5-2013

PREDICTING COMPLEX PHENOTYPE-GENOTYPE RELATIONSHIPS IN GRASSES: A SYSTEMS GENETICS APPROACH

Stephen Ficklin
Clemson University, spficklin@gmail.com

PREDICTING COMPLEX PHENOTYPE-GENOTYPE RELATIONSHIPS IN GRASSES: A SYSTEMS GENETICS APPROACH

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Plant & Environmental Sciences

by
Stephen Patrick Ficklin
May 2013

Accepted by:
Dr. F. Alex Feltus, Committee Chair
Dr. Douglas G. Bielenberg
Dr. Feng Luo
Dr. Hong Luo

Abstract

It is becoming increasingly urgent to identify and understand the mechanisms underlying complex traits. Expected increases in the human population coupled with climate change make this especially urgent for grasses in the *Poaceae* family because these serve as major staples of the human and livestock diets worldwide. In particular, *Oryza sativa* (rice), *Triticum spp.* (wheat), *Zea mays* (maize), and *Saccharum spp.* (sugarcane) are among the top agricultural commodities. Molecular marker tools such as linkage-based Quantitative Trait Loci (QTL) mapping, Genome-Wide Association Studies (GWAS), Multiple Marker Assisted Selection (MMAS), and Genome Selection (GS) techniques offer promise for understanding the mechanisms behind complex traits and to improve breeding programs. These methods have shown some success. Often, however, they cannot identify the causal genes underlying traits nor the biological context in which those genes function. To improve our understanding of complex traits as well improve breeding techniques, additional tools are needed to augment existing methods. This work proposes a knowledge-independent systems-genetic paradigm that integrates results from genetic studies such as QTL mapping, GWAS and mutational insertion lines such as *Tos*17 with gene co-expression networks for grasses—in particular for rice. The techniques described herein attempt to overcome the bias of limited human knowledge by relying solely on the underlying signals within the data to capture a holistic representation of gene interactions for a species. Through integration of gene co-expression networks with genetic signal, modules of genes can be identified with potential effect for a given trait, and the biological function of those interacting genes can be determined.

Acknowledgements

Special thanks are extended to my advisor Dr. F. Alex Feltus for his training, teaching, patience, flexibility and friendship. He has been a wonderful advisor and supported me despite being the atypical student with a full time job and with a large family to support. I would also like to thank the members of my PhD committee, Drs. Doug Bielenberg, Feng Luo and Hong Luo for their willingness to meet, discuss my progress and share their thoughts. This also includes Dr. Bert Abbott who served on the committee during the first few years.

A great amount of gratitude and love is expressed to my wife Beth, who supported my efforts, provided time for me to study and work, and listened to my scientific babblings. Without her support I would not have had this opportunity.

Additional thanks go to my current employer Dr. Dorrie Main. Without her kindness and support in granting a flexible schedule I would not have been able to complete the PhD work. The same thanks are extended to Dr. Richard Hilderman and Dr. Keith Murphy who, during my time at the Clemson University Genomics Institute afforded the same flexibility.

A big thank you is given to the other students in the Feltus Lab, Drs. Josh Vandenbrink and Jacob Spangler who provided comments and insight into this work. I wish them both the best in their future endeavors.

Table of Contents

Page

Table of Contents (Continued)

Table of Contents (Continued)

Page

Table of Contents (Continued)

Page

List of Tables

List of Figures

List of Figures (Continued)

1. Introduction

The Importance of Grasses

The *Poaceae* family of grasses is comprised of flowering monocotyledons and is one of the largest plant families found throughout the world. The family contains cereals such as wheat, maize, rice, sorghum and sugarcane which are some of the most agriculturally and economically important plant species in the world. According to the United Nations Food and Agricultural Organization commodities production statistics of 2011 [1], *Poaceae* species such as sugarcane, maize, rice and wheat comprise the world's most highly produced crops and are also within the top 10 economically valuable crops on the planet (Figure 1.1).

Improvements in nutritional value, grain and total biomass yield, cultivation range, disease resistance, stress tolerance and other complex traits for the cereal crops is important to meet the demands of a growing world population, their livestock, and adaptation in a future of global climate change. Approximately, 1 billion people worldwide are currently undernourished—mostly in underdeveloped countries [2]. Despite major advances in food production owing largely to improvements in irrigation, fertilization, pest management and development of higher yielding crop varieties (the Green Revolution [3]), the estimated pace of worldwide population increase is expected to outpace agricultural production before the year 2050 [2]. Therefore, a higher rate of productivity may be necessary to avoid famines, malnourishment and geo-political turmoil caused by food shortages. Crop varieties that require less water, fewer pesticides and fewer fertilizers are most desired to meet the future global need especially in light of changing climate conditions. Development of new tools to speed and improve efficiency of breeding is becoming urgent.

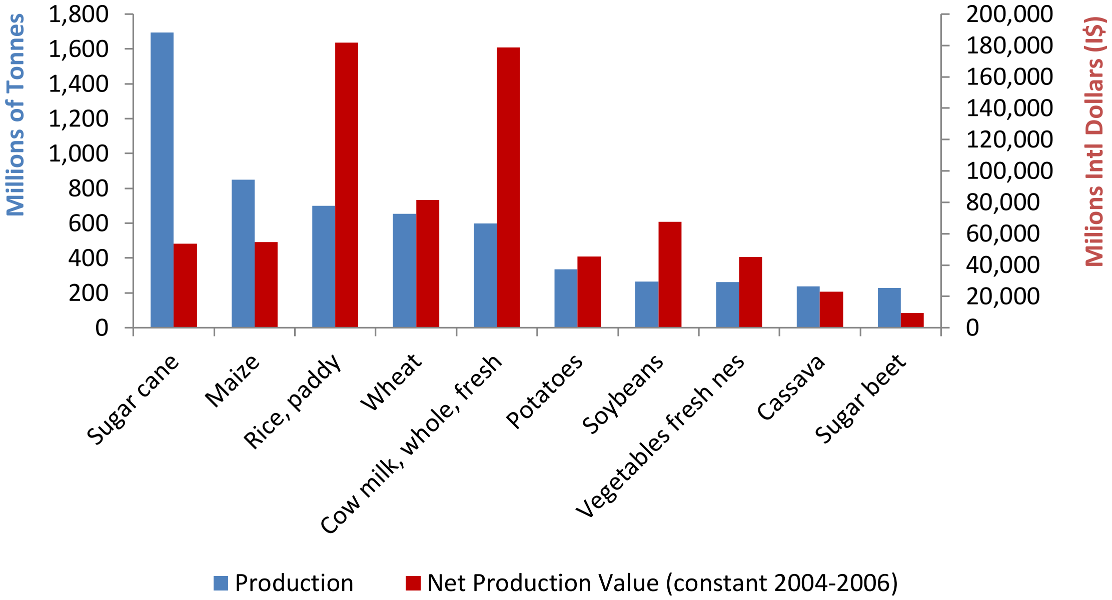


Figure 1.1 The Top 10 Food Commodities.

The total production in tonnage and economic value in millions of international dollars (I$) of the top 10 food commodities as reported by the United Nations Food and Agricultural Organization in 2011 [1] (chart reproduced using publicly available data).

<u>Rice as a Model for the Grasses</u>

Oryza sativa, rice, is a major food commodity in both developed and developing nations. Moreover, rice serves as a major staple food in nations with some of the highest number of poor, particularly in Asian countries (Figure 1.2). Therefore, high productivity and nutrition of rice is essential to maintain adequate food supplies and prices.

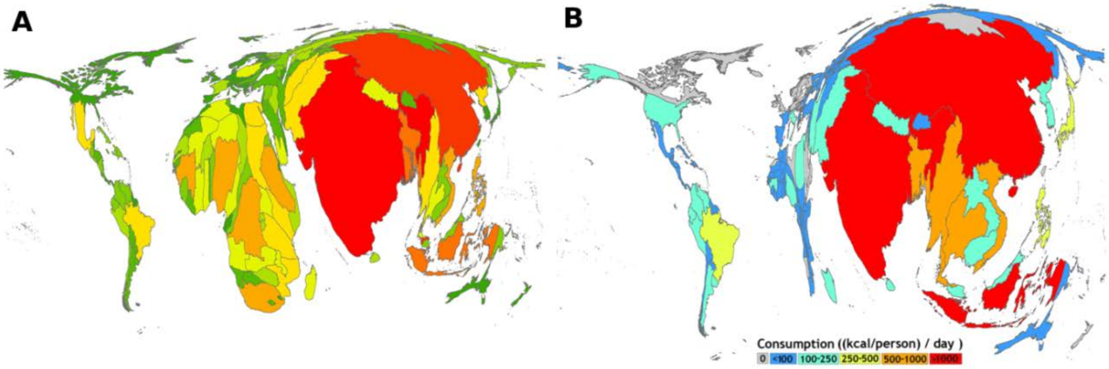


Figure 1.2 Cartograms Representing Poverty and Rice Consumption.

The cartograms shown here visualize a country's size using a specific value rather than land area and allow for gross comparisons between nations. A) Country size indicates the number of people living on less than 1 US dollar (purchasing power adjusted) per day in 2008. B) The amount of rice consumption per country [4, 5]**. (**Both figures A and B are reproduced here under permission of the Creative Commons License. Figures were initially found on the IRRI (http://www.irri.org) website in 2010. Figure 1B was published in *Rice Today* [4])

Beyond the agricultural and economic importance, rice also serves as a model species for investigating complex traits in grasses because of the small diploid genome and wealth of biological resources. These include a well characterized whole genome sequence [6], over 100 genetic and physical maps, 2.5 million molecular markers, over 8,000 QTLs, mutant lines and more [7]. Additionally a wealth of gene expression measurements from both microarray and RNA-seq technologies are available in public repositories such as the NCBI Gene Expression Omnibus (GEO; [8]), Short Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra/), ArrayExpress from the European Bioinformatics Institute [9], and other sources.

It has been shown that gene content and order among the *Poaceae* are highly conserved [10]. Therefore, discovery of gene sets underlying agronomic traits in one grass species should add to the biological knowledge of other grasses. Translation of knowledge between grass species becomes increasingly important for species such as sugarcane which has a large complex polyploid genome and where almost all traits are complex [11]. Due to the resources available for rice, it serves as a logical initiation point for translational studies within the grasses.

Methods for Identifying Genes of Complex Traits

Understanding the genetic mechanisms underlying desirable agricultural traits may prove essential for improvements in productivity and nutrition. Desirable traits often arise from the interactions of alleles from multiple genes—they are polygenic. Complex traits are thus quantitative because the interaction of gene products yields a range of variability in expression. This is opposed to Mendelian traits where phenotypes are qualitative and the result of interacting alleles from a single gene. Examples of important complex traits include disease resistance, nutrition, stress tolerance, plant height and yield. Most recent efforts to identify genes underlying complex traits in rice are concentrated in the techniques of linkage-based Quantitative Trait Loci (QTL) mapping and Genome Wide Association Studies (GWAS). Both methods focus on the use of molecular markers and the accuracy of these methods is highly dependent on the degree of linkage between genes underlying complex traits with the markers. Once markers have been associated with a trait, genes nearby the marker may be identified through methods such as positional cloning, or closer examination of genes in a whole genome assembly if one is available.

QTL mapping is a technique that attempts to locate chromosomal regions that contribute some effect towards a specific trait (QTL set). Traditional linkage-based QTL mapping identifies QTLs using molecular markers, such as SNPs, RFLPs or SSRs, that have a non-random association with the measured trait within a segregating mapping population [12, 13]. The positions of markers on a linkage map are used to estimate the positions of QTLs. There are a variety of methods for obtaining a segregating population needed for QTL mapping. Typically, a population is constructed from a cross between two homozygous, inbred, individuals that demonstrate genetic variability in the set of markers to be used. The first filial, F1, generation of these parental lines are all heterozygous at variant nucleotides which then undergo a crossing schema such as a backcross (cross with one of the parents) to produce a BC1 generation, a sibling cross (cross with another F1 individual) or selfing (F1 crossed with itself) both of which produce a second filial, F2, generation. The progeny from this second round of crosses can then be genotyped and phenotyped for QTL analysis or inbred further to create recombinant inbred lines (RILs) [14] which can then be used for higher resolution QTL analysis. Additionally, in order to increase genetic variability in the mapping population, multiple parental lines can be crossed and their progeny intercrossed and inbred until a population of highly homozygous inbred lines are constructed. Such methods have been used to construct the population of mice known as the Collaborative Cross [15] and the Nested Association Mapping population (NAM) for maize [16].

Currently, for rice alone, over 8,000 QTLs from over 300 different mapping experiments for over 300 different traits are currently available in Gramene [17]. These QTLs have been mapped to physical genomic coordinates by the Gramene curators, allowing for the integration of functional genomic data. As will be discussed, this rich genetic resource makes rice a powerful model for gene discovery in grasses.

QTLs from linkage-based mapping currently present several challenges [18]. First, mapping populations only contain modest recombination histories, therefore the resolution is often quite rough (e.g. 10-30 cM for maize) [19]. Thus, QTLs often specify genomic regions that contain hundreds to thousands of genes, many of which may not contribute to the trait, and makes identification of causal genes difficult. Also, QTLs are identified using the combinations of only two allele sets (one from each parent) at any given locus, thus the full genetic variability of the natural or breeding population may not be accounted for. QTL mapping has been successful at identifying genes that contribute to a trait of interest but these are usually genes with large effects that are highly heritable (low environmental effects) [20, 21]. As a result, QTLs from QTL mapping have yet to be largely useful for breeders [22, 23] especially for traits with potentially hundreds of genes with small effects and lower heritability.

GWAS studies are similar to QTL mapping in that they identify QTLs that non-randomly segregate with a trait. However, GWAS involve significantly more molecular markers—usually thousands to millions of polymorphic SNPs. The mapping population also includes higher genetic and phenotypic variability through inclusion of thousands of individuals across a natural or breeding population with distinct families and identifiable subpopulation structure. Thus a rich recombination history is often present within the population. This is in contrast to linkage-based QTL mapping where the population typically consists of recombinant progeny of two inbred parents. For GWAS, the level of recombination present in the population is estimated through linkage disequilibrium (LD) analysis. A set of markers with a set of alleles that consistently appear together within the population are in LD. Groups of adjacent markers in high LD form LD blocks, and the average size of LD blocks (e.g. 100-300kb in rice [24]) in a population help identify the number of markers needed for the study as well as the possible resolution of the study [18]. This is in contrast to linkage-based QTL mapping where the level of recombination is typically a product of the crossing strategy.

Recently, three GWAS studies in rice have been performed [24-26]. The first by Huang et.al. (2010) studied 517 Chinese land races for 14 different agronomic traits [25]. They performed 1x Illumina re-sequencing of each accession to identify millions of SNPs. Results showed that on average SNPs that significantly co-segregate with traits account for 36% of total phenotypic variability. This same group later sequenced an additional 433 lines (2012) including other global varieties measuring two traits: grain yield and flower time [26]. A second group, Zhao et. al. (2011), performed GWAS analysis of 413 accessions from 82 countries for 34 different traits using a 44,100 SNP array [24]. Both studies report SNPs with significant association to the traits under study.

GWAS also has its own set of challenges. For candidate genes identified through GWAS there often remains the challenge to confirm that implicated genes do in fact contribute to the trait. Also the underlying genetic context within which candidate genes function is not known [27-29]. For example, the regions derived from GWAS studies are many times found in introns and intergenic regions, indicating involvement of regulatory mechanisms. It has been reported that only 12% of SNPs associated with traits in humans are in linkage disequilibrium with protein-coding genes [30]. Perhaps the major challenge however, is that regions identified by GWAS typically only account for a small percentage of the total heritability of a trait such as in the case of human height where a study involving 30000 people identified 40 new loci that only accounted for 5% of the difference in human height [31]. Thus, while GWAS provides improved resolution and can implicate large-effect and some smaller effect genes, it fails to implicate causal genes across the full trait range of expression, and in many cases even across a moderate range of expression. While new, GWAS has had little success in helping with discovery of gene targets for diseases [32] and in plant breeding [33].

Breeding Tools for Complex Traits

In order to meet the demands of a growing population with perhaps less optimal conditions for growing crops, new varieties of cereals are needed. Therefore, breeders are actively searching for new varieties with improvements in water usage, nutrition, yield, disease tolerance and other agronomically important traits. However, the population size needed for current breeding methods can be enormous due to the number of segregating genes, requiring hundreds to thousands of populations and can take from 5 to 10 years to identify an elite variety for rice alone [34]. Therefore, breeders hope to capitalize on recent advancements in molecular technologies to decrease developmental time and the number of required populations. There are two major molecular-based technologies that have been used for assistance in breeding programs of both plants and animals. These include Marker Assisted Selection (MAS) and Genome Selection (GS). Both methods rely on markers with linkage disequilibrium to genes underlying complex traits to assist in breeding.

Marker Assisted Selection (MAS) employ the use of molecular markers such as SNPs, SSRs, AFLPs, RFLPs, etc., to identify individuals with potentially desirable traits [35]. Markers for complex traits are initially identified using methods such as QTL mapping as described previously. The efficacy of MAS is limited to the power of QTL mapping. For example, MAS is useful when the trait heritability is high (low environmental effect), when there are very few QTLs and the population is large (i.e. the trait is controlled by few large-effect genes), and when gene or genes underlying traits are in high linkage disequilibrium to the marker [34]. For traits with potentially hundreds of genes with small effects, MAS becomes woefully inadequate. Unfortunately, there have been very few published results for MAS in plant breeding. Hospital provides a few in a recent review [36], including Patwan, a hard white spring wheat, a rice line with greater resistance to bacterial blight and a few others. In a review by Collard and Mackill, they suggest the lack of published results are because the technology is too new, breeders and large breeding companies release varieties rather than publish the results of MAS techniques, that QTLs in general may be too large to be effective, differences in genetic backgrounds between breeding population and QTL populations make markers from QTL mapping less useful, and other reasons [35].

Genome Selection (GS) is a newer technology that utilizes a large marker data set such as high density SNPs across an entire genome and offers the promise of identifying gene sets with small effects underlying complex traits [37]. The method employs genotyping of large reference breeding population representative of the genetic diversity of the larger population. Genotyping is performed using high density SNP assays and relevant phenotypes are scored across various environmental conditions. Using various statistical methods [33] all SNPs are evaluated at once to identify a set of markers that together represent an approximation of the loci controlling the total heritability of the trait. This is in contrast to MAS where each marker is evaluated individually for the trait. The technique also generates a predictive equation that when evaluated in a separate validation population using genotypes of individuals as input, Breeding Value (BV) are generated for the trait for each individual. Breeding values are an estimate for the amount of heritability in an individual for a specific trait, and individuals with higher BVs are often used when breeding for specific traits. BVs have been used in traditional breeding but are normally generated using carefully recorded pedigree and phenotyping information. After evaluation of the GS derived BVs in the validation population, the predictive equations and SNP sets can be used in future selection population with the objective of reducing phenotyping. It has been shown that use of GS can improve the accuracy of BVs by as much as 167% over MAS, 2-30% in mice using empirical studies and up to 85% in cattle over traditional methods [38]. GS is a new method and still under validation but does seem to offer much promise especially for complex traits with many small effects. However, while successful for breeding programs, the genes underlying complex traits are not identified, nor the context under which these genes operate. Methods that can identify the genes (not just closely linked markers) underlying complex traits would be of great value to understanding the molecular mechanisms of complex traits which in turn could further assist breeders.

Systems Genetics as a Technique for Exploring Complex Traits

The methods and tools discussed previously have helped advance our understanding of complex traits and have been largely successful for traits with a few large-effect genes and high heritability. GS in particular offers great potential for identifying markers associated with small-effect genes and affords improvements to breeders. However, in all of these methods, the genes underlying complex traits are yet to be systematically identified. The interactions between and functions of the relevant gene products are also not known. Additional technologies are needed to help identify these genes and the context by which they interact. Integration of genomics, genetics and systems biology methods into a category known as systems genetics, offers the potential to provide this information.

The goal of systems biology is to use the principles of systems theory to examine interactions of a biological system. Systems theory has been used in disciplines such as engineering [39], ecology [40], and the social sciences [41], to name a few, but is becoming more widespread in biology with the advent of high-throughput assays and sequencing. Also, complex traits are polygenic and genes may at times be multi-functional [42-44]. Because genes do not function alone, an understanding of the genes in isolation, or in isolated pathways does not provide the context for how the gene behaves in the entire system. With the potential for hundreds to potentially thousands of genes working together to yield a physiological or morphological trait the network of interactions should be viewed at a higher "systems" level.

Systems genetics is a subclass of systems biology that integrates the methods of genetics, genomics and systems biology to unravel genotype-phenotype interactions [45]. Systems genetic approaches are relatively new and under active development. They have recently been used to examine the genetic mechanisms behind complex traits in *Drosophila melanogaster* [46, 47]; diseases such as osteoporosis and type 2 diabetes in humans [28, 48]; fear, cancer, HDL-cholesterol, defining genetic interactions in the thalamus, and applications in somatic cell cloning in mice [49-53]; various traits for

Oryza sativa [54]; and Arabidopsis [55, 56] to name a few. Additionally, systems genetics approaches have been suggested for applications in plant and animal breeding [57], disease and drug target discovery [27, 32], and, in-general, understanding of molecular mechanisms underlying complex traits [58].

The "systems" component of a systems genetics study involves the use of networks to model gene and genetic relationships. Networks are comprised of nodes and edges—similar to a road map where nodes are towns or cities and edges are roads connecting them. Examples of networks with nodes and edges can be seen in Figure 1.3 Aa, Bb and Cc. For systems biology, networks can be physical interaction networks where edges represent proteins binding to other proteins (e.g. protein-protein interaction (PPI)) [59], where edges describe co-expression of gene transcripts (co-expression networks) [60], metabolic networks linking metabolites (substrate and products) and reactions [61], or regulatory linking transcription factors and genes they regulate [62]. Biological networks may be the key to discovering specific gene sets and the molecular mechanisms underlying complex traits.

While the type of network may be different, all naturally occurring networks typically exhibit a similar set of properties. These properties include scale-free behavior, small-world, hierarchical and modular structure. These properties are present in many networks such as social networks, cell phone networks, the world-wide web, and biological networks [63-65]. These properties are shown in Figure 1.3 and briefly described here. In network theory, the degree of a node, *k*, is the number of incident edges to the node. A scale-free network is therefore one where the probability distribution, P(*k*), of a node being connected via *k* edges decreases exponentially with increasing *k*. Figure 1.3 shows that random networks do not exhibit scale-free behavior but rather have a Gaussian distribution. Scale-free networks therefore consist of some nodes that are highly connected, called hubs, with most nodes having few connections. As a result of scale-free behavior, networks are typically small-world such that a path between any two nodes in the network traverses only a few other nodes (i.e. six degrees of separation principle). Networks are also modular and those modules are often hierarchical. The clustering co-efficient, *C*, is a property of networks that describes the modularity within it. In modular networks, nodes that are connected tend to be connected to their neighbors as well. A module within a network is a collection of nodes that are more highly connected amongst themselves then with other nodes. These nodes are arranged hierarchically such that interconnections between the modules occur primarily through the hubs [65]. Another property of modular networks is that the distribution of the average clustering co-efficient C(*k*), of any node having degree *k* is non-linear and decreases logarithmically as *k* increases. As seen in Figure 1.3, scale free networks are not required to be modular. However, biological networks do tend to be modular and hierarchical [63].

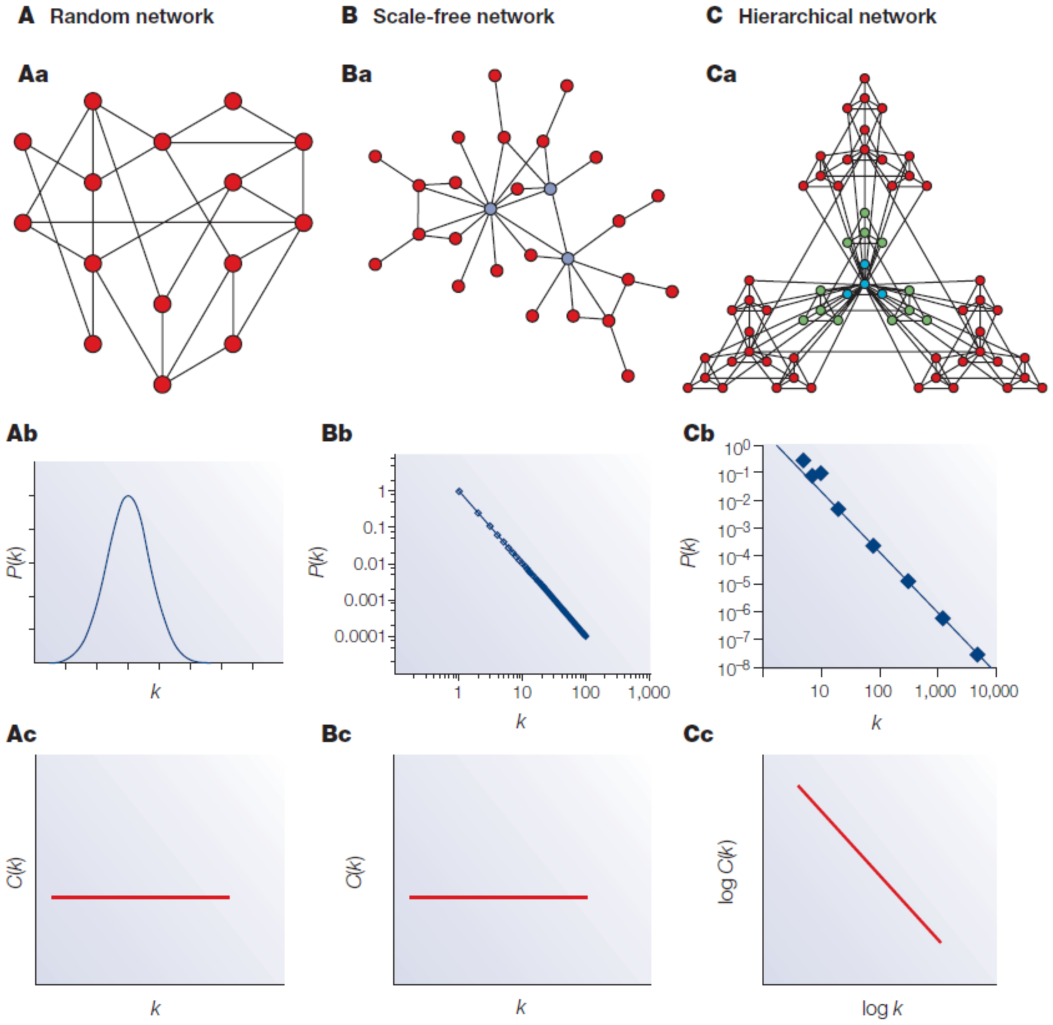


Figure 1.3 Properties of Naturally Occurring Networks.

This figure demonstrates three primary properties of naturally occurring networks, namely scale-free behavior, modularity and a hierarchical order. Column A demonstrates a random network (Aa) where the average degree of any node follows a Gaussian distribution (Ab) with no change in the average clustering coefficient (Ac). Column B demonstrates a scale-free network (Ba) such that the probability degree distribution P(*k*) is logarithmic and decreases with increasing *k* (Bb). Scale-free networks may not exhibit modularity as shown by no change in the average clustering coefficient (Bc). Column C shows a scale-free, modular and hierarchical network (Cc). Both the P(*k*) and C(*k*) are non-linear and decrease logarithmically as *k* increases. This figure has been re-used with permission, see Appendix B [63].

Gene Co-expression Networks

Gene co-expression networks (GCN), also known as transcription networks or relevance networks, are commonly used in systems genetics studies [29, 55, 56, 66, 67]. These networks are constructed using high-throughput expression measurements from microarrays or RNA-seq. A *n* x *n* similarity matrix is constructed using a pair-wise correlation statistic for each pair of genes typically using Pearson correlation, Spearman's [68] or Mutual Information [69]. After calculation of the *n* x *n* similarity matrix, values below a certain threshold are set to zero resulting in the formation of an adjacency matrix. Many methods have been employed for significance thresholding. These include *ad hoc* methods [70-73], permutation testing [74], linear regression [75], rank-based methods [76, 77], Fisher's test of homogeneity [78], spectral graph theory [79], Random Matrix Theory (RMT) [80, 81], Partial Correlation and Information Theory (PCIT) [82], methods that use topological properties [83], and supervised machine learning [84, 85]. The non-zero values of the adjacency matrix that remain after significance thresholding represent the co-expression network.

Next, module detection is performed after network construction. Modules are the basis for analysis of gene co-expression networks. It is through modules that functional units are defined. Genes in modules tend to cooperate in the same biological function and hence guilt-by-association inferences can be made such that genes of unknown function can be hypothesized to be involved with the ascribed function assigned to the module [77]. Because genes that appear together in modules tend to cooperate in the same biological functions, identification of modules with associations to complex traits automatically provides a set of genes, through guilt-by-association, that are potentially causal for the trait.

There are a variety of module detection algorithms including the weighted gene co-expression network analysis package WGCNA [86], link communities [87], Markov clustering [88], Affinity search methods [89], NeMo [90], and MCODE [91] to name a few. Each method uses a different approach and can define modules differently. Changing of input parameters with each algorithm can also affect module size and inclusivity. Therefore, the modules obtained from any of these methods are simply approximations to circumscribe genes that participate in similar function. In reality, real modules that underlie complex traits are probably dynamic and change with different conditions (genotype, environment, etc.). The specificity and sensitivity of module discovery for any given function or trait needs further exploration. However, it is only through modules detected using methods as those listed above that the power of networks is afforded.

As mentioned previously, functional analysis of network modules can be performed to qualify functional processes to which they contribute. Tools such as DAVID [92, 93], EasyGO [94], GOstat [95], FatiGO+ [96], Blast2GO [97], to name a few, are available to identify functional terms that are significantly present in modules more so than in the genomic background. Terms from databases such as Gene Ontology (GO) [98], KEGG pathways [99], Interpro protein domains [100], Plant Ontology (PO) terms [101], AraCyc pathways [102], and others, can all be used as needed for functional enrichment analysis.

Examples with Integration of Genetic Resources and Networks

Armed with networks and functionally annotated modules, data from genetic studies can be integrated to form systems genetics frameworks in an effort to diagnose genotype-phenotype associations. These genetic signals, when mapped to the physical genome, can be derived from QTL Mapping, GWAS, segregating populations with fixed genotypes [103], or mutational insertion lines such as T-DNA or *Tos*17 [104-107]. For example, Ayroles et. al. [103] used 40 inbred lines from *Drosophila melanogaster* which accounted for a large portion of trait variation for six traits being studied. They measured gene expression levels using the Affymetrix *Drosophila* 2.0 platform for flies under various conditions and scored phenotypes of resistance to starvation stress, time to recover from a chill-induced coma, life span, a startle-induced locomotor response and mating speed. Single feature polymorphisms (SFPs) [108] were identified using the microarray results and used to identify genes with polymorphisms between the 40 lines (a total of 3,316 genes). Network modules were constructed from co-expressed transcripts and statistical regression models were used to associate transcripts with phenotypes. They found that several hundred genes within modules were significantly associated with phenotypic variation and that 70% of insertional mutants within the candidate genes did have an effect on the phenotype. Recommendation for future directions included further testing with a higher number of polymorphisms from whole genome sequencing.

As another systems genetic example, polymorphisms from known segregating lines were used to find expression QTLs (eQTLs) which were then integrated with a co-expression network. An eQTL is a genomic region containing polymorphisms that are statistically associated with expression levels of specific genes. Here the trait being examined is the level of gene expression. The implication then is that eQTLs are likely to regulate expression of the gene [109]. Kang et. al. [110] integrated eQTLs with gene co-expression networks in an attempt to identify genes associated with Type 2 diabetes in mice. They selected a set of SNPs that were significant in several recent GWAS studies for Type 2 diabetes. Using an existing database of eQTLs for those SNPs and co-expression networks constructed from over 1000 obese patients, they identified genes that co-localized within co-expression modules that contained genes associated with Type 2 diabetes eQTLs. The study was able to provide several novel candidate genes for Type 2 diabetes.

As another example, the AraNet database integrated a diverse set of 24 data types for *Arabidopsis thaliana* including PPI networks, gene co-expression, metrics for similarity of protein domains and phylogenetic profiles, and gene-gene associations identified from literature mining [56]. These data types were integrated into a single network, called AraNet. Using guilt-by-association inferences they were able to identify modules for seed pigmentation that contained about 200 candidate genes. T-DNA insertional mutants were obtained for these 200 candidate genes, lines were grown and seed pigmentation was scored. Results showed a 10-fold improvement in expression of mutant seed pigmentation over random screens.

Another example, which is presented as Chapter 2 of this dissertation, involves the integration of a gene co-expression network for rice with results from a recent *Tos*17 insertional mutant study [54, 104, 105]. Briefly, a network was constructed using several hundred publically available microarrays and thresholded using the RMT method [80]. Modules were identified using the WGCNA method [86] and functionally annotated using an in-house script for identifying functionally enriched terms. Genes were annotated with phenotypes from the *Tos*17 study and modules were then tested for enrichment of these phenotypes. Modules with significant enrichment contained genes with potential effect towards expression of the trait. Chapter 3 describes an effort to extend this systems genetics approach by examining the translational impact of such studies between closely related species such as maize and rice [66].

A Holistic Approach to Systems Genetics

The systems genetic examples described previously demonstrate the diversity of genomic and genetic resources, as well as systems biology methods that can be successfully applied to identify potential candidate genes underlying complex traits. One challenge however with many systems genetics studies is the effect of experimental bias, such as where variables are tightly controlled: expression profiling and phenotype scoring are taken from the same sample and other experimental conditions are controlled.

As mentioned previously, QTL mapping, GWAS and GS can identify loci but they provide no context for the pathways underlying the regions they identify. In many cases identification of the exact genes is difficult, but also, how do the genes underlying the loci interact to express the trait? What effects do they have in other processes? What other phenotypes might they affect? The network can help provide clues to the genetic interactions at the functional genomic level. But, if the networks are constrained by experimental conditions it is not possible to grasp the full breadth of the interactions, rather, we obtain a glimpse for how the network behaves given a specific set of conditions. Furthermore, interactions cannot be captured if they were never measured such as brief expression of causal genes in a spatially discrete developmental context.

It would be desirable to have a framework where all interactions between any two genes are captured, not just those measured under the experimental conditions. A holistic network that provides relationships across as many conditions as possible (e.g. genotypes, developmental stages, tissues, and environmental conditions) would allow for such a perspective. While we cannot perform expression profiling across all points-in-time for every tissue type, genotype, cell and environmental condition, we can attempt to remove bias and approximate the holistic, multi-dimensional network by incorporating all available data and using knowledge-independent approaches.

Summary

This dissertation, therefore, is an exploration into the possibility of constructing a holistic interaction network in the form of a gene co-expression network and its applicability for understanding genotype-phenotype relationships. The hypothesis is that genotype-phenotype relationships can be identified in holistic networks, and that the context of the interactions can be understood such that side-effects from selecting for a specific genotype can be detected. *Oryza sativa* was selected for this study due to its importance in human nutrition, its agricultural and economic value, and the availability of a wide array of genomic and genetic resources that serve as a reference for translational agriculture to other grasses.

In Chapter 2, this dissertation describes the construction of the first global gene co-expression network for rice using a knowledge-independent approach [54]. This network was composed of all publicly available samples of the Affymetix Rice Genome array obtained from NCBI's Gene Expression Omnibus (GEO; [8]). Samples were not segregated by conditions (no knowledge bias in samples used for correlations) and thresholding of the network used a statistical test to prune low-quality relationships, rather than bait-genes or functional lists. The input samples spanned a variety of genotypes, tissues, environmental conditions, disease states and developmental stages. All samples were used to build the global network in an effort to capture as many interactions across conditions as possible. Additionally, associations between genes and phenotypes provided by the *Tos*17 mutational insertion database [104] were used to identify modules that were enriched for specific phenotypes. Results showed that the enriched function of modules in the global network matched well with expected condition. For example, modules containing processes involved in seed storage were most highly expressed in seed tissue. Additionally, there was significant enrichment of mutant phenotypes in some of the rice modules, implicating those genes in the enriched trait.

Chapter 3 examines the feasibility for using network modules from one species to predict modules with similar function in another species. A global co-expression network for *Zea mays* was constructed using the same methodology as for the global rice network [66]. The two networks were compared to identify modules with similar homology amongst member genes and topology (similar patterns of interconnections among member genes). Results showed a high degree of conservation of network modules between the maize and rice networks indicating the potential for translation of genotype-phenotype knowledge from one species to another.

Chapter 4 describes an analytical exercise to examine networks for robustness [60]. Is the structure of the global networks robust? How variable are the interactions (edges) found in the network? Is the network reproducible if different sample are used? It is important to assess the robustness of the network to ensure that relationships are meaningful. To this end, hundreds of networks were constructed with a randomized input sample set, using an improved implementation of the RMT method [80] in C code developed in collaboration with Dr. Melissa Smith's lab in the Department of Electrical and Computer Engineering at Clemson University. Results show that despite variation in input sample composition, the global networks are highly robust.

One lesson learned from construction of global networks was that an increase in the number of conditions in the dataset leads to increases in unexplained variation of expression, limiting the sensitivity of the thresholding method. Several dynamic thresholding methods have been developed to address this particular issue [82, 85]. Therefore, Chapter 5 describes a new set of rice networks that were constructed using a knowledge-independent pre-clustering approach. The objective was to remove experimental bias using a knowledge-independent approach by grouping input samples by similarity of expression rather than by the bias of annotated conditions. Twenty-five groups of samples were clustered and a distinct co-expression network was constructed for each group. This collection of networks represents the best approximation of a holistic gene co-expression network for rice. Additionally, over 8000 QTLs from QTL mapping studies were integrated with the network as well as significant SNPs from a recent GWAS study. An online web portal was constructed to allow for mining of network modules across the rice networks for significant overlap to traits from genetic studies.

In summary, the approach presented by this dissertation was a deviation from the typical systems genetic approach. The objective was to explore the requirements for creating a framework that could approximate the whole of biological co-expression interactions—not just those captured by a specific set of experimental conditions. Additionally, this work provides the start of a framework needed to test the hypothesis that a holistic network can better predict genotype-phenotype relationships as well as the full context of those relationships, including unexpected side-effects if selecting for specific conditions. The ability to predict such relationships will be of great importance for development of new varieties of rice and other cereals that can help feed and better nourish a growing worldwide population in the midst of unknown climate changes.

2. The Association of Multiple Interacting Genes with Specific Phenotypes In Rice (Oryza sativa) Using Gene Co-Expression Networks

Stephen P. Ficklin[1], Feng Luo[2] and Frank A. Feltus[1,3]

[1]Plant and Environmental Sciences, Clemson University, Clemson, SC 29634, USA.

[2]School of Computing, Clemson University, Clemson, SC 29634, USA.

[3]Dept. of Genetics & Biochemistry, Clemson University, Clemson, SC 29634, USA.

Published in *Plant Physiology*, September 2010

Supplementary figures and tables referenced in this chapter
are available online with the manuscript.

Abstract

Discovering gene sets underlying expression of a given phenotype is of great importance as many phenotypes are the result of complex gene-gene interactions. Gene co-expression networks, built using a set of microarray samples as input, can help elucidate tightly co-expressed gene sets (modules) which are mixed with genes of known and unknown function. Functional enrichment analysis of modules further subdivides the co-expressed gene set into co-functional gene clusters that may co-exist in the module with other functionally related gene clusters. In this study, 45 co-expressed gene modules and 76 co-functional gene clusters were discovered for *Oryza sativa* (rice), using a global, knowledge-independent paradigm and the combination of two network construction methodologies. Some clusters were enriched for previously characterized mutant phenotypes, providing evidence for specific gene sets (and their annotated molecular functions) that underlie specific phenotypes.

Introduction

A current challenge in understanding biological systems, especially those related to multicellular eukaryotic organisms, is the understanding of complex gene product interactions and resulting phenotypes. Integrated studies at a systems biology level are critical for unraveling complex genotype-phenotype relationships. These studies are increasingly feasible with high-throughput microarray assays, next-generation sequencing technologies, proteomics, and the wealth of accumulated functional and structural genomics data across species. *Oryza sativa* (rice) is one of the world's most important food crops, and serves as a model organism for the grass family. An improved understanding of complex interactions among rice genes is of great importance to improve nutritional value, grain yield, cultivation range, disease and stress tolerance of rice and other cereals.

In silico derived networks such as protein-protein interaction, metabolism, transcription, and gene co-expression model real biological interactions and exhibit naturally occurring properties such as small-world, scale-free, modularity and hierarchical characteristics [63, 65]. Barabasi and Oltvai (2004) provide a review of biological networks, and a brief description of relevant network properties can be found in Supplemental Table S1. One type of biological network, the gene co-expression network, is constructed from microarray gene expression profiles [75, 76, 80]. Nodes in the network represent microarray probe sets (or genes), and edges between nodes exist when gene expression profiles are significantly correlated (co-expressed) across all samples. In many cases the microarray samples encompass multiple tissue types, growth stages and experimental variables. Networks constructed from mixed sample sets represent a “global”, or meta-analysis view of gene co-expression.

Gene co-expression networks can be applied to a broad range of biological problems. Examples include those constructed to identify functional gene modules in humans [111], identification of genes involved with cellulose synthase in *Arabidopsis* [75], identification of biomarkers for glycerol kinase deficient mice [112], identification of *cis*-regulatory elements in gene clusters for budding yeast [113], construction of a regulatory network of iron response in *Shewanella oneidensis* [114], and identification of conserved gene clusters across several species [76]. For plants, global co-expression networks have been constructed for *Arabidopsis* [75, 115-119], barley [120], rice [121, 122], and tobacco [123].

Several online resources exist for plant co-expression networks. For *Arabidopsis*, online resources for co-expression networks include the *Arabidopsis* Co-expression Tool (ACT) which allows users to mine genes with similar co-expression patterns as well as functional terms [124], and the *Arabidopsis thaliana trans*-factor and *cis*-elements prediction database (ATTED II) which provides a visualization and online data mining tool for co-expression networks in *Arabidopsis* [125]. The RiceArrayNet (RAN) [121] and STARNET 2 [122] provide similar functionality for rice. An online resource exists for poplar [126] and a similar site named the Coexpressed Biological Processes (CoP) database provides a searchable database of functional associations for co-expression network modules across multiple plant species including rice [127].

Gene co-expression networks do suffer from limitations. First, they cannot provide a full understanding of complex gene-gene interactions because they infer only a single level of interaction: gene co-expression. Also, co-expression can only be measured when genes are consistently co-expressed or when genes are sometimes co-expressed but otherwise consistently silent [73]. Additionally, expression of all genes in every environmental or temporal condition cannot be measured and hence co-expression networks do not capture all possible relationships. Moreover, genes that are not co-expressed, but which may be essential are not captured. Despite these limitations, co-expression networks provide valuable glimpses into complex gene-product interactions.

Once constructed, a gene co-expression network can be examined for sub-networks of co-expressed and possibly co-functional genes. A reduced-bias sub-network discovery method can be performed using knowledge-independent approaches that employ statistical methods to circumscribe non-random gene set interactions. In contrast, gene-guided methods use *a priori* selected “bait” genes to define gene sets consisting of closely connected neighbors [73, 75]. A knowledge-independent approach provides inferences into the interaction set that might be obscured from gene-guided methods which filter genes based on prior assumptions of the biological system under scrutiny. Using a knowledge-independent method, co-expression networks can be subdivided into tightly connected gene modules. Modules are defined as sets of highly correlated (connected) genes that form sub-networks and are often connected to the global network through a few connections.

It has been shown that modules often consist of genes that participate in similar functions [76, 111]. As a result, genes of unknown function or genes not previously known to participate in molecular pathways can be identified through a "guilt-by-association" inference with genes of known function [77]. Alternatively, function-enriched gene clusters within modules can be identified by counting annotated terms, such as Gene Ontology (GO) [128], in a set of genes. Functional enrichment of a given term occurs if the term is significantly more abundant in the module relative to its occurrence in the genome background and implies that the module is associated with the mixture of enriched function. Furthermore, gene subsets within modules can be identified that non-randomly share functional terms (co-functional clusters). Modules may consist of hundreds of nodes with numerous functional terms and multiple co-functional clusters. Publically available tools such as DAVID [92, 93], EASE [129], Fatigo [130] and Blast2GO [131] represent some of the tools that exist for functional enrichment analysis.

Recent studies show that co-expression networks can be used to identify a set of candidate genes underlying specific phenotypes. Mutwil *et al.* demonstrate a novel clustering method for co-expression networks, coupled with associated phenotypic terms, to predict gene sets in *Arabidopsis* for lethality [55]. Lee *et al.* show the predicative power of a network for *Arabidopsis* composed of a diverse set of data (including co-expression data) to predict gene sets associated with lethality and pigmentation [56]. By prioritization of genes through guilt-by-association Lee *et al.* show a ten-fold improvement over screens of random insertion mutants. Both studies demonstrate the applicability this systems genetics approach for predicting biologically meaningfully relationships.

Herein we describe the construction and functional partitioning of a rice gene co-expression network to associate multiple co-expressed gene sets with common molecular function and experimentally verified phenotypes. The underlying implication is that gene sets enriched for known gene lesions may be causal to a specific phenotype, and the molecular functions that are co-enriched for phenotype-associated genes may provide clues to the molecular mechanisms that lead to the phenotype. Each cluster or module is a candidate gene set for studying complex traits where multiple genes may have an effect on phenotypic expression.

Results

The Rice Network

Construction of the rice co-expression network began with a total of 508 Affymetrix rice arrays downloaded from NCBI's Gene Expression Omnibus (GEO) (Supplemental Table S2) which were filtered for outliers and RMA normalized (see Materials and Methods). Pearson correlation between gene expression profiles was used as the underlying metric for co-expression. This study used the strengths of the RMT [80], and WGCNA [86] methods to construct the gene co-expression network. WGCNA was used for module detection and RMT for automatic threshold (signal-to-noise) identification. Figure 2.1 provides a schematic of steps involved in network construction including RMA normalization, outlier detection and removal, calculation of Pearson correlation values, module detection using WGCNA and determination of a threshold value using RMT.

Co-expression network construction yielded 4,528 nodes (mapped to 4,502 rice loci) connected by 43,144 edges within 45 modules, some of which were later removed after thresholding. Supplemental Table S3 provides a listing of all edges in the co-expression network. The network follows the properties of natural biological networks, namely it is small-world, scale-free, modular and hierarchical. The network demonstrates small-world characteristics with an average distance between any two nodes (path length) of 11. Scale-free behavior is indicated by a negative linear correlation between the number of edges, $log(k)$, and the probability of finding a node with k edges, $P(k)$ (Supplemental Figure S1A). A negative correlation between the number of edges, k, and the clustering coefficient for nodes with k edges, $C(k)$, indicates hierarchical and modular behavior (Supplemental Figure S1B) The average clustering coefficient, $<C>$, was 0.318. A graphical representation of the network, generated using Cytoscape [132], can be seen in Figure 2.2. Nodes in the network are color-coded according to the modules.

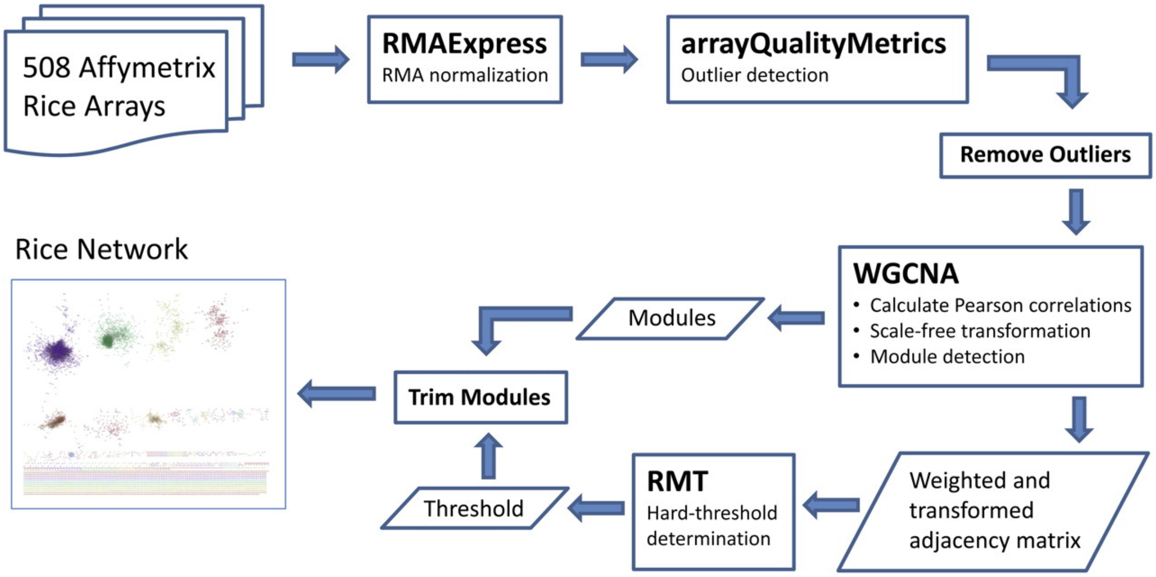

508 Affymetrix Rice Arrays
RMAExpress
RMA normalization
arrayQualityMetrics
Outlier detection
Remove Outliers
WGCNA
• Calculate Pearson correlations
• Scale-free transformation
• Module detection
Modules
Weighted and transformed adjacency matrix
RMT
Hard-threshold determination
Threshold
Trim Modules
Rice Network

Figure 2.1 Network Construction Flow Chart.

The data pipeline for construction of the rice co-expression network involves RMA normalization (Bolstad, 2010), outlier detection and removal (Kauffmann et al., 2009), construction of adjacency matrix and modules using WGCNA (Langfelder and Horvath, 2008), hard-threshold determination using RMT (Luo et al., 2007), and final culling of nodes below the threshold.

In order to explore the relationship between modules, the WGCNA package was used to calculate eigenvectors, or first principle components, for each module. The eigenvector, or eigengene, acts as a representative expression profile for the module and allows for a meta-analytic view of the entire module set. All eigengenes were clustered using WGCNA. Figure 2.3 provides a view of the modules in the form of a dendrogram that indicates “closeness” of expression similarity of the 45 modules. Each module is numbered from zero to 44 and prefixed with ‘ME’, meaning ‘module eigengene’. Adjacent modules are more highly similar in terms of expression. It should be noted that these eigenvectors were computed from WGCNA modules prior to edge removal that were below the RMT-derived hard-threshold.

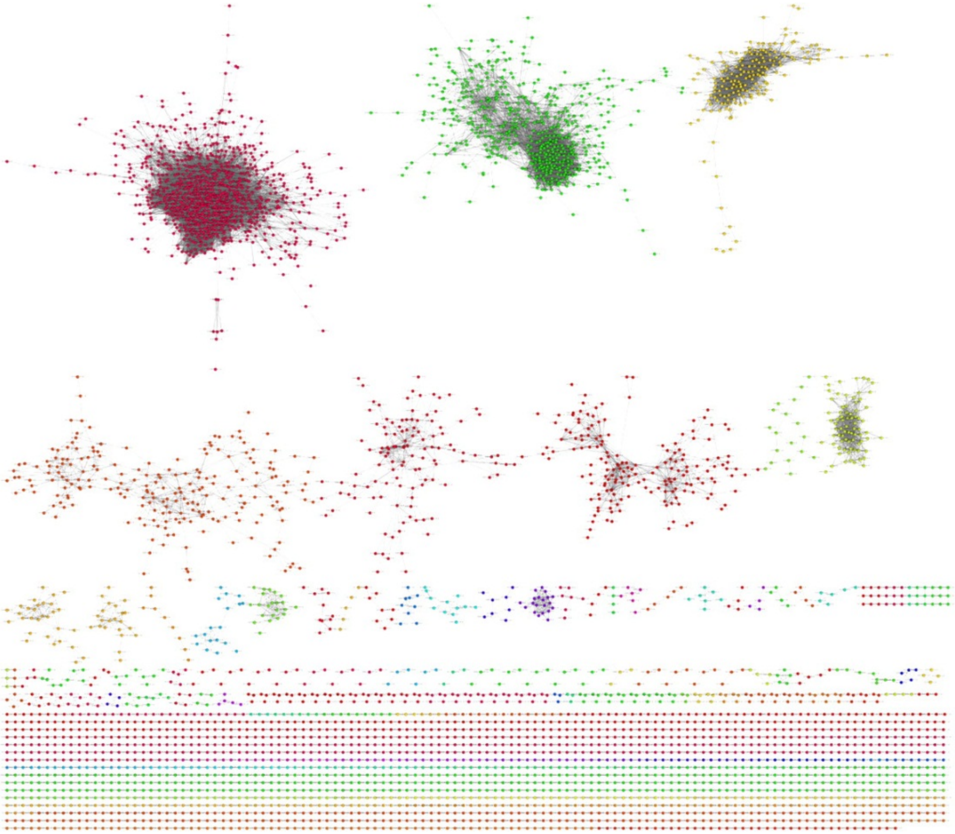


Figure 2.2 Rice Co-expression Network.

The rice network consists of 4,528 nodes, 43,144 edges, and 45 modules. The nodes are color coded by modules.

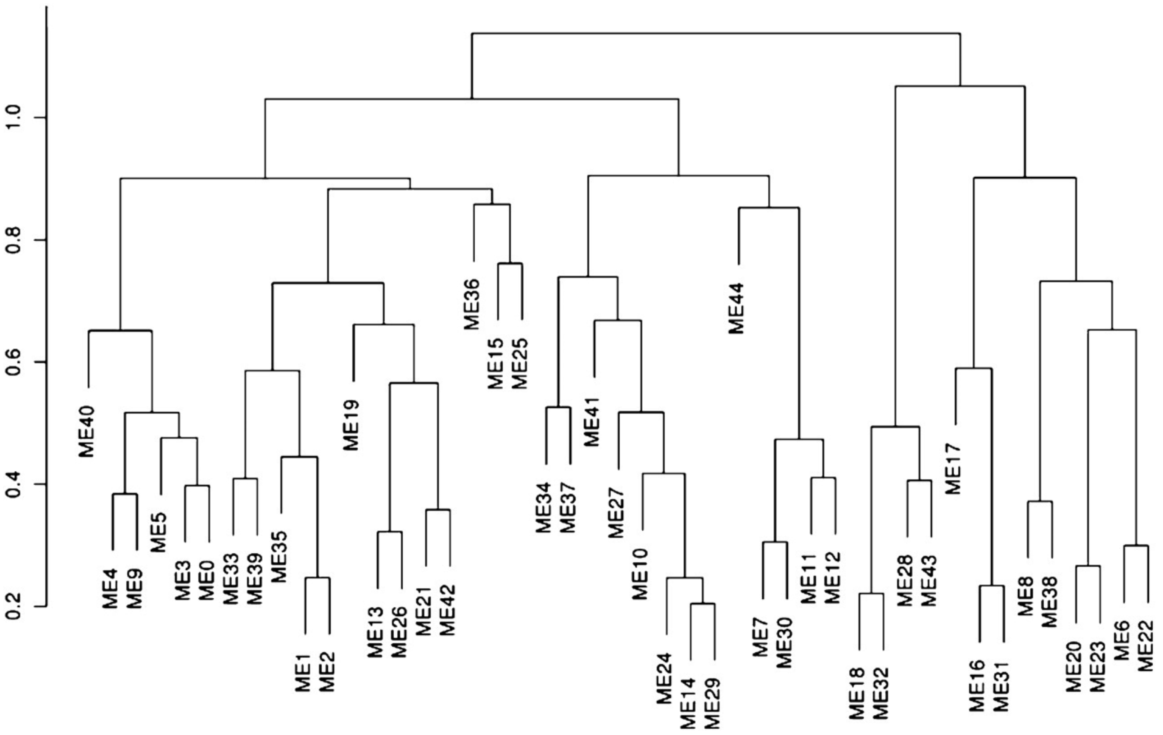


Figure 2.3 Module Eigenvector Clustering.

The rice network consists of 45 modules. The eigenvectors for each module were calculated and clustered using the WGCNA software. The eigenvectors for each module are prefixed with ME in the dendrogram and are calculated prior to thresholding of the network. Adjacent modules are more highly similar in terms of expression.

Mapping of Microarray Probe sets to Rice Loci

Prior to functional enrichment, the mapping of network nodes (microarray probe sets) to annotated rice gene models was necessary to ensure that annotation terms were not over-counted. The Michigan State University (MSU) Rice Genome Annotation version 6.0 contains 56,797 protein coding sequence loci. Of the 57,381 probe sets on the rice microarray, 50,468 mapped to 46,498 loci. Of those mappings, 34,028 probe sets mapped directly with all 11 probes from a single probe set to a gene locus. Of those mappings, 26,382 are unique one-to-one mappings between a probe set and locus. Redundant mappings are those where multiple probe sets map to a single loci. Ambiguous mappings are those where a probe set maps to multiple loci. The distribution of probes, probe sets and loci within the mappings can be observed in the charts of Supplemental Figure S2. There are 17,762 redundant mappings and 4,769 ambiguous mappings. Ambiguity was removed from the mappings, and the remaining redundancy was addressed with a weighted counting method (see Materials and Methods).

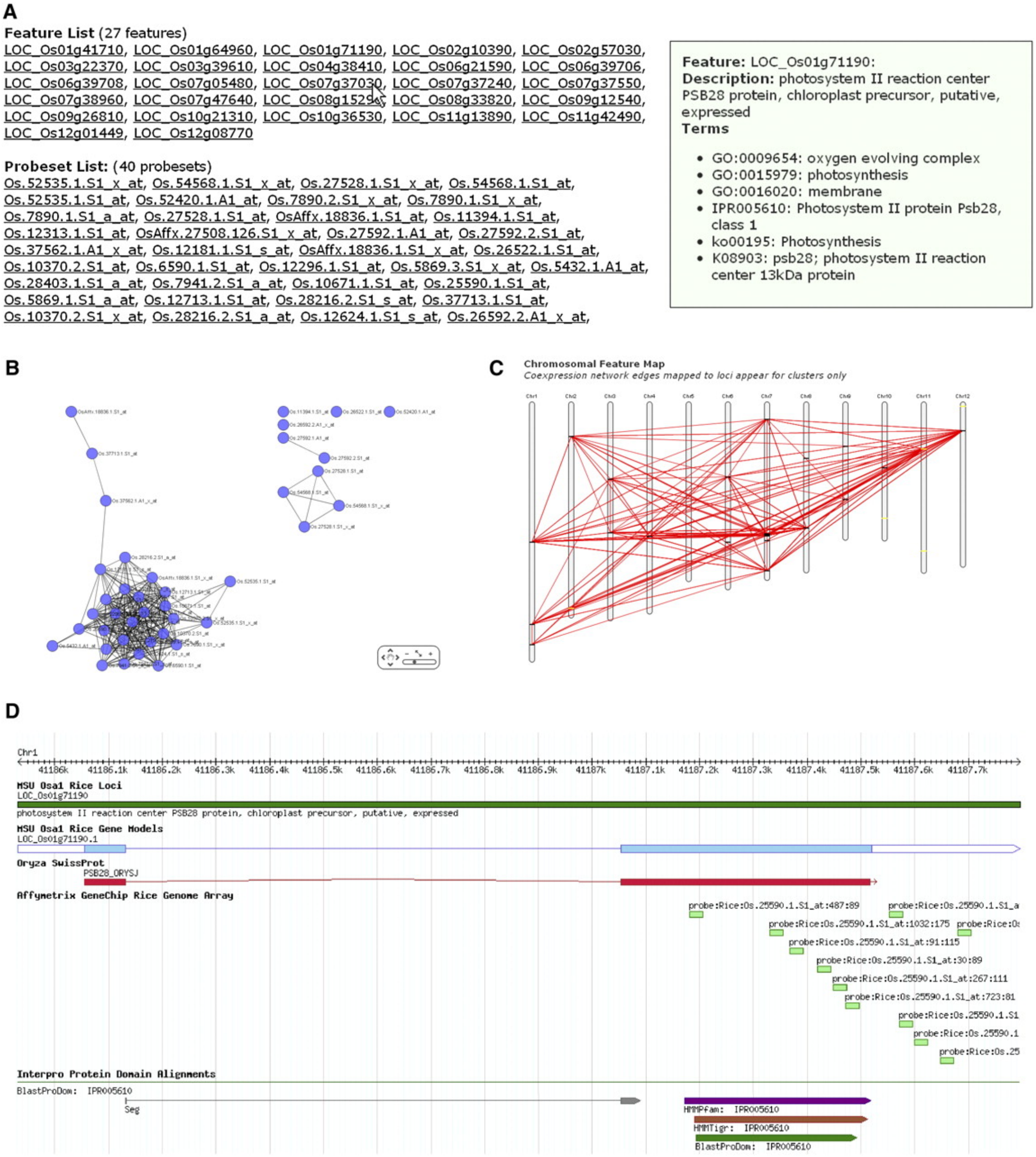
Functional Enrichment and Clustering

A functional enrichment analysis was performed to examine enrichment of annotated terms. After counting GO [128], KEGG [99], InterPro [133], and Tos17 mutant phenotype [105, 134] terms for each module and for the genome background, Fisher's test comparisons were performed for each module to identify functionally enriched terms. Co-functional gene clusters with overlapping function were then identified. Clusters are sub-networks within modules. Nodes in modules are co-expressed and nodes within clusters are both co-expressed and co-functional. Some modules had multiple clusters while others had none. Functional enrichment yielded 2,412 unique enriched terms in all network modules with 939 of these aggregating into clusters. Of the total enriched terms, 21 were unique mutant phenotype terms that associated with 25 clusters. Four mutant phenotype terms were enriched only at the module level (see Supplemental Tables S4, S5, S6, S7).

The average connectivity, <*k*>, was used for ranking clusters and is the average number of connections per node in the cluster sub-network. Additionally, an enrichment score, en-score, was determined which is the inverse log of the geometric mean of the Fisher's *p*-values in each cluster. For easy reference, clusters hereafter are named as follows: M*x*C*y*, where *x* is the module number (e.g. 1 for Module1) and *y* is the cluster number (e.g. 2 for Cluster 2). Modules are named as M*x*. Module numbers originate from WGCNA and cluster numbers are ordered sequentially in descending order of <*k*>.

Online Co-expression Network Browser

An online resource has been created to facilitate co-expression network browsing. The website is available at http://www.clemson.edu/genenetwork. This website allows users to browse the list of probe sets, loci and enriched terms of modules and clusters. Additionally, visualizations are provided for each cluster including free-standing interactive network graphs and cluster networks super-imposed onto the rice genome. Users can search for functional terms, loci, probe sets or other keyword to find modules and clusters that may relate to genes, pathways, functions or phenotypes of interest. The site shows genome alignments for each locus including Interpro domains and alignments with Affymetrix probes. Annotation terms (e.g. GO, Interpro, KEGG terms) link out to external sites. Figure 2.4 shows various screenshots of cluster M6C2 from the website.



Feature List (27 features)
LOC_Os01g41710, LOC_Os01g64960, LOC_Os01g71190, LOC_Os02g10390, LOC_Os02g57030, LOC_Os03g22370, LOC_Os03g39610, LOC_Os04g38410, LOC_Os06g21590, LOC_Os06g39706, LOC_Os06g39708, LOC_Os07g05480, LOC_Os07g37030, LOC_Os07g37240, LOC_Os07g37550, LOC_Os07g38960, LOC_Os07g47640, LOC_Os08g1529, LOC_Os08g33820, LOC_Os09g12540, LOC_Os09g26810, LOC_Os10g21310, LOC_Os10g36530, LOC_Os11g13890, LOC_Os11g42490, LOC_Os12g01449, LOC_Os12g08770

Probeset List: (40 probesets)
Os.52535.1.S1_x_at, Os.54568.1.S1_x_at, Os.27528.1.S1_x_at, Os.54568.1.S1_at, Os.52535.1.S1_at, Os.52420.1.A1_at, Os.7890.2.S1_x_at, Os.7890.1.S1_x_at, Os.7890.1.S1_a_at, Os.27528.1.S1_at, OsAffx.18836.1.S1_at, Os.11394.1.S1_at, Os.12313.1.S1_at, OsAffx.27508.126.S1_x_at, Os.27592.1.A1_at, Os.27592.2.S1_at, Os.37562.1.A1_x_at, Os.12181.1.S1_s_at, OsAffx.18836.1.S1_x_at, Os.26522.1.S1_at, Os.10370.2.S1_at, Os.6590.1.S1_at, Os.12296.1.S1_at, Os.5869.3.S1_x_at, Os.5432.1.A1_at, Os.28403.1.S1_a_at, Os.7941.2.S1_a_at, Os.10671.1.S1_at, Os.25590.1.S1_at, Os.5869.1.S1_a_at, Os.12713.1.S1_at, Os.28216.2.S1_s_at, Os.37713.1.S1_at, Os.10370.2.S1_x_at, Os.28216.2.S1_a_at, Os.12624.1.S1_s_at, Os.26592.2.A1_x_at,

Feature: LOC_Os01g71190:
Description: photosystem II reaction center PSB28 protein, chloroplast precursor, putative, expressed
Terms

- GO:0009654: oxygen evolving complex
- GO:0015979: photosynthesis
- GO:0016020: membrane
- IPR005610: Photosystem II protein Psb28, class 1
- ko00195: Photosynthesis
- K08903: psb28; photosystem II reaction center 13kDa protein

B
C
Chromosomal Feature Map
Coexpression network edges mapped to loci appear for clusters only
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
D
Chr1
41186k
41186.1k
41186.2k
41186.3k
41186.4k
41186.5k
41186.6k
41186.7k
41186.8k
41186.9k
41187k
41187.1k
41187.2k
41187.3k
41187.4k
41187.5k
41187.6k
41187.7k
MSU Osa1 Rice Loci
LOC_Os01g71190
photosystem II reaction center PSB28 protein, chloroplast precursor, putative, expressed
MSU Osa1 Rice Gene Models
LOC_Os01g71190.1
Oryza SwissProt
PSB28_ORYSJ
Affymetrix GeneChip Rice Genome Array
probe:Rice:Os.25590.1.S1_at:487:89
probe:Rice:Os.25590.1.S1_a
probe:Rice:Os.25590.1.S1_at:1032:175
probe:Rice:Os
probe:Rice:Os.25590.1.S1_at:91:115
probe:Rice:Os.25590.1.S1_at:30:89
probe:Rice:Os.25590.1.S1_at:267:111
probe:Rice:Os.25590.1.S1_at:723:81
probe:Rice:Os.25590.1.S1
probe:Rice:Os.25590.1
probe:Rice:Os.25
Interpro Protein Domain Alignments
BlastProDom: IPR005610
Seg
HMMPfam: IPR005610
HMMTigr: IPR005610
BlastProDom: IPR005610

Figure 2.4 Screen Shots of the Online rice co-expression Network Browser

A, The list of loci and probe sets and their mappings. B, The sub network graph with navigation toolbox. C, The sub network graph superimposed on the genome. D, Loci (feature) details including genome alignments from the MSU Rice Genome Browser. The site is located at http://www.clemson.edu/genenetwork.

Functional Significance of Select Modules and Clusters

The largest of the 45 modules is M6 (large module in the top left of Figure 2.2), which consists of 26 clusters. A majority of M6 clusters contain enriched function associated with translation and photosynthesis including carbon fixation and related processes. Many of these clusters are also enriched for terms referencing the plastid, suggesting that M6 consists of genes involved in processes that occur in the chloroplast. For example, cluster M6C1 is ranked highest in average connectivity for the whole network. M6C1 consists of 75 enriched terms, 43 loci, 52 nodes, $<k>$ = 17.54, en-score = 2.61, and 456 edges. The highest ranked (lowest p value) term in this cluster is the GO term for translation (GO:0006412; p = 7.80e-27). Other terms in this cluster include: ribosome, plastid, translation elongation and rRNA binding. Several M6 clusters are enriched with the mutant phenotypic terms 'low tillering', 'extremely dwarf', 'lethal', 'sterile' and 'yellow'. A complete accounting of M6 edges, loci, probe sets, clusters and enriched terms can be found on the co-expression network browser and in Supplemental Tables S3, S4, S5, S6, and S7 respectively. Additionally, this same information is available online with the co-expression network browser. A total of 127 loci are co-expressed in M6 but have no known ascribed function (Supplemental Table S8).

Another interesting cluster is M13C1. M13C1 has the second highest ranked en-score and the second highest $<k>$ indicating that it is highly co-expressed and co-functional. M13C1 consists of 12 enriched terms, 16 loci, 21 nodes, $<k>$ = 12.86, en-score = 10.64 and 135 edges. The highest ranked enriched term is the "Cereal seed allergen/grain softness/trypsin and alpha-amylase inhibitor" protein domain (IPR006106, *p* = 6.28e-23; Table 2.1). Other terms related to lipid transfer and seed storage are also enriched in M13C1. Appropriately, Genevestigator analysis [135], shows high levels of expression in the milk and dough stages as well as in inflorescence, seed and embryo developmental stages (Figure 2.5). Additional heat maps for the top ten connected clusters (excluding M13C1) are available in Supplemental Figure S3. It should be noted that at the time of this study, Genevestigator incorporated approximately 151 samples of the Affymetrix Rice platform from GEO while 508 samples from GEO were used for our network construction. Genevestigator has not incorporated newly available rice arrays. It should be noted that there is a difference in the number of samples for each tissue type between the Genevestigator arrays and the network arrays. However, there are several biological replicates across the various samples for each tissue type and developmental stage in the Genevestigator data set. The only exception is stamen, anther and embryo which have one sample each. Therefore, we expect that Genevestigator results can provide support as to the correctness of the functional clusters in the majority of tissues and stages.

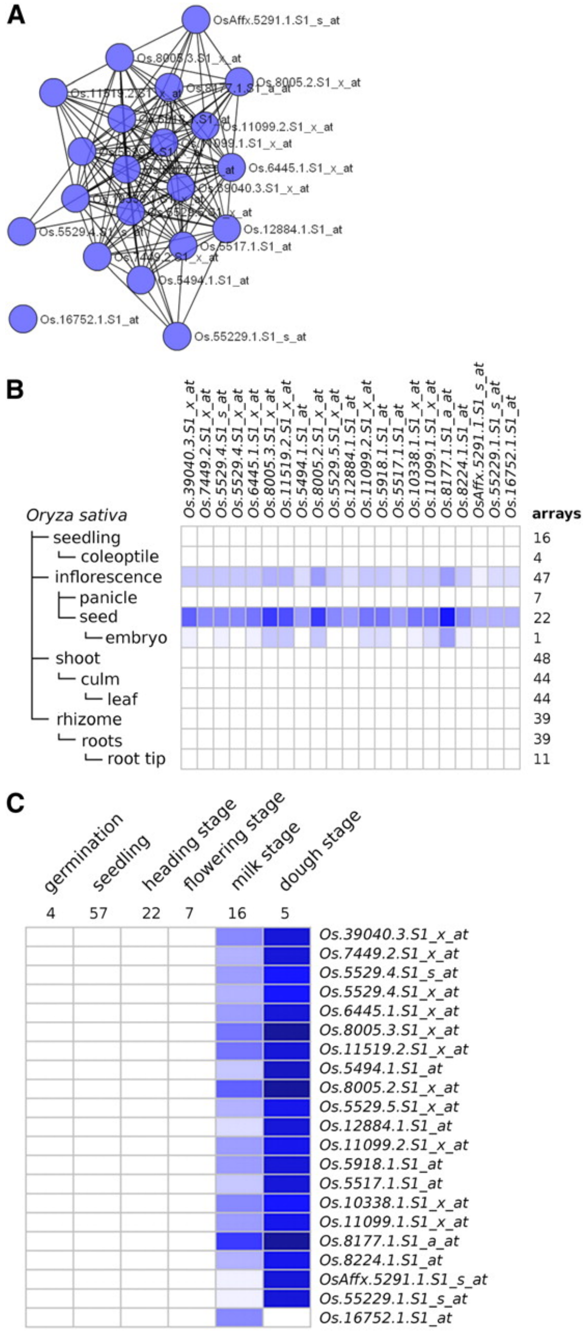


Figure 2.5 Cluster M13C1 with Genevestigator Analysis.

A, The subnetwork for cluster M13C1 (<*k*> = 12.86, en-score = 10.64, 135 edges, 21 nodes, and 16 loci). B, Heat map showing expression levels by anatomical locations. C, The Genevestigator analysis heat map showing expression levels in microarray sets categorized by development stage.

One cluster enriched for phenotypic terms is M2C2 (11 loci, <*k*> = 4.33, en-score = 4.93, 12 nodes and 26 edges). This cluster is enriched with three mutant phenotype terms: ’Sterile’, ‘Dwarf’, and ‘High tillering’. The cluster is also enriched with the ‘Cyclin A/B/D/E’ (IPR014400; *p* = 1.25e-10) and ‘G2/mitotic-specific cyclin A’ (IPR015453; *p* = 9.96e-4) protein domains and other terms related to cyclin in the mitotic cell cycle. Three loci in this cluster have mutant phenotype associations, including one, LOC_Os02g10490, annotated as ‘cyclin, putative, expressed’ and two expressed proteins with no known function, LOC_Os02g35230. All three share the mutant terms ‘dwarf’, ‘low fertile’ and ‘sterile’, and are interconnected.

Table 2.1 Enriched terms from cluster M13C1

(<*k*> = 12.86, en-score = 10.64, 135 edges, 21 nodes and 16 loci)

Term Accession[a]	Description
IPR006106	Cereal seed allergen/grain softness/trypsin and alpha-amylase inhibitor
IPR006105	Cereal seed allergen/trypsin and alpha-amylase inhibitor, conserved site
GO:0004867	Serine-type endopeptidase inhibitor activity
GO:0016068	Type I hypersensitivity
IPR002411	Cereal allergen/alpha-amylase inhibitor, rice-type
IPR016309	Alpha-amylase inhibitor/seed allergen
GO:0005615	Extracellular space
IPR001954	Gliadin/LMW glutenin
IPR013771	Bifunctional trypsin/alpha-amylase inhibitor
IPR003612	Plant lipid transfer protein/seed storage/trypsin-alpha amylase inhibitor
IPR016140	Bifunctional inhibitor/plant lipid transfer protein/seed storage
GO:0045735	Nutrient reservoir activity

[a] GO and IPR accession numbers are from Gene Ontology and Interpro respectively.

Another cluster, M8C1, is enriched for processes related to defense response. This cluster is enriched with multiple mutant terms: ‘extremely dwarf’, ‘late heading’, ‘lazy’, ‘short panicle’, and ‘wide leaf’. The gene products in this cluster include powdery mildew resistance proteins, NBS-LRR proteins, stripe rust resistance proteins, and one protein with unknown function. The protein of unknown function, LOC_Os02g06790, is also enriched for ‘lazy’ and ‘late heading’. Many of the other M8C1 loci are associated with multiple mutant terms that are not enriched.

A list of all clusters enriched for mutant phenotype terms can be found in Table 2.2. Phenotype terms enriched at the module level can be found in Table 2.3. A detailed list of clusters, associated probe sets, gene accessions, clusters and all enriched terms for each module can be found in the Supplemental Data or through the co-expression network browser.

Table 2.2 Complete list of TOS phenotype terms enriched in clusters.

Cluster[a]	<*k*>	en-score	Summarized Function[b]	Phenotype Terms
M2C1	8.22	4.14	Cell cycle/kinesin/cyclin	Dwarf, High tillering, Sterile
M2C2	4.33	4.93	Cell cycle/cyclin	Dwarf, High tillering, Sterile
M2C3	3.69	3.56	DNA replication	Dwarf
M2C4	3.36	3.37	Cell cycle/kinesin	Dwarf, Sterile
M2C7	1.00	4.36	DNA replication	Dwarf
M2C8	0.91	2.71	Cell cycle	Dwarf, Sterile
M2C11	0.75	1.49	Cell cycle	Dwarf
M2C16	0.40	7.36	DNA replication/polymerase	Dwarf
M6C2	12.30	3.12	Photosynthesis/light harvesting	Dwarf, Extremely dwarf, Lethal, Low tillering, Sterile, Yellow
M6C3	9.23	2.54	Electron carrier activity	Dwarf, Extremely dwarf, Lethal, Low tillering, Sterile, Yellow
M6C5	3.60	2.05	Photosynthesis	Dwarf, Extremely dwarf, Lethal, Low tillering, Sterile, Yellow
M6C10	1.60	5.04	Oxidoreductase activity	Dwarf
M6C11	1.56	4.66	Translation	Low tillering
M6C14	1.25	2.22	Glycoside hydrolase	Extremely dwarf
M6C15	1.20	2.47	Carbon fixation	Dwarf
M6C16	1.14	3.94	Translation or Photosynthesis	Dwarf
M6C18	0.80	1.82	Translation or Photosynthesis	Yellow
M6C19	0.73	2.64	Regulation of transcription	Lethal
M6C20	0.50	3.36	Translation or Photosynthesis	Dwarf, Extremely dwarf, Lethal, Low tillering, Sterile, Yellow
M6C22	0.50	3.28	Oxidoreductase activity	Dwarf, Extremely dwarf
M7C1	1.23	2.45	Transporter activity	Pale green leaf
M8C1	0.77	2.50	Defense response	Extremely dwarf, Late heading, Lazy, Short panicle, Wide leaf
M8C2	0.25	1.79	Kinase activity	Stripe
M18C1	1.60	1.63	Lipid binding	Vivipary

[a] Modules are numbered sequentially starting from zero and are prefixed with the letter M. Clusters within a module are numbered sequentially and are prefixed with the letter C. Thus cluster 1 from module 8 is named M8C1.

[b] When function cannot be summarized for the cluster, the modular summarized function is listed.

Table 2.3 List of enriched mutant phenotypes

Module[a]	TOS Term	p-value
M1	Large grain	6.41E-02
M1	Weak	8.68E-02
M2	Dwarf	9.76E-03
M2	High tillering	5.19E-02
M2	Sterile	4.81E-04
M6	Dwarf	5.67E-02
M6	Extremely dwarf	1.07E-03
M6	Lethal	5.84E-03
M6	Low tillering	2.58E-04
M6	Sterile	9.44E-03
M6	Yellow	2.09E-02
M7	Pale green leaf	1.79E-02
M8	Extremely dwarf	3.26E-06
M8	Late heading	1.14E-02
M8	Lazy	3.26E-04
M8	Short panicle	1.01E-03
M8	Stripe	6.98E-02
M8	Wide leaf	9.78E-02
M12	Withering	4.86E-02
M13	Rolled leaf	8.54E-02
M14	Zebra	3.30E-02
M18	Vivipary	1.78E-03
M25	Abnormal shoot	3.78E-02
M32	Germination rate	4.38E-02
M35	Late heading	6.08E-02
M43	Vivipary	3.40E-02

[a] Modules are numbered sequentially from 0 to 57 and are prefixed with the letter M. Thus module 8 is named M8.

Discussion

The major objective for this study was to use a global, meta-analysis, knowledge-independent approach to construct a rice gene co-expression network that predicts clusters of candidate genes involved in complex genotype-phenotype interactions. We hypothesized that tightly co-expressed gene modules, enriched in shared functional annotation, would provide the most fruitful predictions of candidate gene sets that might underlie a given biological process. Using mutant phenotype terms in functional enrichment provides a hypothetical association between phenotype and the gene sets of modules and clusters. Co-enrichment of phenotypes with molecular function terms in a tightly co-expressed gene module suggests a direct association between the functional units carried on genes (e.g. protein domains, GO terms, etc.) and phenotype. When mutant phenotype terms are enriched in a highly connected gene cluster, the phenotypic association can also be extended to the neighboring co-expressed genes within the confines of a given module. Thus, the circumscribed gene sets become candidate factors underlying the expression of complex traits, and their annotated functions provide insight into molecular pathways associated with expression of empirically defined phenotypes. For instance, module M6 contains 127 loci that have no known function. It can be implied that these loci may be involved in some aspect of photosynthesis or translation given the M6 enrichment for photosynthesis/translation related annotations. In the case of the 26 M6 clusters enriched with phenotype terms, it can be predicted through guilt-by-association that other genes in the cluster may also contribute to the enriched phenotypes.

Many clusters in our network can be examined for possible genotype-phenotype interactions. For example, the previously mentioned cluster M2C2 is enriched for cyclin and mitosis as well as mutant phenotypes ‘dwarf’, ‘high tillering’ and ‘sterile’. Two of the genes in this module were shown to have no known function (see Results section). It can therefore be inferred that these two genes are involved in processes related to cyclin and mitotic cell division. Additionally, these two genes also share the ‘dwarf’ and ‘sterile’ mutant terms implicating their role as factors of those phenotypes. These genes are well connected with other nodes in the cluster; therefore, through guilt-by-association we can infer that other genes in the M2C2 cluster are also factors for the enriched phenotypes. Also mentioned previously was cluster M8C1 enriched for defense response terms. This cluster is not as highly connected as M2C2, however inferences can be made that the gene of unknown function in this module participates in defense response, perhaps in an indirect manner, and that all of the genes in the cluster are factors for expression of several phenotypes. Despite lower average connectivity, the nodes all exhibit similar patterns of co-expression. It can be inferred that this unknown gene plays some role related to defense response.

Two different construction methods were integrated to build the co-expression network, namely the WGCNA and RMT methods. These two methods were selected primarily as a means of preserving a knowledge-independent paradigm. A strength of the WGCNA method lies in its ability to detect modules. Module detection in WGCNA follows a knowledge-independent process. However, selection of a threshold for culling the network to limit noise would otherwise rely on functional annotation and empirical judgment [86]. A strength of the RMT method lies in its ability to automatically localize the noise-to-signal threshold without using annotations or empirical judgment. Therefore, we were able to generate a single network by passing the same adjacency matrix (power transformed pair-wise Pearson correlation values) generated by the WGCNA method into the RMT method for threshold detection (see Figure 2.1). This ensured knowledge-independence for meaningful thresholding of the network modules.

Our rice network does not encompass all the gene–gene interactions one would expect from all genes in the genome. The number of nodes in the network is 4,528 whereas the entire genome consists of 56,797 coding sequence loci. The network is therefore not representative of all co-expression relationships for all genes in the network. Co-expression can only be measured when genes are consistently co-expressed or when genes are sometimes co-expressed but otherwise consistently silent [73]. A bias exists in global co-expression networks for relationships that persist across all conditions and tissue types used by the underling microarray samples (e.g. housekeeping processes) or for relationships only expressed in a few tissue types, environmental conditions and developmental stages. The rice network presented here is most noticeably enriched for genes controlling housekeeping processes. Additionally, co-expression relationships that exist primarily in a few tissue types, developmental stages and conditions are not easily identified. The nodes in our network however do have co-expression relationships that are statistically significant across all samples, so each edge in our network is potentially biologically valid. While rice gene space sampling is not complete, the underlying goal was to find highly-connected gene clusters enriched with phenotypic terms. We believe that our approach was successful and inferences of polygenic phenotypic causality for gene sets can be made.

One observation was that some clusters showed significant enrichment in function yet demonstrated very low connectivity within the cluster (e.g. cluster M2C19, $<k> = 0$). The nodes of these clusters were mostly co-expressed through non-clustered intermediaries. Because highly connected genes are more likely to participate in similar function, we ranked clusters by average connectivity, $<k>$. We believe this ranking improves the prediction inferred through guilt-by-association with enriched annotation terms. Therefore, clusters that ranked highest are more likely to yield guilt-by-association inferences for genes of unknown function and as factors for expression of mutant phenotypes. It should not be implied that an absolute $<k>$ cut-off exists as a significance threshold for clusters. Poorly connected clusters may in fact be quite significant and should not be dismissed.

Conclusion

This study describes a set of modules and clusters that can assist with understanding of gene-gene, gene-function, and genotype-phenotype interactions for rice. While the number of enriched phenotype terms is low, the application demonstrates a positive approach for identifying gene sets associated with specific phenotypes. The network provides a set of interesting modules and clusters worthy of further investigation. In the process of investigating the use of co-expression networks we suggest that the RMT and WGCNA network constructions methods can be combined to extend the knowledge-independent approach to the final stages of module discovery. We also propose a cluster ranking method using average connectivity that rewards highly connected clusters with the expectations that highly ranked clusters are most meaningful in the data set. These data can help in the discovery of candidate genes for studies of complex traits in rice as well as a reference for other grass species.

Materials & Methods

Raw Expression Data

The dataset used for construction of the co-expression network was obtained from NCBI's Gene Expression Omnibus (GEO), platform accession number GPL2025. The platform consists of experimental samples from assays using the Affymetrix GeneChip Rice Genome Array (http://www.affymetrix.com/support/technical/byproduct.affx?product=rice). The array consists of 57,381 probe sets derived mostly from TIGR's version 2.0 release of the rice genome and consists of transcripts for both the *japonica* and *indica* cultivars. 550 CEL files were obtained from GEO, and 13 CEL files were removed due to an incorrect Arabidopsis array type. RMA normalization [136] of all microarray samples was performed using the RMAExpress software [137]. . Outliers were detected using the arrayQualityMetrics [138] Bioconductor [139] package, which uses three different statistical tests to identify outliers. Twenty-nine samples failed at least one test and were considered outliers and removed from the dataset. A total of 508 samples remained for network construction. Control probes from the platform were removed from the samples prior to network construction.

Expression Profile Correlation

The expression profile of a gene consists of the set of expression levels across all microarray samples in the study. Initially, construction of the co-expression network requires pair-wise correlation of all gene expression profiles to obtain an *n* x *n* similarity matrix, S:

$S = [s_{ij}]$

$s_{ij} = cor(x_i, x_j)$, where x_i and x_j are the pair of expression profiles for genes *i* and *j*, and *cor(x,y)* represents the Pearson correlation function.

Scale-free Behavior and Module Detection

The WGCNA package [86, 140] provides a robust set of R functions for constructing weighted co-expression networks. The similarity matrix is transformed into an adjacency matrix using a method that employs a power function. This is termed “soft-thresholding.” The result is an adjacency matrix where correlation strength is enhanced for highly correlated genes and correlation information is preserved for module discovery. The values of the adjacency matrix are represented by the following formula:

$a_{ij} = s_{ij}^{\beta}$

The power (β) used to transform the similarity matrix is selected when the resulting network best approximates a scale-free topology. The WGCNA method provides functionality to assist with selection of the power function. For this study a soft-threshold power of 4 was used.

Probe sets with ambiguous mappings to multiple rice loci were removed from the dataset if there were less than 6 probes in the mapping, and remaining probe sets were kept if what remained was a unique or redundant mapping. Of the 4,769 probe sets with ambiguous mappings, 3,223 probe sets were removed. Affymetrix control probes were next removed from the dataset.

The *n* x *n* similarity matrix for the remaining 52,501 probe sets was too large for R which has a 32-bit integer limit on the index size of a matrix. Therefore the algorithm was instructed to break the dataset into 3 blocks with a minimum of 30 probe sets and a maximum of 30,000. The WGCNA package calculates modules of similarly co-expressed genes using a Topological Similarity Matrix (TOM) and a hierarchical clustering method. A value of 0.2 was specified for cutting the resulting dendrogram into distinct modules.

Threshold Selection and Network Analysis

A weighted soft-threshold network maintains edges from all nodes to all nodes with the edge weight indicating the strength of the co-expression. This becomes valuable for module detection. However, selection of a hard-threshold after module detection is required to remove noise. A Random Matrix Theory (RMT) method [80], was used to recognize the boundary between noise and non-noise, and for selection of a hard-threshold of the network. The hard-threshold is determined by the transition of nearest neighbor spacing distribution (NNSD) of the similarity matrix from the Gaussian orthogonal ensemble statistics to the Poisson distribution. The chi-square test (confidence level of 0.001) was used to define the similarity threshold at which NNSD completely follows the Poisson distribution. For the rice co-expression network, a hard-threshold r of the power transformed Pearson correlation matrix at $r = |0.7101|$ was observed. The soft-thresholded, power transformed adjacency matrix is then "hard-thresholded" by setting all values less than the threshold to zero. Nodes with an adjacency value of zero are removed from the modules. Modules with no remaining nodes are discarded. Characterization of the network in terms of scale-free, small-world, modularity and hierarchical behavior was performed using the NetworkAnalyzer package for Cytoscape [141]

Functional Enrichment

The annotation of rice probe sets provided by Affymetrix was derived from TIGR v2.0 gene models. However, more up-to-date annotations were desired. Therefore, annotations were updated using mapping information provided by release version 6.0 of the MSU Rice Genome Annotation Project which maps probe sets to 6.0 gene models [6]. The locus IDs from release 6.0 were then used to provide four classes of function terms, including: Gene Ontology (GO) [128], KEGG [99], InterPro [133], and Tos17 mutant phenotypes [105, 134]. In some cases, such as with the GO and InterPro these annotations were provided by the MSU project. For annotation of KEGG pathways, orthologs and protein families to the locus IDs, the release 6.0 coding sequences were uploaded to the online KEGG Automatic Annotation Server (KAAS) tool [142]. KAAS results were parsed and terms annotated to locus IDs. TOS mutant phenotypic data was associated to locus IDs through BLASTN alignments of Tos17 flanking sequence obtained from NCBI.

Each probe on the microarray was mapped to MSU rice locus IDs by the MSU project and the mappings are made available for download in GFF format (http://www.sanger.ac.uk/Software/formats/GFF/). For functional enrichment, terms from all four classes were counted for the background (entire genome) and for each module in the weighted network. Counting of terms is complicated because multiple probe sets can map to multiple loci and vice versa. Additionally, all 11 probes in a probe set may not map to a single locus. The nature and quantity of these many-to-many mappings are shown in the Supplementary Figure S2.

To account for ambiguity and redundancy when counting, a weighted method was performed. Probe sets that mapped to a locus with fewer than 3 probes were not considered for counting. Probe sets that mapped with more than 11 probes were also not considered for counting. The remaining probe sets contributed a count for each term equal to the following equation:

$$c(t,i,p) = (n_{ip}/11) \times (1/m_i) \times (1/q_p),$$

0 if *t* does not map to *i*,

0 if *p* does not map to *i*

In the equation above, *c(t,i,p)* is the count contributed by a probe set *p* for a given term *t* mapped to locus ID *i*; n_{ip} is the number of probes that map to *i* from probe set *p,* m_i is the total number of probe sets that map to *i,* and q_p is the total number of loci that map to *p*. Perfect one-to-one mappings contribute a count of 1 while all others contribute a value between 0 and 1. The effect of redundancy is accounted for in that the count from multiple probe sets mapping to the same locus never exceeds 1. Ambiguity is reduced by this equation but provides no effect for our purposes as we had removed ambiguity prior to counting.

Once counting was complete, pair-wise Fisher's exact tests were performed using R between the count of terms from each module in the network and the background. Terms with probability values less than 0.1, with a 95% confidence level were considered enriched.

Functional Clustering

Functional clustering was performed using a set of in-house scripts that follow the protocol used by DAVID [92, 93]. Kappa statistics are used to provide a measure of agreement between two (or more) classes of qualitative data. The Kappa *K* value provides a measure of agreement in the range 0 to 1 where 0 indicates no agreement and 1 indicates almost perfect agreement. For this study, a pair-wise kappa score was calculated for each gene using the following contingency matrix:

Locus A

		Terms Present	Terms Not Present	Total
Locus B	Terms Present	C_{11}	C_{10}	$T_{a\_}$
	Terms Not Present	C_{01}	C_{00}	$T_{b\_}$
	Total	$T_{\_a}$	$T_{\_b}$	T_{ab}

Where C_{11} is the number of terms shared by both loci A and B, C_{01} is the number of terms present in locus A but not locus B, C_{10} is the number of terms present in locus B but not A and C_{00} is the number of terms that neither loci share. T_{ab}, which is the sum of either the Total row or column, equals the total number of terms in the module. The Kappa score is calculated using the following equations:

$$\mathrm{K} = \frac{oa - ca}{1 - ca} \text{ , where:}$$

$$oa = \frac{C_{11} + C_{00}}{T_{ab}}$$

$$ca = \frac{T_{\_1} * T_{1\_} + T_{\_0} * T_{0\_}}{T_{ab} * T_{ab}}$$

In the equation above, *oa* is the observed agreement, *ca* is the chance agreement and K is the kappa score.

Clustering of terms consisted of two steps. First seed groups for each module were formed. A seed group was formed for each gene by grouping it with all other genes with which it shares a Kappa score greater than 0.5. Seed groups with less than three genes were not considered. Since probe sets may map to more than one gene the mapping counts described previously were summed and must equal three. Also, seed groups were only considered if 50% or more of the Kappa scores between all group members were greater than 0.5. Second, seed groups of a module were merged through an iterative process that exhaustively compared each group with every other group and merged any two that have 50% similarity. This continued until merging was no longer possible.

Clusters were ranked using two values, the enrichment score (en-score) and average connectivity. The en-score is the negative inverse log of the geometric mean for the Fisher's p-values from all terms in the cluster:

$$s = -\log\left(\left[\prod_{i=1}^{n} a_i\right]^{1/n}\right),$$

where s is the en-score, a_{i} is the Fisher's p-value, and n is the number of terms in the cluster.

The average connectivity, <*k*>, of the cluster is 2*L*/*N*, where *L* is the number of edges and *N* is the number of nodes in the cluster. The average connectivity, <*k*> was used as the primary characteristic for ranking clusters.

3. Gene Co-expression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice

Stephen P. Ficklin[1] and F. Alex Feltus[1,2]

[1]Plant and Environmental Sciences, Clemson University, Clemson, SC 29634, USA.
[2]Dept. of Genetics & Biochemistry, Clemson University, Clemson, SC 29634, USA.

Published in *Plant Physiology*, July 2011

Supplementary figures and tables referenced in this chapter
are available online with the manuscript.

Abstract

One major objective for plant biology is the discovery of molecular sub-systems underlying complex traits. The use of genetic and genomic resources combined in a systems-genetic approach offers a means for approaching this goal. This study describes a maize gene co-expression network built from publicly available expression arrays. The maize network consisted of 2,071 loci which were divided into 34 distinct modules that contained 1,928 enriched functional annotation terms and 35 co-functional gene clusters. Of note, 391 maize genes of unknown function were found to be co-expressed within modules along with genes of known function. A global network alignment was made between this maize network and a previously described rice co-expression network. The IsoRankN tool was used which incorporates both gene homology and network topology for the alignment. 1,173 aligned loci were detected between the two grass networks which condensed into 154 conserved subgraphs that preserved 4,758 co-expression edges in rice and 6,105 co-expression edges in maize. This study provides an early view into maize co-expression space and provides an initial network-based framework for the translation of functional genomic and genetic information between these two vital agricultural species.

Introduction

The combination of genomics, genetics, and systems-level computational methods provides a powerful approach towards insight into complex biological systems. Of particular significance is the discovery of genetic interactions that lead to desirable agricultural and economic traits in the *Poaceae* family (grasses). The *Poaceae* includes valuable crops such as rice (*Oryza sativa*), maize (*Zea mays*), wheat (*Triticum spp.*) and sugarcane (*Saccharum officinarum*) which are globally some of the most agriculturally and economically important crops [143]. Understanding complex interactions underlying agronomic traits within these species is therefore of great significance, in particular to help with crop improvements to meet the challenges of plant and human health, but also for basic understanding of complex biological systems.

In addition to their pivotal role in agriculture, grasses offer a powerful model system in that their genomes are closely conserved and functional genomic knowledge gained in one species can be hypothesized to occur in another syntenic region (translational functional genomics [144]). In cases of grass species with poorly resolved, polyploid genomes such as sugarcane where genomic resources are not as far progressed as other grasses (e.g. rice, sorghum, maize, etc.), translational functional genomics methods may be the most cost-effective strategy for crop improvement as well as unraveling the functional consequences of polyploidy. Additionally, crops rich in genetically mapped loci deposited in sites like Gramene [7] provide a rich source of systems genetic hypotheses that could in principle accelerate the translation of interacting gene sets associated with complex traits into grasses with poor genetic resources [103, 145].

One method of identifying interacting gene sets is through the construction of a gene co-expression network, which is constructed through the discovery of non-random gene-gene expression dependencies measured across multiple transcriptome perturbations, often derived from a collection of microarray data sets. During co-expression network construction, the tendency of *m* transcripts to exhibit similar (or not) expression patterns across a set of *n* microarrays is determined. In the case where dependency is determined via a correlation metric (e.g. Pearson's r), a comprehensive *m* x *m* matrix of correlation values is generated, which represents expression similarity. The "similarity matrix" is then thresholded to form an "adjacency matrix" which represents an undirected graph where edges (co-expression) exist between two nodes (transcripts) when a correlation value in the matrix is above the significance threshold. Computational methods are then applied to circumscribe groups of network nodes that are highly connected (co-expressed gene *modules*) [86, 89, 90, 146, 147]. It has been shown that genes in these modules participate in similar biological processes, and therefore guilt-by-association inferences can be applied to module genes with no known function that are connected to module genes of known function [73, 77].

Global co-expression networks are those that incorporate expression data from a variety of tissues, developmental stages and environmental conditions into a single network—the goal being to capture stable co-expression relationships across a diverse collection of experimental perturbations. Global gene co-expression networks maintain similar properties as other naturally occurring networks, such as human social networks and protein-protein interaction networks. These networks tend to be scale-free, small-world, modular and hierarchical [63, 65]. Detailed descriptions of these properties can be found in the report by Barabasi and Oltvai (Barabasi and Oltvai, 2004).

Plant co-expression networks have previously been constructed for *Arabidopsis* [55, 56, 75, 115, 117-119, 148], barley [120], rice [121, 149], poplar [126], and tobacco [123]. Several online plant resources also exist for searching co-expression relationships within and sometimes between these species, as well as incorporating functional and other data types. These include the *Arabidopsis* Co-expression Toolkit (ACT) [124], STARNET 2 [122], RiceArrayNet (PlantArraynet) [121], ATTED-II [125], Co-expressed biological Processes (CoP) database [127], AtCOECis [150], The Gene Co-expression Network Browser[149], AraNet [56] and a second AraNet [55]. Clearly, there is a burgeoning interest in using a network approach to discover gene-gene dependencies across the field of plant biology.

Given the recent and rapid increase of available biological networks, an important method is the identification of common patterns of connectivity between two networks. Inter-network comparisons are used for several purposes, including improved identification of functional orthologs between species [151], and identification of evolutionarily conserved subgraphs, or sets of highly-connected genes which demonstrate conserved function [76]. Several different network comparison methods exist which perform either local or global comparisons. Local network alignments (LNA) attempt to align small subsets of nodes between multiple networks, whereas global network alignments (GNA) attempt to find the best alignment of all nodes in one network with another [152]. Various heuristics exist for global alignment of two or more networks and typically these methods first use homology to prioritize alignment of nodes and then incorporate a measure of topology to refine alignments [153-158]. Some methods strictly use topology to guide alignments [159] given that network motifs are often conserved in functionally related systems [160, 161]. The majority of network alignment methods have been used to align protein-protein interaction networks, whereas one method has recently been published for alignment of gene co-expression networks [162].

This study adds to the growing compendium of systems-level knowledge for plants by first describing a maize gene co-expression network, and then through a global network alignment with a rice co-expression network [149], we identified common subgraphs of co-expressed gene sets between the two grass species. For network alignment, we applied a tool, IsoRankN [154], which incorporates both gene homology and network topology in its alignment algorithm. The use of homology contributes conservation of sequence, and topology contributes conservation of co-expression—both of which are associated with functional relatedness. We describe the discovery of multiple sets of modules between rice and maize that are both enriched for similar functional terms and that are potentially evolutionarily conserved between the two grasses. This functional similarity between modules in maize and rice seem to agree with idea that function may be translated through the aligned nodes of two networks. This may serve as a method for identifying functional modules in other grass species. Phenotypic associations available in the rice network may also provide an initial glimpse at the possibilities of translational systems genetics from rice to maize and other cereals. In practice, this method may assist with prioritization of genes for future mutational studies.

Results

Maize Co-Expression Network Construction

The maize co-expression network was constructed using 253 Affymetrix Maize GeneChip Genome Array microarray samples obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository. A listing of these array accessions and the experimental conditions under which the transcriptome was measured can be found in Supplemental Table S1. Construction of the maize network was performed using the same method as published previously for rice [149]. Maize microarray datasets were RMA normalized [136], and 40 outlier arrays were removed using the R/arrayQualityMetrics package [138]. Upon inspection, these outliers seemed to be a result of low-quality hybridizations or non-standard experimental conditions and did not appear to derive from a common biological system. Next, all pair-wise gene expression correlations were determined (Pearson's r). The resulting correlation (similarity) matrix was used as input into both the WGCNA soft-threshold [86] and RMT hard-threshold [80] methods for network construction. The WGCNA method identified a power of 6 to power-raise the similarity matrix and later divided the network into 34 distinct gene modules, whereas 45 modules were detected for rice (Table 3.1). The relationship between maize modules in terms of similarity of expression is shown in Supplemental Figure S1. The RMT method provided a hard threshold cut-off value of 0.5781 for the WGCNA power-raised matrix. This is the point where the nearest-neighbor spacing distribution within the network transitions from what would appear as random noise to non-random signal (χ^2 *p*-value > 0.001). The final maize network consisted of 31,983 edges between 2,708 probe sets (2,071 gene models) which corresponds to 15.4% of the original probe sets on the array (Table 3.1). A global view of the maize co-expression network can be seen in Figure 3.1, where individual modules are distinctly colored. A detailed list of edges for the maize network can be found in Supplemental Table S2. The maize network is available online, along with the previously described rice network, for browsing and searching at http://www.clemson.edu/genenetwork. Network properties, such as node-degree and clustering co-efficient distributions can be found in Supplemental Figure S2.

Table 3.1 Characteristics of the Rice and Maize Networks

Characteristic	Rice Network	Maize Network
Array	Affymetrix Rice GeneChip	Affymetrix Maize Gene Chip
NCBI GEO accession for array	GPL2025	GPL4032
Probesets on array	54,168	17,555
Genomic loci mapped to probesets	46,499	14,792
Microarray Samples [a]	508	253
WGCNA selected power threshold	4	6
WGCNA module dendrogram cutoff	0.20	0.20
RMT hard threshold	0.7101	0.5781
Probesets in network	4,528	2,708
Edges in probeset network	43,144	31,983
Loci in network	2,257	2,071
Edges in loci network	32,820	33,397
Modules	45	34
Enriched terms	2,373	1,928
Functional clusters	76	35
Clustered terms	960	596
Enriched phenotypic terms	17	N/A

[a] Number of samples remaining after outlier detection and removal.

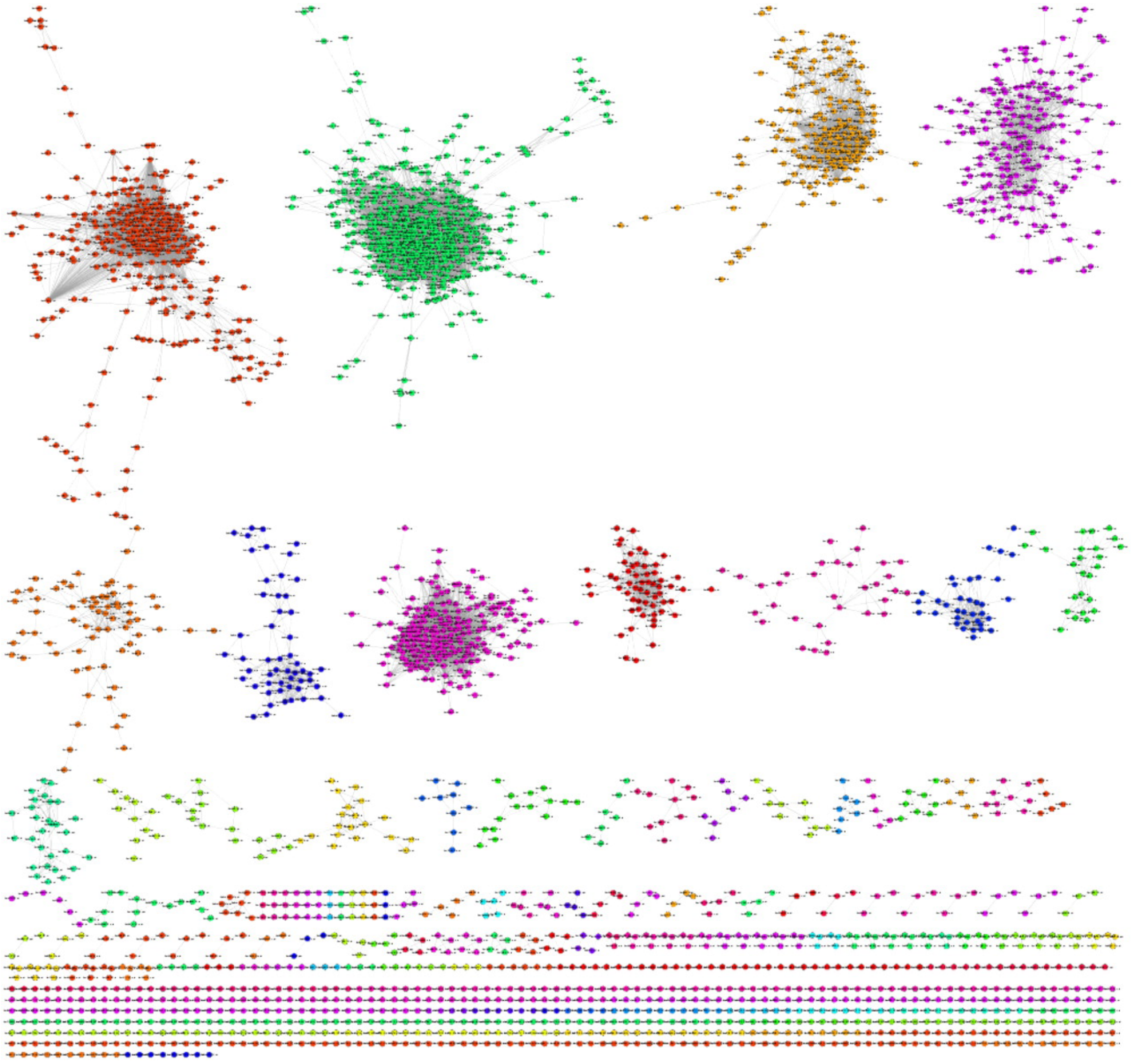


Figure 3.1 Maize Co-expression Network.

Nodes are probe sets from the Affymetrix GeneChip Maize Genome Array. Edges indicate significant co-expression between probe sets above a hard threshold. The various colors indicate the different modules of the network.

Functional Enrichment and Clustering of Co-Expressed Maize Gene Modules

Functional enrichment was performed for each of the 34 modules identified by WGCNA using annotation terms from the Gene Ontology (GO) [128], InterPro [133] and KEGG [99] using an in-house method similar to the tool DAVID [92, 93]. A total of 1,928 unique annotation terms were found to be enriched in the maize modules (Fisher's exact test; $p < 0.1$). Co-functional clusters, or subsets of nodes within a module that share enriched functional annotation, were identified using the DAVID approach. The identified clusters were sorted first by average connectivity, $<k>$, and second by enrichment score (e-score), the geometric mean of enrichment *p*-value. A total of 35 co-functional gene clusters identified from 596 enriched terms were found within 10 modules. Detailed lists of probesets and genomic loci within modules and co-functional clusters, as well as enriched annotation terms can be found in Supplemental Tables S3-S6. A total of 383 maize loci are represented in the network with no known functional annotation (Supplemental Table S7). Of these, approximately 50% or 193 of the 391 genes of unknown function have 3,092 co-expressed edges with genes in co-functional clusters (Supplemental Table S8). Therefore, it may be possible to infer function for these loci using the principle of guilt-by-association.

Interestingly, co-functional clusters ordered first by $<k>$ in both the maize and rice networks seem quite similar. A list of the top-ten ordered clusters in both networks can be seen in Table 3.2. For example, the highest ordered maize cluster by $<k>$ was enriched for functional terms related to the ribosome and translation (module ZmM6C25; <k> = 20.2; e-score= 2.62). Similarly, the corresponding rice cluster was also enriched for terms related to the ribosome and translation (module OsM6C25; <k> = 17.0; e-score = 1.98). The second highest maize cluster was enriched for seed storage activity (ZmM5C1; <k> = 10.0; e-score = 3.38), as was the third highest rice cluster (OsM13C1; <k> = 12.9; e-score = 12.22). Additional top-10 ordered clusters with similar annotation terms in both maize and rice, although not in the same order, include clusters enriched for photosynthesis, glycolysis, and microtubule activity.

Table 3.2 Side-by-side functional comparison of top-10 maize and rice co-functional clusters, ordered by average connectivity.

Maize Cluster[a]	<*k*>[b]	E-score[c]	Summarized Function	Rice Cluster[a]	<k>[b]	E-Score[c]	Summarized Function
ZmM2C1	20.2	2.62	Ribosome/Translation	OsM6C25	17.0	1.98	Ribosome/Translation
ZmM5C1	10.0	3.38	Seed storage	OsM6C4	14.1	4.13	Photosynthesis/Light harvesting
ZmM9C1	10.0	7.25	Histone/DNA Binding	OsM13C1	12.9	12.22	Seed storage
ZmM1C1	10.0	1.87	Photosynthesis/Light harvesting	OsM6C23	10.5	2.04	Carbon fixation/Carotenoid biosynthesis
ZmM4C1	8.8	4.07	Ribosome/Translation	OsM6C16	9.2	3.14	Photosynthesis
ZmM2C2	5.7	2.43	Translation elongation	OsM2C2	7.4	5.13	Kinesin/Microtubule motor activity
ZmM9C2	5.4	4.31	Histone/DNA Binding	OsM6C14	5.4	3.19	Glycolysis
ZmM11C1	5.3	3.05	Kinesin/Microtubule motor activity	OsM13C5	5.0	11.24	Transcription factor activity
ZmM1C3	4.0	2.89	Glycolysis	OsM13C2	4.6	3.57	Nutrient reservoir activity
ZmM19C1	3.9	5.38	Transcription factor activity	OsM6C11	3.7	4.08	Ribosome binding/Protein folding

[a]Modules are numbered sequentially starting from zero and are prefixed with a letter M. Clusters within a module are numbered sequentially and are prefixed with the letter C. Modules and clusters are prefixed with a species abbreviation: 'Os' for rice and 'Zm' for maize. Thus, cluster 1 from module 8 in rice is named OsM8C1. [b]<k> is the average connectivity of the nodes in the cluster. [c] E-score is the enrichment score, or geometric mean of the Fisher's test enrichment p-values of the cluster.

Rice and Maize Co-expression Network Alignment

Using the constructed maize network, a comparison with the existing rice network was performed with the goal of identifying evolutionarily conserved co-expression patterns. A comparative summary of the statistics for both the rice and maize networks can be seen in Table 3.1. To allow for direct comparison of various network alignment methods, the probe set based networks were first condensed into a locus based network that contained 2,071 loci for maize and 2,257 loci for rice (Supplemental Tables S9-S10). IsoRankN [154] was used to perform global alignments between the maize and rice network. To help clarify the meaning of the various modules, clusters and subgraphs constructed with this analysis, we provide definitions as well as naming conventions (Table 3.3).

Table 3.3 Synopsis of subgraph definition and naming convention.

Term	Definition	Naming Schema[a]
Subgraph	Any collection of nodes and edges that form a subset of the global network.	
Module	A subgraph within the global network that consists of highly-connected groups of nodes. For this study, modules are determined using the WGCNA method which groups nodes by measures of similarity. See Supplemental Figure S1.	*Sp*M*x*
Functional Cluster	A functional cluster is a subgraph within a module where the nodes have a high degree of similarity in functional terms (e.g. Gene Ontology (GO), Interpro and KEGG terms).	*Sp*M*x*Cy
Conserved Subgraph	A subgraph which is present in one network and has a corresponding subgraph in another network. These subgraphs share nodes which have been locally aligned using a network alignment tool.	subgraph_*z*

[a] For naming, 'Sp' indicates a two-letter species abbreviation; the 'M' in M*x* indicates the subgraph is a module where *x* is the module number; C indicates a cluster followed by the cluster number, *y*; *z* is a four digit number given to uniquely identify conserved subgraphs.

IsoRankN provides three input parameters that affect the results of the network alignments. These parameters include an iteration parameter, a threshold parameter and an alpha value. Documentation for IsoRankN indicates that the iteration parameter should vary between 10 to 30; the threshold between 1e-3 to 1e-5; and the alpha value between 0 and 1. The alpha value controls the contribution of homology and topology, and a value of 0 would strictly use homology for alignments whereas 1 would strictly use topology. A value of 0.5 would weight equally the contribution of both homology and topology. Therefore, to identify an adequate set of parameters for IsoRankN for alignment of the rice and maize networks, these parameters were varied from one extreme to the other within the suggested documented ranges. In total, 189 tests were performed. To measure the change of the biological signal caused by varying these parameters, we used Kappa statistics to provide a measure of functional similarity between conserved subgraphs. Functional enrichment was performed for each conserved subgraph in both maize and rice. The similarity of terms enriched in corresponding conserved subgraphs of maize and rice is measured by a Kappa score, where a value greater than 0 indicates that the conserved subgraph in rice is similar, more than could be expected by chance, to the corresponding subgraph in maize. A value of 1 indicates the two are identical in terms of enriched terms. A plot of average Kappa scores and subgraph counts across 20 alpha values, for an iteration value of 30 and threshold value of 1e-4 is shown in Figure 3.2. Aside from the extreme alpha values near 0 and 1, the functional similarity of conserved subgraphs is relatively consistent across the alpha values. The graph in Figure 3.2 was effectively identical for each combination of iteration and threshold we tested. This similarity indicates that convergence of the alignment occurs at low stringency, and that selection of almost any parameter blend for those we selected for testing would be effective. The Kappa score (or functional similarity) of the subgraphs in rice and maize at an alpha value of 0 is very high; however, the number of conserved subgraphs at that value is very low. The opposite is true for an alpha value of 1. It seemed most parameter sets, aside from the extreme alpha values, would generate an adequate set of subgraphs with a reasonably high average similarity (average kappa), so we selected conserved subgraphs derived from alignments from IsoRankN using an alpha value of 0.8, iteration value of 30 and a threshold of 1e-4 because this particular combination of parameters seemed to provide the highest average Kappa. Because average Kappa score and subgraph count were very similar across all parameter variations we only present here a single representative result set. Using these parameters, we detected 1,173 aligned loci, which were later connected into 154 conserved subgraphs. These subgraphs preserved 4,758 edges in rice and 6,105 edges in maize (Supplemental Tables S11-12). Functional enrichment and clustering, identical to that performed for the network modules, was performed for these subgraphs as well. The co-functional clusters of these conserved subgraphs can be found in Supplemental Tables S13-S14.

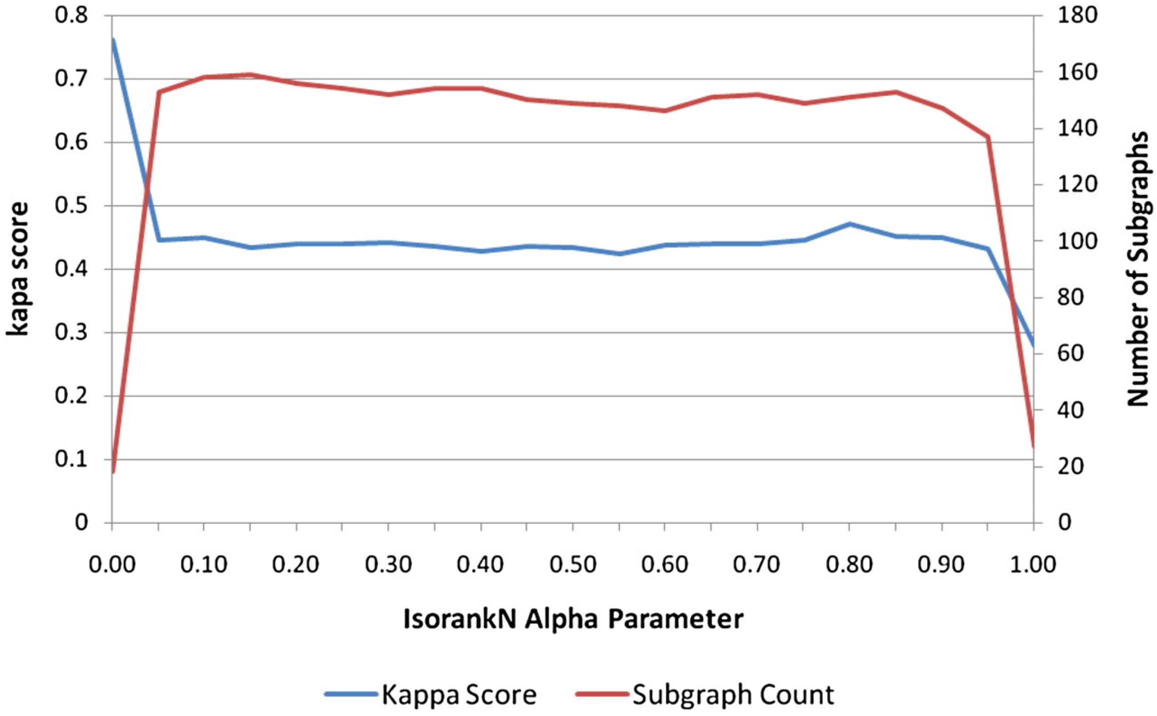


Figure 3.2 Varying the homology-to-topology ratio has little effect on conserved maize-rice subgraph discovery.

This graph shows the distribution of the average κ scores (blue line) and the number of conserved subgraphs (red line) across 20 α values for IsoRankN at an iteration setting of 30 and a threshold value of 1e-4. This graph is a representative plot for 189 trials of IsoRankN where the iteration parameters varied at 10, 20, and 30 and the threshold parameter varied at 1e-3, 1e-4, and 1e-5. Other combinations yielded almost identical graphs.

Common Function in Conserved Maize-Rice Subgraphs

For the IsoRankN alignments, functional enrichment and Kappa analysis of the conserved subgraphs yielded nine subgraphs with a perfect Kappa score of 1 indicating an identical set of enriched terms. These include subgraphs enriched for early nodulin 93 proteins, hydrolase activities, DNA binding, peptidase, nucleosome assembly, transcription factor activity and others (see Supplemental Table S16). However, these subgraphs are relatively small with two to five edges. The four largest conserved subgraphs are enriched for terms involved in photosynthesis, DNA replication, the ribosome, and starch synthase (Table 3.4), all of which have a Kappa score greater than 0.5. Incidentally, these four classes of enriched terms are also present in the top 10 list of enriched clusters as seen in Table 3.2.

Table 3.4 Top 10 largest conserved subgraphs by size.

Subgraph	Kappa Score	Maize nodes	Rice nodes	Top Enriched KEGG / GO Term for Maize Conserved Subgraph	Top Enriched KEGG GO Term For Rice Conserved Subgraph
subgraph_0107	0.64	323	278	GO:0009765 photosynthesis, light harvesting	GO:0015979 photosynthesis
subgraph_0067	0.59	120	95	GO:0000786 nucleosome	GO:0003777 microtubule motor activity
subgraph_0034	0.73	57	35	GO:0005840 ribosome	GO:0005840 ribosome
subgraph_0282	0.51	49	45	K13679 granule-bound starch synthase	K00703, glgA; starch synthase
subgraph_0624	0.15	11	2	GO:0015934 large ribosomal subunit	GO:0005840 ribosome
subgraph_0341	0.09	11	3	K02634 petA; apocytochrome f	K02709 psbH; photosystem II PsbH protein
subgraph_0046	0.87	9	3	GO:0005840 ribosome	GO:0005840 ribosome
subgraph_0031	0.72	9	7	K10999 CESA; cellulose synthase A	K10999 CESA; cellulose synthase A
subgraph_0033	0.10	9	4	GO:0005773 vacuole	GO:0003676 nucleic acid binding
subgraph_0035	0.91	8	2	GO:0015934 large ribosomal subunit	GO:0015934 large ribosomal subunit

For the rice network, phenotypic terms derived from the *Tos*17 retrotransposon insertion mutation studies [105, 134] were mapped to loci and included in the functional enrichment and clustering of network modules [149]. Of the 154 conserved subgraphs, 20 conserved subgraphs from rice are enriched for Tos17 phenotypic terms, which include phenotypes such as ‘sterile’, ‘pale yellow leaf’, ‘high tillering’, ‘vivipary’ and more. A listing of these 20 conserved subgraphs can be seen in Table 3.5. The subgraphs are ranked by descending order of Kappa score, which indicates the similarity of functional annotations between the rice and maize conserved subgraphs. Several subgraphs have a Kappa score of 1.0, indicating identical functional similarity, but overall, a high degree of similarity between most of the subgraphs is evident. Figure 3.3 shows the relationship between the global rice and maize networks (Figure 3.3 A and C respectively) with the conserved subgraphs of each (Figure 3 B and D respectively) as constructed using IsoRankN node alignments. Light grey lines between the global rice and maize networks simply map the location of nodes with their conserved counterparts. Light grey lines between the two conserved subgraph networks show node alignments provided by IsoRankN. Light red lines indicate node alignments with phenotypic associations in rice. Figure 3.4 shows a close-up view of conserved subgraph ‘subgraph_0107’.

Table 3.5 Top functional term for conserved subgraphs derived from IsoRankN in maize and rice with phenotypic associations in rice

Subgraph ID	Kappa	Maize Genes	Rice Genes	Rice Phenotypes	Maize Top Enriched GO/IPR Term	Rice Top Enriched GO/IPR Term
subgraph_0065	1.00	4	5	Pale green leaf, Long culm, Albino, Drooping leaf, Yellow, Low tillering	IPR005050 Early nodulin 93 ENOD93 protein	IPR005050 Early nodulin 93 ENOD93 protein
subgraph_0908	1.00	2	2	Pale green leaf	GO:0016787 hydrolase activity	GO:0016787 hydrolase activity
subgraph_0060	1.00	5	4	Late heading	GO:0043565 sequence-specific DNA binding	GO:0043565 sequence-specific DNA binding
subgraph_0005	0.89	2	2	Germination rate	GO:0016788 hydrolase activity, acting on ester bonds	GO:0016788 hydrolase activity, acting on ester bonds
subgraph_0046	0.87	9	3	Spl/Lesion mimic	GO:0005840 ribosome	GO:0005840 ribosome
subgraph_0907	0.67	2	2	Others	IPR005516 Remorin, C-terminal region	IPR005516 Remorin, C-terminal region
subgraph_0107	0.64	323	278	Pale green leaf	GO:0009765 photosynthesis	GO:0015979 photosynthesis
subgraph_0777	0.62	2	2	Long culm	IPR010525 Auxin response factor	IPR010525 Auxin response factor
subgraph_0067	0.59	120	95	Lamina joint, Thick culm, Lax panicle, High tillering	GO:0000786 nucleosome	GO:0003777 microtubule motor activity
subgraph_0649	0.57	5	2	Vivipary	GO:0005783 endoplasmic reticulum	GO:0005783 endoplasmic reticulum
subgraph_0727	0.56	2	2	Yellow, Narrow leaf	GO:0003899 DNA-directed RNA polymerase activity	GO:0004197 cysteine-type endopeptidase activity
subgraph_0092	0.47	2	2	Vivipary, Yellow	IPR001944 Glycoside hydrolase, family 35	IPR000922 D-galactoside/L-rhamnose binding SUEL lectin
subgraph_0029	0.44	8	4	Short panicle, Dense panicle	GO:0043687 post-translational protein modification	GO:0005840 ribosome
subgraph_0218	0.20	2	2	Virescent	GO:0006754 ATP biosynthetic	GO:0005524 ATP binding

					process	
subgraph_0621	0.18	4	3	Abnormal shoot	GO:0045735 nutrient reservoir activity	GO:0005215 transporter activity
subgraph_0893	0.17	2	3	Sterile, Stripe	GO:0006464 protein modification process	GO:0004197 cysteine-type endopeptidase activity
subgraph_0006	0.17	2	2	Vivipary	GO:0009289 fimbrium	GO:0015079 potassium ion transmembrane transporter activity
subgraph_0624	0.15	11	2	Short panicle, Abnormal panicle shape, Small grain	GO:0015934 large ribosomal subunit	GO:0005840 ribosome
subgraph_0105	0.12	4	3	Rolled leaf	GO:0016857 racemase and epimerase activity	GO:0016020 membrane
subgraph_0599	0.11	3	3	Rolled leaf, Pale green leaf	GO:0004871 signal transducer activity	GO:0007155 cell adhesion

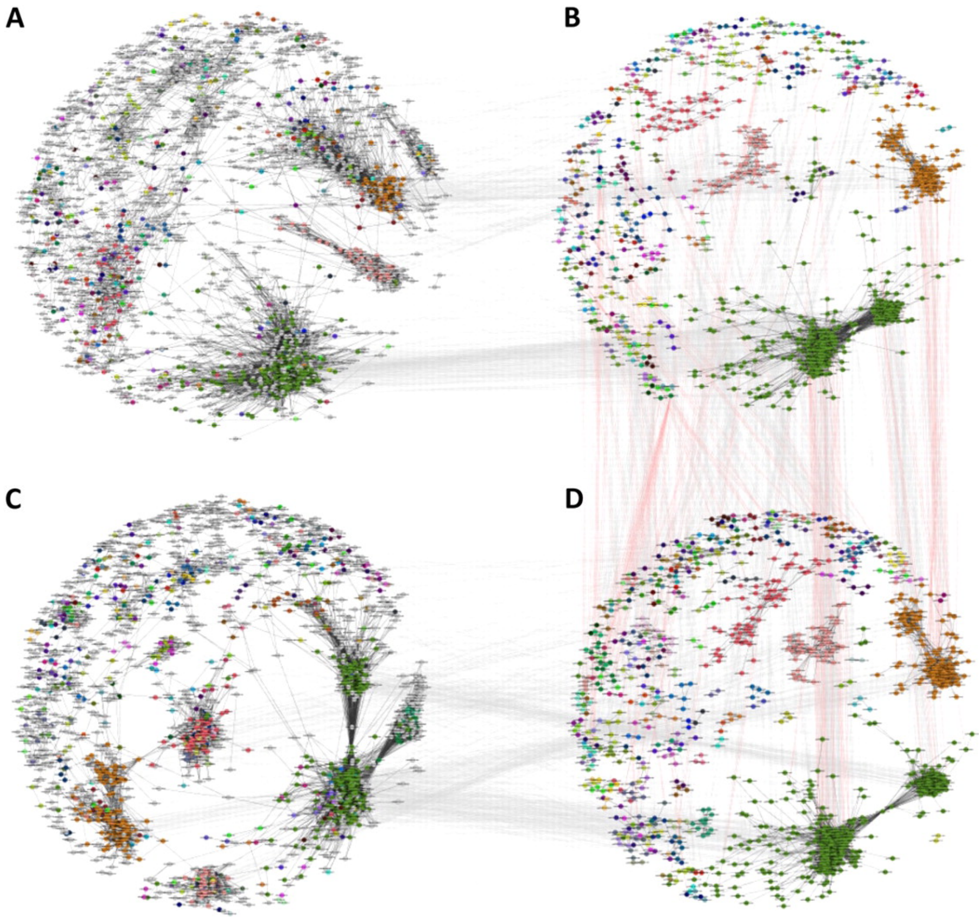


Figure 3.3 Conserved subgraphs between rice and maize.

A, The global locus-based network for rice. B, The conserved network for rice with colored subgraphs. C, The global locus-based network for maize. D, The conserved network for maize with colored subgraphs. Nodes in B and D are color coded according to the conserved subgraphs to which they belong. The same colored nodes in B belong to the same conserved subgraph in D. These same nodes are colored identically in the global networks to show global placement. Nodes colored gray in the global networks are not assigned to a conserved subgraph. Dark-colored edges in the global and conserved subgraphs represent co-expression edges. Lightly colored lines between the global networks in A and C and the conserved subgraphs in B and D simply indicate the positions of the same nodes in both types of networks. Lightly colored gray lines between the conserved subgraphs of rice and maize in B and D show the locations of aligned nodes as indicated by IsoRankN. Lightly colored red lines between B and D originate from the rice conserved subgraph in B and indicate known phenotypic associations in rice with possible translation to maize.

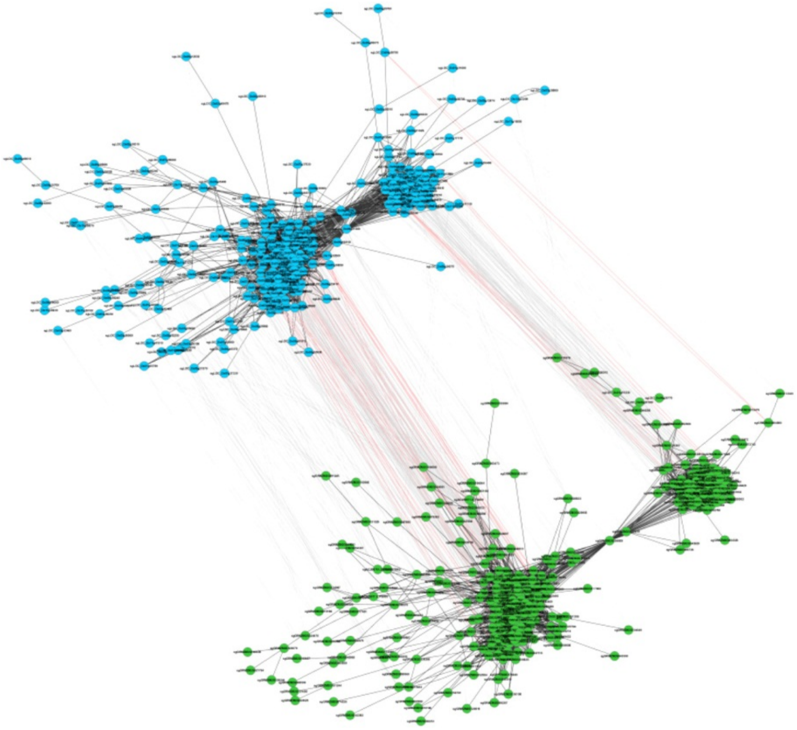


Figure 3.4 Largest conserved subgraph with implied phenotypic associations.

Shown is the largest conserved subgraph, subgraph_0107, between rice (blue nodes) and maize (green nodes). Dark edges in the subgraph are co-expression relationships. Light edges indicate alignments between the two subgraphs determined using IsoRankN. Light red edges indicate phenotypic associations with nodes in rice that are aligned to nodes in maize.

Discussion

The purpose of this study was to identify conserved, co-expressed gene sets between two vital agricultural species: rice and maize. To identify these gene sets, we first constructed a maize co-expression network, *de novo,* and aligned it to a previously described rice co-expression network [149]. Our hypothesis was that the discovery of conserved network nodes (genes) and edges would provide an initial framework for the translation of complex functional genomic and genetic knowledge from one species to another. This strategy is complementary to traditional comparative genomic approaches where known function is translated between taxa via homology and/or synteny. Additionally, the WGCNA and RMT tools were selected to preserve a knowledge-independent approach. The networks were thresholded (using RMT) and modules were constructed (using WGCNA) without prior knowledge of the underlying gene functions.

The Global Maize Gene Co-expression Network

Here we provide the first known maize gene co-expression network. This network facilitates research in maize by providing lists of interacting genes annotated for specific biological processes which provide clues to candidate gene (known and novel) involved in those processes. Additionally, 391 genes with unknown function (Supplemental Tables S7-S8) are co-expressed within modules and 194 of those un-annotated genes are interconnected within 32 different co-functional modules. For example, Cluster ZmM5C1 is the fourth highest ordered co-functional cluster by average connectivity, <*k*> and contains nine loci. However, there are 24 directly connected neighboring genes that have no ascribed GO, KEGG or Interpro function. The enriched functional terms for this cluster include seed storage activities. Guilt-by-association inferences would suggest that the 24 genes of unknown function in ZmM5C1 may be involved in seed storage or related processes. Therefore, these genes make interesting, perhaps novel, candidates for understanding the biological process associated with seed storage. In total, 194 genes of unknown function, through 3,092 edges, now suggest inference for the biological processes summarized by 33 different co-functional clusters (Supplemental Table S8).

The Small Size of Global Co-expression Networks

The maize network is small in comparison to the number of loci mapped to the probe sets present on the microarray. Using 32,540 gene models from the ZmB73 4a.53 release of the maize genome [163], 14,792 (45%) of the known maize loci were measured on the microarray platform. Of those, only 2,071 loci (14%) were present in the global maize network. We observed a similar phenomenon in rice, where almost 86% of known rice transcripts mapped to the probe sets on the microarray platform, but similarly a low fraction of the measured loci (10%) were present in the final network. Therefore, in regards to the number of potential co-expression relationships, both networks are relatively small to what we would expect across the organism’s life cycle. The Random Matrix Theory (RMT) method [80] was specifically used to define the threshold to reduce random noise from the final network to ensure that the detected co-expression relationships were strong. Therefore, the small size of the network is most likely caused by relationships lost within the “noise” of the dataset, combined with the fact that not all co-expression relationships from all conditions are represented by the dataset.

Is it possible to boost the biological signal and increase the gene space fraction captured in co-expression networks? Lowering the significance threshold, even using reasonable methods designed to limit the number of false positives, would increase the number of loci in the network, but could reduce the overall quality of the biological signal and possibly confound the interpretation of modules [79]. Usadel et al., discusses several reasons that significant co-expression correlations can be lost [164]. These include sample selection, complex interaction types, and selection of normalization and correlation methods. Our data suggests that the rice and maize networks consist primarily of co-expression relationships derived from basal biological processes whose expression is most common across the samples used to build the network. It would seem that the global co-expression networks currently available for plants, including the rice and maize networks we have generated, are immediately useful for these common processes but lack representation of less frequent processes and other subtle interactions. For maize and rice, it may be that more significant co-expression relationships would be detected if A) additional transcriptome measurements are made from tissue systems not present in the current network which would increase the sampling frequency and probability of detecting rarer co-expression relationships; B) overlap from multiple tissue-specific transcriptomes on a single sample are reduced by segregating datasets to be tissue/condition specific; C) additional statistical methods are employed to identify co-expression relationships specific to unique tissues, conditions or developmental stages, essentially dissecting the input data into subsystems. It should be noted that the detection of co-expression relationships between highly homologous transcripts including gene variants may require extensive transcriptome measurements from a non-hybridization based platform (e.g. RNAseq) before the full potential of global co-expression networks, measured in the observed number of co-expression relationships, can be realized.

Conservation Between Rice and Maize Co-expression Networks

From a qualitative perspective, the apparent collective role of genes in co-functional clusters from both rice and maize networks, when ordered by average connectivity, were quite similar (Table 3.2). Co-functional clusters were ordered by connectivity under the premise that highly co-expressed genes are more likely involved in similar biological processes. As mentioned previously, functional terms from processes such as translation, seed storage, glycolysis, photosynthesis and the cell cycle are all enriched in the top 10 functional clusters of both networks, and provide good indication that the two co-expression networks, derived from independent microarray samples for two different species, demonstrate conservation in terms of connectivity of co-expressed genes for common biological processes.

The apparent conservation of co-expression patterns between rice and maize is further bolstered through a formal global alignment of the two networks via IsoRankN and identification of conserved subgraphs. Many of the conserved subgraphs between rice and maize show a high degree of similarity of enriched functional terms indicating a high level of conservation, which we quantified using Kappa statistics (Table 3.4 and Supplemental Table S15). For example, the function of the top 10 conserved subgraphs by size is shown in Table 3.4. Many of these share similar function, especially when Kappa scores are closest to 1. This is notable because the likelihood that nodes in the conserved subgraph would be significantly co-expressed, aligned together based on topology and homology and have non-random chance of similarity between their respective enriched function is low. Moreover, modules and co-functional clusters in maize also align to modules and co-functional clusters in rice that have similar functional annotations. For example, Figure 3.5 shows the fourth largest conserved subgraph “subgraph_0282” with 49 nodes from maize (lower left) and 45 from rice (upper right). Both the maize and rice loci from this subgraph are enriched for terms involved in seed storage, nutrient reservoir activity and starch synthase with a Kappa score of 0.51. Not only are the functional enrichments similar between aligned nodes but module co-expression relationships are also maintained. The majority of maize genes in subgraph_0282 belong to module ZmM5 (with only 3 from ZmM19) and all of the genes from rice belong to module OsM13. Also, co-functional clusters show evidence of alignment as well. Within this same conserved subgraph, the orange nodes from rice in Figure 3.5 and the purple nodes from maize are from co-functional clusters OsM13C1 and ZmM5C1 respectively. Both of these clusters are enriched for seed storage activities, and the nodes from these two clusters have direct alignments with the other.

It should be noted that low Kappa scores between enriched terms of the rice and maize networks do not indicate that the node alignments are weak. Kappa scores are based on functional similarity and are dependent on the underlying functional annotation of the loci. For instance, conserved orthologous loci in two genomes may not have been annotated identically, yet are aligned due to sequence homology and network topology. Also, similar function may be annotated in somewhat equivalent yet different functional terms. Therefore, a high functional similarity between conserved subgraphs was used to help validate the network alignments, but a lack of functional similarity does not indicate a poor alignment.

Translation of Function and Phenotype

As mentioned previously, conservation between co-expression networks is a powerful tool for validating the correctness of each aligned networks. In essence, conservation reduces the noise within the network because it provides another layer of evidence for co-expression [165]. Moreover, the alignment between species strengthens the guilt-by-association inferences made for genes of unknown function. For example, cluster ZmM5C1 was described previously as containing co-expression with 24 genes of unknown function. Seven of the loci from ZmM5C1 appear in conserved subgraph_0282 (purple nodes of Figure 3.5). Guilt-by-association inferences may be applied to these genes of unknown function, however, the inference is made stronger because the co-expression relationships are conserved.

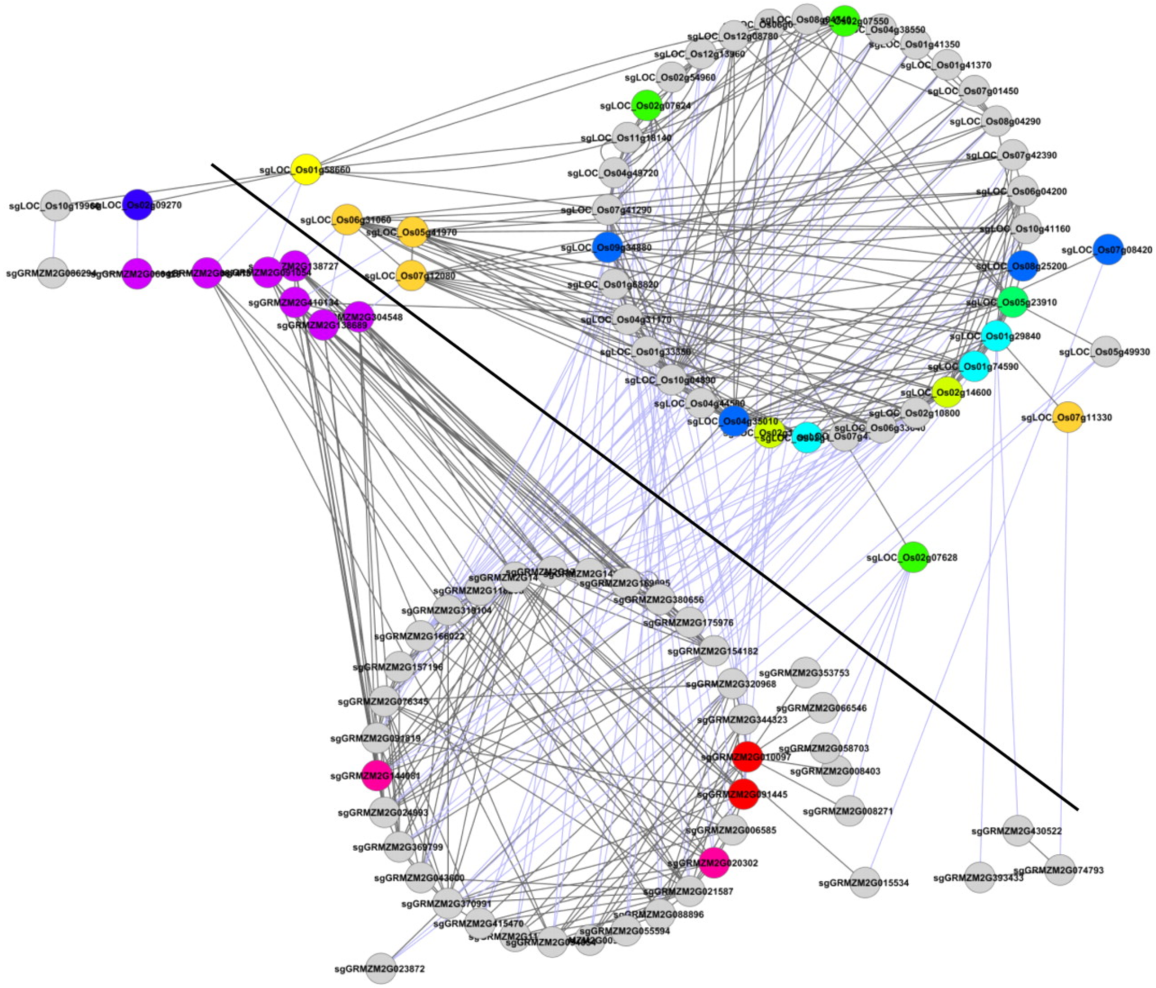


Figure 3.5 Subgraph_0282 from rice and maize.

This subnetwork shows the co-expression edges and conserved alignments for all nodes of subgraph_0282 between maize and rice. Co-expression edges are gray lines, and network alignments are light blue lines. Nodes below the heavy diagonal line are from the maize network, and nodes above it are from rice. All of the rice nodes belong to module OsM13, and the majority of the maize nodes are from module ZmM5, with the exception of the three rightmost nodes in the bottom half, which belong to module Zm19. Yellow nodes in maize are for loci of unknown function. Purple nodes in maize are from cluster ZmM5C1, annotated for nutrient reservoir/seed storage activity. Orange nodes in rice are from cluster OsM13C1, also annotated for nutrient reservoir/seed storage activity. Nodes of other colors belong to the other functional clusters within the module. Gray nodes belong within the subgraph but are not part of a co-functional cluster.

A powerful application of alignments between rice and maize networks is the potential to translate gene sets with enriched phenotypes from rice to maize through conserved subgraphs. For instance, Table 3.5 provides a list of conserved subgraphs that have enriched phenotypes in rice. These terms are not only present in annotations of genes in the subgraph but enriched. Of particular note is subgraph_0065 with 5 genes in rice, 4 genes in maize and 6 phenotypic terms: Pale green leaf, Long culm, Albino, Drooping leaf, Yellow, Low tillering. This subgraph has a Kappa score of 1 indicating perfect similarity between annotated terms and is annotated as Early Nodulin 93 protein. It may be that this high level of similarity is due in part to the fact that sequence homology was employed in network alignment, and sequence homology is often used to transfer functional annotation from one species to another. However, network topology based on co-expression edges was weighted more strongly in the IsoRankN alignment, indicating that co-expression relationships are also maintained between rice and maize alignments. Therefore, it seems appropriate that these phenotypic associations from rice can be inferred to the four maize genes as well as connected neighbors in the subgraph.

Conclusion

Gene co-expression network alignments coupled with genetic and functional genomic data provides a method for translation of gene function and genotype-phenotype associations between species, and is especially useful for species with limited genetic resources. Experimental evidence will be needed to determine the true predictive power of co-expression relationships (intra-network and inter-network), but the functional similarity we observed in conserved subgraphs seems quite promising. Still, better quantitative measures of biological signal are needed to validate the co-expression relationships. If an in silico metric for biological signal can be identified, it would provide a means to calculate Type I and Type II error under alternate network construction protocols. However, given that gene co-expression networks have already been used to successfully identify candidate genes for specific traits [55, 56] it is natural to conclude that function and phenotype can also be transferred across species to help identify genes involved in complex traits. The power of this translational systems-genetics approach will be increasingly more useful as more genetic data is made available for grasses, especially in the form of genome-wide association studies. In particular the translation of function and phenotype into large polyploidy species, such as sugarcane, would be especially powerful because the capture of genetic associations can be difficult, expensive and genome resources tend to lag behind less complex species.

Materials and Methods

Maize Network Construction

The method used for construction of the maize gene co-expression network was identical to that previously described for the rice gene co-expression network [149]. A total of 293 samples from the Affymetrix Maize GeneChip Genome Array microarray were obtained from NCBI's GEO repository. RMA normalization [136] using the software package RMAExpress [166] and outlier detection using the arrayQualityMetrics [138] tool for

Bioconductor [167] were used to remove outlier samples. Arrays that failed all three outlier tests were excluded from further analysis. Then, a similarity matrix was constructed by performing pair-wise Pearson correlations for every probe set across all samples. We selected Pearson correlations because it was commonly supported by both the WGCNA and RMT tools. Next, the WGCNA package [86] was used to convert the similarity matrix into an adjacency matrix by raising the similarity matrix to a power of 6. The power chosen is one that best approximates scale free behavior in the resulting network and is selected by the software. Finally, the RMT algorithm [80] was used to select a hard threshold which limits the noise in the resulting network.

Functional Enrichment & Clustering

The gene models used for this study were from the maize B73 genome [163] version 4.53a obtained from the maizesequence.org website. Gene Ontology (GO) [128], InterPro [133] and KEGG [99] terms were used for functional annotation of these gene models. In the case of GO and InterPro terms, these were obtained directly from the maizesequence.org website. KEGG terms were obtained by uploading maize coding sequences (CDS) to the KEGG/KAAS server which maps KEGG terms using a homology-based method [142]. An in-house tool similar to the online DAVID tool [92, 93] was used to perform functional enrichment using a Fisher's Exact test against each network module and the genome background. Modules were further subdivided into functional clusters using pair-wise Kappa statistics between all genes. Functional clusters were ordered by the geometric mean of the Fisher's *p*-values, the en-score, or by the average clustering coefficient, <*k*>, which provides a measure of interconnectedness of the nodes in the functional cluster.

Maize-Rice Network Comparison

The maize network was compared with the previously described rice network [149]. The maize and rice networks, as well as functional enrichment and cluster discovery, were constructed with an identical protocol. However, as a result of improvements to the in-house scripts that perform functional enrichment, the functional enrichment and clustering was performed again for the rice network before comparison. The maize and rice networks, including both the original and updated functional enrichment results for rice are all available online at http://www.clemson.edu/genenetwork.

Before network comparisons were performed, nodes in both the rice and maize network were converted from microarray probe sets to genomic loci. In some cases these were one-to-one mappings between probesets and genes. However, some microarray probe sets map to more than one genomic loci, and vice-versa. These mappings are ambiguous, but were retained with the assumption that a significant edge to these nodes could be informative because one or more mapped genes would be producing the correlated transcript. During conversion from a probe set to a loci-based network, edges were placed between two loci whenever they mapped to connected probe sets. Edges were also preserved in cases where a single locus mapped to more than one probe set in a different module.

Network comparisons between the rice and maize gene co-expression networks were performed using IsoRankN [154], which provided a mixed topology and homology-based global alignment methodology. First, the maize and rice protein sequence datasets were obtained from the MSU v6.0 assembly for rice [6] and the maize B73 genome [163] version 4.53a and were aligned against one another using BLASTP (parameters: -e 1e-6 -F 'm S', -s T, -m 8, and -b 1) following the recommendations given in the publication by Moreno-Hagelsieb and Latimer for selecting blast parameters for reciprocal best-hits [168]. The homolog scores derived from blast, and the network edges list were used as input to IsoRankN. Several iterations were performed by varying the parameters for the software. IsorankN's own iteration parameter was adjusted at values 10, 20 and 30. The threshold parameter was adjusted at values 1e-3, 1e-4 and 1e-5, and the alpha value which controls the contribution of topology versus homology in aligning the networks was varied from 0.0 to 1.0 in 0.1 increments. In total, 189 iterations were performed for IsoRankN. IsoRankN generates sets of one-to-many mappings where in some cases multiple aligned loci are in a single set. Each pair or group was referred to as an alignment set. Conserved subgraphs were generated using these output files with an in-house Perl script. Conserved subgraphs were constructed in a two-step method. The first step selected edges that were conserved between the two networks and the second step identified subgraphs of interconnected loci.

The process for selecting preserved edges was performed by comparing two loci from two different alignment sets in one network, with two loci from the same alignment sets from the other network. If an edge existed in both networks using the four selected loci, then both edges were marked as conserved. The following pseudo code describes the process:

```
S = the array of aligned sets of loci
for each set in S as $s_i$
  for each set in S as $s_j$ where $s_i$ is not $s_j$
    for each locus $l_i$ in $s_i$
      for each locus $l_j$, in $s_j$
        for each locus $k_i$ in $s_i$ where $k_i$ is not $l_i$
          for each locus $k_j$ in $s_i$ where $k_i$ is not $k_j$
            if  $l_i$ and $l_j$ are in the same network and connected  and
              $k_i$ and $k_j$ are in the same network and connected
            then mark both edges as conserved.
```

Edges and nodes that were not marked as conserved were discarded and the remaining networks, one for rice and the other for maize, became the "conserved" sub networks.

Finally, conserved subgraphs within the conserved networks were identified by first selecting an edge from one conserved network to serve as a seed for the subgraph. The aligned loci in the other conserved network were also used as a seed. Thus, the process of defining subgraphs was performed in parallel in both networks. Next, the edges of all of the connected neighbors of the seed were added to the subgraph. The process was continued by iterating recursively through the neighbors and adding their edges until all possible edges are exhausted. Then, a new edge, which has not yet been added to a subgraph was selected to act as the next seed until all edges in the conserved networks are placed in subgraphs. These subgraphs were labeled numerically and a label for a subgraph in rice is the same for the corresponding conserved subgraph in maize, and vice-versa.

Functional enrichment was then performed for each subgraph using the same method described previously for modules in the global network. Subgraphs were then compared using Kappa statistics. As described previously, Kappa statistics were used to provide a measure of similarity between the functionally enriched terms of genes in a network module. Here, Kappa statistics are used to provide a measure of similarity between the two conserved subgraphs of maize and rice that have the same label. Subgraphs are then ranked by Kappa score from greatest to smallest. Conserved subgraphs are given a four digit unique number prefixed with the word ‘subgraph’.

4. Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory

Scott M. Gibson[1†], Stephen P. Ficklin[2†], Sven Isaacson[3], Feng Luo[5],
F. Alex Feltus[2,4*], Melissa C. Smith[1]

[1] Holcombe Dept. of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA
[2] Plant and Environmental Sciences, Clemson University, Clemson, SC 29634, USA
[3] Dept. of Computer Science, Wittenberg University, Springfield, OH 45504, USA
[4] Dept. of Genetics & Biochemistry, Clemson University, Clemson, SC 29634, USA
[5] School of Computing, Clemson University, Clemson, SC 29634, USA

[†]Equal contribution to study.

Published in *Plos One,* Feb 2013.

Supplementary figures and tables referenced in this chapter
are available online with the manuscript.

Note: Only portions of the published manuscript related to work performed by myself are included in this chapter. The description and results regarding scalability performed by Scott Gibson are not included but are available in the published paper.

Abstract

The study of gene relationships and their effect on biological function and phenotype is a focal point in systems biology. Gene co-expression networks built using microarray expression profiles are one technique for discovering and interpreting gene relationships. A knowledge-independent thresholding technique, such as Random Matrix Theory (RMT), is useful for identifying meaningful relationships. Highly connected genes in the thresholded network are then grouped into modules that provide insight into their collective functionality. While it has been shown that co-expression networks are biologically relevant, it has not been determined to what extent any given network is functionally robust given perturbations in the input sample set. For such a test, hundreds of networks are needed and hence a tool to rapidly construct these networks. To examine functional robustness of networks with varying input we enhance an existing RMT implementation for improved scalability and test functional robustness of human (*Homo sapiens*), rice (*Oryza sativa*) and budding yeast (*Saccharomyces cerevisiae*). We demonstrate dramatic decrease in network construction time and computational requirements and show that despite some variation in global properties between networks, functional similarity remains high. Moreover, the biological function captured by co-expression networks thresholded by RMT is highly robust.

Background

Analyzing gene expression across one or more biological systems is a complex challenge for experimental design, computational resource requirements, and biological interpretation. The objective is a detailed understanding of complex gene interactions underlying biological function. A number of methods have emerged for accumulating gene co-expression relationships into networks using microarray expression profiling experiments to concomitantly measure gene activity of thousands of genes [112, 169-171]. In co-expression networks, nodes represent gene products (e.g. mRNA transcripts) and edges indicate a significant correlation of expression between a gene pair (co-expression). Groups of nodes that are highly connected (and thus correlated) indicate a biological relationship and can be separated into co-functional gene interaction modules.

Several methods have been used to construct RNA co-expression networks and all methods require an *n*-transcript by *m*-sample expression matrix as input. For example, Weighted Gene Co-expression Network Analysis (WGCNA) is a popular method that uses the network property of scale-free topology [63, 172] to identify co-expression edges without determining a specific significance threshold. The result is an undirected graph where the weight of each connection represents the strength of correlation between a pair of genes [86]. Another method is Random Matrix Theory (RMT) taken from the field of particle physics [173] which is used in a number of applications that require separating noise from disorder in a complex systems. RMT is used to determine a significance threshold and has been employed for studying wireless communication channels [174], the stock market [175], and gene co-expression networks [80]. The RMT-based approach is a reliable method for generating networks across a wide range of datasets. For example, the RMT method has been used to generate biologically meaningful networks for *E. coli*, yeast, *Arabidopsis*, maize, rice, *Drosophila*, mouse, and human [66, 80, 149].

Many methods for construction of co-expression networks compare gene expression measurements from samples across multiple experimental conditions using a correlation statistic. The most common and widely studied metric is Pearson's correlation coefficient—an appropriate measure of correlation when data is linear and follows a normal distribution. For non-normal input, the Spearman and Kendall rank correlations are common alternatives that tend to be weaker indicators in many cases but more resistant to outliers [176]. When the behavior of input data does not match these correlation methods, mutual information functions (MI) can be calculated to determine the relationships among genes. Although MI is powerful, it is significantly more computationally intensive than traditional correlation metrics, making it less attractive for large network analysis [177]. Once a statistical method has been chosen, a matrix of correlation values is computed—pairwise for each gene across all samples. These correlation values are analyzed to determine a significance threshold for separating biologically meaningful correlations from weak correlations and random noise in the system.

Despite the biological relevance of co-expression networks thresholded using RMT, an in-depth exploration into the functional robustness of the network has not been undertaken. Do changes in the number and source of input samples have an effect on the biological function represented in the network? What is the effect on capture of biological function as transcript number is decreased? One reason for the lack of detailed study on functional robustness may be that testing on a mass scale with construction of hundreds of networks across thousands of genes using existing techniques would require excessive computation time and data storage requirements.

To explore network functional robustness and algorithm scalability we describe the construction of co-expression networks from three very different organisms: *Oryza sativa* (rice), *Homo sapiens* (human) and *Saccharomyces cerevisiae* (yeast). Using real mRNA expression profiles, a series of expression matrices of varied sample and transcript measurements (microarray probe sets) were generated by randomly removing samples and probe sets from the original input dataset. The RMT-based algorithm was then employed for network construction over this wide range of input dimensions and the resulting network properties were compared with the original (non-varied) network as indicators of functional robustness. We implemented an improved version of RMT in the C programming language and demonstrate that it is highly scalable and can construct networks at an unprecedented 10^3 scale thereby enabling high-throughput network construction and analysis such as the robustness analysis we describe.

Results and **Discussion**

Network Robustness Tests

Gene co-expression networks have been shown to be useful for finding relevant gene interactions [55, 56, 75, 76, 111-113, 115, 117-121, 123, 126, 148, 149]. In some cases, gene expression data from public repositories such as NCBI GEO [8] are combined for an organism to glean as many interactions across tissue types, experimental conditions, genotypes, developmental stage or time series in order to approximate a more holistic representation of an organism’s interactome. It is not currently possible to measure expression levels of every gene in every point in time and space; therefore, it is useful to determine how missing data affects the functional robustness of the network. As new samples are added or removed, how will the significant biological relationships represented in the network change? Can any given network be considered biologically relevant or do changes in sample composition alter that relevance?

Table 4.1 Microarray samples used for network construction

Organism	NCBI GEO Platform	Samples Used	Probe Sets[a]	Genome Assembly Version	Transcripts in Genome Assembly	Genes Measured by Platform[b]	Genes in Global Network[c]
Human	GPL570	2,000	40,685	hg19	1,962,491	18,509	828 (4%)
Rice	GPL2025	1,360	52,489	MSU v6.0	67,393	37,151	2660 (7%)
Yeast	GPL2529	1,701	10,359	S288C	6,717	5,750	805 (14%)

[a] Total probe sets after removal of control probe sets, ambiguous and outlier probe sets.

[b] Only includes genes that map unambiguously to probe sets with no differentiation between splice variants.

[c] Percentage is in terms of measurable genes

Table 4.2 Conservation of relationships between global and perturbed networks

Species	Percent Samples/ Probe sets	Global Edges	Edges[a]	Shared Edges[b]	Edges Lost	New Edges	Modules	Average Kappa[c]
Human	75/100	3,111	2,763	2,622 (84%)	489	141	129	0.72
Rice	75/100	34,470	36,210	32,530 (94%)	1,940	3,680	748	0.82
Yeast	75/100	8,643	8,758	8,240 (95%)	403	518	179	0.73
Human	50/100	3,111	2,542	2,326 (75%)	785	216	117	0.66
Rice	50/100	34,470	38,620	31,720 (92%)	2,750	6,900	786	0.78
Yeast	50/100	8,643	8,559	7,869 (91%)	774	690	180	0.67
Human	25/100	3,111	2,538	2,096 (67%)	1,015	442	124	0.59
Rice	25/100	34,470	34,530	28,080 (81%)	6,390	6,450	710	0.71
Yeast	25/100	8,643	8,583	7,437 (86%)	1,206	1,146	171	0.65

[a] The average number of edges in network with samples removed

[b] Edges in common between the perturbed network and the global network

[c] Kappa = 1 indicates perfect similarity, Kappa > 0 is non-significant.

Our improved RMT software, called RMTGeneNet, allowed for mass construction of test networks using a knowledge-independent thresholding technique to check for robustness as data composition was varied. In total, 528 total networks were constructed from NCBI GEO datasets for human, rice, and yeast (see Table 4.1 for microarray platform accession). Input datasets were derived from 2,000 randomly selected human samples, 1,360 rice samples (all available at the time of study), and 1,701 yeast samples (all available at the time of study). Prior to network construction, outlier samples were removed and the normalized expression matrices were reduced by randomly removing 25%, 50%, and 75% of the original samples and/or probe sets thereby mimicking the effects of A) variable transcriptome sampling and B) variably interrogated gene space. We refer to the network with 100% probe sets and 100% samples as the "global" network. Networks with randomly removed sample and probe sets are referred to as "perturbed" networks. Topological and functional properties of the perturbed networks were each compared to the relevant global network to examine the effects of input dataset variability.

Topology Robustness Results

Most naturally occurring networks, including biological networks, maintain certain topological characteristics [63]. We measured some of these characteristics by counting nodes, edges, nodes and edges in common (or shared) with the global network, the average degree ($<k>$), clustering co-efficient, and scale-free behavior (γ) of each network. By measuring changes in topology we examined when variation in sample and probe set size creates networks that cease to look normal relative to the global network. Shared node and edge counts for the human network can be found in Figure 4.1 A and B, respectively. Boxplots for rice and yeast were similar and can be found in Supplemental Figures S6B, S6C, S7B and S7C. The non-perturbed human global network consisted of 3,111 edges and 828 nodes (Table 4.1). Randomly removing samples at 25%, 50% and 75% showed a decrease in node and edge counts. As probe sets were randomly removed, the number of edges and nodes decreased further to about one-half the nodes and one-third of edges at 25% probe sets. A similar decrease held true for both rice and yeast networks, although the effect was less pronounced for yeast (Supplemental Figures S4B, S4C, S5B, S5C). Summary statistics for all topological properties tested for human, rice and yeast can be found in Supplemental Tables S1-S10.

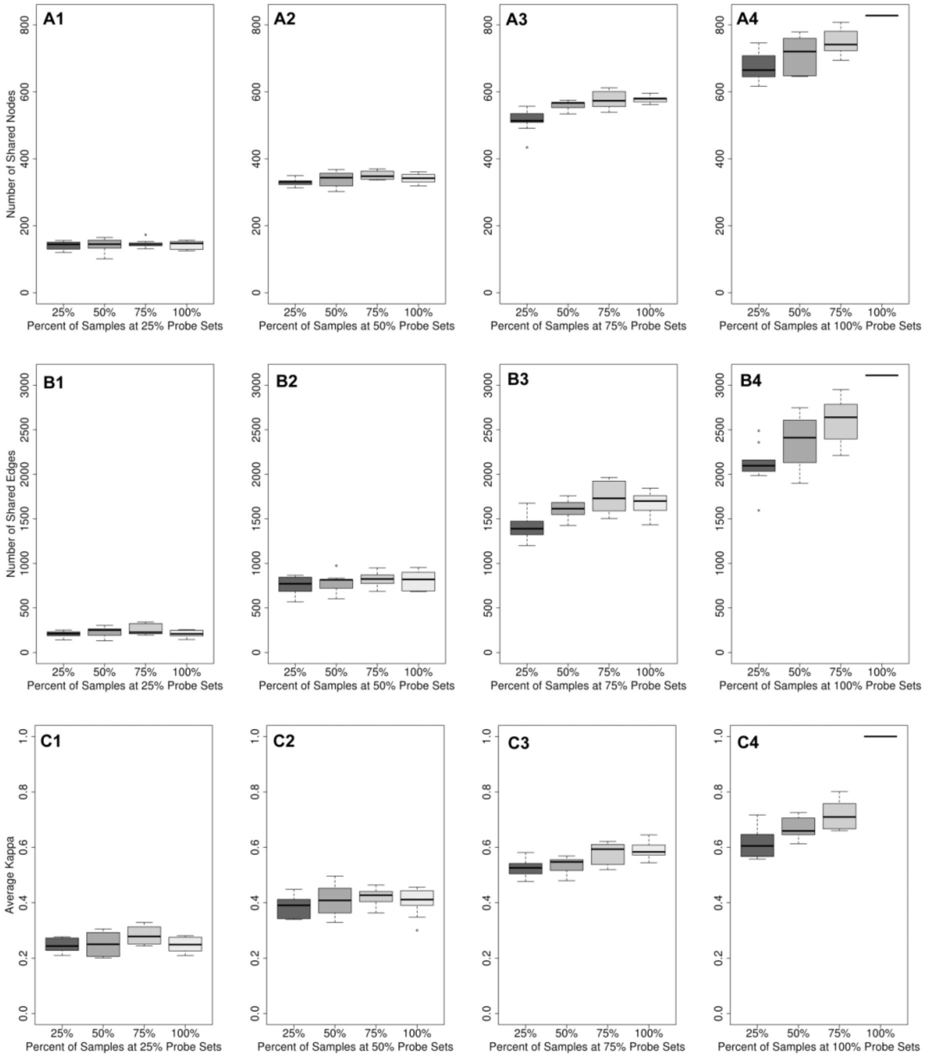


Figure 4.1 Topological and functional properties of the human networks with randomly removed samples and probe sets.

A) The number of nodes shared with the global network for each perturbed network is shown at various sample removal rates (x-axis) when probe sets were retained at rates of 25% (A1), 50% (A2), 75% (A3) and 100% (A4); B) The number of edges shared with the global network for each perturbed network is shown at various sample removal rates (x-axis) when probe sets were retained at rates of 25% (B1), 50% (B2), 75% (B3), and 100% (B4); C) The average Kappa, κ, (functional similarity) between modules in the perturbed network with modules in the global network is shown at various sample removal rates (x-axis) when probe sets were retained at rates of 25% (C1), 50% (C2), 75% (C3), and 100% (C4). The single line in the far right of plots A4, B4 and C4 represents the global network.

While edge and node counts remained relatively high when all probe sets were kept, the number of similar (or shared) nodes and edges with that of the global network quantified how interactions in the perturbed networks were consistent with the original global network. Results show that as samples were removed, the number of similar or shared nodes and edges also remained relatively high (Figure 4.1 A and 1B), but there was loss (Supplemental Figures S6-S7). In human, at 25% samples input, the average number of nodes was 751 (a loss of 77 nodes or ~10% of the global network). The average number of shared nodes was 671 indicating a further loss of 80 nodes (a total of 18% lost nodes) with a gain of 80 new nodes that were not in the global network. For edges, at 25% samples, 573 edges or 18% were lost and an additional 442 edges (14%) were not shared, indicating a total average loss of 1,015 edges (32%) with a gain of 442 new edges. Conservation of edges (relationships) for human, rice and yeast can be seen in Table 4.2 . It seems, therefore that variations in sample quantity, even at 25% samples, did not affect the majority of relationships that appeared in the network.

The 2,000 samples used as input for the human global network were randomly selected from over 48,000 candidate NCBI GEO samples and therefore should represent a blend of measurements from disparate tissues, conditions, stages and genotypes. Our results indicated that with 25% of the original samples (approximately 500 experiments) the relationships captured (shared edges) in the human perturbed network looked very similar (67%) to that of the global network. Because there were fewer samples for both rice and yeast in NCBI GEO (1,360 and 1,701 respectively) we did not randomly select from those, but used all samples for global network construction. The percent difference between the global network for rice and yeast with only 25% samples (340 samples for rice and 425 for yeast) was 18% and 13% respectively—fewer differences than for human. The fact that we saw fewer differences for rice and yeast may be because we did not randomly sample from the dataset pool as we did for human. If any given condition is over-represented in its co-expression relationships, it should suffer less effect from a decrease in number of samples.

From our results, we can expect that a sample size of near 300-500 samples would result in a network with a high number of robust relationships. An additional 1,500 samples did add a significant number of new interactions, but there were diminishing returns. For sample sets that are more random in time and space, such as the human dataset, the difference is greatest but a diminishing return was still evident.

Also, varying the number of samples had another effect—that of adding new relationships. As mentioned above, 442 new edges appeared on average in the 25% sample networks for human. Also, in some cases, such as for rice, the number of edges was greater than the global (Table 4.2). We suspect these new relationships slightly missed the RMT threshold for the global network but passed the threshold in the perturbed networks.

Removal of probe sets simulated an array platform with diminished capture of the total transcriptome. As would be expected, measuring fewer genes results in smaller networks. Loss of probe sets that measure hub nodes would create a greater loss than non-hubs, and the number of lost relationships would be dependent on the scale-free distribution: $P(k) = ck^{-\gamma}$, where $P(k)$ is the probability of any node having k connections, c

being a normalization constant and γ the power. We found that reducing probe sets by half reduces edges in the network by 45% for human, 41% for rice and 33% for yeast, and shared edges by 73% for human, 70% for rice and 73% for yeast. Therefore, a platform with reduced capacity to measure expression of all transcripts, as well as the fact that global networks only capture a small number of genes (4-14%), severely restricted the network from approximating a holistic representation of gene product interactions. These results may help qualify the amount of expected loss of co-expression space capture.

Other topological properties such as scaling exponent (γ) and clustering coefficient were measured. Supplemental Figure S8 shows an average γ that stays relatively unchanged across all levels of samples and probe sets for all three species. The estimate of γ was calculated by fitting each network to a Kronecker scale-free graph model [178] and all networks exhibited a γ of 1.3-1.6—well within the expected range for a scale-free network. For clustering coefficient, seen in Supplemental Figure S9, the value remained relatively constant across all changes in samples and probe sets—all within 0.5-0.6. These results indicate that despite changes in sample and probe set composition, all networks generated using the Random Matrix Theory (RMT) thresholding method exhibit characteristics of typical naturally occurring networks.

Functional Robustness Results

To test for change in biological function, we examined the number of link-community [87, 179] modules found in the networks. We assumed that decreases in the number of modules would result from a loss of biological relationships in the network. Similarly, a loss of modules would decrease the ability to identify functional units in a network—lowering applicability of the network (or functional robustness). Decreases in the number of shared nodes and edges indicate loss of captured relationships, which affects module detection and functional classification of modules. To measure functional similarity, terms from the Gene Ontology (GO) [128], InterPro [100, 133], KEGG [99] and Pfam [180] databases were tested for enrichment in modules. Only terms that were enriched (occurred more often than by random chance alone, $p <= 0.001$) were considered.

We also compared functional similarity of each perturbed network with the global network using Kappa statistics [181]. The average Kappa (κ) is the average of all κ from a pair-wise comparison of the modules of a perturbed network with the global network. A κ value of 1 indicates perfect functional similarity between the two networks and a value of 0 indicates no significant functional similarity. While a κ score greater than 0 indicates a significant similarity, in practice a higher κ value is typically used to threshold meaningful comparisons. We chose a stringent κ value of 0.6 as a meaningful threshold for examining biological robustness.

Functional similarity was measured by counting the number of modules (the number of co-functional groups of genes) and using Kappa statistics to measure similarity between modules. When samples were varied and probe sets remained at 100% the number of modules varied only slightly (Table 4.2), which may indicate that genes that are lost typically do not play a critical role in maintaining module structure. Kappa testing was then used to identify to what degree modules in perturbed networks were new constructs or were conserved with the global network. The average κ across all pairwise module comparisons between the perturbed networks and the global was very high for all levels of sample variation, ranging from 0.59-0.72 for human (Table 4.2, Figure 4.1C) and similar for yeast and rice (Supplemental Figure S12; Supplemental Table S9). These results indicate that networks, even with 25% of samples, are in general functionally conserved with networks that have 3 times the number of samples. Random removal of samples has little effect on the functional representations in the network. This functional consistency supports the idea that the relationships lost by a decrease in samples are primarily from genes that do not serve as hub nodes or that belong to highly-connected modules that can maintain structure despite loss of some constituents.

RMT Threshold Robustness

Finally, we were interested to identify how the RMT threshold changed as samples and probe sets were randomly removed. A rise in threshold would indicate an increase in variability of the gene expression pairwise correlations. One important characteristic of global networks thresholded using a knowledge-independent approach is that they tend to be quite small. As described previously, the human, rice and yeast global networks contained only 4%, 7% and 14% respectively of the measurable genes of their microarray platforms. This low gene count in the network is a side-effect of high-variability in the dataset. This variability is most likely a result of combining measurements from disparate tissues, conditions, developmental stages and genotypes. For the human, rice and yeast networks, there did seem to be a slight upward trend in the threshold as samples were removed, and a downward trend as probe sets were removed (Supplemental Figure S13; Supplemental Table S10). However, the changes were minimal and potentially non-significant. The results do seem to show that as probe sets are removed the variability of the dataset decreases. This stability is to be expected as probe sets are removed and cannot contribute to the correlations.

Conclusions

Results show that the RMT construction method that employs a knowledge-independent thresholding strategy is able to create networks with a high degree of robust relationships and modules. Where samples are randomly distributed across tissues, developmental stages, genotypes, etc., (such as our human dataset) networks were 67% similar despite only 25% of samples with a high degree of functional similarity (0.59κ). The robustness of networks where samples were over-representations of certain conditions, tissues, stages or genotypes, such as expected in the yeast and rice networks, exhibited even higher similarity. We conclude therefore that all of the networks where only samples varied (probe sets remained at 100%) are moderately robust. However, due to the diminishing return of adding more samples, global networks cannot serve as a mechanism for capturing and representing the entire interactome of an organism, or even at least the entire interactome measured by the collection of samples used to construct the network.

Also, the improved code exhibited approximately 29x speedup over existing methods and data storage enabling the construction of hundreds of networks for applications such as our robustness analysis. Network construction execution time was shown to scale linearly with the number of samples per probe set and exponentially with the total number of probe sets. Data storage size also scaled exponentially with the total number of probe sets indicating that future research on larger datasets will require more sophisticated computing systems with increased parallelization or algorithms optimized for many-core multi-node architectures to produce the most biologically significant data in a reasonable amount of time.

Methods

Construction of RMTGeneNet Software Package

The Random Matrix Theory (RMT) algorithm [80] used in this study was previously written in Java—a high-level programming language that excels in simplicity and portability with a wide range of pre-programmed libraries. However, it has been demonstrated that languages like C and FORTRAN generally provide better overall performance and greater optimizations because of their lower level access to computer system resources. Thus, a C implementation of the RMT algorithm was written using the GNU Scientific Library [182] and Intel® Math Kernel Library [183] to test for performance improvement and address potential optimizations. RMTGeneNet consists of three software components: 'ccm' for performing Pearson correlations of probe set expression profiles, 'rmm' for performing RMT to identify a network cutoff threshold and a Perl script ‘parse_pearson_bin.pl’ which generates a network edge list. RMTGeneNet is freely available in a GitHub repository at https://github.com/spficklin/RMTGeneNet.

Construction of Global Co-Expression Networks

Global gene co-expression networks were constructed for human (*Homo sapiens*), rice (*Oryza sativa*) and yeast (*Saccharomyces cerevisiae*). First, Affymetrix® microarray samples were obtained from NCBI GEO [8]. For the human network, a random selection of 2,000 samples was obtained from the tens-of-thousands available from the Human Genome U133 Plus 2.0 Array platform (GPL570). For rice, 1,360 samples were obtained from the Rice Genome Array platform (GPL2025) and 1,701 samples from the Yeast Genome 2.0 Array platform (GPL2529). Next, samples were RMA normalized [136] for each organism respectively using the command-line interface for the RMAExpress software [184]. After normalization, outliers were detected using the arrayQualityMetrics [138] package provided by BioConductor [167]. Samples indicated as outliers in two of three outlier tests were removed from the dataset. Ambiguous probe sets that could potentially hybridize with multiple gene products were removed from the expression data. Ambiguous probe sets were determined by mapping probe sets to genes and filtering those mapping to multiple genes. The mapping of probe sets to human genes was obtained directly using the Table Browser of the UCSC Genome Browser [185, 186] for the hg19 build of the human genome. For rice, the mappings were obtained directly from the Michigan State University (MSU) Rice Genome Annotation Project [6] for the rice genome v6.0. For yeast, the mappings were obtained by using

NCBI megablast (parameters: -W 25 -F F -D 3) to align probe sequences to the Saccharomyces cerevisiae S288C genome [187]. Next, a similarity matrix was constructed using the ccm software of the in-house RMTGeneNet package. The similarity matrix contained Pearson correlations of probe set expression profiles across all non-outlier samples. Random Matrix Theory (RMT) was then used for knowledge-independence identification of a signal-to-noise threshold for culling the similarity matrix. The rmm software of the RMTGeneNet package was used for RMT thresholding. Finally, a flat file edge list was constructed by providing the RMT threshold and the similarity matrix to the parse_pearson_bin.pl Perl script of the RMTGeneNet package. The edge list for each organism served as the final global co-expression network respectively.

Randomization of Samples and Probe sets

In order to test for network robustness, a percentage of samples and probe sets in the human, rice, and yeast datasets were randomly removed at 25%, 50% and 75% from the expression matrix: columns are samples, rows are probe sets, matrix cells are expression values. This removal process was repeated at least 10 times for each combination of samples/probe sets removed. A new network was constructed for each perturbed dataset using the RMTGeneNet package and each network was then tested using various metrics to measure robustness. Networks were constructed in parallel on the heterogeneous Palmetto computational cluster housed at Clemson University.

Author contributions

SG and SI converted the RMT Java code to C. SG generated perturbed networks for robustness testing and measured scalability. SPF constructed the global co-expression networks and measured robustness of perturbed networks. FAF and MCS directed the project. FL was the original developer of the RMT Java package and advised the group migrating to C. All authors reviewed and contributed to the content of the paper.

4.1. A Systems-Genetics Approach and Data Mining Tool For the Discovery of Genes Underlying Complex Traits in *Oryza Sativa*

Stephen P. Ficklin[1] and F. Alex Feltus[1,2]

[1]Plant and Environmental Sciences, Clemson University, Clemson, SC 29634, USA.

[2]Dept. of Genetics & Biochemistry, Clemson University, Clemson, SC 29634, USA.

Submitted for publication Feb, 2013

Supplementary figures and tables referenced in this chapter
are available in Appendix A of this dissertation.

Abstract

Many traits of biological and agronomic significance in plants are controlled in a complex manner where multiple genes and environmental signals affect the expression of the phenotype. In Oryza sativa (rice), thousands of quantitative genetic signals have been mapped to the rice genome. In parallel, thousands of gene expression profiles have been generated across many experimental conditions. Through the discovery of networks with real gene co-expression relationships, it is possible to identify co-localized genetic and gene expression signals that implicate complex genotype-phenotype relationships. In this work, we used a knowledge-independent, systems genetics approach, to discover a high-quality set of co-expression networks, termed Gene Interaction Layers (GILs). Twenty-two GILs were constructed from 1,306 Affymetrix microarray rice expression profiles that were pre-clustered to allow for improved capture of gene co-expression relationships. Functional genomic and genetic data, including over 8,000 QTLs and 766 phenotype-tagged SNPs (p-value <= 0.001) from genome-wide association studies, both covering over 230 different rice traits were integrated with the GILs. An online systems genetics data-mining resource, the GeneNet Engine, was constructed to enable dynamic discovery of gene sets (i.e. network modules) that overlap with genetic traits. Through the evidence of gene-marker correspondence, functional enrichment and genes already associated with the trait, site visitors can quickly identify genes with potential shared causality for a trait. A set of 2 million SNPs was incorporated into the database and serve as testable biomarkers for genes in modules that overlap with genetic traits. Herein, we

describe two modules found using GeneNet Engine, one with significant overlap with the trait amylose content and another with significant overlap with blast disease resistance.

Introduction

The past century has seen major advances in our understanding of genotype-phenotype relationships underlying Mendelian and complex traits controlled primarily by large-effect genes. However, methods for discovery of the genetic factors controlling complex traits are not fully mature, limiting our ability to use genetic-based methods for understanding some diseases and for breeding of certain traits in plants and animals. In plants such as *Oryza sativa* (rice), quantitative trait loci (QTL) analysis has been a key method for identifying genomic positions associated with traits of interest. While QTL analysis has been successful in associating some traits with large-effect genes [20, 21], it has failed to identify the genetic factors for traits comprised primarily of small-effect genes. In a 2009 review on the status of QTL analysis for rice, Yamamoto *et. al.* suggest the need for integration of genomics-based methods to improve the sensitivity for discovery of small-effect genes [188]. For example, gene co-expression networks, integrated with genetic and functional genomic information, offer the potential to identify large-effect and small-effect gene sets underlying complex traits. This combination of network biology, genetics and genomics data is a recent area of study and is known as systems genetics [45, 189].

Gene co-expression networks, or relevance networks [73, 76], are increasingly common tools that describe complex gene expression relationships. Co-expression networks consist of a set of nodes interconnected by edges where the presence of an edge indicates significant dependence (e.g. Pearson's correlation coefficient (PCC)) between two genes across the set of input expression profiles. Co-expression networks have specific topological properties similar to most naturally occurring networks: they are often scale-free, hierarchical and small world [64]. Typically, construction of gene co-expression networks uses microarray-derived expression profiles as input, although RNA-seq datasets have recently been used [190, 191]. A wealth of publicly available expression datasets are currently available in repositories such as the NCBI Gene Expression Omnibus (GEO; [8]), Short Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra/), ArrayExpress from the European Bioinformatics Institute [9], and other sources. The samples submitted to these repositories include a record of the experimental conditions (i.e. genotype, environment, tissue, developmental stage). After network construction, highly-connected genes are circumscribed into gene modules which tend to be involved in similar biological processes. Modules that contain genes with no known function can be ascribed putative function through "guilt-by-association" inferences [73, 77]. Many co-expression networks for plants are currently available [54-56, 75, 115, 117-121, 123, 126, 148, 192]. Also, the utility of co-expression networks has spurred development of numerous online web resources available for exploration of gene interaction relationships in plants [54-56, 121, 122, 124, 125, 127, 150].

A deepening view of gene output captured in public expression profiles can be mined to build as holistic a view as possible of gene interaction for an organism. Typically when co-expression networks are constructed input samples are either segregated using a knowledge-dependent method [192, 193] or combined into a single input set [54, 55, 150]. However, there are limitations to both approaches for maximal discovery of an organism's interactome. Segregating samples using a knowledge-dependent approach relies on human knowledge, and sometimes imprecise and inconsistent vocabularies to identify conditions. Even for highly controlled experiments, unknown variables in each sample set increase noise within the dataset, thus limiting capture of co-expression relationships. Combining all samples into a single compendium exacerbates the problem, especially as the sample set contains measurements from a highly diverse set of conditions [82]. While a completely holistic, "pan" co-expression network is not possible (as we cannot measure every gene in every experimental condition), improved, knowledge-independent methods are needed to detect co-expression relationships for all conditions using smarter dataset sorting approaches.

Therefore, the objective of this work was to build a high resolution series of rice gene co-expression networks using an optimized RMTGeneNet network construction pipeline [60]to bring a high-level, holistic view of the interaction space of rice—one of the most important staple food crops in the world. Knowledge-independent methods for network construction and module discovery were employed to overcome knowledge-bias in the detection of rice gene interaction. Prior to co-expression network construction, we used K-means clustering of input microarray samples to maximize capture of gene interactions that otherwise would be hidden in noise of typical network construction methods. Our approach generated multiple co-expression networks from the full set of Affymetrix GeneChip® Rice Genome arrays available in NCBI GEO at that time–one for each K-means cluster. We refer to each network as a Gene Interaction Layer (GIL). Using this improved capture of gene co-expression in the GIL collection, we aimed to integrate genetic data from QTL and Genome Wide Association Studies (GWAS) to highlight network modules with potential quantitative phenotype association. To help explore the rice GIL collection and associated genetic signals, we created a new online data mining resource called GeneNet Engine for exploration of network modules with potential association to genetic traits. Genes within significant network modules serve as potential candidates underlying complex genetic traits and potentially contain small effect genes.

Results

Network Construction

Prior to network construction, 1,306 microarray samples were downloaded from NCBI GEO [8] and pre-processed including normalization, outlier detection and removal of control and ambiguous probesets. Ambiguous probesets are those that map to more than one locus on the rice genome. In total, 123 control probesets and 4,772 ambiguous probesets were removed, as well as 19 outlier samples. Microarray samples were then clustered into 25 groups with similar expression using K-means clustering. A network for each K-means cluster was then constructed using the RMTGeneNet package [60]. RMTGeneNet first generates pair-wise Pearson Correlation Coefficients (PCC) for all genes and then uses Random Matrix Theory (RMT) [80] to identify an optimal threshold for culling PCC values. Of the 25 clusters, the RMT method generated 22 co-expression networks, or Gene Interaction Layers (GILs). The three dataset groups that failed to generate networks had very high expression similarity across all probesets. The number of input samples per GIL ranged from 19 to 231 with an average size of 53.8 and a median of 39 (Table 5.1). The probesets of the input samples of each GIL were mapped to 46,498 of the 57,133 genes (81%) on the Michigan State University's (MSU) v6.0 Rice genome [54]. The collection of GILs contains 282,484 edges among 16,664 nodes (genes) and together captures 35% of the measurable genes of the array and 29% of the total genes of the MSU v6.0 genome. For all GILs, the PCC threshold was quite high, ranging from 0.91 to 0.99 indicating that all relationships (edges) are highly co-expressed.

Table 5.1 Network Details from k-means clustered microarray samples into 25 groups.

Network	Input Samples	Outlier Samples	Total Edges	Total Nodes	RMT Threshold	<*k*>[a]	Modules
G0001	22	0	*Failed to construct*				
G0002	81	0	19,338	2,890	0.91	13.38	569
G0003	73	2	36,346	3,155	0.92	23.04	676
G0004	14		*Failed to construct*				
G0005	26	0	14,641	1,991	0.97	14.71	290
G0006	32	0	38,331	1,914	0.97	40.05	355
G0007	90	0	12,059	2,806	0.91	8.60	370
G0008	65	3	18,383	3,276	0.91	11.22	476
G0009	25	0	3,579	1,366	0.99	5.24	173
G0010	16	1	*Failed to construct*				
G0011	74	0	10,411	1,624	0.96	12.82	397
G0012	40	0	29,971	3,622	0.93	16.55	522
G0013	37	0	2,366	896	0.98	5.28	150
G0014	118	3	6,738	1,963	0.90	6.87	330
G0015	21	0	9,374	1,034	0.98	18.13	256
G0016	21	1	2,358	1,670	0.97	2.82	129
G0017	24	0	8,688	1,607	0.96	10.81	194
G0018	36	0	8,434	1,660	0.95	10.16	216
G0019	19	1	4,689	3,022	0.97	3.10	234
G0020	73	3	6,268	2,308	0.91	5.43	260
G0021	231	2	227	204	0.98	2.23	24
G0022	58	0	7,516	2,007	0.92	7.49	279
G0023	54	3	34,398	1,596	0.94	43.11	302
G0024	39	0	5,880	2,512	0.95	4.68	325
G0025	57	0	2,489	1,167	0.98	4.27	135
Total	1,346	19	282,484	n/a[b]			6,662

[a] The average degree of a GIL.

[b] The total number of nodes is 16,664 across all GILs and nodes may be present in multiple GILs.

Gene Module Detection and Co-Similarity

The link community method [87] was used to find modules of highly interconnected nodes within the GILs. In total, 6692 link community modules (LCM) were discovered. Modules were named using a three-part schema separated by an underscore (e.g. OsK25v1.0_G0011_LCM020), where the first part 'OsK25v1.0' represents the *O. sativa* GIL collection version 1.0 (derived from presorting with K-means 25), a second part prefixed with the letter ‘G’ indicates the GIL to which the module belongs and the third part prefixed by ‘LCM’ indicates the unique module within the GIL. The average number of modules per GIL was 302.8 and the median 284.5 (Figure 5.1). The collection of GILs represents interactions between 35% of the measurable genes and some of those genes are present in more than one GIL. As shown in Figure 5.1A, the majority of nodes are present in only a single GIL (6,608 nodes, 40%), and the number of times a node appears in multiple GILs decreases. Edges tend to be more unique per GIL as 201,121 (71%) are only found in a single GIL and the number of times an edge appears in more than one GIL is significantly less (Figure 5.1B).

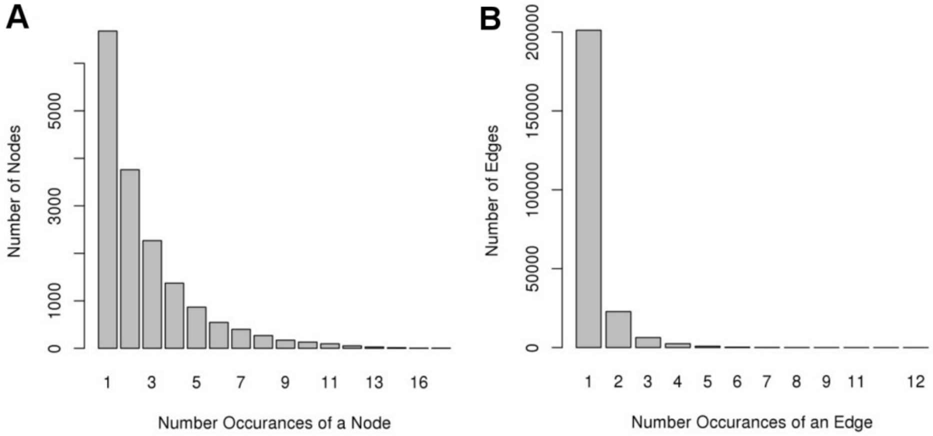


Figure 5.1 Redundant Edges and Nodes.

The number of times that a A) rice gene (node) or B) co-expressed gene pair (edge) appears in different GILs.

To obtain a measure of similarity between modules across all GILs, a correlation between Kappa scores (measuring functional similarity between two modules) and

Jaccard indices (measuring similarity of node composition) was performed. First, functional enrichment analysis of the modules was performed using terms from the Gene Ontology (GO; [128]), InterPro [100] and KEGG [99]. Only terms enriched within a module with a Fisher's *p*-value of 0.01 or less were considered enriched. Next, full pair-wise comparisons between modules with 30 or more nodes from all GILs were performed using both Kappa statistics and a Jaccard similarity test. Only enriched functional terms were used with the Kappa test. Kappa scores range from -1 to 1 with values less than 0 indicating no significant similarity of function and a score of 1 indicating identical similarity of function. A Jaccard index ranges from 0 (indicating no nodes in common) to 1 (all nodes in common). Figure 5.2 shows a scatterplot of Jaccard similarity coefficients versus Kappa scores with $R^2 = 0.5$ (*p*-value < 2.2e-16) indicating a good degree of correlation between the node composition of modules and the enriched function of modules.

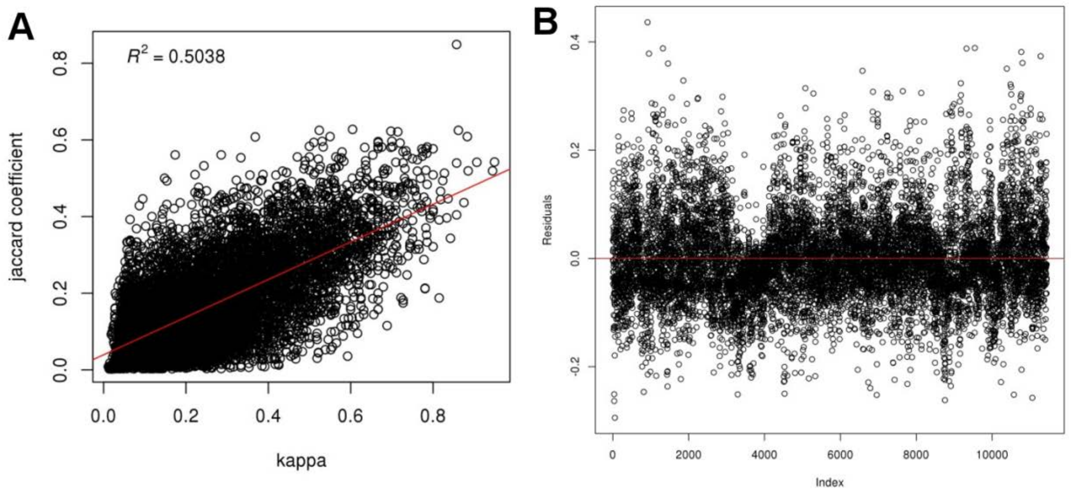


Figure 5.2 Jaccard vs Kappa Scatterplot.

Jaccard (similarity of node composition) and Kappa (similarity of functional annotation) statistics were performed, pair-wise, for all modules across all GILs. A) The scatterplot of Jaccard coefficient vs Kappa κ for all modules with 30 or more nodes. B) Residual plot of Jaccard coefficient vs Kappa κ.

A meta-network of LCM modules was then created using the similarity scores as described previously. In theory, a Kappa score greater than 0 can be considered meaningful however in practice higher values are often used for greater stringency. We used a Kappa score threshold of 0.5, which corresponds to a Jaccard score of approximately 0.3 in the scatterplot of Figure 5.2. Edges were added to the meta-network between pairs of modules with a Kappa score of 0.5 or greater. Figure 5.3 shows a diagram of the LCM module meta-network. In this network, the nodes are LCM modules and edges indicate a high degree of similarity (Kappa > 0.5 and Jaccard > 0.3). The edges are color-coded according to the GIL to which the modules belong. If two modules from different GILs shared an edge, then the edge was black. The meta-network contains 13,578 edges across 4,965 LCM modules (75% of all LCM modules). The number of edges in the meta-network that connect LCM modules of the same GIL is 12,253 (90%) with 1,325 (10%) connecting two different GILs.

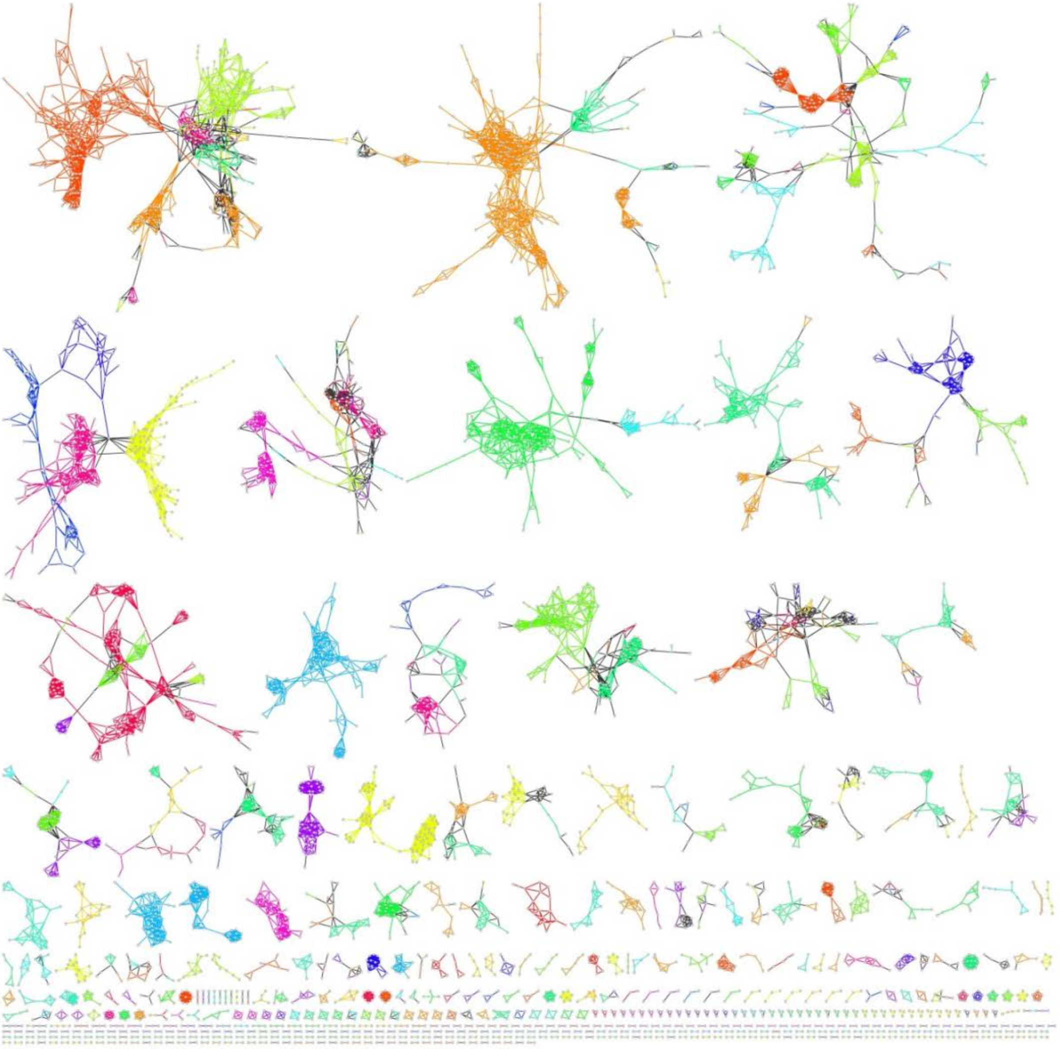


Figure 5.3 Gene Module "meta-network."

The nodes in the meta-network are LCM modules from all GILs that have a pair-wise Kappa score >= 0.5 and Jaccard coefficient >=0.3. Edges are colored if both nodes in the edge belong to the same GIL. Each GIL is assigned a unique color. Edges where each node belongs to a different GIL are black. Nodes are grey.

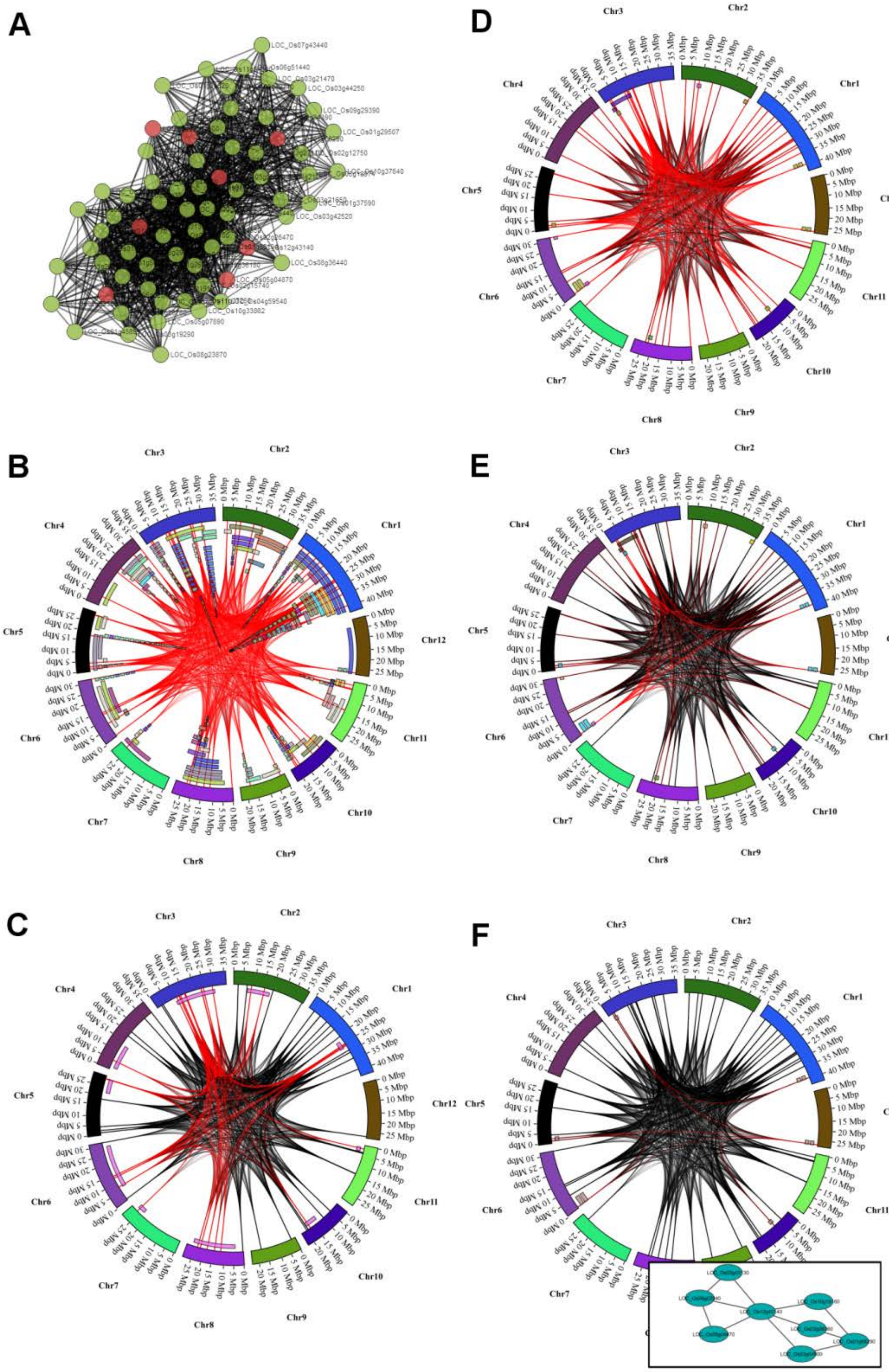
Interactive Systems-Genetics Exploration Tool

To integrate genetic data with GILs, and to construct an online resource for exploration of genotype-phenotype relationships, the physical positions of significant genetic data from QTLs and GWAS studies were obtained. Over 8,000 QTL intervals, along with their corresponding genomic coordinates were downloaded from Gramene's QTL database [17]. Also a 300kb LD window surrounding significant SNPs (*p*-value < 0.0001) from a recent GWAS study by Zhao et al were integrated [24]. Genes overlapping both QTL and GWAS SNP intervals were putatively assigned the trait. These associations as well as all GILs were input into an online database called GeneNet Engine which is available online at http://sysbio.genome.clemson.edu. The data is housed in a Chado database schema [194] with custom tables and visualized using Tripal [195]. Next all available rice SNPs from NCBI's dbSNP [196] database were uniquely mapped to the rice genome and loaded into the database so that an end-user can identify proximal biomarkers for genotype-phenotype hypothesis testing. Users can query the database using a locus name, module name, functional term, or trait of interest to examine the possibility that one or more modules may play a role in a particular function. Supplemental Figure S1 provides a screen shot of the search engine.

The GeneNet Engine also provides a module explorer. The module explorer (Supplemental Figure S5) consists of a set of tabs that provides network visualization ('Module View' tab), a genome network visualization ('Genome View' tab), lists of module nodes, edges, functionally enriched terms, a form for specifying traits to select ('Filter by Trait' tab), a list of all overlapping traits and genetic features, and a form for generating a list of potential SNP biomarkers that flank highlighted nodes within a specified window size. In the network module view, an interactive module is provided using Cytoscape Web [197]. Users are presented a network module with which they can move nodes, and zoom in and out. Clicking a node will provide functional annotations about the node (locus details box in Supplemental Figure S2). In the 'Filter by Trait' tab, users can dynamically alter the module view or genome view by selecting one or more specific traits, a genetic feature type (e.g. QTL or GWAS SNP) and by limiting the number of overlapping traits an edge must pass through to be highlighted (Supplemental Figure S3). Additionally, circular plots are available in the 'Genome View' tab allowing visitors to visualize the network within the context of the chromosomal coordinates as well as visualization of QTL or GWAS SNP regions that overlap with nodes in the module. Examples of circular plots for the module OsK25v1.0_G0002_LCM0431 can be seen in Figure 5.4. For reference, the module view is present in Figure 5.4A. Figure 5.4B-F highlight changes in the genome view as filtering parameters are changed. Figure 5.4B shows overlapping edges with QTLs for plant height. Edges with at least one node within a QTL region are colored red. In cases where there are large QTLs or where QTLs cover large swaths of the genome, almost all of the edges are red. Figure 5.4C shows the same plot but only with a single QTL set for plant height. These QTL are all from the same genetic map and fewer overlaps are present. Figure 5.4D shows the same module overlapping genetic features for the trait amylose content. While not as dense as QTLs for plant height they do overlap a large portion of the module. Therefore, a limit that an edge must pass through at least 3 different genetic features was imposed for the image in Figure 5.4E. Figure 5.4F contains the plot for amylose content overlapping QTLs from a single genetic map. Users can obtain *p*-values for the filters they employ by looking on the 'Genetic Features' tab of the Module Explorer.

Figure 5.4 Circular Genome Plots of Network Module OsK25v1.0_G0002_LCM0431.

The chromosomes of rice are shown as the outer circle. Gray arcs are edges of the module. Endpoints of each edge are fixed on the physical location in the genome where the node (gene) is found. Red arcs are edges that overlap a genetic feature. The colored tiles along the chromosomes represent genetic features (e.g. QTLs or regions around significant SNPs in GWAS). A) Network view of the module. Red nodes overlap with genetic traits for amylose content, green nodes do not. B) Circular plot of the module with all genetic features for plant height. C) Plot with QTLs from a single genetic map (Cornell 9024/LH422 RI QTL 1996) and edges highlighted red where an edge overlaps at least two QTL. D) Plot of the all genetic features for amylose content where edges overlap with at least 1 genetic feature. E) Plot of all genetic features for amylose content with overlap of at least 3 genetic features. F) Plot of module edges with QTLs from a single genetic map (CNHAU Zhen97/H94 QTL 2005) with overlap of at least 3 genetic features. The inset graph shows the connectivity of the overlapping nodes.


A
LOC_Os07g43440
LOC_Os06g51440
LOC_Os03g21470
LOC_Os03g44250
LOC_Os09g29390
LOC_Os01g29507
LOC_Os02g12750
LOC_Os10g37640
LOC_Os03g21650
LOC_Os01g37590
LOC_Os03g42520
Os02g26470
Os12g43140
LOC_Os08g36440
LOC_Os05g04870
Os02g15740
Os04g59540
LOC_Os10g33882
LOC_Os05g07890
Os09g19290
LOC_Os08g23870
B
C
D
E
F
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
0 Mbp
5 Mbp
10 Mbp
15 Mbp
20 Mbp
25 Mbp
30 Mbp
35 Mbp
40 Mbp

Discussion

The primary objectives of this project were three-fold. The first objective was to use all publicly available microarray-based RNA expression profiling data in NCBI GEO to generate co-expression networks for *O. sativa* that could capture as many gene interactions as possible. Second was to integrate, on a massive scale, network nodes with results from genetic analyses such as QTL and GWAS studies with the expectation that network modules could serve as a genome reduction strategy for finding genes that may be associated with a given trait. The final objective was to construct a systems genetics data mining platform for discovery of relationships between network modules and genetic traits and the reagents that could be readily used for hypothesis testing.

One major challenge mentioned in the Introduction was that of overcoming an increase in noise as disparate samples from various conditions are used for network construction. Performing a gene pair-wise correlation across all samples only allows for genes that are similarly expressed across all conditions to be found. Gene correlations expressed in only a few samples will not be found due to dilution. A larger and more diverse input dataset would result in a smaller network [82]. Additionally, thresholding methods such as *ad hoc* methods [70-73] have been used to allow for flexible thresholding but, they provided little statistical guidance and can incorporate non-significant relationships. To capture all relationships in the dataset, we avoided methods that require bait genes, such as linear regression [75]. Rank-based methods [76, 77] did offer an attractive feature in that they allow for dynamic thresholding. Dynamic thresholding does not apply a constant threshold across the entire set of PCC values, but rather examines the neighborhood around each gene to determine the threshold. Partial Correlation and Information Theory (PCIT) [82] and supervised machine learning [84, 85] also generate high-quality networks with dynamic thresholding, but were not currently adaptable to our network pipeline. By pre-clustering of samples based on gene expression pattern alone, we are able to use Random Matrix Theory (RMT) to provide thresholding for a highly significant set of relationships for each GIL. A unique RMT threshold is determined for each GIL, thus our approach behaves similarly to a dynamic thresholding method but without dependence on global PCC values such as the case with rank-based methods. Because RMT is knowledge-independent and is not biased towards prior and possibly incomplete knowledge, we were able to capture a very high quality set of relationships derived solely on the underlying expression values. While we used K-means clustering for pre-sorting, one benefit to our approach is that any number of data clustering methods could be used.

The availability of rich genetic data for rice was a key motivation for this study. We used approximately 8,000 genome mapped QTLs from Gramene. The Gramene curators painstakingly mapped markers for all QTLs to the MSU v6.0 genome assembly, thus providing genomic coordinates for the QTLs. The precise genes causal for many of the traits underlying these QTLs are unknown. Therefore, we simply assigned the QTL trait to all genes underlying the QTL intervals. Given the imprecision of QTL mapping, and our assigning a trait to all genes underlying a SNP or QTL region, we introduce many false positive gene-phenotype associations. The visualizations and lists provided on the GeneNet Engine (Figure 5.5 and Figure 5.7) will highlight all genes and edges from a network module that overlap with a QTL or GWAS SNP, but most likely will include false positives by random chance alone. The probability that a network module could contain a gene underlying a region for a genetic feature can be quite high, especially in the case of large QTLs, many QTLs for the same trait or where the module is large. Additionally, other factors such as tandem array genes (TAGs) can bias correspondence *p*-values due to overlap redundancy. TAGs typically are involved in similar function or pathways and hence would be co-expressed and typically present in the same module. TAGs therefore would bias *p*-values calculations that expect a normal distribution. Despite these challenges we simply provide a Fisher's test as a probability metric for false positives. However, we caution that this is only meant as a guide for filtering modules of interest, and further work is needed to identify an appropriate method for *p*-value calculation.

Because we mixed samples from a variety of *O. sativa* genotypes we obtained network relationships for the species as a whole and not specifically for a single genotype. Therefore, it may be possible that a network module may represent pathways specific to an individual or subspecies, and other modules could be specific to other subspecies. Moreover, a module could be a conglomeration of interactions across a set of individuals or subspecies. As evidence for this, a linear relationship exists between the square root of the number of QTLs (across all studies) and the amount of genome space they cover (Figure 5.5). This seems to confirm the notion that hundreds (or potentially thousands) of genes may contribute to a trait, and as more genotypes are analyzed, the more genes that are captured by QTLs. The GWAS study by Zhao *et. al.* also suggests that different groups of genes control the same trait in different subpopulations [24]. Therefore, it would seem that the collection of all QTLs for a given trait becomes an approximation of a pan-QTL set for the species. Similarly, the GIL collection is an approximation of a pan co-expression network.

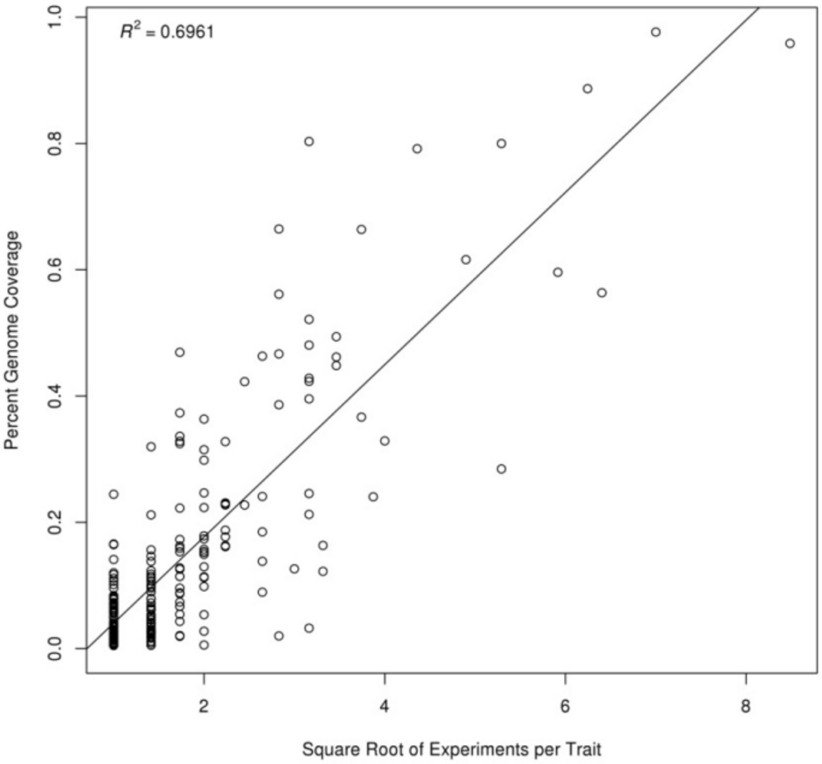


Figure 5.5 Number of QTLs per Trait vs Genome Coverage.

The scatterplot shows the relationship between the total percent covered of the physical genome versus the square root of the number of experiments per trait for QTL data from Gramene. Inset shows plot of residuals.

To demonstrate the use of the GeneNet Engine, we use as an example the trait amylose content. It is well understood that the Waxy gene (Wx) plays a major role in amylose content [198]. This gene resides on chromosome 6 of *Oryza sativa* and is at locus LOC_Os06g04200 on the MSU v6.0 genome. A recent study of 171 rice accessions shows that two SNPs in the Waxy gene account for 86.7% of the variation in amylose content [199], indicating it is a large effect gene. Recently, Zhao *et. al.* included amylose content as a trait in their GWAS study and significantly identified 68 SNPs associated with amylose content with a mixed model *p*-value < 1e-4 [24]. In an effort to find small effect loci that may affect variation in amylose content, a search was performed using the GeneNet Engine. Using the search page a filter was entered that provided the Waxy gene locus, LOC_Os06g04200, as well as overlap with the amylose content trait. In this case, the genetic feature was limited to a 'GWAS SNP'. The result yielded 6 modules from the Rice GIL collection and one from a previous global rice network [54] which has also been added to the GeneNet Explorer. Most of the network modules were small (between 5–15 nodes). In the GIL collection, the largest module was OsK25v1.0_G0023_LCM0301, with 30 nodes, and it had the largest average connectivity ($<k>$ = 17.47) indicating that the nodes were more highly interconnected than the other 5 modules. The GeneNet Engine provides a Fisher's *p*-value as a simple means for filtering modules that may have a high probability of false positives. As mentioned previously, this *p*-value is simply a guide and does not necessarily imply a high probability of causality for the trait. The top enriched functional terms for all 7 modules included seed storage protein (IPR006044), alpha-amylase inhibitor (IPR013771), and transcription factor CBF/NF-Y (IPR003958). All 6 GIL collection modules were present in GIL G0023 except for one (enriched for Transcription factor CBF/NY-Y) which was present in GIL G0003. Starch synthase (K00703) was also enriched in all 7 modules. All 6 of the Rice GIL modules overlapped with only 1 or 2 GWAS SNPs, with *p*-values quite high (from 0.2 to 0.03), indicating a high probability of false positives. However, after including overlapping genes underlying QTLs using the 'Filter by Trait' tab in the Module Explorer, the *p*-values were all lower and the most highly connected GIL module, OsK25v1.0_G0023_LCM0301, overlapped with 13 QTLs and 2 GWAS SNPs (15 genetic features) with a *p*-value of 1.9e-4 (Figure 5.6). The module from the global network was much larger, overlapped 4 GWAS SNPs and 34 QTLs but had a high probability of false positives (*p*-value = 0.03). While p-values were not significant for some of the smaller modules, it would seem that any of these modules could be potential candidates to explore small-effect variation in amylose content. Potentially, combining several of these modules may provide, as a group, a set of possible small-effect candidate genes. The OsK25v1.0_G0023_LCM0301 module seemed most suited for exploration as it is relatively small (only 30 genes) had a significant *p*-value (1.9e-4) and all nodes were highly connected indicating a high degree of cooperation. The effects of these genes may be examined through additional lab experiments, such as where plants with mutations can be grown and phenotyped. As a direct means for verification through experimentation, GeneNet Explorer can provide a list of SNPs that could potentially serve as biomarkers for breeding. For module OsK25v1.0_G0023_LCM0301, over 4200

SNPs were obtained, all within 50kb of genes that overlapped genetic features for amylose content.

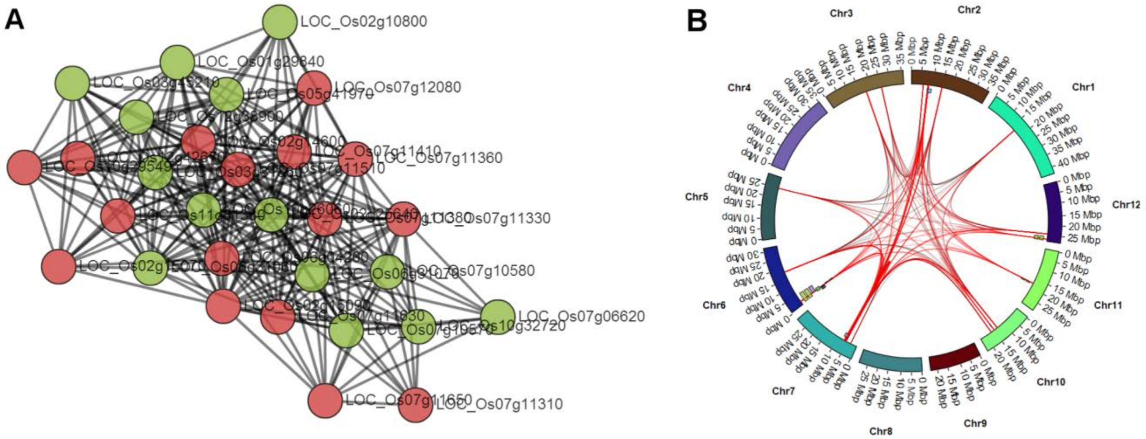


Figure 5.6 A Significant Module for Amylose Content.

Module OsK25v1.0_G0023_LCM0301 significantly overlaps with 15 different genetic features (2 SNPs, 13 QTLs, p-value=1.9e-4) and is significantly enriched for Bifunctional trypsin/alpha-amylase inhibitor helical domain and starch synthase. A) Red circles indicate nodes that overlap with genetic features and green nodes do not. B) The distribution of module edges along the genomic chromosomes. GWAS SNPs are barely visible as tick marks whereas QTLs are visible as small colored blocks along the chromosomes. Edges are red if one node lies within the region of a genetic feature.

As a second example we use the trait for blast disease resistance. The Pi-ta gene is known to be associated with blast resistance [200]. The locus for this gene on the MSU v6.0 genome is LOC_Os12g18360, but unlike the example for amylose content, it does not appear in any network modules. Additionally, 200 QTLs are present for blast disease resistance which covers a large portion of the genome. Therefore, the chance that any module would overlap with the set of QTLs for blast disease resistance is high.

However, only 16 GWAS SNPs were associated (mixed model *p*-value < 0.0001). Therefore, a search was entered into GeneNet Engine to find modules overlapping with the blast disease resistance trait but only that overlapped with GWAS SNPs. Additionally, a limit of 10 nodes was included to limit the appearance of smaller modules. A total of 242 matching modules were returned. Results were sorted by an increasing node size and examined to find modules overlapping more SNPs than other modules of similar size. The module OsK25-v1.0_G0008_LCM0015 had a module size of 25 nodes and overlapped with 3 SNPs while others of similar size overlapped with 1 or 2. Figure 5.7 shows the network view and genome plot for this module which has a false positive Fisher's *p*-value of 5.9e-4. The module is enriched primarily for an Ankyrin repeat (IPR002110), but also for Syntaxin (IPR006011, IPR006012, SNARE proteins) and for disease resistance protein (IPR000767). There are several Ankyrin repeat containing proteins that are involved in many biological processes but they are also known to participate in disease resistance, such as in the case of the *OsBIANK1* gene which is expressed during infection of Magnaporthe grisea, the blast disease fungus [201, 202]. Additionally, Syntaxin SNARE has been shown to participate in resistance to pathogens through membrane-vesicle fusion in the delivery of anti-pathogen compounds [200]. The evidence provided by the overlap of 3 module nodes with 3 of the 16 SNPs (p-value = 5.9e-4) associated with blast disease resistance in addition to the functional annotations make the OsK25-v1.0_G0008_LCM0015 a good candidate for further study of potential genes that participate in resistance to blast fungus infection. Any of the genes in this module could potentially serve as small effectors of the trait.

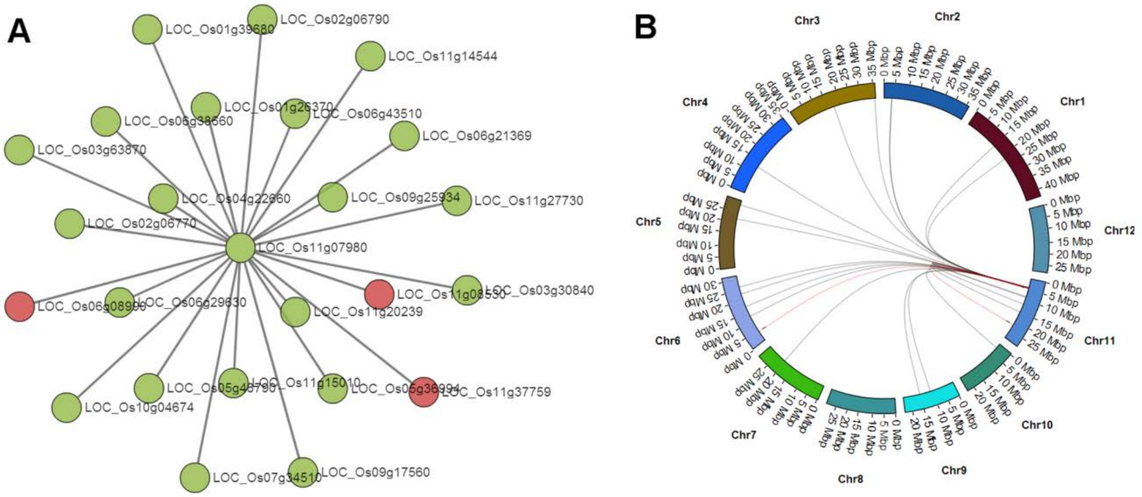


Figure 5.7 A Significant Module for Blast Disease Resistance.

Module OsK25-v1.0_G0008_LCM0015 significantly overlaps with 3 different GWAS SNPs (p-value = 5.9e-4) and is functionally enriched for Ankyrin, Syntaxin and disease resistance protein. A) Red circles indicate nodes that overlap with genetic features and green nodes do not. B) The distribution of module edges along the genomic chromosomes. GWAS SNPs are barely visible as tick marks and edges are red if one node overlaps the region surround a GWAS SNP.

The methods for discovery of significant modules for both examples above were somewhat different. In the first example a known gene was used to guide discovery of interesting modules, whereas for the second a significant module was found by browsing through a few hundred results. As seen in Figure 5.4, a module can overlap with many genetic features from multiple traits (e.g. plant height and amylose content). This should be expected naturally as genes are known to be multi-functional, but most likely many of these overlaps are false positives. Therefore, the GeneNet Engine will calculate *p*-values for false positives dynamically as users change filtering parameters in the Module

Explorer, thus allowing users to explore different filters. Also, as mentioned previously, the more experiments across genotypes the more likely the QTLs will cover more of the genome, creating more false positives raising *p*-values for all modules that overlap with the trait. In these cases, users may want to focus on modules that overlap with individual genetic maps. Users can filter by genetic map in the ‘Filter by Trait’ tab of the GeneNet Explorer (Supplemental Figure S3). Therefore, it may be necessary to apply various searching approaches to find modules of interest for a specific trait, but as demonstrated in the two examples, interesting modules for further testing can be found.

The rice K-means 25 GIL collection and the GeneNet Engine are the first release of a large-scale, integrated systems-genetic resource for plants to help with prediction of genes underlying complex traits. However, several improvements can be made. The choice of a *K* value of 25 was selected by using the common “rule of thumb” function of $k = \sqrt{(n/2)}$. However, we were only able to capture 35% of the measurable genes on the Affymetrix GeneChip array. This fell short of our goal to capture near 100% of the measurable genes; however this level of coverage is possible. In another study where the approach of pre-clustering was applied to *Arabidopsis thaliana*, approximately 98% of genes were capture in the GIL set (unpublished data). For that study, a *K* value was selected by iterating through different *K* sizes to maximize gene capture. It would be beneficial to find a more appropriate value of *K* for constructing a rice GIL collection that captured relationships from more genes in the array, with the potential of capturing all of them. Alternatively, other more dynamic pre-clustering methods may be used other than *K*-means to improve interaction capture.

Additionally, it may be beneficial to augment module detection to take into account overlap with genetic features. For this project we used the link community method for module discovery [87]. This method and many others rely on parameter settings that can be more or less inclusive. Therefore, network modules are a function of not only the underlying connectivity but the parameters used during execution of the algorithm. Generating modules that optimally capture a specific biological process is challenging and one set of parameters may capture well some processes but not others. In the first example for amylose content, all 6 modules overlapped with genetic traits for amylose content, had the Waxy gene and all had similar functional enrichment. All modules were relatively small with the exception of the largest module, OsK25v1.0_G0023_LCM0301, and all modules, except one, came from the GIL G0023. This concurs with the fact that GILs tend have modules of similar function. As seen in the scatterplot of Figure 5.2 and the meta-network of Figure 5.3, network modules tend to be most similar to other modules within the same GIL. It would seem, therefore, that the module detection algorithm could potentially take advantage of genetic and functional relatedness to stitch together potentially more significant modules. But, in summary, a more flexible and dynamic module creation method may improve the creation and identification of gene sets underlying complex traits.

Conclusion

Here we present the Rice GIL collection of networks that are a first attempt at using pre-clustering of *O. sativa* RNA expression profiles to capture all co-expression relationships measured by the full compendium of publicly available microarray samples at NCBI GEO. Our goal has been to guide network construction and module discovery solely through the evidence of gene expression. The knowledge-independent approach reduces bias towards our limited knowledge of the underlying biological processes. We integrate experimentally validated genetic data from over 8,000 rice QTLs from Gramene and significant SNPs from a recent rice GWAS study to create a platform for discovery of network modules that may be associated with trait causality. The platform is made available in the form of an interactive website named GeneNet Engine found at http://sysbio.genome.clemson.edu. The value in this approach is two-fold. First, it brings to light potentially small-effect genes (those that are connected in the module) and serves as a filtering technique to locate genes that underlie genetic features for complex traits such as QTLs. We anticipate that significant or interesting modules from GeneNet Engine can be used for further lab-based experimentation which can translate to quicker discovery of genes underling complex traits and further application in rice breeding.

Materials and Methods

Construction of the Rice GIL Networks

Before construction of the Rice GIL networks, all available samples from the Affymetrix GeneChip® Rice Genome array were obtained from NCBI GEO [8]. At the time, 1306 samples were retrieved. All samples were then pre-processed with RMA normalization [136] using RMAExpress [184] and sample outliers were detected using the arrayQualityMetrics package [138] for BioConductor [167]. Samples that failed at least two of the three outlier test were removed. The output consisted of an m x n expression matrix where m is the number of samples and n is the number of probesets on the array. Next, control probes were removed from the matrix as well as ambiguous probes that mapped to more than one gene.

After pre-processing the samples in the expression matrix were then grouped. The *kmeans* function of R was used to segregate samples into sets of similar overall expression. A value of k = 25 was determined using the common "rule of thumb" function of $k = \sqrt{(n/2)}$, and hence 25 clusters of samples were generated. Twenty-two separate networks were then constructed by first passing each group through the same pre-processing, quality control pipeline described previously: samples within a group were normalized, outliers were removed and control and ambiguous probesets were removed. Three K-means clusters did not construct networks and were removed. The list of microarray samples and the *K*-means cluster (and GIL) are provided in Supplemental Table S1.

Next, the co-expression network for each *k*-means group was constructed using the RMTGeneNet software package [60]. RMTGeneNet is a software package written in the C programming language that quickly generate correlation matrices and network adjacency matrices. RMTGeneNet first performs pair-wise correlation analysis for every probeset on the array, generating an *m* x *m* similarity matrix of correlation values ranging from -1 to 1. Next, it employs Random Matrix Theory (RMT) [80] to find an optimal threshold. According to RMT, the more random a matrix, the more the nearest-neighbor spacing distribution (NNSD) of eigenvalues appears Gaussian. The less random, the more Poisson-like it appears. RMT determines a threshold for the similarity matrix by measuring when the NNSD ceases to appear Poisson (*p*-value = 0.001). An adjacency matrix is constructed by setting all values less than the threshold to zero. In total, 22 adjacency matrices were produced: one for each K-means cluster. Finally, probesets were mapped to genes in the MSU Rice v6.0 [6] assembly of the *Oryza sativa* genome, and 22 gene co-expression networks, or Gene Interaction Layers (GILS), were constructed. GILs were generated in parallel using Clemson University's Palmetto computation cluster.

Module Discovery

After construction of the 22 GILs, modules were determined using the link-community method [87]. This approach allows a gene to be present in multiple modules. This approach is more reasonable for multi-functional genes and does not restrict genes to a single module such as other methods (e.g. MCL [88]). We used the *linkcomm* function for R [179] to generate LCM modules for all 22 GILs.

Functional Enrichment

All modules from all GILs underwent functional enrichment analysis to look for significantly over-represented terms in relation to the genomic background. Terms from the Gene Ontology [128], and InterPro [100] databases mapped to genes were obtained directly from the MSU website and KEGG [99] terms were mapped to genes using the KEGG Automatic Annotation Server [142]. Functional enrichment was performed using a DAVID-like [92, 93] Perl script developed in-house. Terms enriched with a Fisher's test *p*-value < 0.01 where kept.

Genome Mapping of Genetic Data

Genetic data from the Gramene QTL database [7, 17], and from a recent GWAS study [24] were used in this study for associating traits with network modules. Gramene curators used marker information to map over 8,000 QTL regions from various studies to positions on the *Oryza sativa* MSU v6.0 genome sequence. We then putatively associated all genes underlying the QTL regions the QTL trait. QTLs that only mapped to a single marker and were therefore smaller than 5bp were enlarged to 2Mb. For the GWAS study, only significant SNPs (*p*-value < 0.0001) from the mixed model analysis were used. Genes within a 300kb window around the SNP were putatively associated with the SNP trait. The range of 300kb flanking was used because this was the estimated average linkage disequilibrium for *Oryza sativa japonica* reported in the GWAS study (the largest of the three subspecies). The trait names used for both the QTLs and SNPs are from Gramene's Trait Ontology (TO) [203]. The TO terms used for both QTLs and SNPs were provided by Gramene. Additionally, traits from the *Tos*17 retrotransposon study [104, 105] were also included in this study but were associated to network modules using the same process as for functional enrichment described previously. The process was the same as described for the global network for *Oryza sativa* [54]. The gene assignments to *Tos*17 phenotypes as well as enrichment are present in the GeneNet Engine but are not discussed in this manuscript.

Data Storage and Visualization

All genomic, genetic and network data was stored within a Chado database [194]. Custom tables were created for storing network data (nodes, edges, and modules). Materialized views were constructed to enable faster searching. Visualization of genomic, genetic and network data was implemented using Tripal [195], an open-source publicly available construction toolkit for online genomic and genetic databases. A custom Tripal extension module was written specifically for this project and used for display of network data, as this functionality was not already part of Tripal. Cytoscape Web [197] was used for the network module visualization and the d3 JavaScript library (http://d3js.org) was used for drawing the circular genome plots. Network modules from all 22 GILs are searchable on the GeneNet Engine v0.9 site at http://sysbio.genome.clemson.edu. The Tripal Network extension module is freely available, but is under active development and is therefore available upon request.

Use of SNPs

SNPs from NCBI’s dbSNP database [196] for *Oryza sativa* were obtained through bulk download from dbSNP’s FTP site. SNPs were then mapped to the MSU v6.0 build of the *Oryza sativa* genome using blat [204]. Only SNPs that mapped once to the genome with a minimum percent identity of 0.98 across the full length of the SNP flanking sequence were kept (approximately 2.8 million). These SNPs were loaded into the database and are intended to serve as potential biomarkers.

Conclusion

This dissertation presents a knowledge-independent systems-genetic approach to identify gene sets underlying complex traits and their associated biological functions. The first effort was to construct a global co-expression network for rice and integrate that network with genetic data from *Tos*17 mutational insertion studies. The rice network was constructed to be a “global” network such that all samples were used in the construction. Several modules were identified that had significant enrichment of phenotypes from the *Tos*17 study. The implication is that those modules could contain the genes, not known before, to have an effect on expression of the trait.

Next, the ability to translate knowledge gained from one systems-genetics analysis in one species (rice) to another closely related (maize) was examined. The expectation was that gene sets that underlie complex traits in one species could be used as predictors of the same trait in closely related species. The study showed a high-degree of topological and gene sequence similarity between modules identified in the rice co-expression network and a novel maize co-expression network. This suggests that network modules in one species can be used to identify modules in another species that underlying a complex trait.

Third, the robustness of the co-expression network itself was examined to identify the level that interactions observed were side-effects of some sample bias. Were the functional modules in the global network stable? Results showed that despite some minimal variability in node and interaction composition, the modules were in fact quite stable, thus reinforcing the results obtained in the previous two studies. Additionally, the collaboration with the Smith Lab yielded the RMTGeneNet software package capable of quickly producing RMT thresholded networks and afforded the ability to generate hundreds of networks in a short period of time that were used to test robustness.

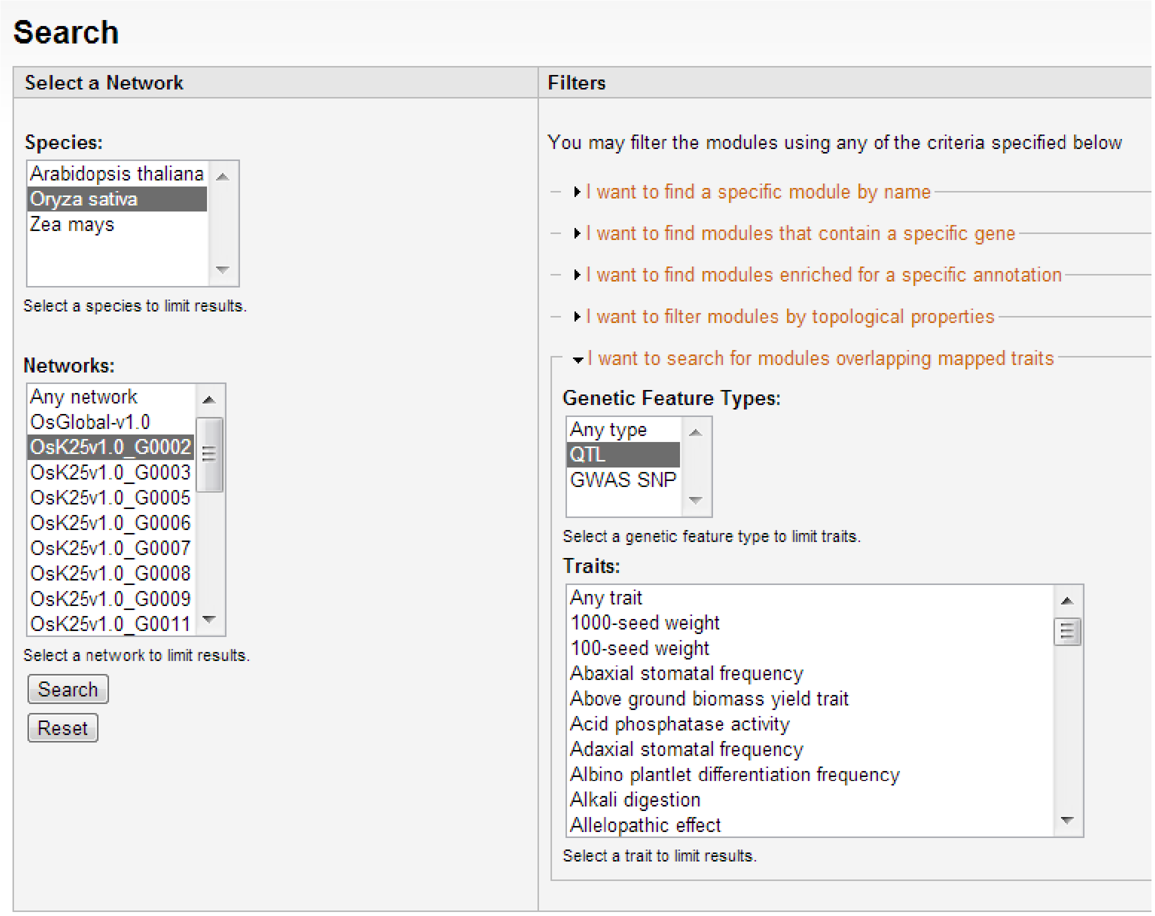
Fourth results from a GWAS study of 34 agronomic traits in rice and over 8000 QTLs in rice were integrated with a new set of co-expression networks for rice. These co-expression networks form a collection termed the Gene Interaction Layer (GIL) collection. The rice GIL collection captured over 30% of the measurable genes of the rice microarray platform and greatly increased the number of interactions. The goal was to approximate a more holistic view of rice co-expression relationships while avoiding experimental bias. Examples of modules from the GIL collection that had significant overlap with loci from GWAS and QTL studies were presented. A new online exploration tool called the GeneNet Engine (http://sysbio.genome.clemson.edu) was also introduced.

As described in the Introduction (Chapter 1) of this dissertation, the goal of this work was to identify the gene sets underlying complex traits in grasses as well as the their functional context. QTL Mapping, GWAS and Genome Selection methods could identify regions in linkage disequilibrium with causal genes but in many cases could not identify the genes themselves. Also, the functional role and interactions of those genes was unknown. Did the work presented in this dissertation answer the initially stated goal? The answer is yes and no. First, systems-genetic analysis and results have been lacking in the grasses. So, the work performed here has added a wealth of new resources and tools for exploring gene relationships, functional gene modules and potential genotype-phenotype relationships for a set of very important plants. Gene sets, or network modules, have been identified with significant association to phenotypes derived from existing genetic studies, and a mechanism to look for new associations has been provided to the community through the GeneNet Engine exploration tool. Examination of network robustness and the ability to translate systems-genetic knowledge between closely related species has added to the greater understanding of networks both in general and within the grasses. Therefore, a portion of the goal was met, in that progress has been made. However, validation of these modules with significant association to complex traits needs to be performed. While modules have been identified, it is not known the amount of heritability of the trait that these modules explain. Are they applicable to only a few genotypes? Does the module discovery method adequately circumscribe all the necessary small and large-effect genes for the trait? How many genes are present in the modules that may have little or no effect on the trait, and would they need to be removed before the module could be used in a breeding strategy? What side-effects are there if one of these significant modules were used with a breeding strategy? Are there underlying environmental effect that limits the success of these modules as a selection tool? Many questions remain unanswered and while results show promise, much work remains to be done.

In conclusion, systems-genetics offers a promising avenue as a technique that can augment existing strategies to unravel the mechanism underlying complex traits. The tools and results presented here offer a start to such studies for the grasses. With continued improvements to the analysis methods and expansion of this work into lab and field-based validation the systems genetics approach should prove beneficial for improving important agronomic traits in grasses with the continued objective to improve the quality and availability of such an important group of species.

Appendices

Appendix A Supplemental Figures and Tables From Chapter 5



Search

Select a Network

Species:

Arabidopsis thaliana
Oryza sativa
Zea mays

Select a species to limit results.

Networks:

Any network
OsGlobal-v1.0
OsK25v1.0_G0002
OsK25v1.0_G0003
OsK25v1.0_G0005
OsK25v1.0_G0006
OsK25v1.0_G0007
OsK25v1.0_G0008
OsK25v1.0_G0009
OsK25v1.0_G0011

Select a network to limit results.

Search

Reset

Filters

You may filter the modules using any of the criteria specified below

- ▸ I want to find a specific module by name
- ▸ I want to find modules that contain a specific gene
- ▸ I want to find modules enriched for a specific annotation
- ▸ I want to filter modules by topological properties
- ▾ I want to search for modules overlapping mapped traits

Genetic Feature Types:

Any type
QTL
GWAS SNP

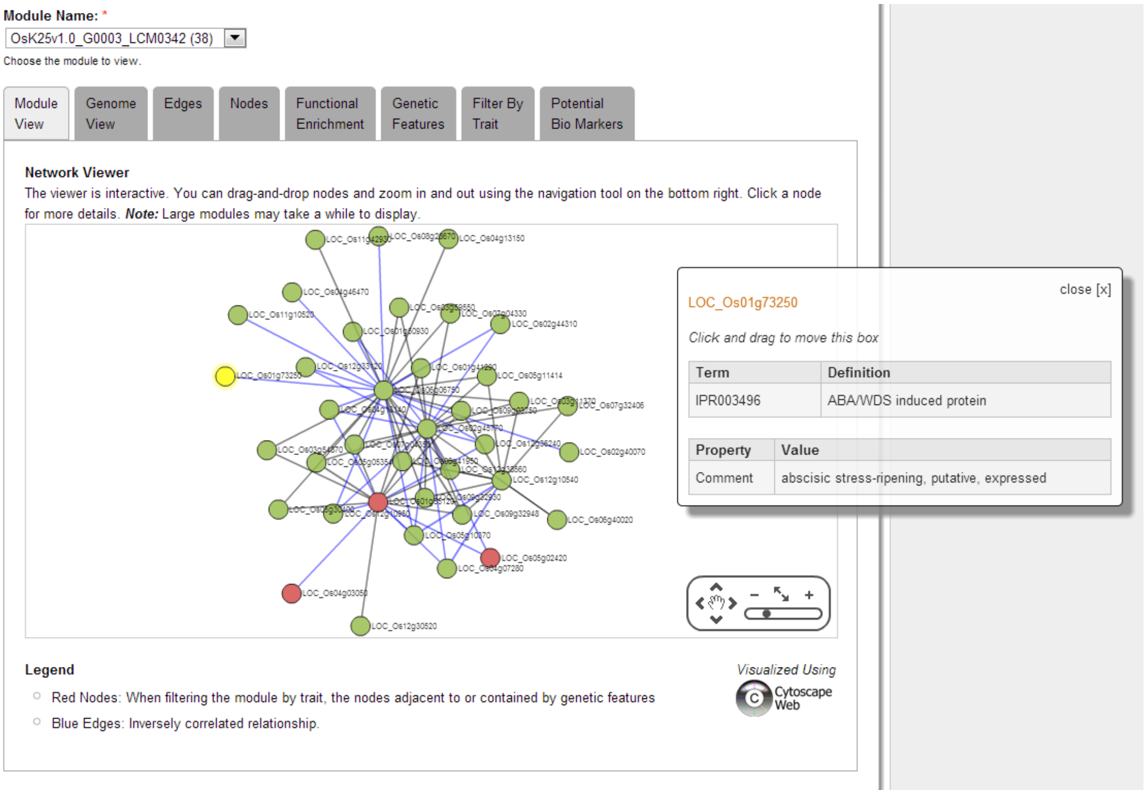
Select a genetic feature type to limit traits.

Traits:

Any trait
1000-seed weight
100-seed weight
Abaxial stomatal frequency
Above ground biomass yield trait
Acid phosphatase activity
Adaxial stomatal frequency
Albino plantlet differentiation frequency
Alkali digestion
Allelopathic effect

Select a trait to limit results.

Supplemental Figure S1. The GeneNet Engine v0.9 Search Form. The search form can be used to locate network modules by species, network name, module name, specific gene, functional annotation terms, traits, and simple topology.



Network Viewer

The viewer is interactive. You can drag-and-drop nodes and zoom in and out using the navigation tool on the bottom right. Click a node for more details. ***Note:*** Large modules may take a while to display.

LOC_Os01g73250

Click and drag to move this box

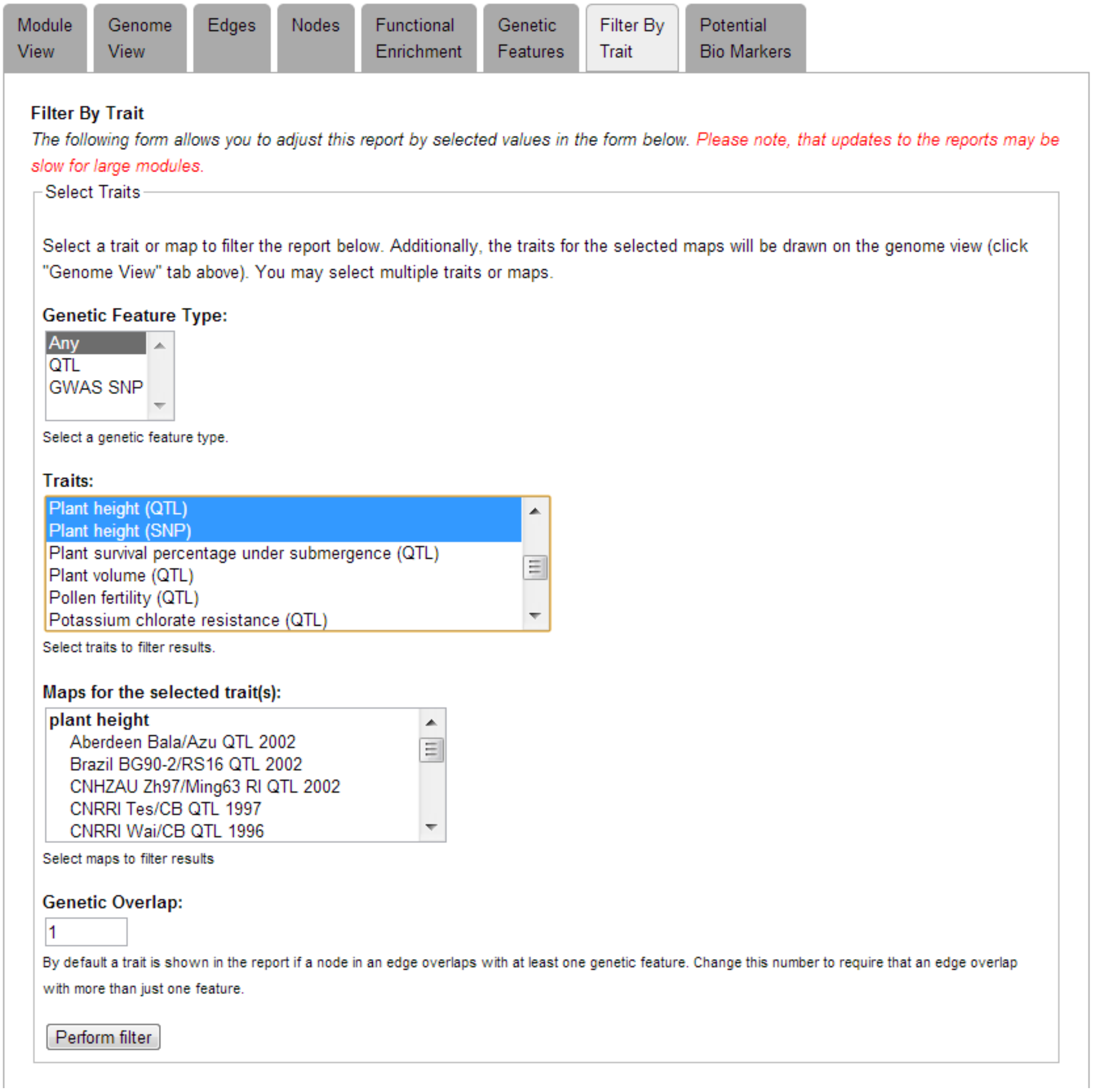
Term	Definition
IPR003496	ABA/WDS induced protein

Property	Value
Comment	abscisic stress-ripening, putative, expressed

Legend

- Red Nodes: When filtering the module by trait, the nodes adjacent to or contained by genetic features
- Blue Edges: Inversely correlated relationship.

Supplemental Figure S2. The GeneNet Engine Module Explorer. Contains an interactive module viewer, a genome viewer with circular plots of the module, edge and node lists, functional enrichment report and trait selection tool to filter reports and views by specific genetic traits.



Filter By Trait

The following form allows you to adjust this report by selected values in the form below. Please note, that updates to the reports may be slow for large modules.

Select Traits

Select a trait or map to filter the report below. Additionally, the traits for the selected maps will be drawn on the genome view (click "Genome View" tab above). You may select multiple traits or maps.

Genetic Feature Type:

Any
QTL
GWAS SNP

Select a genetic feature type.

Traits:

Plant height (QTL)
Plant height (SNP)
Plant survival percentage under submergence (QTL)
Plant volume (QTL)
Pollen fertility (QTL)
Potassium chlorate resistance (QTL)

Select traits to filter results.

Maps for the selected trait(s):

plant height
Aberdeen Bala/Azu QTL 2002
Brazil BG90-2/RS16 QTL 2002
CNHZAU Zh97/Ming63 RI QTL 2002
CNRRI Tes/CB QTL 1997
CNRRI Wai/CB QTL 1996

Select maps to filter results

Genetic Overlap:

1

By default a trait is shown in the report if a node in an edge overlaps with at least one genetic feature. Change this number to require that an edge overlap with more than just one feature.

Supplemental Figure S3. Filter by Trait Tab of the Module Explorer. Users can alter the module explorer to identify edges overlapping genetic features. Users can select features by trait name, genetic feature type, genetic maps (if applicable) and specify the amount of overlap required.

Table S1. *Microarray samples within each GIL*

NCBI Sample Accession	K-means Cluster (GIL)
GSM302922	1
GSM302923	1
GSM304390	1
GSM304394	1
GSM304395	1
GSM304397	1
GSM304478	1
GSM304485	1
GSM304497	1
GSM304646	1
GSM304653	1
GSM304654	1
GSM304664	1
GSM304669	1
GSM304671	1
GSM304677	1
GSM764300	1
GSM764301	1
GSM764302	1
GSM764303	1
GSM764304	1
GSM764305	1
GSM100443	2
GSM100444	2
GSM100445	2
GSM100446	2
GSM149411	2
GSM149412	2
GSM159177	2
GSM159178	2
GSM159179	2
GSM345235	2
GSM345236	2
GSM345237	2
GSM345238	2
GSM345239	2
GSM345240	2
GSM345241	2
GSM345242	2
GSM345243	2
GSM345244	2
GSM345245	2
GSM359902	2
GSM359903	2
GSM359904	2
GSM359905	2
GSM359906	2
GSM359907	2
GSM359908	2
GSM359909	2
GSM359910	2
GSM359911	2
GSM359912	2
GSM359913	2
GSM359914	2
GSM359915	2
GSM359916	2
GSM359917	2
GSM359918	2
GSM359919	2
GSM359920	2
GSM359921	2
GSM359922	2
GSM359923	2
GSM359924	2
GSM470620	2
GSM470621	2
GSM470624	2
GSM470625	2
GSM470634	2
GSM470635	2
GSM470718	2
GSM470719	2
GSM470722	2
GSM470723	2
GSM470732	2
GSM470733	2
GSM619236	2
GSM619237	2

GSM619238	2	GSM357615	3	GSM357647	3
GSM619239	2	GSM357616	3	GSM357648	3
GSM619240	2	GSM357617	3	GSM357649	3
GSM619241	2	GSM357618	3	GSM357650	3
GSM619242	2	GSM357620	3	GSM357651	3
GSM619244	2	GSM357621	3	GSM357653	3
GSM619245	2	GSM357622	3	GSM357654	3
GSM619246	2	GSM357623	3	GSM357655	3
GSM619247	2	GSM357624	3	GSM357656	3
GSM619248	2	GSM357625	3	GSM357657	3
GSM619249	2	GSM357626	3	GSM357658	3
GSM619250	2	GSM357627	3	GSM357659	3
GSM645328	2	GSM357628	3	GSM357660	3
GSM645329	2	GSM357629	3	GSM357661	3
GSM645330	2	GSM357630	3	GSM357662	3
GSM645331	2	GSM357631	3	GSM357682	3
GSM645332	2	GSM357632	3	GSM357683	3
GSM645333	2	GSM357633	3	GSM357684	3
GSM645340	2	GSM357636	3	GSM409771	3
GSM645341	2	GSM357637	3	GSM409772	3
GSM645342	2	GSM357638	3	GSM409773	3
GSM645343	2	GSM357639	3	GSM409774	3
GSM645344	2	GSM357640	3	GSM409781	3
GSM645345	2	GSM357641	3	GSM630939	3
GSM195218	3	GSM357642	3	GSM789503	3
GSM195219	3	GSM357643	3	GSM789504	3
GSM195220	3	GSM357644	3	GSM789505	3
GSM351447	3	GSM357645	3	GSM789506	3
GSM357614	3	GSM357646	3	GSM789507	3

GSM789508	3
GSM789509	3
GSM789510	3
GSM789511	3
GSM789512	3
GSM789513	3
GSM789514	3
GSM789515	3
GSM789516	3
GSM789517	3
GSM387556	4
GSM387557	4
GSM387558	4
GSM387559	4
GSM387560	4
GSM387561	4
GSM387562	4
GSM387563	4
GSM387564	4
GSM468795	4
GSM468796	4
GSM468799	4
GSM468800	4
GSM195221	5
GSM195222	5
GSM195223	5
GSM351427	5
GSM351428	5
GSM351429	5
GSM351430	5
GSM351431	5
GSM351432	5
GSM351433	5
GSM351434	5
GSM351435	5
GSM351436	5
GSM357619	5
GSM357652	5
GSM409421	5
GSM409422	5
GSM409427	5
GSM409428	5
GSM409429	5
GSM515490	5
GSM515491	5
GSM515492	5
GSM515493	5
GSM515494	5
GSM515495	5
GSM159189	6
GSM159190	6
GSM159191	6
GSM159192	6
GSM159193	6
GSM159194	6
GSM159195	6
GSM159196	6
GSM159197	6
GSM159198	6
GSM159199	6
GSM159200	6
GSM351437	6
GSM351438	6
GSM351439	6
GSM351440	6
GSM351441	6
GSM351442	6
GSM351443	6
GSM351444	6
GSM351445	6
GSM351446	6
GSM429984	6
GSM686458	6
GSM686459	6
GSM686460	6
GSM686461	6
GSM686462	6
GSM686463	6
GSM686464	6
GSM686465	6
GSM686466	6
GSM100440	7
GSM100442	7
GSM149409	7
GSM149410	7
GSM154829	7
GSM154831	7

GSM154832	7
GSM154833	7
GSM154937	7
GSM154938	7
GSM154939	7
GSM154940	7
GSM154941	7
GSM154942	7
GSM154943	7
GSM154944	7
GSM159172	7
GSM159173	7
GSM195226	7
GSM261998	7
GSM261999	7
GSM262000	7
GSM262001	7
GSM262002	7
GSM262003	7
GSM262004	7
GSM262005	7
GSM262006	7
GSM262007	7
GSM262008	7
GSM262009	7
GSM262010	7
GSM262011	7
GSM262012	7
GSM262013	7
GSM262014	7
GSM262015	7
GSM262016	7
GSM262017	7
GSM262018	7
GSM262019	7
GSM262020	7
GSM262021	7
GSM302920	7
GSM302921	7
GSM357679	7
GSM357680	7
GSM357681	7
GSM431925	7
GSM431926	7
GSM431927	7
GSM431928	7
GSM431929	7
GSM431930	7
GSM431931	7
GSM431932	7
GSM458175	7
GSM458176	7
GSM458177	7
GSM458178	7
GSM458179	7
GSM458180	7
GSM458181	7
GSM458182	7
GSM458183	7
GSM458184	7
GSM458185	7
GSM458186	7
GSM461481	7
GSM461482	7
GSM461528	7
GSM461534	7
GSM542564	7
GSM542565	7
GSM542566	7
GSM542567	7
GSM542568	7
GSM542569	7
GSM545318	7
GSM545319	7
GSM545320	7
GSM545321	7
GSM545322	7
GSM545323	7
GSM545324	7
GSM545325	7
GSM545326	7
GSM692533	7
GSM692534	7
GSM692535	7
GSM159180	8
GSM159181	8
GSM159182	8

GSM159183	8
GSM159184	8
GSM159185	8
GSM203238	8
GSM203239	8
GSM203240	8
GSM203241	8
GSM203242	8
GSM203243	8
GSM470636	8
GSM470637	8
GSM470638	8
GSM470639	8
GSM470640	8
GSM470641	8
GSM470643	8
GSM470648	8
GSM470649	8
GSM470666	8
GSM470667	8
GSM470734	8
GSM470738	8
GSM470739	8
GSM470746	8
GSM470747	8
GSM470765	8
GSM470766	8
GSM591764	8
GSM591765	8
GSM591766	8
GSM591767	8
GSM591768	8
GSM591769	8
GSM591770	8
GSM591771	8
GSM591772	8
GSM595945	8
GSM595946	8
GSM595947	8
GSM645322	8
GSM645323	8
GSM645324	8
GSM645325	8
GSM645326	8
GSM645327	8
GSM645334	8
GSM645336	8
GSM645337	8
GSM645338	8
GSM645339	8
GSM645346	8
GSM645347	8
GSM645348	8
GSM645349	8
GSM645350	8
GSM645351	8
GSM822094	8
GSM822095	8
GSM822096	8
GSM822097	8
GSM822098	8
GSM822099	8
GSM195227	9
GSM351448	9
GSM351449	9
GSM351450	9
GSM351451	9
GSM351452	9
GSM470656	9
GSM470657	9
GSM470754	9
GSM470755	9
GSM686467	9
GSM686468	9
GSM686469	9
GSM692539	9
GSM692540	9
GSM692541	9
GSM692542	9
GSM692543	9
GSM692544	9
GSM692545	9
GSM692546	9
GSM692547	9
GSM692548	9
GSM692549	9
GSM692550	9

GSM100441	10
GSM275405	10
GSM275406	10
GSM275407	10
GSM275408	10
GSM275409	10
GSM275410	10
GSM275411	10
GSM275412	10
GSM275413	10
GSM275414	10
GSM275415	10
GSM275416	10
GSM692536	10
GSM692537	10
GSM692538	10
GSM174883	11
GSM174884	11
GSM174885	11
GSM174887	11
GSM174888	11
GSM421667	11
GSM421668	11
GSM421669	11
GSM421670	11
GSM421671	11
GSM421672	11
GSM421673	11
GSM421674	11
GSM421675	11
GSM421676	11
GSM421677	11
GSM421678	11
GSM421679	11
GSM421680	11
GSM421681	11
GSM421682	11
GSM421683	11
GSM421684	11
GSM421685	11
GSM421686	11
GSM421687	11
GSM421688	11
GSM421689	11
GSM421690	11
GSM421691	11
GSM421692	11
GSM421693	11
GSM421694	11
GSM421695	11
GSM421696	11
GSM421697	11
GSM421699	11
GSM421700	11
GSM421701	11
GSM421702	11
GSM421703	11
GSM421704	11
GSM421705	11
GSM421706	11
GSM421707	11
GSM421708	11
GSM421709	11
GSM421710	11
GSM421711	11
GSM421712	11
GSM421713	11
GSM421714	11
GSM421715	11
GSM421716	11
GSM421717	11
GSM421718	11
GSM421719	11
GSM421720	11
GSM421721	11
GSM421722	11
GSM421723	11
GSM421724	11
GSM421725	11
GSM421726	11
GSM468793	11
GSM468794	11
GSM468797	11
GSM468798	11
GSM822064	11
GSM822065	11
GSM822066	11

GSM822067	11
GSM822068	11
GSM822069	11
GSM159201	12
GSM159202	12
GSM159203	12
GSM159204	12
GSM159205	12
GSM159206	12
GSM159207	12
GSM159208	12
GSM159209	12
GSM429985	12
GSM470642	12
GSM470646	12
GSM470647	12
GSM470650	12
GSM470651	12
GSM470652	12
GSM470653	12
GSM470654	12
GSM470655	12
GSM470658	12
GSM470659	12
GSM470736	12
GSM470740	12
GSM470741	12
GSM470744	12
GSM470745	12
GSM470748	12
GSM470749	12
GSM470750	12
GSM470751	12
GSM470752	12
GSM470753	12
GSM470756	12
GSM470757	12
GSM645352	12
GSM645353	12
GSM645354	12
GSM645355	12
GSM645356	12
GSM645357	12
GSM377070	13
GSM377071	13
GSM377072	13
GSM377073	13
GSM377074	13
GSM377075	13
GSM377076	13
GSM377077	13
GSM377078	13
GSM377079	13
GSM377080	13
GSM377081	13
GSM377082	13
GSM377083	13
GSM377084	13
GSM377085	13
GSM377086	13
GSM409780	13
GSM409782	13
GSM409783	13
GSM409784	13
GSM409785	13
GSM409786	13
GSM409787	13
GSM409788	13
GSM409789	13
GSM409790	13
GSM422672	13
GSM422674	13
GSM422676	13
GSM506394	13
GSM506395	13
GSM506396	13
GSM506397	13
GSM506398	13
GSM506399	13
GSM506400	13
GSM154945	14
GSM154946	14
GSM154947	14
GSM154948	14
GSM154949	14
GSM154950	14
GSM154951	14

GSM154952	14	GSM365260	14	GSM495739	14
GSM154953	14	GSM365262	14	GSM495740	14
GSM154954	14	GSM365263	14	GSM495741	14
GSM154955	14	GSM365266	14	GSM495742	14
GSM154956	14	GSM365267	14	GSM495743	14
GSM174886	14	GSM365268	14	GSM495744	14
GSM174889	14	GSM366813	14	GSM563421	14
GSM174890	14	GSM366818	14	GSM563422	14
GSM207558	14	GSM366819	14	GSM563423	14
GSM207559	14	GSM368873	14	GSM570988	14
GSM207560	14	GSM378729	14	GSM570989	14
GSM207562	14	GSM378730	14	GSM570990	14
GSM207563	14	GSM378731	14	GSM570991	14
GSM207564	14	GSM421698	14	GSM570992	14
GSM207565	14	GSM476769	14	GSM570993	14
GSM207566	14	GSM476770	14	GSM570994	14
GSM207567	14	GSM476771	14	GSM570995	14
GSM267998	14	GSM476772	14	GSM570996	14
GSM267999	14	GSM476773	14	GSM570997	14
GSM357122	14	GSM476774	14	GSM570998	14
GSM357133	14	GSM476775	14	GSM570999	14
GSM357134	14	GSM476776	14	GSM591761	14
GSM357135	14	GSM476777	14	GSM591762	14
GSM357136	14	GSM476778	14	GSM591763	14
GSM357137	14	GSM476779	14	GSM647655	14
GSM357685	14	GSM476780	14	GSM647656	14
GSM357686	14	GSM495736	14	GSM647657	14
GSM357687	14	GSM495737	14	GSM696667	14
GSM357688	14	GSM495738	14	GSM696668	14

GSM696669	14
GSM696670	14
GSM696671	14
GSM696672	14
GSM696673	14
GSM696674	14
GSM696675	14
GSM696676	14
GSM696677	14
GSM696678	14
GSM696679	14
GSM696680	14
GSM696681	14
GSM696682	14
GSM696683	14
GSM696684	14
GSM696685	14
GSM696686	14
GSM698482	14
GSM698483	14
GSM698484	14
GSM698485	14
GSM698486	14
GSM698487	14
GSM409775	15
GSM409776	15
GSM409777	15
GSM409778	15
GSM409779	15
GSM470622	15
GSM470623	15
GSM470721	15
GSM595933	15
GSM595934	15
GSM595935	15
GSM595936	15
GSM595937	15
GSM595938	15
GSM595939	15
GSM595940	15
GSM595941	15
GSM595942	15
GSM595943	15
GSM595944	15
GSM630938	15
GSM195225	16
GSM696647	16
GSM696648	16
GSM696649	16
GSM696650	16
GSM696651	16
GSM696652	16
GSM696653	16
GSM696654	16
GSM696655	16
GSM696656	16
GSM696657	16
GSM696658	16
GSM696659	16
GSM696660	16
GSM696661	16
GSM696662	16
GSM696663	16
GSM696664	16
GSM696665	16
GSM696666	16
GSM173080	17
GSM173086	17
GSM173089	17
GSM173091	17
GSM173093	17
GSM173094	17
GSM278844	17
GSM278845	17
GSM278846	17
GSM278847	17
GSM278848	17
GSM278849	17
GSM278850	17
GSM278851	17
GSM278852	17
GSM278853	17
GSM278854	17
GSM278855	17
GSM403000	17
GSM470735	17
GSM470737	17

GSM692530	17
GSM692531	17
GSM692532	17
GSM159210	18
GSM159211	18
GSM159212	18
GSM159213	18
GSM159214	18
GSM159215	18
GSM159216	18
GSM159217	18
GSM159218	18
GSM159219	18
GSM159220	18
GSM159221	18
GSM195229	18
GSM195230	18
GSM240994	18
GSM240995	18
GSM240996	18
GSM240997	18
GSM240998	18
GSM240999	18
GSM302916	18
GSM302917	18
GSM470660	18
GSM470661	18
GSM470662	18
GSM470663	18
GSM470664	18
GSM470665	18
GSM470758	18
GSM470759	18
GSM470760	18
GSM470761	18
GSM470762	18
GSM470763	18
GSM630940	18
GSM630941	18
GSM254091	19
GSM254092	19
GSM254093	19
GSM254095	19
GSM254096	19
GSM254097	19
GSM302918	19
GSM302919	19
GSM366811	19
GSM366812	19
GSM366814	19
GSM366815	19
GSM366816	19
GSM366817	19
GSM366820	19
GSM366821	19
GSM366822	19
GSM615979	19
GSM692770	19
GSM116195	20
GSM116398	20
GSM116399	20
GSM116400	20
GSM116401	20
GSM116402	20
GSM159259	20
GSM159260	20
GSM159261	20
GSM159262	20
GSM159263	20
GSM159264	20
GSM159265	20
GSM159266	20
GSM159267	20
GSM159268	20
GSM159269	20
GSM159270	20
GSM402997	20
GSM402998	20
GSM402999	20
GSM403001	20
GSM403002	20
GSM403003	20
GSM403004	20
GSM403005	20
GSM403006	20
GSM403007	20
GSM403008	20

GSM403009	20
GSM403010	20
GSM403011	20
GSM403012	20
GSM470626	20
GSM470627	20
GSM470628	20
GSM470629	20
GSM470630	20
GSM470631	20
GSM470632	20
GSM470633	20
GSM470668	20
GSM470669	20
GSM470670	20
GSM470671	20
GSM470672	20
GSM470673	20
GSM470674	20
GSM470675	20
GSM470676	20
GSM470677	20
GSM470678	20
GSM470679	20
GSM470724	20
GSM470725	20
GSM470726	20
GSM470727	20
GSM470728	20
GSM470729	20
GSM470730	20
GSM470731	20
GSM470767	20
GSM470768	20
GSM470769	20
GSM470770	20
GSM470771	20
GSM470772	20
GSM470773	20
GSM470774	20
GSM470775	20
GSM470776	20
GSM470777	20
GSM470778	20
GSM470618	21
GSM470619	21
GSM470716	21
GSM470717	21
GSM470720	21
GSM559869	21
GSM559870	21
GSM559871	21
GSM559872	21
GSM559873	21
GSM559874	21
GSM559875	21
GSM559876	21
GSM559877	21
GSM559878	21
GSM559879	21
GSM559880	21
GSM559881	21
GSM559882	21
GSM559883	21
GSM559884	21
GSM559885	21
GSM559886	21
GSM559887	21
GSM559888	21
GSM559889	21
GSM559890	21
GSM559891	21
GSM559892	21
GSM559893	21
GSM559894	21
GSM559895	21
GSM559896	21
GSM559897	21
GSM559898	21
GSM559899	21
GSM559900	21
GSM559901	21
GSM559902	21
GSM559903	21
GSM559904	21
GSM559905	21
GSM559906	21

GSM559907	21	GSM559936	21	GSM559965	21
GSM559908	21	GSM559937	21	GSM559966	21
GSM559909	21	GSM559938	21	GSM559967	21
GSM559910	21	GSM559939	21	GSM559968	21
GSM559911	21	GSM559940	21	GSM559969	21
GSM559912	21	GSM559941	21	GSM559970	21
GSM559913	21	GSM559942	21	GSM559971	21
GSM559914	21	GSM559943	21	GSM559972	21
GSM559915	21	GSM559944	21	GSM559973	21
GSM559916	21	GSM559945	21	GSM559974	21
GSM559917	21	GSM559946	21	GSM559975	21
GSM559918	21	GSM559947	21	GSM559976	21
GSM559919	21	GSM559948	21	GSM559977	21
GSM559920	21	GSM559949	21	GSM559978	21
GSM559921	21	GSM559950	21	GSM559979	21
GSM559922	21	GSM559951	21	GSM559980	21
GSM559923	21	GSM559952	21	GSM559981	21
GSM559924	21	GSM559953	21	GSM559982	21
GSM559925	21	GSM559954	21	GSM559983	21
GSM559926	21	GSM559955	21	GSM559984	21
GSM559927	21	GSM559956	21	GSM559985	21
GSM559928	21	GSM559957	21	GSM559986	21
GSM559929	21	GSM559958	21	GSM559987	21
GSM559930	21	GSM559959	21	GSM559988	21
GSM559931	21	GSM559960	21	GSM559989	21
GSM559932	21	GSM559961	21	GSM559990	21
GSM559933	21	GSM559962	21	GSM559991	21
GSM559934	21	GSM559963	21	GSM559992	21
GSM559935	21	GSM559964	21	GSM559993	21

GSM559994	21
GSM559995	21
GSM559996	21
GSM559997	21
GSM559998	21
GSM559999	21
GSM560000	21
GSM560001	21
GSM560002	21
GSM560003	21
GSM560004	21
GSM560005	21
GSM560006	21
GSM560007	21
GSM560008	21
GSM560009	21
GSM560010	21
GSM560011	21
GSM560012	21
GSM560013	21
GSM560014	21
GSM560015	21
GSM560016	21
GSM560017	21
GSM560018	21
GSM560019	21
GSM560020	21
GSM560021	21
GSM560022	21
GSM560023	21
GSM560024	21
GSM560025	21
GSM560026	21
GSM560027	21
GSM560028	21
GSM560029	21
GSM560030	21
GSM560031	21
GSM560032	21
GSM560033	21
GSM560034	21
GSM560035	21
GSM560036	21
GSM560037	21
GSM560038	21
GSM560039	21
GSM560040	21
GSM560041	21
GSM560042	21
GSM560043	21
GSM560044	21
GSM560045	21
GSM560046	21
GSM560047	21
GSM560048	21
GSM560049	21
GSM560050	21
GSM560051	21
GSM560052	21
GSM560053	21
GSM560054	21
GSM560055	21
GSM560056	21
GSM560057	21
GSM560058	21
GSM560059	21
GSM560060	21
GSM560061	21
GSM560062	21
GSM560063	21
GSM560064	21
GSM560065	21
GSM560066	21
GSM560067	21
GSM560068	21
GSM560069	21
GSM560070	21
GSM560071	21
GSM560072	21
GSM560073	21
GSM560074	21
GSM560075	21
GSM560076	21
GSM560077	21
GSM560078	21
GSM560079	21
GSM560080	21

GSM560081	21
GSM560082	21
GSM560083	21
GSM560084	21
GSM560085	21
GSM560086	21
GSM560087	21
GSM560088	21
GSM560089	21
GSM560090	21
GSM560091	21
GSM560092	21
GSM560093	21
GSM560094	21
GSM159186	22
GSM159187	22
GSM159188	22
GSM399470	22
GSM399471	22
GSM399472	22
GSM399473	22
GSM409430	22
GSM409431	22
GSM409432	22
GSM409433	22
GSM409434	22
GSM429482	22
GSM429483	22
GSM429484	22
GSM429485	22
GSM429982	22
GSM429983	22
GSM470644	22
GSM470645	22
GSM470680	22
GSM470681	22
GSM470682	22
GSM470683	22
GSM470684	22
GSM470685	22
GSM470686	22
GSM470687	22
GSM470688	22
GSM470689	22
GSM470690	22
GSM470691	22
GSM470692	22
GSM470693	22
GSM470694	22
GSM470695	22
GSM470696	22
GSM470697	22
GSM470742	22
GSM470743	22
GSM470779	22
GSM470780	22
GSM470781	22
GSM470782	22
GSM470783	22
GSM470784	22
GSM470785	22
GSM470786	22
GSM470787	22
GSM470788	22
GSM470789	22
GSM470790	22
GSM470791	22
GSM470792	22
GSM470793	22
GSM470794	22
GSM470795	22
GSM470796	22
GSM159271	23
GSM159272	23
GSM195224	23
GSM195228	23
GSM302914	23
GSM302915	23
GSM357664	23
GSM357665	23
GSM357666	23
GSM357667	23
GSM357668	23
GSM357669	23
GSM357670	23
GSM357671	23
GSM357672	23

GSM357673	23	GSM470709	23	GSM431940	24
GSM357674	23	GSM470710	23	GSM615977	24
GSM357675	23	GSM470711	23	GSM615978	24
GSM357676	23	GSM470712	23	GSM615980	24
GSM357677	23	GSM470713	23	GSM615981	24
GSM357678	23	GSM470714	23	GSM615982	24
GSM470606	23	GSM470715	23	GSM692771	24
GSM470607	23	GSM692527	23	GSM692772	24
GSM470608	23	GSM692528	23	GSM740867	24
GSM470609	23	GSM692529	23	GSM740868	24
GSM470610	23	GSM154957	24	GSM740869	24
GSM470611	23	GSM154958	24	GSM740870	24
GSM470612	23	GSM154959	24	GSM740871	24
GSM470613	23	GSM154960	24	GSM740872	24
GSM470614	23	GSM255762	24	GSM740873	24
GSM470615	23	GSM255763	24	GSM740874	24
GSM470616	23	GSM281583	24	GSM740875	24
GSM470617	23	GSM281584	24	GSM740876	24
GSM470698	23	GSM281585	24	GSM740877	24
GSM470699	23	GSM281586	24	GSM740878	24
GSM470700	23	GSM281587	24	GSM645335	25
GSM470701	23	GSM281588	24	GSM698669	25
GSM470702	23	GSM431933	24	GSM698670	25
GSM470703	23	GSM431934	24	GSM822070	25
GSM470704	23	GSM431935	24	GSM822071	25
GSM470705	23	GSM431936	24	GSM822072	25
GSM470706	23	GSM431937	24	GSM822073	25
GSM470707	23	GSM431938	24	GSM822074	25
GSM470708	23	GSM431939	24	GSM822075	25

GSM822076	25
GSM822077	25
GSM822078	25
GSM822079	25
GSM822080	25
GSM822081	25
GSM822082	25
GSM822083	25
GSM822084	25
GSM822085	25
GSM822086	25
GSM822087	25
GSM822088	25
GSM822089	25
GSM822090	25
GSM822091	25
GSM822092	25
GSM822093	25
GSM822100	25
GSM822101	25
GSM822102	25
GSM822103	25
GSM822104	25
GSM822105	25
GSM822106	25
GSM822107	25
GSM822108	25
GSM822109	25
GSM822110	25
GSM822111	25
GSM822112	25
GSM822113	25
GSM822114	25
GSM822115	25
GSM822116	25
GSM822117	25
GSM822118	25
GSM822119	25
GSM822120	25
GSM822121	25
GSM822122	25
GSM822123	25
GSM822124	25
GSM822125	25
GSM822126	25
GSM822127	25
GSM822128	25
GSM822129	25

Appendix B Permissions for Reprinting

The following text demonstrates permissions to include the two Plant Physiology articles that comprise chapters 2 and 3 of this dissertation. The order details shown below were cut-and-pasted from http://www.copyright.com:

Confirmation Number: 11025116
Order Date: 08/29/2012
Customer Information
Customer: Stephen Ficklin
Invoiced: CCOM361638

Plant Physiology

- **Order detail ID:** 62869749
- **ISSN:** 0032-0889
- **Publication year:** 2010
- **Publication Type:** Journal
- **Publisher:** American Society of Plant Physiologists
- **Rightsholder:** AMERICAN SOCIETY OF PLANT BIOLOGISTS
- **Author/Editor:** Stephen P. Ficklin
- **Permission Status:** ✓ **Granted**
- **Permission type:** Republish or display content
- **Type of use:** Republish in a dissertation
- **Requested use:** Dissertation
- **Republishing organization:** Clemson University
- **Organization status:** Non-profit 501(c)(3)
- **Republication date:** 12/15/2012
- **Circulation/ Distribution:** 10
- **Type of content:** Full article/chapter
- **Description of requested content:** The Association of Multiple Interacting Genes with Specific Phenotypes in Rice Using Gene Coexpression Networks
- **Page range(s):** 13-24
- **Translating to:** No Translation
- **Requested content's publication date:** 09/01/2010

Plant Physiology

- **Order detail ID:** 62869750
- **ISSN:** 0032-0889
- **Publication year:** 2011
- **Publication Type:** Journal
- **Publisher:** American Society of Plant Physiologists
- **Rightsholder:** AMERICAN SOCIETY OF PLANT BIOLOGISTS
- **Author/Editor:** Stephen P. Ficklin
- **Permission Status:** ✓ **Granted**
- **Permission type:** Republish or display content
- **Type of use:** Republish in a dissertation
- **Requested use:** Dissertation
- **Republishing organization:** Clemson University
- **Organization status:** Non-profit 501(c)(3)
- **Republication date:** 12/15/2012
- **Circulation/ Distribution:** 10
- **Type of content:** Full article/chapter
- **Description of requested content:** Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice
- **Page range(s):** 1244-1256
- **Translating to:** No Translation
- **Requested content's publication date:** 07/01/2011

Nature Reviews Genetics

- **Order detail ID:** 63544135
- **ISSN:** 1471-0056
- **Publication Type:** Journal
- **Volume:**
- **Issue:**
- **Start page:**
- **Publisher:** NATURE PUBLISHING GROUP
- **Permission Status:** ✓ **Granted**
- **Permission type:** Republish or display content
- **Type of use:** Republish in a thesis/dissertation

Order License Id: 3105470132929

Requestor type	Academic institution
Format	Print, Electronic
Portion	chart/graph/table/figure
Number of charts/graphs/tables/figures	1
Title or numeric reference of the portion(s)	Network biology: understanding the cell's functional organization. Figure in Box 2.

Editor of portion(s)	N/A
Author of portion(s)	Albert-László Barabási & Zoltán N. Oltvai
Volume of serial or monograph	5
Page range of portion	105
Publication date of portion	Feb 2004
Rights for	Main product
Duration of use	Life of current edition
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	0 to 499
Made available in the following markets	Education (PhD dissertation)
The requesting person/organization	Stephen Ficklin
Order reference number	
Author/Editor	Stephen Ficklin
The standard identifier	Ficklin-2013
Title	Predicting Complex Phenotype-Genotype Relationships in Grasses: A Systems Genetics Approach
Publisher	Clemson University
Expected publication date	May 2013
Estimated size (pages)	200

References

1. FAOSTAT. *Food and Agricultural Organization of the United Nations, Commodities Production Statistics. http://faostat.fao.org/.* 2011;

2. *Food: The growing problem.* Nature, 2010. **466**(7306): p. 546-7.

3. Khush, G.S., *Green revolution: preparing for the 21st century.* Genome, 1999. **42**(4): p. 646-55.

4. Hijmans, R., *Cartograms: distortion for a better view.* Rice Today, 2008. **7**(1): p. 12.

5. Institute, I.R.R., *http://www.irri.org/gis/cartograms/cartograms.htm* Accessed Aug 2010.

6. Ouyang, S., et al., *The TIGR Rice Genome Annotation Resource: improvements and new features.* Nucleic Acids Res, 2007. **35**(Database issue): p. D883-7.

7. Jaiswal, P., *Gramene database: a hub for comparative plant genomics.* Methods Mol Biol, 2011. **678**: p. 247-75.

8. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--10 years on.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1005-10.

9. Parkinson, H., et al., *ArrayExpress update-an archive of microarray and high-throughput sequencing-based functional genomics experiments.* Nucleic Acids Research, 2011. **39**: p. D1002-D1004.

10. Devos, K.M. and M.D. Gale, *Genome relationships: the grass model in current research.* Plant Cell, 2000. **12**(5): p. 637-46.

11. Grivet, L. and P. Arruda, *Sugarcane genomics: depicting the complex genome of an important tropical crop.* Curr Opin Plant Biol, 2002. **5**(2): p. 122-7.

12. Lander, E.S. and D. Botstein, *Mapping Mendelian Factors Underlying Quantitative Traits Using Rflp Linkage Maps.* Genetics, 1989. **121**(1): p. 185-199.

13. Miles, C.M. and M. Wayne, *Quantitative Trait Locus (QTL) Analysis.* Nature Education, 2008. **1**(1).

14. Crow, J.F., *Haldane, Bailey, Taylor and recombinant-inbred lines.* Genetics, 2007. **176**(2): p. 729-32.

15. Chesler, E.J., et al., *The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics.* Mamm Genome, 2008. **19**(6): p. 382-9.

16. McMullen, M.D., et al., *Genetic properties of the maize nested association mapping population.* Science, 2009. **325**(5941): p. 737-40.

17. Ni, J., et al., *Gramene QTL database: development, content and applications.* Database (Oxford), 2009. **2009**: p. bap005.

18. Mackay, T.F., E.A. Stone, and J.F. Ayroles, *The genetics of quantitative traits: challenges and prospects.* Nat Rev Genet, 2009. **10**(8): p. 565-77.

19. Flint-Garcia, S.A., et al., *Maize association population: a high-resolution platform for quantitative trait locus dissection.* Plant J, 2005. **44**(6): p. 1054-64.

20. Fan, C., et al., *GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein.* Theor Appl Genet, 2006. **112**(6): p. 1164-71.

21. Wissuwa, M., et al., *Substitution mapping of Pup1: a major QTL increasing phosphorus uptake of rice from a phosphorus-deficient soil.* Theor Appl Genet, 2002. **105**(6-7): p. 890-897.

22. Silvio, S. and R. Tuberosa, *To clone or not to clone plant QTLs: present and future challenges.* Trends in Plant Science, 2005. **10**(16): p. 297-304.

23. Long, Y., C. Zhang, and J. Meng, *Challenges for QTL Analysis in Crops.* Journal of Crop Science and Biotechnology, 2008. **11**(1): p. 7-12.

24. Zhao, K., et al., *Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa.* Nat Commun, 2011. **2**: p. 467.

25. Huang, X., et al., *Genome-wide association studies of 14 agronomic traits in rice landraces.* Nat Genet, 2010. **42**(11): p. 961-7.

26. Huang, X., et al., *Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm.* Nat Genet, 2012. **44**(1): p. 32-9.

27. Schadt, E.E., B. Zhang, and J. Zhu, *Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments.* Genetica, 2009. **136**(2): p. 259-269.

28. Farber, C.R., *Systems genetics: a novel approach to dissect the genetic basis of osteoporosis.* Current osteoporosis reports, 2012. **10**(3): p. 228-35.

29. Farber, C.R. and A.J. Lusis, *Future of osteoporosis genetics: enhancing genome-wide association studies.* J Bone Miner Res, 2009. **24**(12): p. 1937-42.

30. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.

31. Maher, B., *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.

32. Penrod, N.M., R. Cowper-Sal-lari, and J.H. Moore, *Systems genetics for drug target discovery.* Trends Pharmacol Sci, 2011. **32**(10): p. 623-30.

33. Jannink, J.L., A.J. Lorenz, and H. Iwata, *Genomic selection in plant breeding: from theory to practice.* Brief Funct Genomics, 2010. **9**(2): p. 166-77.

34. Akhtar, S., et al., *Marker Assisted Selection in Rice.* Journal of Phytology, 2010. **2**(10): p. 66.

35. Collard, B.C. and D.J. Mackill, *Marker-assisted selection: an approach for precision plant breeding in the twenty-first century.* Philos Trans R Soc Lond B Biol Sci, 2008. **363**(1491): p. 557-72.

36. Hospital, F., *Challenges for effective marker-assisted selection in plants.* Genetica, 2009. **136**(2): p. 303-10.

37. Goddard, M.E. and B.J. Hayes, *Genomic selection.* J Anim Breed Genet, 2007. **124**(6): p. 323-30.

38. Guo, Z., et al., *Evaluation of genome-wide selection efficiency in maize nested association mapping populations.* Theor Appl Genet, 2012. **124**(2): p. 261-75.

39. Blanchard, B.S., W.J. Fabrycky, and W.J. Fabrycky, *Systems engineering and analysis*. Vol. 4. 1990: Prentice Hall Englewood Cliffs, New Jersey.

40. Odum, H.T., *Systems Ecology; an introduction.* 1983.

41. Langlois, R.N., *Systems theory, knowledge, and the social sciences.* The Study of Information, Wiley, New York, 1983.

42. Dudley, A.M., et al., *A global view of pleiotropy and phenotypically derived gene function in yeast.* Mol Syst Biol, 2005. **1**: p. 2005 0001.

43. Gillis, J. and P. Pavlidis, *The Impact of Multifunctional Genes on "Guilt by Association" Analysis.* PLoS One, 2011. **6**(2).

44. Stearns, F.W., *One hundred years of pleiotropy: a retrospective.* Genetics, 2010. **186**(3): p. 767-73.

45. Nadeau, J.H. and A.M. Dudley, *Genetics. Systems genetics.* Science (New York, N.Y.), 2011. **331**(6020): p. 1015-6.

46. Swami, M., *SYSTEMS GENETICS Networking complex traits.* Nature Reviews Genetics, 2009. **10**(4): p. 219-219.

47. Jumbo-Lucioni, P., et al., *Systems genetics analysis of body weight and energy metabolism traits in Drosophila melanogaster.* BMC Genomics, 2010. **11**: p. 297.

48. Taneera, J., et al., *A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets.* Cell metabolism, 2012. **16**(1): p. 122-34.

49. Park, C.C., et al., *Gene networks associated with conditional fear in mice identified using a systems genetics approach.* BMC Syst Biol, 2011. **5**: p. 43.

50. Quigley, D. and A. Balmain, *Systems genetics analysis of cancer susceptibility: from mouse models to humans.* Nat Rev Genet, 2009. **10**(9): p. 651-7.

51. Leduc, M.S., et al., *Using bioinformatics and systems genetics to dissect HDL-cholesterol genetics in an MRL/MpJ x SM/J intercross.* Journal of lipid research, 2012. **53**(6): p. 1163-75.

52. Morahan, G., et al., *Systems genetics can provide new insights in to immune regulation and autoimmunity.* J Autoimmun, 2008. **31**(3): p. 233-6.

53. Cheng, Y., et al., *Systems Genetics Implicates Cytoskeletal Genes in Oocyte Control of Cloned Embryo Quality.* Genetics, 2013.

54. Ficklin, S.P., F. Luo, and F.A. Feltus, *The Association of Multiple Interacting Genes with Specific Phenotypes In Rice (Oryza sativa) Using Gene Co-Expression Networks.* Plant Physiol, 2010.

55. Mutwil, M., et al., *Assembly of an Interactive Correlation Network for the Arabidopsis Genome Using a Novel Heuristic Clustering Algorithm.* Plant Physiology, 2010. **152**(1): p. 29-43.

56. Lee, I., et al., *Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana.* Nature Biotechnology, 2010. **28**(2): p. 149-U14.

57. Kadarmideen, H.N., P. von Rohr, and L.L. Janss, *From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding.* Mamm Genome, 2006. **17**(6): p. 548-64.

58. Li, H., *Systems genetics in "-omics" era: current and future development.* Theory Biosci, 2013. **132**(1): p. 1-16.

59. De Las Rivas, J. and C. Fontanillo, *Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.* PLoS Comput Biol, 2010. **6**(6): p. e1000807.

60. Gibson, S., et al., *Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory*. PLoS One, 2013. **8**(2): p. e55871.

61. Hatzimanikatis, V., et al., *Metabolic networks: enzyme function and metabolite structure.* Curr Opin Struct Biol, 2004. **14**(3): p. 300-6.

62. Babu, M.M., et al., *Structure and evolution of transcriptional regulatory networks.* Curr Opin Struct Biol, 2004. **14**(3): p. 283-91.

63. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.

64. Barabasi, A.-L., et al. *Scale-Free and Hierarchical Structures in Complex Networks*. 2003. Granada (Spain): AIP.

65. Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks.* Science, 2002. **297**(5586): p. 1551-5.

66. Ficklin, S.P. and F.A. Feltus, *Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice*. Plant Physiology, 2011. **156**(3): p. 1244-56.

67. Jellen, L.C., J.L. Beard, and B.C. Jones, *Systems genetics analysis of iron regulation in the brain.* Biochimie, 2009. **91**(10): p. 1255-9.

68. Dobrin, R., et al., *Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease.* Genome Biol, 2009. **10**(5): p. R55.

69. Steuer, R., et al., *The mutual information: Detecting and evaluating dependencies between variables.* Bioinformatics, 2002. **18**: p. S231-S240.

70. Tsaparas, P., et al., *Global similarity and local divergence in human and mouse gene co-expression networks.* BMC Evol Biol, 2006. **6**: p. 70.

71. Jordan, I.K., et al., *Conservation and coevolution in the scale-free human gene coexpression network.* Mol Biol Evol, 2004. **21**(11): p. 2058-70.

72. Reverter, A., et al., *Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer.* Bioinformatics, 2006. **22**(19): p. 2396-404.

73. Aoki, K., Y. Ogata, and D. Shibata, *Approaches for extracting practical information from gene co-expression networks in plant biology.* Plant Cell Physiol, 2007. **48**(3): p. 381-90.

74. Carter, S.L., et al., *Gene co-expression network topology provides a framework for molecular characterization of cellular state.* Bioinformatics, 2004. **20**(14): p. 2242-50.

75. Persson, S., et al., *Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets.* Proc Natl Acad Sci U S A, 2005. **102**(24): p. 8633-8.

76. Stuart, J., et al., *A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.* Science, 2003. **302**(5643): p. 249-255.

77. Wolfe, C.J., I.S. Kohane, and A.J. Butte, *Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.* BMC Bioinformatics, 2005. **6**: p. 227.

78. Nayak, R.R., et al., *Coexpression network based on natural variation in human gene expression reveals gene interactions and functions.* Genome Res, 2009. **19**(11): p. 1953-62.

79. Perkins, A.D. and M.A. Langston, *Threshold selection in gene co-expression networks using spectral graph theory techniques.* BMC Bioinformatics, 2009. **10 Suppl 11**: p. S4.

80. Luo, F., et al., *Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory.* BMC Bioinformatics, 2007. **8**: p. 299.

81. Jalan, S., et al., *Random matrix analysis of localization properties of gene coexpression network.* Phys Rev E Stat Nonlin Soft Matter Phys, 2010. **81**(4 Pt 2): p. 046118.

82. Reverter, A. and E.K. Chan, *Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks.* Bioinformatics, 2008. **24**(21): p. 2491-7.

83. Elo, L.L., et al., *Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process.* Bioinformatics, 2007. **23**(16): p. 2096-103.

84. Puelma, T., R.A. Gutierrez, and A. Soto, *Discriminative local subspaces in gene expression data for effective gene function prediction.* Bioinformatics, 2012. **28**(17): p. 2256-64.

85. Bassel, G.W., et al., *Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets.* Plant Cell, 2011. **23**(9): p. 3101-3116.

86. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.

87. Ahn, Y.Y., J.P. Bagrow, and S. Lehmann, *Link communities reveal multiscale complexity in networks.* Nature, 2010. **466**(7307): p. 761-4.

88. Hwang, W., et al., *A novel functional module detection algorithm for protein-protein interaction networks.* Algorithms Mol Biol, 2006. **1**: p. 24.

89. Li, A. and S. Horvath, *Network module detection: Affinity search technique with the multi-node topological overlap measure.* BMC Res Notes, 2009. **2**: p. 142.

90. Rivera, C.G., R. Vakil, and J.S. Bader, *NeMo: Network Module identification in Cytoscape.* BMC Bioinformatics, 2010. **11 Suppl 1**: p. S61.

91. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* BMC Bioinformatics, 2003. **4**: p. 2.

92. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.

93. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

94. Zhou, X. and Z. Su, *EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species.* BMC Genomics, 2007. **8**: p. 246.

95. Beissbarth, T. and T.P. Speed, *GOstat: find statistically overrepresented Gene Ontologies within a group of genes.* Bioinformatics, 2004. **20**(9): p. 1464-1465.

96. Al-Shahrour, F., et al., *FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.* Nucleic Acids Research, 2007. **35**: p. W91-W96.

97. Conesa, A. and S. Gotz, *Blast2GO: A comprehensive suite for functional analysis in plant genomics.* Int J Plant Genomics, 2008. **2008**: p. 619832.

98. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature Genetics, 2000. **25**(1): p. 25-9.

99. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2008. **36**(Database issue): p. D480-4.

100. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.

101. Avraham, S., et al., *The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations.* Nucleic Acids Res, 2008. **36**(Database issue): p. D449-54.

102. Mueller, L.A., P. Zhang, and S.Y. Rhee, *AraCyc: a biochemical pathway database for Arabidopsis.* Plant Physiol, 2003. **132**(2): p. 453-60.

103. Ayroles, J.F., et al., *Systems genetics of complex traits in Drosophila melanogaster.* Nat Genet, 2009. **41**(3): p. 299-307.

104. Miyao, A., et al., *A large-scale collection of phenotypic data describing an insertional mutant population to facilitate functional analysis of rice genes.* Plant Mol Biol, 2007. **63**(5): p. 625-35.

105. Miyao, A., et al., *Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome.* Plant Cell, 2003. **15**(8): p. 1771-80.

106. Jeon, J.S., et al., *T-DNA insertional mutagenesis for functional genomics in rice.* Plant J, 2000. **22**(6): p. 561-70.

107. Kuromori, T., et al., *A collection of 11 800 single-copy Ds transposon insertion lines in Arabidopsis.* Plant J, 2004. **37**(6): p. 897-905.

108. Kumar, R., et al., *Single feature polymorphism discovery in rice.* PLoS One, 2007. **2**(3): p. e284.

109. Li, H. and H. Deng, *Systems genetics, bioinformatics and eQTL mapping.* Genetica, 2010. **138**(9-10): p. 915-24.

110. Kang, H.P., et al., *Integration of disease-specific single nucleotide polymorphisms, expression quantitative trait loci and coexpression networks reveal novel candidate genes for type 2 diabetes.* Diabetologia, 2012. **55**(8): p. 2205-13.

111. Lee, H., et al., *Coexpression Analysis of Human Genes Across Many Microarray Data Sets.* Genome Research, 2004. **14**(6): p. 1085-1094.

112. MacLennan, N.K., et al., *Weighted gene co-expression network analysis identifies biomarkers in glycerol kinase deficient mice.* Mol Genet Metab, 2009. **98**(1-2): p. 203-14.

113. Mariño-Ramírez, L., et al., *Identification of cis-Regulatory Elements in Gene Co-expression Networks Using A-GLAM*. 2009. p. 1-20.

114. Yang, Y., et al., *Snapshot of iron response in Shewanella oneidensis by gene network reconstruction.* BMC Genomics, 2009. **10**: p. 131.

115. Atias, O., B. Chor, and D.A. Chamovitz, *Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network.* BMC Syst Biol, 2009. **3**: p. 86.

116. Mao, L., et al., *Arabidopsis gene co-expression network and its functional modules.* BMC Bioinformatics, 2009. **10**: p. 346.

117. Wang, Y., et al., *Function Annotation of an SBP-box Gene in Arabidopsis Based on Analysis of Co-expression Networks and Promoters.* Int J Mol Sci, 2009. **10**(1): p. 116-32.

118. Mentzen, W.I., et al., *Articulation of three core metabolic processes in Arabidopsis: fatty acid biosynthesis, leucine catabolism and starch metabolism.* BMC Plant Biol, 2008. **8**: p. 76.

119. Wei, H., et al., *Transcriptional coordination of the metabolic network in Arabidopsis.* Plant Physiol, 2006. **142**(2): p. 762-74.

120. Faccioli, P., et al., *From single genes to co-expression networks: extracting knowledge from barley functional genomics.* Plant Mol Biol, 2005. **58**(5): p. 739-50.

121. Lee, T.H., et al., *RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice.* Plant Physiol, 2009. **151**(1): p. 16-33.

122. Jupiter, D., H. Chen, and V. VanBuren, *STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data.* BMC Bioinformatics, 2009. **10**: p. 332.

123. Edwards, K.D., et al., *TobEA: an atlas of tobacco gene expression from seed to senescence.* BMC Genomics, 2010. **11**: p. 142.

124. Manfield, I.W., et al., *Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W504-9.

125. Obayashi, T., et al., *ATTED-II provides coexpressed gene networks for Arabidopsis.* Nucleic Acids Res, 2009. **37**(Database issue): p. D987-91.

126. Ogata, Y., H. Suzuki, and D. Shibata, *A database for poplar gene co-expression analysis for systematic understanding of biological processes, including stress responses.* Journal of Wood Science, 2009. **55**(6): p. 395-400.

127. Ogata, Y., et al., *CoP: a database for characterizing co-expressed gene modules with biological information in plants.* Bioinformatics, 2010. **26**(9): p. 1267-8.

128. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

129. Hosack, D.A., et al., *Identifying biological themes within lists of genes with EASE.* Genome Biol, 2003. **4**(10): p. R70.

130. Al-Shahrour, F., et al., *FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W91-6.

131. Gotz, S., et al., *High-throughput functional annotation and data mining with the Blast2GO suite.* Nucleic Acids Res, 2008. **36**(10): p. 3420-35.

132. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

133. Apweiler, R., et al., *The InterPro database, an integrated documentation resource for protein families, domains and functional sites.* Nucleic Acids Res, 2001. **29**(1): p. 37-40.

134. Hirochika, H., et al., *Retrotransposons of rice involved in mutations induced by tissue culture.* Proc Natl Acad Sci U S A, 1996. **93**(15): p. 7783-8.

135. Hruz T, L.O., Szabo G, Wessendrop F, Bleuler S, Oertle L, Widmayer P, Gruissem W and Zimmermann P, *Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes.* Advances in Bioinformatics, 2008. **2008**: p. 5.

136. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.

137. Bolstad, B. *RMAExpress*. 2009 2009 [cited 2010; Available from: http://rmaexpress.bmbolstad.com/.

138. Kauffmann, A., R. Gentleman, and W. Huber, *arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.* Bioinformatics, 2009. **25**(3): p. 415-6.

139. Gentleman, R.C., et al., *Bioconductor: Open software development for computational biology and bioinformatics.* Genome Biology, 2004. **5**: p. R80.

140. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Statistical Applications in Genetics and Molecular Biology, 2005. **4**: p. -.

141. Assenov, Y., et al., *Computing topological parameters of biological networks.* Bioinformatics, 2008. **24**(2): p. 282-4.

142. Moriya, Y., et al., *KAAS: an automatic genome annotation and pathway reconstruction server.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W182-5.

143. FAOSTAT. *Food and Agricultural Organization of the United Nations, Commodities Production Statistics. http://faostat.fao.org/site/339/default.aspx*. 2007; Available from: http://faostat.fao.org/site/339/default.aspx.

144. Paterson, A.H., et al., *Comparative Genomics of Grasses Promises a Bountiful Harvest.* Plant Physiology, 2009. **149**(1): p. 125-131.

145. Wang, K., M. Li, and H. Hakonarson, *Analysing biological pathways in genome-wide association studies.* Nat Rev Genet, 2010. **11**(12): p. 843-54.

146. Xu, G., et al., *Module detection in complex networks using integer optimisation.* Algorithms for Molecular Biology, 2010. **5**: p. -.

147. Chang, R.L., et al., *Deterministic graph-theoretic algorithm for detecting modules in biological interaction networks.* Int J Bioinform Res Appl, 2010. **6**(2): p. 101-19.

148. Mao, L., et al., *Arabidopsis gene co-expression network and its functional modules.* BMC Bioinformatics, 2009. **10**(1): p. 346.

149. Ficklin, S.P., F. Luo, and F.A. Feltus, *The association of multiple interacting genes with specific phenotypes in rice using gene coexpression networks.* Plant Physiol, 2010. **154**(1): p. 13-24.

150. Vandepoele, K., et al., *Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks.* Plant Physiology, 2009. **150**(2): p. 535-546.

151. Bandyopadhyay, S., R. Sharan, and T. Ideker, *Systematic identification of functional orthologs based on protein network comparison.* Genome Res, 2006. **16**(3): p. 428-35.

152. Singh, R., J. Xu, and B. Berger, *Global alignment of multiple protein interaction networks.* Pac Symp Biocomput, 2008: p. 303-14.

153. Kalaev, M., V. Bafna, and R. Sharan, *Fast and accurate alignment of multiple protein networks.* J Comput Biol, 2009. **16**(8): p. 989-99.

154. Liao, C.S., et al., *IsoRankN: spectral methods for global alignment of multiple protein networks.* Bioinformatics, 2009. **25**(12): p. I253-I258.

155. Zaslavskiy, M., F. Bach, and J.P. Vert, *Global alignment of protein-protein interaction networks by graph matching methods.* Bioinformatics, 2009. **25**(12): p. i259-67.

156. Flannick, J., et al., *Automatic parameter learning for multiple local network alignment.* J Comput Biol, 2009. **16**(8): p. 1001-22.

157. Hu, H., et al., *Mining coherent dense subgraphs across massive biological networks for functional discovery.* Bioinformatics, 2005. **21 Suppl 1**: p. i213-21.

158. Chindelevitch, L., C.S. Liao, and B. Berger, *Local optimization for global alignment of protein interaction networks.* Pac Symp Biocomput, 2010: p. 123-32.

159. Kuchaiev, O., et al., *Topological network alignment uncovers biological function and phylogeny*. J R Soc Interface, 2010. **7**(50): p. 1341-54.

160. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nature Genetics, 2002. **31**(1): p. 64-68.

161. Milo, R., et al., *Network motifs: Simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-827.

162. Zarrineh, P., et al., *COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms.* Nucleic Acids Res, 2010. **Epub ahead of print**.

163. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.

164. Usadel, B., et al., *Co-expression tools for plant biology: opportunities for hypothesis generation and caveats.* Plant Cell Environ, 2009. **32**(12): p. 1633-51.

165. Obayashi, T. and K. Kinoshita, *COXPRESdb: a database to compare gene coexpression in seven model animals.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1016-22.

166. Bolstad, B.M. *RMAExpress*. 2010 Dec, 13 2010].

167. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.

168. Moreno-Hagelsieb, G. and K. Latimer, *Choosing BLAST options for better detection of orthologs as reciprocal best hits.* Bioinformatics, 2008. **24**(3): p. 319-324.

169. De Smet, F., et al., *Adaptive quality-based clustering of gene expression profiles.* Bioinformatics, 2002. **18**(5): p. 735-46.

170. Yeung, K.Y., R.E. Bumgarner, and A.E. Raftery, *Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data.* Bioinformatics, 2005. **21**(10): p. 2394-402.

171. Butte, A.J., et al., *Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12182-6.

172. Barabasi, A.L., E. Ravasz, and T. Vicsek, *Deterministic scale-free networks.* Physica A, 2001. **299**: p. 559-564.

173. Wigner, E.P., *Random Matrices in Physics.* SIAM Review, 1967. **9**(1): p. 1-23.

174. Tulino, A.M. and S. Verdú, *Random matrix theory and wireless communications*. Foundations and trends in communications and information theory. 2004, Hanover, MA: Now. vi, 184 p.

175. Plerou, V., et al., *Random matrix approach to cross correlations in financial data.* Phys Rev E Stat Nonlin Soft Matter Phys, 2002. **65**(6 Pt 2): p. 066126.

176. Chok, N.S., *Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data*, in *Graduate School of Public Health*. 2010, University of Pittsburgh.

177. Wang, H., et al., *Towards patterns tree of gene coexpression in eukaryotic species.* Bioinformatics, 2008. **24**(11): p. 1367-73.

178. Leskovec, J., et al., *Kronecker Graphs: An Approach to Modeling Networks.* Journal of Machine Learning Research, 2010. **11**: p. 985-1042.

179. Kalinka, A.T. and P. Tomancak, *linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type.* Bioinformatics, 2011. **27**(14): p. 2011-2.

180. Punta, M., et al., *The Pfam protein families database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D290-301.

181. Cohen, J., *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **20**(1): p. 10.

182. Galassi, M., et al., *Gnu Scientific Library: Reference Manual*. 2003: Network Theory Ltd.

183. *Intel® Math Kernel Library*. 2012; Available from: http://software.intel.com/en-us/articles/intel-mkl/.

184. Bolstad, B.M. *RMAExpress*. 2012 July, 2012]; http://rmaexpress.bmbolstad.com/:[Available from: http://rmaexpress.bmbolstad.com/.

185. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

186. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D773-9.

187. Goffeau, A., et al., *Life with 6000 genes.* Science, 1996. **274**(5287): p. 546, 563-7.

188. Yamamoto, T., J. Yonemaru, and M. Yano, *Towards the understanding of complex traits in rice: substantially or superficially?* DNA research : an international journal for rapid publication of reports on genes and genomes, 2009. **16**(3): p. 141-54.

189. Li, H. and P. Zhang, *Systems genetics: challenges and developing strategies.* Biologia, 2012. **67**(3): p. 435-446.

190. Massa, A.N., et al., *The transcriptome of the reference potato genome Solanum tuberosum Group Phureja clone DM1-3 516R44.* PLoS One, 2011. **6**(10): p. e26801.

191. Iancu, O.D., et al., *Utilizing RNA-Seq data for de novo coexpression network inference.* Bioinformatics, 2012. **28**(12): p. 1592-7.

192. Spangler, J.B., et al., *Conserved Non-Coding Regulatory Signatures in Arabidopsis Co-Expressed Gene Modules.* PLoS One, 2012. **7**(9).

193. Presson, A.P., et al., *Integrated Weighted Gene Co-expression Network Analysis with an Application to Chronic Fatigue Syndrome.* Bmc Systems Biology, 2009. **2**.

194. Mungall, C.J. and D.B. Emmert, *A Chado case study: an ontology-based modular schema for representing genome-associated biological information.* Bioinformatics, 2007. **23**(13): p. i337-46.

195. Ficklin, S.P., et al., *Tripal: a construction toolkit for online genome databases.* Database (Oxford), 2011. **2011**: p. bar044.

196. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.

197. Lopes, C.T., et al., *Cytoscape Web: an interactive web-based network browser.* Bioinformatics, 2010. **26**(18): p. 2347-8.

198. Wang, Z.Y., et al., *The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene.* Plant J, 1995. **7**(4): p. 613-22.

199. Chen, M.H., et al., *Waxy gene haplotypes: Associations with apparent amylose content and the effect by the environment in an international rice germplasm collection.* Journal of Cereal Science, 2008. **47**(3): p. 536-545.

200. Ayliffe, M.A. and E.S. Lagudah, *Molecular genetics of disease resistance in cereals.* Annals of Botany, 2004. **94**(6): p. 765-773.

201. Li, D.Y., et al., *Ectopic Expression of Rice OsBIANK1, Encoding an Ankyrin Repeat-Containing Protein, in Arabidopsis Confers Enhanced Disease Resistance to Botrytis cinerea and Pseudomonas syringae.* Journal of Phytopathology, 2013. **161**(1): p. 27-34.

202. Zhang, X.C., et al., *Molecular characterization of rice OsBIANK1, encoding a plasma membrane-anchored ankyrin repeat protein, and its inducible expression in defense responses.* Molecular Biology Reports, 2010. **37**(2): p. 653-660.

203. Jaiswal, P., et al., *Gramene: development and integration of trait and gene ontologies for rice.* Comp Funct Genomics, 2002. **3**(2): p. 132-6.

204. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.