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

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