Plant Science 223 (2014) 45–48

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Review

Systems genetics: A paradigm to improve discovery of candidate genes and mechanisms underlying complex traits

F. Alex Feltus*

Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

ARTICLE INFO

Article history:
Received 15 December 2013
Received in revised form 18 February 2014
Accepted 2 March 2014
Available online 13 March 2014

Keywords:
Systems genetics
eQTL
Co-expression network
Genotype–phenotype

ABSTRACT

Understanding the control of any trait optimally requires the detection of causal genes, gene interaction, and mechanism of action to discover and model the biochemical pathways underlying the expressed phenotype. Functional genomics techniques, including RNA expression profiling via microarray and high-throughput DNA sequencing, allow for the precise genome localization of biological information. Powerful genetic approaches, including quantitative trait locus (QTL) and genome-wide association study mapping, link phenotype with genome positions, yet genetics is less precise in localizing the relevant mechanistic information encoded in DNA. The coupling of salient functional genomic signals with genetically mapped positions is an appealing approach to discover meaningful gene–phenotype relationships. Techniques used to define this genetic–genomic convergence comprise the field of systems genetics. This short review will address an application of systems genetics where RNA profiles are associated with genetically mapped genome positions of individual genes (eQTL mapping) or as gene sets (co-expression network modules). Both approaches can be applied for knowledge independent selection of candidate genes (and possible control mechanisms) underlying complex traits where multiple, likely unlinked, genomic regions might control specific complex traits.

Contents

1. Introduction 45
2. Systems genetics via eQTL mapping 46
3. Systems genetics via co-expression network mapping to genetic positions 46
4. Conclusions and future prospects 47
References 48

1. Introduction

No gene acts in isolation. Biological information encoded in DNA, for example, must first be transcribed into RNA for which steady-state concentrations are controlled through the complex biochemistry of distal gene products including *trans*-acting regulatory proteins, chromatin remodeling machinery, the RNA polymerase complex, RNA splicing factors, RNA transport proteins, and RNA degradation factors. Prior to influencing steady-state RNA levels of a given gene, each protein factor may have undergone functional modification at the level of protein translation, post-translational modification, sequestration, or conformational change. Thus, each and every gene product at birth is interacting with hundreds of gene products prior to translation or performing its function as native RNA. Of course this simple example does not begin to describe the complex interaction of mature gene products in the biochemical pathways that control qualitative and quantitative traits with a range of heritability.

It is now common to measure gene output on a genome scale for all known genes in an organism to address the complexity of real-world gene expression. Individual transcript RNA concentrations are determined with global detection techniques such as RNA hybridization to microarrays or through the conversion of RNA into DNA and direct sequencing with high-throughput next-generation sequencers. In the next-generation RNAseq method, specific transcript concentrations are determined by mapping reads back to

* Correspondence to: Department of Genetics and Biochemistry, Clemson University, 105 Collings Street, Room #302C, Clemson, SC 29634, USA.
Tel.: +1 864 656 3231; fax: +1 864 656 6879.
E-mail address: ffeltus@clemson.edu

http://dx.doi.org/10.1016/j.plantsci.2014.03.003
0168-9452/

a reference genome or transcript assembly, and then counting molecule occurrence. While more challenging, it is also possible to profile gene expression at the protein level using proteomic profiling methods. Through comparison of biologically relevant sample groups, it is routine to identify differentially expressed genes that are associated with a change in gene expression state. In this way, information encoded at specific genome positions (i.e. functional genes) can be associated with relevant biological conditions.

Of course, gene expression varies for each individual in a population. Gene expression is initialized by the genetic and epigenetic background of an individual organism and heavily influenced by the regulatory context within a cell as well as by external environmental factors. It is sometimes possible to associate causal or nearby polymorphic markers with heritable, quantitative traits. Quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) use linkage analysis and population genetics, respectively, to identify genome intervals associated with expression of phenotype. QTL mapping and GWAS, however, merely narrow down the genome position to near where the causal variation is located and rarely identify the causal variation. From a genomics perspective, genetics reduces the genome to a reasonable fraction for the discovery of candidate sequences encoding relevant functional information.

Once the genome fraction controlling the trait is genetically tagged, the researcher often turns to laborious positional cloning experiments or selects proximal candidate genes via prior knowledge and intuition. Since some or all of the functions of an individual gene may not be known, the candidate gene approach is often unsuccessful or tempts the researcher to continue to try to fit the gene into a causal hypothesis, which can waste time and resources. If successful, the detection of a causal gene might be relevant only in the mapping population where it was discovered and rarely provides context of how this genetically relevant genome position (e.g. large effect QTL) interacts with other genes leading to expression of a complex trait. Furthermore, it is also necessary fill in the "missing data" of genetically undetectable genes involved in phenotype expression to truly understand the underlying biochemistry underlying a phenotype. Ideally, the selection of candidate gene options should be identified in a knowledge independent manner that maintains gene dependency context, even for those genes that are "genetically invisible" for which there is not enough power to measure an effect in a given mapping population.

A subfield of systems biology, *systems genetics*, provides a powerful approach to merge genomic and genetic data to discover not only candidate genes underlying the expressed phenotype but also ascertain the mechanistic context of a gene or gene interaction module [1,2]. Systems genetics involves the analysis of high-dimensional genomic data, thousands of measurements often in a matrix format, such as RNA expression levels for tens of thousands of genes in an organism. Gene expression is mapped to specific genome positions and coded for biological context. These specific positions can then be phased into genetically derived genome positions to generate 'candidate mechanism' hypotheses in a monogenic or polygenic context. Two systems genetics methods that illuminate this powerful approach are described in the following sections.

2. Systems genetics via eQTL mapping

One method to merge functional genomic data with genetic signal is through expression quantitative trait locus (eQTL) mapping [3]. In this approach, applied early in yeast [4], transcriptomes are profiled using microarrays or direct sequencing (RNAseq) in a well genotyped, segregating population. RNA expression levels, a collection of quantitative "traits", are associated with polymorphic markers identifying *cis*- and *trans*-acting positions affecting specific gene expression. Using the eQTL approach, segregating gene expression patterns are pinpointed empirically and clues to mechanism affecting gene expression are revealed. In *Arabidopsis* for example, thousands of eQTLs were identified in a recombinant inbred line mapping population [5]. In a rice study, the eQTL approach has been used to identify over 16,000 eQTL control points, a subset of which corresponded with biomass yield [6]. In a separate rice eQTL analysis, eQTL hotspots were associated with oxidative stress [7]. A systems genetics study by Faraji et al. [8] provides an excellent example of the power of eQTL mapping. They analyzed mRNA and miRNA expression profile data from tumors from mice progeny segregating for tumor metastatic potential. Following co-expression network construction and miRNA eQTL analysis, they were able to discover specific miRNA controllers of transcriptional networks underlying metastasis potential in their system. Furthermore, they were able to validate their findings empirically. The eQTL method points to specific regulatory mechanisms at specific genome positions (i.e. genome control points of steady state-RNA levels of Gene X) that may be responsible for specific traits. When eQTLs are identified using a population segregating for a trait of interest, the regulatory mechanisms pointed to by the eQTL can be extrapolated to understand gene output at the level of steady-state mRNA.

While extremely powerful, the eQTL approach does have limitations. First, these experiments are very expensive. Each individual must be phenotyped (RNA profiled) and genotyped in order to map the eQTL. In the future, it may be possible to use next-generation sequencing techniques to cheaply profile the RNA from any sample, but there will still be a heavy cost in terms of computational resources to process these Big Data collections. Fortunately, scalable computational solutions exist such as iPlant, a computational discovery environment specifically geared toward solving plant biology problems [9]. Another limitation is that if the relevant tissue or developmental time point with high impact on phenotypic expression is not sampled, then the causal eQTLs will not be identified. This issue can be addressed by including more tissue and time course measurements in the experimental design phase albeit with a significant increase in cost. Finally, eQTLs are determined individually for each transcript and do not immediately identify gene–gene dependency, a key concern for complex traits, unless a common control *trans*-acting control point is mapped for several loci. Is there another systems genetics approach to couple gene (co-)expression and with genetically mapped loci?

3. Systems genetics via co-expression network mapping to genetic positions

An alternate, possibly parallel, approach is to determine what gene–gene relationships are possible in an organism by building gene interaction networks from public (or private) gene expression profiles, even if the data that was obtained from a genetically undefined system. These gene dependencies can then be tested for correspondence with genetic networks obtained from rigorous genetic analyses. For example, gene dependencies can be identified through the construction of gene co-expression networks (GCNs) [10]. RNA profiles have been generated under myriad of experimental conditions and genetic backgrounds for numerous plants. As of this writing, there are over 71,000 Gene Expression Omnibus public dataset records for green plants (Viridiplantae; taxonomy ID 33090 [11]). On a per-organism basis, these RNA profiling experiments can be repurposed to identify gene co-expression relationships in the form of GCNs.

Plant GCNs and protein interaction networks (e.g. [12]) have been constructed for numerous species resulting in numerous

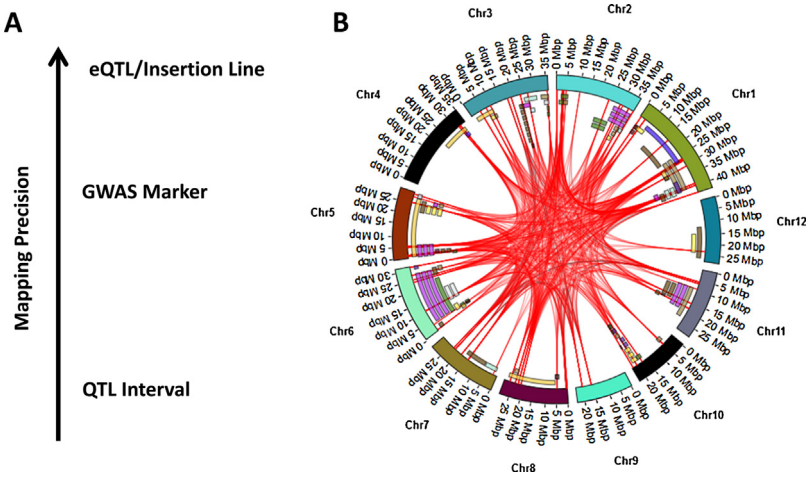

Fig. 1. Coupling genetic signal with candidate gene positions. (A) The mapping precision of identifying the causal DNA information impacting phenotypic expression improves by mapping technique. (B) Coupling gene interactions with known genetic positions should increase the selection of candidate genes controlling phenotypic expression. The 12 rice chromosomes are shown in a circle with a module of co-expressed genes overlapping genetically mapped intervals linked by redlines. The boxes indicate mapped QTL intervals for a specific trait. Boxes of the same color are from the same mapping experiment. Genome alignments of GCN modules, GWAS SNVs, and QTLs with references for several plants can be found in the Gene NetEngine browser at http://sysbio.genome.clemson.edu.

collections of gene interaction patterns. For example, we constructed GCNs for *Arabidopsis thaliana* using 7105 RNA profiles obtained across numerous experimental conditions and tissue sources [13]. In an effort to reduce confounding noise from multiple RNA sources, we used a knowledge-independent approach to sort RNA expression profiles into groups based on overall expression patterns prior to network construction. This approach resulted in 86 *Arabidopsis* GCNs that encompass 558,022 unique co-expression relationships between 19,588 genes. This 86 GCN compendium is a gene interaction framework map for *Arabidopsis*. While it is easier to capture RNA co-expression relationships, it may be important to identify protein co-expression and protein–protein interaction networks given that some functionally relevant gene interaction modules might not be exist as correlated steady-state RNA. One way to look at a gene interaction network is a list of pre-mapped genome interaction points that can be associated with genetic positions and applied to systems genetics problems.

Once a gene interaction network is constructed, it can be dissected into communities of genes that interact tightly in the network termed gene *modules*. Each module is hypothesized to be co-functional in that the co-expressed genes are involved in similar biological processes. Discrete function can be assigned by counting the number of times a label (e.g. Gene Ontology term) is present in a module and if that count differs significantly relative to the genome background. Functional profiling can be applied to any label including reverse genetics datasets where disruption lines have been phenotyped and the gene lesion mapped. For example, we constructed a knowledge-independent compendium of twenty-two *Oryza sativa* GCNs from 1306 Affymetrix microarray profiles downloaded from the NCBI Gene Expression Omnibus. Functional profiling was performed on genes mapped to phenotypes measured in rice retrotransposon Tos17 insertion lines for each module obtained from this GCN compendium [14]. This approach revealed numerous modules significantly enriched for specific rice traits and suggests that genes in the phenotype-enriched network module that where not found to be phenotype-associated could play a role in phenotypic expression in alternate genetic backgrounds. In essence, module-phenotype assignment is a genome reduction strategy that provides rationale for the module genes being candidate genes without the bias of prior knowledge of what should control a trait.

Gene disruption experiments are often designed so that the DNA sequence alteration can be pinpointed on a chromosome and the candidate gene identified. In addition to module enrichment of discrete reverse genetics derived phenotypes, the co-functionality of genes in a network module can also be associated with longer genetically mapped genome intervals that may contain numerous candidate genes (e.g. QTL intervals or GWAS-tagged linkage disequilibrium blocks). For example, a collection of trait-associated single nucleotide variants (SNVs) in a GWAS study or a set of QTL intervals in a single study points to specific genetic positions near controlling genes (or *trans* eQTLs). The genomic resolution of significant GWAS SNVs is often higher than that of QTL intervals (Fig. 1A). If the assumption is made that the collection of genetic signals are partially controlled at the level of RNA co-expression, then significant overlap with these genetic positions would be evidence for possible control of phenotypic expression. In a model study in *Oryza sativa*, we aligned the modules from 22 rice GCNs to the genome as well as phenotypically significant GWAS SNVs and over 8000 rice QTLs. An example module aligned to several arbitrary sets of QTLs from the same trait is shown in Fig. 1B. Using this approach, an investigator can constrain a candidate gene list to genes proximal to a set of genetic signals as well as co-expressed across multiple experimental conditions, possibly within the same populations where the genetic measurements were obtained.

4. Conclusions and future prospects

Systems genetics can be used to identify candidate genes and molecular mechanisms underlying complex phenotypes using approaches including eQTL mapping, functional profiling of reverse genetic phenotypes, and association of gene modules with genetically mapped positions. These approaches are in an early stage, but a growing number of promising studies support their utility across the tree of life [1,15–17]. Given the growing number of genome positions (i.e. genes) mapped by GWAS for a given complex trait, a systems biology approach may be necessary to integrate and interpret the significance of these signals. For example, a Genetic Association Database search for genes identified with GWAS and associated with Type II diabetes resulted in more than 400 unique loci [18]. Robust models of this many potential genetic interactions will be necessary to dissect causality if this many genes are involved in phenotypic expression across a relevant population.

As with most genomics studies, the reduction of measurement cost will enhance the systems genetics approach. If next-generation sequencing platforms continue to capture more sequence for less

money, then measuring RNA samples from more tissues in more mapping populations may be feasible as a standard mechanistic phenotyping procedure. A powerful application would be to determine RNA expression profiles across multiple organs from pre-genotyped individuals in genetic diversity panels and nested association mapping populations such as those developed for maize [19] and sorghum [20]. These genetically defined populations with hundreds to thousands of individuals have been extensively phenotyped for morphological and other traits across environments. Coupling these genetically mapped positions discovered in these mapping populations with eQTLs would provide powerful mechanistic insight into specific traits as demonstrated by the mouse study discussed above [8]. Similarly, GCNs derived from these same transcriptome measurements could help pinpoint gene dependencies when network modules that overlap genetically defined intervals. This technique could be especially fruitful if the genes in a network module are controlling phenotypic expression in the mapping population as they would be more likely to correspond with genetically derived positions.

References

[1] C.R. Farber, Systems genetics: a novel approach to dissect the genetic basis of osteoporosis, Curr. Osteoporos. Rep. 10 (2012) 228–235.
[2] J.H. Nadeau, A.M. Dudley, Systems genetics, Science 331 (2011) 1015–1016.
[3] H. Li, H. Deng, Systems genetics, bioinformatics and eQTL mapping, Genetica 138 (2010) 915–924.
[4] R.B. Brem, G. Yvert, R. Clinton, L. Kruglyak, Genetic dissection of transcriptional regulation in budding yeast, Science 296 (2002) 752–755.
[5] M.A. West, et al., Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*, Genetics 175 (2007) 1441–1450.
[6] J. Wang, et al., A global analysis of QTLs for expression variations in rice shoots at the early seedling stage, Plant J. 63 (2010) 1063–1074.
[7] F. Liu, et al., Gene expression profiles deciphering rice phenotypic variation between Nipponbare (Japonica) and 93-11 (Indica) during oxidative stress, PLoS ONE 5 (2010) e8632.
[8] F. Faraji, et al., An integrated systems genetics screen reveals the transcriptional structure of inherited predisposition to metastatic disease, Genome Res. 24 (2014) 227–240.
[9] S.A. Goff, et al., The iPlant collaborative: cyberinfrastructure for plant biology, Front. Plant Sci. 2 (2011) 34.
[10] A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, I.S. Kohane, Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 12182–12186.
[11] T. Barrett, et al., NCBI GEO: archive for functional genomics data sets – 10 years on, Nucleic Acids Res. 39 (2011) D1005–D1010.
[12] Arabidopsis Interactome Mapping Consortium, Evidence for network evolution in an *Arabidopsis interactome* map, Science 333 (2011) 601–607.
[13] F.A. Feltus, S.P. Ficklin, S.M. Gibson, M.C. Smith, Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an *Arabidopsis* case study, BMC Syst. Biol. 7 (2013) 44.
[14] A. Miyao, et al., A large-scale collection of phenotypic data describing an insertional mutant population to facilitate functional analysis of rice genes, Plant Mol. Biol. 63 (2007) 625–635.
[15] J.F. Ayroles, et al., Systems genetics of complex traits in *Drosophila melanogaster*, Nat. Genet. 41 (2009) 299–307.
[16] S.P. Ficklin, F.A. Feltus, A systems genetics approach and data mining tool to assist in the discovery of genes underlying complex traits in *Oryza sativa*, PLoS ONE 8 (2013) e68551.
[17] C.L. Plaisier, et al., A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia, PLoS Genet. 5 (2009) e1000642.
[18] Y. Zhang, et al., Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information, BMC Med. Genomics 3 (2010) 1.
[19] M.D. McMullen, et al., Genetic properties of the maize nested association mapping population, Science 325 (2009) 737–740.
[20] E.S. Mace, C.H. Hunt, D.R. Jordan, Supermodels: sorghum and maize provide mutual insight into the genetics of flowering time, Theor. Appl. Genet. 126 (2013) 1377–1395.