5**B-C**), which did not differentiate the rate at which co-expression signal decayed.

In cases where true candidate genes identified by GWAS are strongly co-expressed, as simulated here, a substantial number of false positive SNPs or an introduction of false candidate genes through uncertainty in SNP-to-gene mapping can be tolerated and network metrics and detected using our method. These results indicate that true GWAS scenarios, where the majority of SNPs do not perfectly resolve candidate genes, can be systematically integrated with co-expression networks to efficiently filter out false candidates introduced by SNP-to-gene mapping using the parameters explored here. More so, instances where several interleaving genes exist between associated SNPs and true candidates can be detected using co-expression networks as a functional filter for candidate gene identification.

The potential for using this approach, however, is highly dependent on the linkage disequilibrium of the organism in question as well as the genetic architecture of the trait being studied. Simulations provide both insight on the feasibility of using Camoco to evaluate overlap between co-expression networks and GWA studies as well as a survey of the SNP-to-gene mapping parameters that should be used when utilizing this approach (See Discussion for more detail).

In the context of maize, simulations performed here suggest that systematic integration of co-expression networks to interpret GWAS results should prove to be a powerful tool in identifying the causal genes associated with quantitative traits in true GWAS scenarios.

## Application of Camoco to prioritize causal genes driving elemental accumulation in maize grain

Identifying the biological processes underlying the elemental composition, also known as the ionome, of plant tissues can lead to a better understanding of plant adaptation as well as improved crops (Baxter & Dilkes 2012). High throughput analytic approaches such as inductively coupled plasma-mass spectrometry (ICP-MS) are capable of measuring elemental concentrations for multiple elements and scalable to thousands of samples a week. Using ICP-MS, we measured 20 elemental concentrations in maize kernels from the nested association mapping (NAM) population (Buckler et al. 2009) from grow-outs from 4 different geographic locations described in depth by Zielger et al. (Cite Ziegler et al. Preprint). To reduce environmental specific factors, only SNPs that were significantly associated with elements combined across all locations reported by Ziegler et al. were considered.

### Table 3

#### Maize grain ionome SNP-to-gene mapping results

Significant SNPS associated with the maize grain Ionome were mapped to candidate genes by collapsing SNPs within window sizes down to effective SNPs then taking genes upstream and downstream of the effective SNP up to the flank limit.

Briefly, the maize nested association mapping (NAM) panel was used to identify joint linkage intervals. Approximately 30 million SNPs and small copy-number variants were projected onto the association panel and used to perform a GWAS for each of the 17 elements. Over four thousand SNPs were significantly associated with elemental accumulation (Table 3; see Materials and Methods). These SNPs were used as input to Camoco to generate candidate genes for each element using a range of SNP-to-gene mapping parameters: 50kb, 100kb, and 500kb windows (up/downstream) limited to 1, 2 or 5 flanking genes (upstream and downstream of SNP; See Fig. 1A). In total, 4,243 statistically significant SNPs were associated with maize grain ionome traits, implicating between 5,272 and 22,927 unique genes depending on the SNP-to-gene mapping parameter used. On average for each trait, significantly associated SNPs identified 118 non-overlapping windows across the ten chromosomes of maize (i.e. effective loci, see Materials and Methods) and contain 612 positional candidate genes.

### Camoco identifies high priority candidate causal genes under ionomic GWAS loci

Given the large number of candidate genes associated with elemental accumulation, we used Camoco to integrate effective loci identified by the ionome GWAS with the three co-expression networks described above (ZmPAN, ZmRoot and ZmSAM) to identify high priority candidate genes driving elemental accumulation in maize (See Fig. 1C). Using our approach, we assessed the overlap between each of the three co-expression networks paired with each of the 17 ionomic traits. For each network-trait combination, Camoco reported a rank list of prioritized candidate causal genes, each associated with an FDR that reflects the significance of co-expression connecting the candidate gene to genes near other loci associated with the given trait (Supp. Table 4). We defined a set of high-confidence discoveries by reporting candidates that were discovered at a false discovery rate (FDR) < 30% based on at least two SNP-to-gene mapping parameter settings (e.g.: 50kb/1Flank and 100kb/1Flank), a set we called the high priority overlap (HPO) set (see Supp. Table 5 and Materials and Methods).

Indeed, we found strong evidence of co-expression among genes linked to GWAS peaks across the ionomic traits (Table 4). For example, for cadmium, 630 GWAS SNPs were analyzed using the ZmRoot network. Camoco reported 209 HPO genes at FDR < 30% based on strong co-expression relationships with genes linked to other loci associated with cadmium. The number of candidate causal genes discovered varied significantly across the ionomic traits we examined. For example, when using the root co-expression network, high-confidence candidates were discovered for 15 of 17 different elements we examined (Al, As, B, Ca, Cd, Cu, Fe, K, Mg, Mn, Mo, P, Se, Sr, Zn), but the total number of candidate causal genes varied from 1 to 126 HPO genes, with Camoco reporting only 1 for Mg and Mo and 126 for Cadmium. This likely reflects variability in the functional coherence of the genes that are associated with each trait. Traits with association signal spread across several co-regulated genes are amenable to Camoco gene prioritization while traits for which a broad spectrum of genes are associated will be more difficult to interpret using this co-expression approach.

### Table 4

#### Maize Grain Ionome High Priority Candidate Genes

Gene specific density and locality metrics were compared to (n=1000) random sets of genes of the same size to establish a 30% False Discovery rate. Genes were considered candidates if they were observed at 2 or more SNP-to-gene mappings (i.e. HPO). Candidates in the "Either" column are HPO genes discovered by either density or locality in any network. The number of genes discovered for each element is further broken down by co-expression method (density, locality, both) and by network (ZmPAN, ZmSAM, ZmRoot). Candidates in the "Both" column were either discovered by density and locality in the same network or in different networks (Any). Note: zero elements had HPO genes using “Both” methods in the ZmPAN or ZmSAM networks.

### Co-expression networks derived from variation across genotypically diverse accessions support stronger candidate gene discoveries

Strikingly, the number of candidate causal genes discovered by Camoco varied substantially depending on which co-expression network was used as the basis for discovery. The ZmRoot co-expression network provided the strongest input, discovering genes for 15 of the 17 elements for a total of 335 HPO genes, ranging from 1 to 126 per trait (Supp. Table 5). In contrast, the ZmSAM network, which was constructed based on tissue and developmental expression atlas collected exclusively from the B73 accession, supported the discovery of candidate genes for just 8 elements (B, Ca, K, Mg, Ni, P, Rb and Se) for a total of 74 HPO genes, ranging from 1 to 52 per trait (Supp. Table 5). The ZmPAN network, which was constructed from seedling tissue across 503 different accessions provided intermediate results, supporting high-confidence candidate discoveries for 10 elements (Al, As, Cd, Mg, Mn, Mo, Se, Sr and Zn) for a total of 228 HPO genes, ranging from 1 to 97 per trait (Supp. Table 5). The relative strength of the different networks for discovering candidate causal genes was consistent even at stricter FDR thresholds (e.g. FDR ≤ 0.10; Supp. Table 5).

### Fig. 6

#### HPO Genes for Cd and Se in the ZmRoot Network

The strongest 100,000 interactions were used to layout genes (n=7,844) in the ZmRoot network in order to visualize clustering. A force directed algorithm positioned genes (A; green nodes) showing approximate boundaries (dotted black circles) of the top 10 MCL clusters (Supp. Table 2). The ZmRoot network view was filtered to possible candidate genes (blue nodes) identified from SNP-to-gene mapping (Table 3) for Cd and Se (B and C respectively). Network edges were removed from the visualization, though MCL cluster boundaries were preserved. HPO genes for each element (highlighted in red) co-localize to specific clusters.

Fig. 6 shows an example view of the discovery process for Cd and Se HPO genes in the ZmRoot network. We organized a global view of the ZmRoot network using the strongest 100,000 interactions with a force directed layout algorithm to show high level clustering. While a large number of possible candidate genes from SNP-to-gene mapping for each element (Fig. 6B-C, blue nodes) span many of the MCL clusters identified in the network (dotted ellipses), HPO genes discovered by density and locality for each element co-localize to only a small number of MCL clusters (red nodes).

### Density and locality network metrics provide complementary information

As part of the Camoco pipeline, we implemented two different network metrics for measuring the strength of a given candidate causal gene’s co-expression relationships with genes in other GWAS-identified loci: density and locality (See Eq.3 and Eq.4). Density simply measures the fraction of observed co-expression interactions to total possible co-expression relationships between the candidate gene and genes linked to other GWAS-identified loci, while locality is normalized to account for the proportion of interactions between the candidate gene and the rest of the genome (i.e. genes not near a GWAS locus). Overall, density identified more HPO genes than locality. For example, across all traits and networks, 391 HPO candidate genes were discovered using density while 247 HPO candidate genes were discovered using locality (See Table 4, Density:Any and Locality:Any). Interestingly, the high-confidence genes discovered were largely complementary, both in terms of which traits they produced results on and for which network. Among the two sets of genes (391 and 247 genes, respectively), 26 HPO genes were discovered in common (Table 4: Both:Any). While this overlap is statistically significant (p ≤ 1.5e-13; hypergeometric), the large number of uniquely discovered genes suggests the two measures are capturing largely complementary co-expression subnetworks. Indeed, when we measured the direct correlation of gene-specific density and locality measures across several GWAS traits and GO, we observed a shift toward positive but weak correlations (Supp. Figure 6), confirming why largely non-overlapping genes are discovered by the two metrics. Density produced the strongest results when using the ZmRoot network (326 HPO genes using density vs. 11 HPO gene using locality), while locality provided stronger results on the ZmPAN network (228 HPO genes using locality and 0 HPO genes using density). We did observe that the utility of the locality metric appeared to be linked to the number of accessions used to construct the network (Supp. Table 6). For example, both the ZmPAN and ZmRoot networks were rebuilt using the common set of 20 accessions from the 503 ZmPAN and 46 ZmRoot accessions. In both instances, substantially fewer HPO genes were discovered compared to the full sets. The ZmPAN network was also built using the common set of 20 accessions as well as an additional 26 random or CML biased lines to simulate the number of accessions used in the ZmRoot network. In this case, there was a small increase in the number of HPO genes compared to the 20 accession networks, but still substantially fewer HPO genes discovered than the full ZmPAN network.

### Most candidate causal genes are trait-specific

One important question is the extent to which putative causal genes overlap across different ionomic traits. It is plausible that some mechanisms affecting elemental accumulation are modulated by common biological processes, which could have been detected with our Camoco pipeline. We compiled the complete set of HPO genes discovered for each element and assessed overlap in putative causal genes across the complete set of 17 elements (Table 5). Broadly, most of the discovered candidate causal genes are element specific, with relatively little overlap between elements (Table 5). However, a limited number of element pairs did exhibit statistically significant overlap including Cd, which shared significant overlap with 7 other elements (Al, Cu, K, Mg, Mo, Se and Sr), and Se, which shared significant overlap with 3 other elements (As, Cd and Mg), and Mo, which shared significant overlap with 2 other elements (Al, Cd). These candidate genes represent important potential modulators of elemental composition and are worthy of further study (Supp. Table 7).

### Table 5

#### Element HPO candidate gene overlap

Overlap between the 610 HPO genes discovered between different elements by either density or locality and in any network. The diagonal shows the number of HPO genes discovered for each element. Values in the upper triangular (green) show the number of genes that overlap between elements. The values in the lower triangular designate the p-values (hypergeometric) for overlap between the two sets of HPO genes. Red cells indicate significance with Bonferroni correction.

### Enrichment analysis of putative causal genes

To explore the broader bioprocesses represented among the high-confidence candidate causal genes from our analysis of ionomic traits, we performed Gene Ontology enrichment analysis on the list of candidates produced. GO enrichment on the set of HPO genes discovered for each individual did reveal enrichments for some elements (Supp. Table 8). For example, Sr was enriched for “anion transport” (GO: 0006820; p≤0.008) and “metal ion transmembrane transporter activity” (GO:0046873; p≤0.015). However, these enrichment results were relatively limited, possibly due to the lack of functional annotations for the maize genome, and none passed strict multiple-test correction (Bonferroni). To compensate for the sparsity of annotations, we started with the HPO gene set discovered for each trait and identified an expanded set of highly connected co-expression network neighbors, designated the HPO+ sets, which were formed based on genes’ aggregate connectedness to the HPO set (see Materials and Methods). The HPO+ sets for several of the ionomic traits showed strong GO enrichment, many of which had terms that passed strict multiple-test correction, including Al, As, Cd, Cu, Fe, K, P, Se, Sr, and Zn (Supp. Table 9). Several of the enriched GO terms were common across HPO+ sets for different elements (Fig. 7). For example, we found enrichment for a collection of GO terms related to ion transport (GO:0006811) including anion transport (GO: 0006820) and potassium ion transport (GO:0006813) and others (GO:0015849, GO:0015711, GO:0046942, GO:0006835), which were supported by enrichments from multiple elements (Al, Cd, Fe, Sr) (See Fig. 7; “Transport” cluster). We also observed a set of 6 elements whose HPO+ sets (Al, Cd, Cu, K, Se, Sr) were enriched for GO terms related to chromatin organization (e.g. GO:0006325, GO:0071824, GO:0034728, GO:0006334, See Fig. 7; “Chromatin Organization” cluster), which is interesting because <I think Brian had an explanation about heavy metals and chromatin that would be nice to put in here>.

Several of the observed GO enrichments were trait-specific including collections of GO terms reflecting “Chemical Response” (Se), “Microtubule Movement” (As), “Adhesion” (Cu), and “Saccharide Metabolism” (P). For example, the “Saccharide Metabolism” collection of GO term enrichments was driven by 5 HPO+ genes for P, one of which was tgd1 (GRMZM2G044027; See Supp. Table 9). Mutations on the Arabidopsis ortholog for tgd1 caused the accumulation of triacylglycerols and oligogalactolipids and showed a decreased ability to incorporate phosphatidate into galactolipids (Fan et al. 2015). Tgd1 has been shown to function as a ATP-binding cassette (ABC) transporter known to transport other substrates, including inorganic and organic cations and anions (Roston et al. 2012). Tgd1 was an original candidate causal gene (HPO set), and four other genes were identified as strongly connected neighbors in the co-expression network, two with unknown function, GRMZM2G018241 and GRMZM2G030673, and two involved in cellulose synthase, GRMZM2G122277 and GRMZM2G177631. These genes play a role in glucan and polysaccharide metabolism and biosynthesis pathways, which are likely to impact phosphorus allocation and accumulation.

We should note that there were several other enriched GO terms that demonstrated the deficiencies of automated annotation approaches, including terms related to “blood coagulation” and “regulation of body fluid levels”, which were likely due to annotations translated to maize genes on the basis of protein sequence homology. While, at face value, these term descriptions are inapplicable in plant species, the fact that these terms contained HPO genes as well as strong network co-expression lends itself to cases where annotations assigned through orthology could be further refined given co-expression evidence (see Discussion).

In general, this approach of using the co-expression networks to expand the neighborhood of the high-confidence candidate causal genes and then assessing the entire set for functional coherence through GO enrichment is a productive strategy for gaining insight into what processes are represented. Yet, this approach is particularly in the challenging context of the annotation-sparse maize genome, where only ~1% of genes have mutant phenotypes (Lawrence et al. 2004). Descriptions were high level or insufficient to fully identify causal genes. However, the enriched terms discovered here fit previously described pathways known to impact elemental traits. With bolstered confidence that subnetworks containing HPO genes contained coherent biological information, we refined our analysis by curating target HPO genes for their involvement in specific biological processes, namely those that contribute to the transport, storage and utilization of the ionome in maize seeds.

### Fig. 7

#### Gene Ontology Biological Process Enrichment for the Ionome

The HPO gene sets along with strongest co-expressed neighbors (HPO+) were analyzed for GO enrichment in the “Biological Process” namespace. Each node represents a GO term organized hierarchically in a tree with directed edges designating parent terms. Shaded terms were enriched for HPO+ genes (p < 0.05; hypergeometric). Dotted circles represent curated functional terms describing the enriched nodes in clade of the tree. Each clade is annotated with the ionomic terms that were represented in the GO enrichment.

### Putative causal genes have strong literature support for their involvement in ionomic traits

In addition to the GO enrichment analysis of high-confidence candidate causal genes identified by our approach, we manually examined the literature evidence supporting the association of these genes with ionomic traits. Indeed, this examination revealed genes and pathways known to affect the ionome. In addition to genes with known roles in elemental homeostasis, HPO genes for some ionomic traits included multiple genes encoding known members of the same pathway or protein complex. This strongly suggests that true biological signal was enriched by our novel combination of expression level polymorphisms and GWAS and provides strong evidence of novel associations between multiple pathways and elemental homeostasis.

For example, one gene with highly pleiotropic effects on the maize kernel ionome is sugary1 (su1; GRMZM2G138060) (Baxter et al. 2014). Genetic polymorphisms that affect seed compartment proportions or the major storage constituents are expected to contribute disproportionately to variation in seed ionomic contents. For this reason, six IL14H RIL’s that were still segregating for the recessive su1 allele were previously tested for ionomic effects (cite). This demonstrated that segregation for a loss of function allele at su1, on the cob, affected the levels of P, S, K, Ca, Mn, Fe, As, Se, and Rb in the seed (Baxter et al. 2014). The su1 gene was present among the HPO genes for Se accumulation (Supp. Table 5) based on the root co-expression network. The su1 locus was only identified in the HPO set for the element Se, but was linked to significant NAM GWAS SNPs for the elements P, K, and As. Thus, of the eight elements that were identified as co segregating with the su1 locus in the IL14H RIL population that were also measured in the NAM panel, four were affected by su1 in the GWAS analysis. It is formally possible that su1, which is expressed in multiple plant compartments including the roots, might also affect the seed ionome through effects beyond a dramatic loss of seed starch. This may result from coordinate regulation of the encoded isoamylase and other root-expressed determinants of S and Se metabolism, a role for starch in root architecture or S and Se assimilation, or a result of unexpected coordination between root and seed expression networks. The finding that HPO network neighbors for P were enriched among carbohydrate biosynthetic enzymes favors the former two of these hypotheses.

Our analysis of the GWAS loci for the element Se produced among the largest high-confidence candidate causal genes. Several genes with known effects on the ionome, or known to be impacted by the ionome, were identified within this HPO set. For example, one candidate gene, GRMZM2G327406, encodes an Adenylyl-sulfate kinase (adenosine-5'-phosphosulfate (APS) kinase 3), which is a key component of the sulfur and selenium assimilation pathway and plays a role in the formation of the substrate for protein and metabolite sulfation. Downstream from this gene, and also critical for the metabolism of sulfur amino acids and the biosynthesis of the 21st amino acid, selenocysteine, Camoco also identified a gene encoding a cysteine desulfurulyase (GRMZM2G581155).

Following the work of Chao et al., alterations in cell size and cell division in the root are expected to have effects on K accumulation in the leaves (Chao et al. 2011). Two of the four subunits of polycomb repressive complex 2 (PRC2), known to act on the cell cycle via the retinoblastoma-related proteins (RBRs), were identified as HPO genes for the K analog Rb. Both MSI1 (GRMZM2G090217) and FIE2 (GRMZM2G148924; ZmSAM-Rb), members of the PRC2, are co-expressed in the ZmSAM network. The RBR-binding E2F-like transcription factor (GRMZM2G361659; ZmSAM-Rb) was also found, a further indication that cell cycle regulation, via these proteins’ interactions in maize could provide a common mechanism for these associations. Histone deacetylases from the RPD3 family are known to interact with RBR proteins as well. The RPD3-like histone deacetylase 2 from maize was identified in the same HPO set (GRMZM2G136067; ZmSAM-Rb). The Arabidopsis homologs of both MSI and this histone deacetylase have known roles as histone chaperones, and the latter directly binds Histone H2B. Remarkably, histone H2B (GRMZM2G401147; SAM-Rb) was also an HPO hit. Lastly, an Actin utilizing SNF2-like chromatin regulator18 (GRMZM2G126774 ZmSAM-Rb) was identified as yet another SAM-Rb hit. Taken together these demonstrate a strong enrichment for known protein-protein interactors important for chromatin regulation and cell cycle control among the HPO set for the K analog Rb.

### Transporters

A number of transporters with known roles in ionome homeostasis were also identified among the HPO genes. Among these were: a P type ATPase transporter of the ACA P2B subfamily 4 (GRMZM2G140328; ZmRoot-Sr) encoding a homolog of known plasma membrane localized Ca transporters in multiple species (fucking citing Ivan), an ABC transporter homolog of the family involved in organic acid secretion in the roots from the As HPO set (GRMZM2G415529; ZmRoot-As) (Badri et al. 2007), and a pyrophosphate energized pump (GRMZM2G090718; ZmPAN-Cd) [cite]. While not specific transporters for the elements they were linked to, several annotated transporters were identified in the HPO sets for multiple elements: a sulfate transporter (GRMZM2G444801, ZmRoot-K), a cationic amino acid transporter (AC207755.3\_FG005; ZmPAN-Cd, ZmPAN-Mo), and an inositol transporter (GRMZM2G142063; ZmRoot-Fe, ZmRoot-Cd, ZmRoot-Sr).

Cadmium was well measured and affected by substantial genetic variance (Zeigler et al., 2017). We detected the largest number of HPO candidate genes for Cd (209 genes; See Table 4). Among these were the maize glossy2 gene (GRMZM2G098239, ZmPAN-Cd), which is responsible for a step in the biosynthesis of hydrophobic barriers [citation]. This implicates the biosynthesis and deposition of hydrophobic molecules in the root in accumulation of ions and may point to root processes, rather than epicuticular waxes deposition, as the primary mode by which these genes may affect water dynamics. ARR1-like gene GRMZM2G067702 was also an HPO gene associated with Cd. Previous work has shown that ARR genes from Arabidopsis are present in the stele where they regulate the activity of HKT1 (Schfenberger, Tested Salted Caramel and Ghridelli, 2010 Plant Journal). This gene was expressed at the highest level in the stele at 3 days after seeding (DAS)[cite].

### Multi ionomic hits

One of the high-confidence candidate genes, which appeared in the HPO sets for Cd, is the Gibberellin (GA) signaling component and DELLA and GRAS-domain transcription factor Dwarf9 (GRMZM2G024973, D9) and was included in the HPO set for Cd in the ZmRoot network (Winkler & Freeling 1994). D9 is one of two DELLA paralogs in the maize genome (the other is Dwarf8, (D8), both of which can be mutated to dominant negative forms that display dwarf phenotypes and dramatic suppression of GA response. D9 was discovered among the high-confidence candidates for Cd while D8 was not, though both D8 and D9 are present in the root based co-expression network (ZmRoot). Given the indistinguishable phenotypes of the known D8 and D9 mutants, the most likely explanation for this result is that there was allelic variation for D9 in the GWAS panel, but not D8. The GA biosynthetic enzyme Ent Kaurene synthase (GRMZM2G093603) encoding the dwarf5 locus (Fu et al. 2016) also affected the concentration of seed Cd and appeared among the HPO genes for Sr. However, there was only moderate, but positive co-expression between dwarf5 and d9 (ZmRoot: Z=1.03; ZmPAN:: Z=1.04). This gene is required for the biosynthesis of bioactive GA via ent-Kaurene, strongly suggesting that GA signaling in the roots shapes the ionome and alters the accumulation of Cd in seeds, with potential impacts on human health.

### Fig. 8

#### Ionomic profiles of D8 and D9 mutants

Explanation of boxplots

We used dominant GA-insensitive mutants of both of the DELLA domain genes to provide single-locus tests of the role of GA signaling in determining the seed ionome. Seeds from D9 and D8 and their null segregating sibling lines were grown in the field and kernels’ element profiles analyzed. The dominant D8-mpl and D9-XX alleles have roughly equivalent effects on above-ground plant growth and similar GA-insensitivity phenotypes (Winkler & Freeling 1994). Both mutants were obtained from the maize genetics co-op and crossed three times to inbred B73 to generate BC2F1 families segregating 1:1 for the dwarf phenotype. Ears from phenotypically dwarf and phenotypically wild type siblings were collected and processed for single-seed ionomic profiling (Fig. ). Both dwarf lines had significantly different elemental compositions from their wild type siblings. A joint analysis revealed that Cu, Fe, P and Sr were higher in the dwarf lines (p < xxx, test). In addition to the elements that were different in the joint model, D9 was also significantly different for S. D8, which is expressed in roots at lower levels than D9 but many fold higher levels in shoots [cite], also was significantly greater in Mo accumulation. It is possible that D8 has a shoot-driven effect on Mo accumulation in the seed but we note that previous work (Asaro et al. 2016) identified a large effect QTL affecting Mo and containing the Mot1 gene a mere 22 Mb away from D8, the allele of which is unknown in the original D8 donor that was used to make the BC2 population analyzed here. As such, linkage drag carrying a Mot1 allele cannot be ruled out. This dominant-negative allele of D9 did not recapitulate the Cd accumulation effect of the linked GWAS QTL that was the basis for its discovery as a high-confidence candidate gene by Camoco, but our data suggests that both D8 and D9 have the potential for broad effects on ionomic phenotypes.

### Fig. 9

#### Co-expression network for D9 and cadmium HPO genes

Co-expression interactions among high priority candidate (HPO) genes were identified in the ZmRoot network for Cadmium and visualized at several levels. Panel **A** shows local interactions among the 126 cadmium HPO genes (red nodes). Genes are positioned based on chromosomal position and are arranged in a circle to visualize inter-locus interactions. Interactions among HPO genes and Della9 (D9; GRMZM2G024973) are highlighted in yellow. Panel **B** shows a force directed layout of d9 with HPO neighbors.

Additional genes co-expressed with D9 were investigated to determine which among these were associated with ionomic traits, and in particular, seed Cd levels. In the ZmRoot network, D9 had strong co-expression interactions with 38 other HPO genes (Fig. A). Among these were two other GRAS domain transcription factors including one of the maize Shortroot paralogs (GRMZM2G132794) and a second GRAS domain transcription factor (GRMZM2G079470). Both of these, as well as the presence of many cell cycle genes among the co-expressed genes and ionomics trait affecting genes, raised the possibility that, like in Arabidopsis (REF FIT IRT), DELLA-dependent processes responsive to GA both shape the architecture of the root and the ionome [Cite]. In Arabidopsis, DELLA expression disrupts Fe uptake and loss of DELLA prevents some of Fe-deficiency mediated root growth suppression. Our finding that constitutive DELLA activity in the roots, affected by the D9-1 and D8-mpl mutants, resulted in excess Fe points to a conserved role for the DELLA domain transcription factors and GA signaling in Fe homeostasis in maize, a plant with an entirely different Fe uptake system than *A. thaliana*. However, the direction of the effect was opposite to that observed in *A. thaliana*. Future research into the targets of the DELLA proteins in maize will be required to further address these differences.

Remarkably, the HPO co-expression network associated with D9 also contained three genes with expected roles in the biosynthesis and polymerization of phenylpropanoids. The successive steps of CCR1 (GRMZM2G131205) and the maize LigB paralog (GRMZM2G078500) as well as a laccase paralog were co-expressed with D9 (GRMZM2G336337). LigB, which in Angiosperms is only known to be required for the formation of a pioneer specialized metabolite of no known function in *A. thaliana*, was linked to QTL for multiple ions including Cd, Mn, Zn, and Ni. CCR1, however, was only found for Cd with a very strong effect if summed across multiple SNPs. The Laccase 12 gene (GRMZM2G336337) was also a multi-ionomic hit with linked SNPs affecting Cd, Fe, and P.

Patterns of co-expression for D9 also were found in the ZmPAN network. Consistent with the hypothesis that maize DELLA regulated the type II iron uptake mechanism used by grasses, the nicotianamine syntase3 gene (GRMZM2G439195, ZmPAN-Cd, ??-Al), which is required for making the type II iron chelators, was both a Cd GWAS hit and substantially co-expressed with D9 in the ZmPAN network such that it contributed to the identification of D9 as an HPO gene for Cd.

# Discussion

We demonstrated that gene expression data, through the construction of co-expression networks, can be used as an effective basis for prioritizing causal genes from GWAS. Our approach addresses a challenging bottleneck in the process of translating large sets of statistically associated loci into more mechanistic understanding of the trait. Marker SNPs identified by a GWA study provide an initial lead on a region of interest, but due to linkage disequilibrium, the candidate region can be quite broad, which implicates many potentially causal genes. In addition to LD, many SNPs identified by GWAS studies lie in regulatory regions quite far from their target genes (Clark et al. 2006; Castelletti et al. 2014; Louwers et al. 2009). Additionally, factors such as bias due to cis- co-expression and different co-expression metrics and networks need to be considered in order to identify co-expression signal.

These factors can result in several plausible candidate genes even where a locus is identified by GWAS with high confidence. A common approach to interpreting such a locus is through manual inspection of the genome region of interest with a goal of identifying candidate genes whose function is consistent with the phenotype of interest, which can introduce bias in the discovery process and completely ignores uncharacterized genes. For non-human and non-model species, like maize, this situation is especially challenging because the large majority of the genome remains functionally uncharacterized. Our approach leverages the powerful collections of gene expression data, which can be readily collected now for most species of interest, to add an important interpretation and prioritization filter to the output of a GWA study.

We demonstrate that Camoco was able to identify subsets of genes that were linked to candidate SNPs and also exhibit strong co-expression with genes near other candidate SNPs. The resulting prioritized gene sets reflect sets of co-regulated genes that can potentially be used to infer a broader biological process in which genetic variation can affect the phenotype of interest. Indeed, using Camoco, we found strong evidence for such gene sets in 13 of 17 elemental accumulation phenotypes we examined. These small sets of genes represent a small fraction of the candidates implicated by the GWAS for each phenotype, and we note that the large majority of our HPO genes were not the closest genes to the identified SNPs, and thus, would not have been identified using the simple approach of identifying the genes closest to each marker SNP (cite supp evidence?). It is important to note that our approach will necessarily fail at identifying some real causal genes. Phenotypes that are caused by genetic variation in a single or small number of genes, or conversely, that are caused by a diverse set of otherwise functionally unrelated genes are not good candidates for our approach. The core assumption on which Camoco is designed is that there are multiple genetic variants in different genes involved in a common biological process that individually cause phenotypic variation. We expect that this is often true of phenotypes (and this is supported by the fact that we have discovered strong candidates for the majority of traits examined), but we expect there are exceptional traits, and individual causal genes, for which our approach will not work. <optional: point out some of the obvious example candidate genes that we miss>

In completing the evaluation of our approach based on Gene Ontology, we observed a trend worth noting. We used sets of genes annotated to the same GO term as a gold standard for groups of functionally coherent genes. We simulated the effect of imperfect SNP-to-gene mapping by assuming that subsets of these GO terms were identified by a simulated GWAS trait in which the neighboring genes (encoded nearby on the genome) were added to simulate the scenario where we could not resolve the causal gene from linked neighboring genes. This analysis was useful as it establishes the boundaries of possibility for our approach, i.e. how much noise in terms of false candidate genes can be tolerated before the entire premise of our approach fails. As described in Fig. 5, this analysis suggests a sensitivity of ~40% using a SNP-to-gene mapping rule of +/-500 kb and up to 2 flanking genes (nearly 75% false candidates due to SNP-to-gene mapping). This result suggests that we would not be likely to discover processes as coherent as GO terms if linkage extended beyond this point. Intriguingly, when we actually went to apply our Camoco system to GWAS-identified loci for the ionomic traits, we observed some surprising results. At the same window/flank parameter setting noted above, we were able to make significant discoveries for xxx of yyy elements using the xxx metric based on the yyy network. This success rate is substantially higher than predicted by our GO simulations at the same window/flanking gene parameter setting. We speculate that this conservative estimation of Camoco’s performance is most likely due to the relatively poor quality of the GO annotations we used for our evaluation. There are likely a large number of processes for which the genes simply are not functionally coherent. Our analysis with Camoco suggests that the loci implicated by these ionomic GWAS loci are actually more coherent than maize genes co-annotated to many GO terms. Indeed one of the key motivations of our approach was that crop genomes like maize have limited GO annotation, and this result emphasizes the extent of this limitation. This reflects both limitations in our understanding of the molecular genetics of maize, but also a lack of species-specific curation, which is not surprising given the limited funding available for such efforts.

Beyond highlighting the challenges of a genome lacking functional annotation, these results also suggest an interesting direction for future work. Despite the limited GO annotation for maize genes, there have been a wealth of genome-wide association studies, many of them enabled by the powerful mapping populations that have been constructed (e.g. NAM (McMullen et al. 2009)). Our results suggest that these sets of loci, and a proper mapping to the genes they represent, could serve as a powerful resource for characterizing gene function. More systematic efforts to curate the results from such genome-wide association studies, filter gene sets with tools such as Camoco, and provide public access in convenient forms would be worthwhile. Maize is somewhat exceptional in this regard, but there are several other crop species with rich population genetic resources with limited genome functional annotation that could benefit from this approach as well.

## Co-expression context matters

Using our Camoco approach, we evaluated 17 ionomic traits for overlap with three different co-expression networks. While the networks exhibited similar enrichment for co-expression among genes within GO terms (Table 1), suggesting they were of similar quality, the performance of these networks varied substantially when used to prioritize candidate genes from the set of maize grain ionome GWAS. The two co-expression networks generated from expression profiles collected across a diverse set of individuals (ZmRoot, ZmPAN) performed substantially better than the ZmSAM network, which was based on a large collection of expression profiles across different tissues and developmental stages derived from a single reference line (B73). We emphasize that this result is not a reflection of the data quality or even the general utility of the co-expression network derived the tissue/developmental atlas. Evaluations of this network suggested a similar level of enrichment for co-expression relationships among genes involved in the same biological processes. Our results indicate that the processes underlying the genotypic variation associated with traits captured by GWAS are better captured by transcriptional variation observed across genetically diverse individuals. Indeed, we saw that although the networks captured similar levels of GO term enrichment Table 1, the actual GO terms that drove that enrichment are quite different (Supp. Table 1), which is consistent with our previous analysis of co-expression networks (Robert J. Schaefer, Briskine, Springer & Myers 2014). Between the two co-expression networks based on expression variation across individuals, we also observed interesting differences depending on which tissues were profiles. Our co-expression network derived from sampling of root tissue across a diverse set of individuals (ZmRoot) provided the best performance at the FDR we analyzed (Table 4), producing a total of 335 (326 from density and 11 from locality, 2 in both) HPO candidate genes as compared to 228 (all from locality) HPO candidate genes produced by the ZmPAN network, which was derived from expression profiles of whole seedlings. This result confirms our original motivation for collecting tissue-specific gene expression profiles— we expected that processes occurring in the roots would be central to element accumulation phenotype (which were measured in kernels). The difference between the performance of these two networks was modest, however, and much less significant than the difference between the developmental/tissue atlas-derived network and the diverse genotype-derived network. We also note that the performance of the ZmRoot versus the ZmPAN network was quite different depending on which network metric we used. Specifically, performance of the ZmRoot network was dependent on the density metric while performance of the ZmPAN network was dependent on the locality metric (Table 4), which is worth future investigation. In general, our results strongly suggest that co-expression networks derived from expression profiling of genetically diverse individuals as opposed to deep expression atlases derived from focus on a single reference genotype, will be more powerful for interpreting candidate genetic loci associated with phenotypes. Furthermore, our findings suggest that where it is possible to identify relevant tissues for a phenotype of interest, tissue-specific expression profiling across genetically diverse individuals is an effective strategy. This has important implications for data generation efforts in other species.

# Conclusion

Here, we integrate co-expression network with loci associated with elemental accumulation in maize grain. We built three different co-expression networks and simulated their ability to detect co-expression using GO terms. We then use these networks to identify patterns of co-expression in a set of measuring association for 17 different elemental traits, which resulted in the discovery of 610 high-confidence candidate causal genes. These candidate gene sets were enriched for bioprocesses related to the ionome. Although the large majority of the high-confidence candidate genes are uncharacterized and worth further study, we did find specific candidate genes that have previously been described to affect the plant ionome. We validated our approach through mutant analysis, and confirmed that variants at the D9 locus broadly impacted the plants’ elemental profiles. Our approach successfully prioritizes causal genes underlying GWAS-identified loci based solely on gene expression data and establishes a basis for functional interpretation of otherwise uncharacterized genes associated with complex traits.

# Materials and Methods

## Software implementation of Camoco

Camoco (Co-analysis of Molecular Components) is a python library including a suite of command line tools that inter-relates and co-analyzes different layers of genomic data. Namely, it integrates genes present near and around GWAS loci with functional information derived from gene co-expression networks. Camoco was developed to build and analyze co-expression networks from gene transcript expression data (i.e. RNA-Seq) but can also be utilized on other expression data such as metabolite or protein abundance as well as legacy expression data such as from micro-arrays.

This software implements three main routines: 1) construction and validation of co-expression networks from a counts or abundance matrix, 2) Mapping SNPs (or other loci) to genes, and 3) an algorithm that assesses the ‘overlap’ of co-expression among candidate genes near significant GWAS peaks.

Camoco is open source and freely available under the MIT license. Full source code, software examples as well as instructions on how to install and run Camoco are available at:

<http://github.com/schae234/Camoco>.

## Construction quality control of co-expression networks

### ZmPAN: A genotypically diverse, PAN genome co-expression network

Camoco was used to process the FPKM table reported by Hirsh et al. and to build a co-expression network. The raw gene expression data were passed through the quality control pipeline in Camoco Supp. File 1. After QC, 24,756 genes were used to build the network. For each pairwise combination of genes, a Pearson Correlation Coefficient (PCC) was calculated across normalized FPKM profiles (Supp. Fig. 1A) then mean centered and standard normalized (Z-score hereafter) to allow cross network comparison (Supp. Fig. 1B). A global significance threshold of Z ≥ 3 was set on co-expression interactions in order to calculate gene degree and other conventional network measures.

To assess overall network health, several approaches were taken. First, a Z-score of edges between genes co-annotated in the maize Gene Ontology (GO) terms was compared to edges in 1000 random terms containing the same number genes. Supp. Fig. 1C shows the distribution of p-values compared to empirical Z-score of edges within a GO term. With a nominal p-value cutoff of 0.05, the PAN co-expression network had 11.9 fold more GO terms than expected with a p-value ≤ 0.05, suggesting that edges within this co-expression network capture meaningful biological variation. Degree distribution is also as expected within the network (i.e. the number of interactions any given gene has). Supp. Fig. 1D shows empirical degree distributions compared to the power law, exponential, and truncated power law distributions. Typically, the degree distributions of biological networks are best fit by a truncated power law distribution, which is consistent with the ZmPAN genome co-expression network (Ghazalpour et al. 2006).

### ZmSAM: A Maize RNA-Seq Tissue Atlas co-expression network

Publicly available gene expression data was generated from downloaded from Stelpflug et. al (Stelpflug et al. 2015). In total, 22,691 genes passed quality control metrics specified in Supp. File 1. Similar to the *ZmPAN* network described above, gene interactions were calculated between each pairwise combination of genes to produce ~257 million network edges. A global significance threshold of Z ≥ 3 was set on co-expression interactions in order to differentiate significantly co-expressed gene pairs.

Supp. Fig. 2A shows before network edge score normalization. Network interaction scores were mean centered and standard normalized (Supp. Fig. 2B). The ZmSAM network shows a 10.8-fold enrichment for strong edge scores (p ≤ 0.05) between genes annotated in Gene Ontology terms (Supp. Fig. 2C). A final network health check shows that the empirical degree distribution of the ZmSAM network is consistent with previously characterized biological networks (Supp. Fig. 2D).

### ZmRoot: A Genotypically Diverse Maize Root Co-Expression Network

Root RNA was extracted and sequenced from 48 diverse maize lines using TruSeq stranded RNA library prep and Illumina HiSeq 100bp paired end RNA Sequencing (RNASeq) reads. Raw reads were deposited into the short read archive (SRA) under Project number PRJNA304663. Reads were pre-processed using a standard mapping pipeline. Raw reads were passed through quality control using the program AdapterRemoval (Lindgreen 2012) which collapses overlapping reads into high quality, single reads while also trimming residual PCR adapters. Reads were mapped to the Maize 5b reference genome using BWA (Li & Durbin 2009; Schubert et al. 2014) PCR duplicates were detected and removed, and then realignment was performed across detected insertions and deletions resulting in between 14 and 30 million high-quality, unique nuclear reads per sample. Two samples were dropped due to low coverage bringing the total number of samples to 46.

Quantification of gene expression levels into fragments per kilobase per million reads (FPKM) was done using a modified version of HTSeq that quantifies both paired- and unpaired-end reads (Anders et al. 2014) available at http://github.com/schae234/MixedHTSeq. Raw FPKM tables were imported into Camoco and passed through the quality control pipeline. After QC steps (Supp. File 1), 25,260 genes were included in co-expression network construction containing ~319 million interactions. Supp. Fig. 3A shows raw PCC scores while Supp. Fig. 3B shows Z-scores after standard normal transformation. Similar to ZmPAN and ZmSAM co-expression among GO terms was compared to random gene sets of the same size as GO terms (1000 instances) showing a 13.5-fold enrichment for significantly co-expressed GO Terms (Supp. Fig. 3C). The degree distribution of the ZmRoot network closely follows a truncated power law similar to the other networks build here (Supp. Fig. 3D).

## SNP-to-Gene mapping and effective loci

Two parameters are used within Camoco during SNP-to-gene mapping: candidate window size and maximum number of flanking genes. Windows were calculated both upstream and downstream of input SNPs. SNPs having overlapping windows were collapsed down into *effective loci*, resulting in a list of non-overlapping loci. SNPs with the highest resample model inclusion probability (RMIP), which is a statistical measure of the strength of association with the trait of interest and is an output of the GWAS (Valdar et al. 2009), were retained for further analysis. Remaining effective loci were cross referenced with the Maize reference genome (Maize 5b) to convert effective SNPs to candidate gene sets that include up to a certain number of flanking genes both upstream and downstream from the effective SNP.

## Calculating subnetwork density and locality

Co-expression was measured among candidate genes using two metrics: density and locality. Network *density* is formulated as the average interaction strength between *all* (un-thresholded) pairwise combinations of gene-gene interactions that occur between input genes normalized for the number of total pairs among input genes:

### Eq.1

Where X-bar is the mean sub-network interaction score, E(X) is the expected network interaction score, σ(X) is the standard deviation of network interactions, and N is the number of interactions in the sub-network.

Network *locality* assesses the proportion of significant co-expression interactions that are locally connected to other subnetwork genes compared to the number of global network interactions. To quantify network locality, both local and global degree is calculated for each gene within a sub-network. To account for degree bias, where genes with a high global degree are more likely to have more local interactions, a linear regression is calculated on local degree using global degree (local ~ global) and regression residuals for each gene are analyzed:

### Eq.2

Gene specific density is calculated by considering subnetwork interactions on a per-gene basis:

### Eq.3

Gene locality residuals can be interpreted independently to identify gene specific locality:

### Eq.4

We note that, here, interactions among genes that originate from the same effective GWAS SNP (i.e. cis-interactions) were removed from density and locality calculations. During SNP-to-gene mapping, candidate genes retained information containing a reference back to the parental GWAS SNP. A software flag within Camoco allows from interactions derived from the same parental SNP to be discarded from co-expression score calculations. Statistical significance of subnetwork density and locality was assessed by comparing subnetwork scores to 1000 random sets of candidate genes, conserving the number of input genes. Using these randomizations, p-values for calculated for co-expression scores.

## Simulating GWA studies using Gene Ontology (GO) terms

GO (Harris et al. 2004) annotations were downloaded for maize genes from http://ftp.maizesequence.org/release-4a.53/functional\_annotations/. Co-annotated genes within a GO term were treated as true causal genes identified by a hypothetical GWA study. Terms between 50 and 100 genes were included to simulate the genetic architecture of a multi-genic trait. In each co-expression network, terms having genes with significant co-expression (p-value ≤ 0.05; density or locality) were retained for further analysis. Noise introduced by imperfect GWAS was simulated using two different methods to decompose how noise affects significantly co-expressed networks.

### Missing Candidate Rate

### Eq. 6

### False Candidate Rate

### Eq. 7

### Simulating Missing candidate gene rate (MCR):

The effects of MCR were evaluated by subjecting GO terms with significant co-expression (p<0.05; described above) to varying levels of missing candidate rates. True GO term genes were replaced with random genes at varying rates (MCR: 0%, 10%, 20%, 50%, 80%, 90%, 100%). The effect of MCR was evaluated by assessing the number of GO terms that retained significant co-expression (compared to n=1000 randomizations) at each level of MCR.

### Adding false candidate genes by expanding SNP-to-gene mapping parameters

To determine how false candidates due to imperfect SNP-to-gene mapping affected the ability to detect co-expressed candidate genes linked to a GWAS trait, significantly co-expressed GO terms were assessed at the different SNP-to-gene mapping parameters (50kb 100kb, 500kb and 1, 2, or 5 flanking genes). Effective FCR at each SNP-to-gene mapping parameter setting was calculated by dividing the number of true GO genes with candidates identified after SNP-to-gene mapping. Since varying SNP-to-gene mapping parameters changes the number of candidate genes considered within a term, each term was considered independently for each parameter combination.

## Maize Ionome GWAS

Elemental concentrations were measured for 19 different elements and one elemental ratio in the maize kernel using inductively coupled plasma mass spectrometry (ICP-MS). Outliers were removed from single seed measurements using median absolute deviation (Davies & Gather 2012). Basic linear unbiased predictors (BLUPs) for each elemental concentration were calculated across different environments and estimate variance components (Hung et al. 2012). Joint linkage analysis was run using TASSEL version 3.0 (Bradbury et al. 2007) with over 7,000 SNPs obtained by genotype-by-sequencing (GBS) (Elshire et al. 2011). An empirical p-value cutoff was determined by performing 1,000 permutations in which the BLUP phenotype data was shuffled within each NAM family before joint-linkage analysis was performed. The p-value corresponding to a 5% false discovery rate was used for inclusion of a QTL in the joint linkage model.

Genome wide association was performed using stepwise forward regression implemented in TASSEL version 4.0 similar to other studies (Wallace et al. 2014; Cook et al. 2012; Tian et al. 2011). Briefly, genome wide association was performed on a chromosomal-by-chromosome basis. To account for variance explained by QTL on other chromosomes, the phenotypes used were the residuals from each chromosome calculated from the joint-linkage model fit with all significant joint-linkage QTL except those on the given chromosome. Association analysis for each trait was performed 100 times by randomly sampling, without replacement, 80% of the lines from each population.

The final input SNP dataset contained 28.9 million SNPs obtained from the maize HapMap1 (Gore et al. 2009), the maize HapMap2 (Chia et al. 2012), as well as an additional ~800,000 putative copy-number variants from analysis of read depth counts in HapMap2 (Wallace et al. 2014; Chia et al. 2012). These ~30 million markers were projected onto all 5,000 lines in the NAM population using low density markers obtained through GBS. A cutoff p-value value (p ≤ 1e-6) was used from inclusion in the final model. SNPs associated with elemental concentrations were considered significant in they were selected in more than 5 of the 100 models (Resample model inclusion probability (RMIP)) (Valdar et al. 2009)

## Identifying ionome high priority overlap (HPO) genes and HPO+ genes

Gene specific density and locality were calculated for candidate genes identified from ionome GWAS as well as for 1000 random sets of genes of the same size. Gene specific metrics were converted to the standard normal scale (Z-score) by subtracting the average gene specific score from the randomized set and dividing by the average randomized standard deviation. A false discovery rate was established by incrementally evaluating the number of GWAS candidates discovered at a Z-score threshold compared to the average number discovered in the random sets.

High priority overlap (HPO) candidate genes for each element were identified by requiring candidate genes to have an co-expression FDR ≤ 30% in two or more SNP-to-gene mapping scenarios in the same co-expression network using the same co-expression metric (i.e. density or locality).

HPO+ candidate gene sets were identified by taking the number of HPO genes discovered in each element (n genes) and querying each co-expression network for the set of (n) genes that had the strongest aggregate co-expression.

# Acknowledgements

# References

Anders, S., Pyl, P.T. & Huber, W., 2014. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2), pp.166–169. Available at: http://bioinformatics.oxfordjournals.org/content/31/2/166 [Accessed September 29, 2014].

Andorf, C.M. et al., 2015. MaizeGDB update: new tools, data and interface for the maize model organism database. *Nucleic acids research*, p.gkv1007. Available at: http://nar.oxfordjournals.org/content/early/2015/10/01/nar.gkv1007.full.

Asaro, A. et al., 2016. The Interaction of Genotype and Environment Determines Variation in the Maize Kernel Ionome. *G3&amp;#58; Genes|Genomes|Genetics*, 6(December), pp.4175–4183. Available at: http://g3journal.org/cgi/doi/10.1534/g3.116.034827.

Badri, D. V. et al., 2007. Altered Profile of Secondary Metabolites in the Root Exudates of Arabidopsis ATP-Binding Cassette Transporter Mutants. *Plant Physiology*, 146(2), pp.762–771. Available at: http://www.plantphysiol.org/cgi/doi/10.1104/pp.107.109587.

Baxter, I., 2010. Ionomics: The functional genomics of elements. *Briefings in functional genomics*, 9(2), pp.149–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20081216 [Accessed October 29, 2012].

Baxter, I. & Dilkes, B.P., 2012. Elemental profiles reflect plant adaptations to the environment. *Science (New York, N.Y.)*, 336(6089), pp.1661–3. Available at: http://www.sciencemag.org/content/336/6089/1661.abstract [Accessed October 4, 2015].

Baxter, I.R. et al., 2014. Single-kernel ionomic profiles are highly heritable indicators of genetic and environmental influences on elemental accumulation in maize grain (Zea mays). *PLoS ONE*, 9(1).

Baxter, I.R. et al., 2008. The leaf ionome as a multivariable system to detect a plant’s physiological status. *Proceedings of the National Academy of Sciences of the United States of America*, 105(33), pp.12081–6. Available at: http://www.pnas.org/content/105/33/12081.abstract [Accessed October 2, 2015].

Bradbury, P.J. et al., 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics (Oxford, England)*, 23(19), pp.2633–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17586829 [Accessed July 12, 2014].

Buckler, E.S. et al., 2009. The genetic architecture of maize flowering time. *Science (New York, N.Y.)*, 325(5941), pp.714–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19661422 [Accessed October 29, 2012].

Bunyavanich, S. et al., 2014. Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. *BMC Medical Genomics*, 7(1), p.48.

Calabrese, G.M. et al., 2017. Integrating GWAS and Co-expression Network Data Identifies Bone Mineral Density Genes SPTBN1 and MARK3 and an Osteoblast Functional Module. *Cell Systems*, 4(1), p.46–59.e4.

Caldwell, K.S. et al., 2006. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, Hordeum vulgare. *Genetics*, 172(1), pp.557–567.

Castelletti, S. et al., 2014. A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL Vgt1. *G3 (Bethesda, Md.)*, 4(5), pp.805–12. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4025479&tool=pmcentrez&rendertype=abstract [Accessed September 15, 2015].

Chao, D.-Y. et al., 2011. Sphingolipids in the Root Play an Important Role in Regulating the Leaf Ionome in Arabidopsis thaliana. *The Plant Cell*, 23(3), pp.1061–1081. Available at: http://www.plantcell.org/cgi/doi/10.1105/tpc.110.079095.

Chia, J.-M. et al., 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics*, 44(7), pp.803–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22660545 [Accessed October 9, 2012].

Clark, R.M. et al., 2006. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nature genetics*, 38(5), pp.594–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16642024 [Accessed August 6, 2015].

Cook, J.P. et al., 2012. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant physiology*, 158(2), pp.824–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22135431 [Accessed October 5, 2012].

Davies, L. & Gather, U., 2012. The Identification of Multiple Outliers. *Journal of the American Statistical Association*. Available at: http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476339 [Accessed November 20, 2015].

Dongen, S. van, 2000. MCL: A Cluster Algoithm for Graphs.

Elshire, R.J. et al., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5), p.e19379. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Fan, J. et al., 2015. Arabidopsis TRIGALACTOSYLDIACYLGLYCEROL5 Interacts with TGD1, TGD2, and TGD4 to Facilitate Lipid Transfer from the Endoplasmic Reticulum to Plastids. *The Plant Cell*, 27(October), p.tpc.15.00394. Available at: http://www.plantcell.org/lookup/doi/10.1105/tpc.15.00394.

Fu, J. et al., 2016. A Tandem Array of *ent* -Kaurene Synthases in Maize with Roles in Gibberellin and More Specialized Metabolism. *Plant Physiology*, 170(2), pp.742–751.

Ghazalpour, A. et al., 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. G. Gibson, ed. *PLoS genetics*, 2(8), p.e130. Available at: http://dx.plos.org/10.1371/journal.pgen.0020130 [Accessed April 29, 2014].

Gore, M. a et al., 2009. A first-generation haplotype map of maize. *Science (New York, N.Y.)*, 326(5956), pp.1115–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19965431 [Accessed June 10, 2011].

Guerinot, M. Lou & Salt, D.E., 2017. Fortified Foods and Phytoremediation . Two Sides of the Same Coin 1. , 3755.

Harris, M. a et al., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, 32(Database issue), pp.D258-61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308770&tool=pmcentrez&rendertype=abstract [Accessed October 10, 2012].

Hirsch, C.N. et al., 2014. Insights into the maize pan-genome and pan-transcriptome. *The Plant cell*, 26(1), pp.121–35. Available at: http://www.plantcell.org/content/early/2014/01/31/tpc.113.119982.abstract [Accessed July 14, 2014].

Hung, H.-Y. et al., 2012. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, 108(5), pp.490–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3330692&tool=pmcentrez&rendertype=abstract [Accessed November 20, 2015].

Kim, D., 2015. Methods of integrating data to uncover genotype – phenotype interactions. *Nature Publishing Group*, 16(2), pp.85–97. Available at: http://dx.doi.org/10.1038/nrg3868.

Kump, K.L. et al., 2011. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics*, 43(2), pp.163–168. Available at: http://www.nature.com/doifinder/10.1038/ng.747.

Lawrence, C.J. et al., 2004. MaizeGDB, the community database for maize genetics and genomics. *Nucleic acids research*, 32(Database issue), pp.D393-7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308746&tool=pmcentrez&rendertype=abstract [Accessed October 30, 2012].

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–60. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Li, M. et al., 2008. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC bioinformatics*, 9(1), p.398. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-398 [Accessed April 28, 2016].

Lindgreen, S., 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes*, 5(1), p.337. Available at: http://www.biomedcentral.com/1756-0500/5/337 [Accessed September 4, 2015].

Louwers, M. et al., 2009. Tissue- and expression level-specific chromatin looping at maize b1 epialleles. *The Plant cell*, 21(3), pp.832–42. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2671708&tool=pmcentrez&rendertype=abstract [Accessed September 13, 2015].

McMullen, M.D. et al., 2009. Genetic properties of the maize nested association mapping population. *Science (New York, N.Y.)*, 325(5941), pp.737–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19661427 [Accessed October 29, 2012].

Mochida, K. et al., 2011. Global landscape of a co-expressed gene network in barley and its application to gene discovery in Triticeae crops. *Plant & cell physiology*, 52(5), pp.785–803. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3093127&tool=pmcentrez&rendertype=abstract [Accessed August 15, 2011].

Morrell, P.L. et al., 2005. Low levels of linkage disequilibrium in wild barley (Hordeum vulgare ssp. spontaneum) despite high rates of self-fertilization. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7), pp.2442–2447.

Obayashi, T. et al., 2014. ATTED-II in 2014: Evaluation of Gene Coexpression in Agriculturally Important Plants. *Plant and Cell Physiology*, 55(1), pp.e6–e6. Available at: http://pcp.oxfordjournals.org/cgi/doi/10.1093/pcp/pct178.

Ozaki, S. et al., 2010. Coexpression analysis of tomato genes and experimental verification of coordinated expression of genes found in a functionally enriched coexpression module. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 17(2), pp.105–16. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2853382&tool=pmcentrez&rendertype=abstract [Accessed April 16, 2016].

Peiffer, J.A. et al., 2014. The Genetic Architecture of Maize Height. *Genetics*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24514905 [Accessed March 19, 2014].

Roston, R.L. et al., 2012. TGD1, -2, and -3 proteins involved in lipid trafficking form ATP-binding cassette (ABC) transporter with multiple substrate-binding proteins. *Journal of Biological Chemistry*, 287(25), pp.21406–21415.

Sarkar, N.K., Kim, Y.-K. & Grover, A., 2014. Coexpression network analysis associated with call of rice seedlings for encountering heat stress. *Plant Molecular Biology*, 84(1–2), pp.125–143.

Schaefer, R.J. et al., 2014. Discovering functional modules across diverse maize transcriptomes using COB, the co-expression browser. *PLoS ONE*, 9(6).

Schaefer, R.J., Briskine, R., Springer, N.M. & Myers, C.L., 2014. Discovering functional modules across diverse maize transcriptomes using COB, the co-expression browser. *PLoS ONE*, 9(6), p.99193.

Schaefer, R.J., Briskine, R., Springer, N.M., Myers, C.C.L., et al., 2014. Discovering functional modules across diverse maize transcriptomes using COB, the co-expression browser F. Börnke, ed. *PLoS ONE*, 9(6), p.99193. Available at: http://dx.plos.org/10.1371/journal.pone.0099193 [Accessed June 7, 2016].

Schaefer, R.J., Michno, J.-M. & Myers, C.L., 2016. Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*.

Schubert, M. et al., 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature protocols*, 9(5), pp.1056–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24722405.

Stelpflug, S.C. et al., 2015. An expanded maize gene expression atlas based on RNA-sequencing and its use to explore root development. *The Plant Genome*, (608), pp.314–362.

Swanson-Wagner, R. et al., 2012. Reshaping of the maize transcriptome by domestication. *PNAS*, 109(29), pp.11878–11883. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1201961109 [Accessed June 7, 2016].

Taşan, M. et al., 2014. Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. , 12(2).

Tian, F. et al., 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics*, 43(2), pp.159–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21217756 [Accessed October 29, 2012].

USDA, 2016. *Crop Production 2015 Summary*,

Valdar, W. et al., 2009. Mapping in structured populations by resample model averaging. *Genetics*, 182(4), pp.1263–77. Available at: http://www.genetics.org/content/182/4/1263.long [Accessed August 6, 2015].

Wallace, J.G. et al., 2014. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS genetics*, 10(12), p.e1004845. Available at: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004845 [Accessed September 24, 2015].

Winkler, R.G. & Freeling, M., 1994. Physiological genetics of the dominant gibberellin-nonresponsive maize dwarfs, Dwart8 and Dwart9. *Planta*, 193, pp.341–348.

Wolfe, C.J., Kohane, I.S. & Butte, A.J., 2005. Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC bioinformatics*, 6, p.227. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1239911&tool=pmcentrez&rendertype=abstract [Accessed April 8, 2016].

Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics*, 8(3), pp.206–16. Available at: http://dx.doi.org/10.1038/nrg2063 [Accessed July 11, 2014].

Zheng, Z.-L. & Zhao, Y., 2013. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to “Candidatus Liberibacter asiaticus” infection. *BMC genomics*, 14(1), p.27.

# Supplementary Figures

## Supp. Fig. 1

### ZmPAN Network Health

Global network health of the maize PAN genome (ZmPAN) co-expression network. **(A)** Raw Pearson correlation coefficient distribution of all co-expression interactions. **(B)** Fisher transformed, variance stabilized and mean centered network interaction Z-score distributions. **(C)** Empirical density plot for genes in each GO term compared to the corresponding p-value derived from measuring density in 1000 random gene sets of the same size. **(D)** Degree distribution of ZmPAN genome co-expression network compared to power law, exponential, and truncated power law distributions.

## Supp. Fig. 2

### ZmSAM Network Health

Global network health of the maize ZmSAM co-expression network. **(A)** Raw Pearson correlation coefficient distribution of all co-expression interactions. **(B)** Variance stabilized and mean centered network interactions. **(C)** Empirical density plot for genes in each GO term compared to the corresponding p-value derived from measuring density in 1000 random gene sets of the same size. **(D)** Degree distribution of Tissue/Developmental co-expression network compared to power law, exponential, and truncated power law distributions.

## Supp. Fig. 3

### ZmRoot Network Health

Global network health of the maize ZmRoot co-expression network. **(A)** Raw Pearson correlation coefficient distribution of all co-expression interactions. **(B)** Variance stabilized and mean centered network interactions. **(C)** Empirical density plot for genes in each GO term compared to the corresponding p-value derived from measuring density in 1000 random gene sets of the same size. **(D)** Degree distribution of ZmRoot co-expression network compared to power law, exponential, and truncated power law distributions.

## Supp. Fig. 4

### MCR supplemental figure

Panel (A) shows that absolute number of GO terms that remain significantly co-expressed at varying levels of MCR in each network. Red curves show all GO terms with an initial co-expression p-value ≤ 0.05. Blue and purple curves show GO terms with either moderate or strong initial co-expression (at MCR=0). Panels (B-C) shows the percent and absolute number of GO terms that remain significantly co-expressed at varying levels of MCR. The red curves show small GO terms (50 ≤ n < 65), the blue curve shows medium sized GO terms (65 ≤ n < 80) and the purple curve shows large terms (80 ≤ n < 100).

## Supp. Fig. 5

### FCR supplemental figure

Panel (A) shows that absolute number of GO terms that remain significantly co-expressed at varying levels of FCR in each network. Red curves show all GO terms with an initial co-expression p-value ≤ 0.05. Blue and purple curves show GO terms with either moderate or strong initial co-expression. Panels (B-C) shows the percent and absolute number of GO terms that remain significantly co-expressed at varying levels of FCR. The red curves show small GO terms (50 ≤ n < 65), the blue curve shows medium sized GO terms (65 ≤ n < 80) and the purple curve shows large terms (80 ≤ n < 100).

## Supp. Figure 6

### Histograms of Pearson correlation coefficients between gene specific density and locality

Pearson correlation was measured between gene specific density and locality in each network. PCCs between metrics were compared by grouping sets of genes in either Ionome elements (e.g. Al, Fe, etc.) or GO Terms at 50, 100 and 500 kb Window Size and 1,2, and 5 gene flank limits.

# Supplementary Files

## Supp. File 1

### Quality control and Co-expression networks overview

This file contains log information, quality control parameters, and network build parameters for gene co-expression networks.

# Supplementary Tables

## Supp. Table 1

### Full gene ontology term density and locality p-values

Density and locality scores were measured between genes within each GO Term. Subnetwork p-values were generated for both density and locality by comparing each term’s metric to n=1000 randomized gene sets of the same size.

## Supp. Table 2

### Network MCL cluster gene assignments

Clusters in all three networks were identified using the MCL algorithm. Genes in each network were assigned to cluster IDs. Lower cluster IDs have a larger number of genes.

## Supp. Table 3

### Network signal of GO terms with various levels of MCR/FCR.

Co-expression among co-annotated GO terms genes was compared to random gene sets of the same size to generate a p-values. Noise was introduced by varying the missing candidate rate (MCR) or false candidate rate (FCR). Missing candidates were removed in proportion to the values in the table, while false candidates were introduced using SNP-to-gene mapping values (see WindowSize and FlankLimit). FCR values are reported as averages across 10 percent quantiles (see Fig. 5).

## Supp. Table 4

### Maize Grain Ionome GWAS Network Overlap Candidate Genes

Candidate genes were identified by extracting candidate genes in each co-expression network (ZmSAM, ZmPAN or ZmRoot) using SNP-to-gene mapping for each Element (using WindowSize and FlankLimit). Co-expression (density or locality) among all genes within a subnetwork was compared to randomized gene sets of the same size to establish p-values. Gene specific z-scores were computed by comparing the empirical gene-specific density (Eq.3) or locality (Eq.4) to the average density or locality observed in randomized gene sets, then correcting for standard deviation. False discovery rates (FDR) were calculated for candidate genes with positive gene-specific co-expression values by comparing the number of genes discovered at a z-score cutoff to the average number of genes discovered in randomized sets.

## Supp. Table 5

### Maize Grain Ionome GWAS High Priority Overlap (HPO) Candidate Genes

High priority overlap (HPO) genes were identified by calculating gene-specific density or locality (Method column) for each element at different SNP-to-gene mapping parameters (see WindowSize and FlankLimit columns). At a FDR cutoff of 30%, genes were defined as HPO if they were observed at two or more SNP-to-gene mapping parameters.

## Supp. Table 6

### HPO genes discovered with networks built from accessions subsets

The number of HPO genes discovered in full ZmPAN (503 accessions) and ZmRoot (46 accessions) networks were compared to networks built with a subset of accessions. Both ZmPAN and ZmRoot networks were re-built using a common set of 20 accessions. The ZmPAN network was re-built using 46 accessions consisting of the 20 common accessions and either 26 random or 26 CML biases accessions to simulate the number used in the full 46 accession ZmRoot network. Each network analyzed for HPO genes in the 17 GWAS elements.

## Supp. Table 7

### Multiple Element HPO gene list

The number commonly discovered HPO genes, hypergeometric p-values of set overlap, and GRMZM IDs across multiple elements.

## Supp. Table 8

### Element Gene Ontology Enrichment

HPO genes for each element were tested for enrichment among genes co-annotated for Gene Ontology (GO) terms (hypergeometric test). Bonferroni correction is included as a column, treating each GO term as an independent test.

## Supp. Table 9

### HPO plus neighbors Gene Ontology Enrichment

Elemental HPO gene sets were supplemented with an additional set of highly connected neighbors equal to the number of genes in the HPO set. These HPO+ gene sets were tested for enrichment among genes annotated for Gene Ontology terms (hypergeometric test).