Running Head: Systems Biology and Natural Genetic Variation

* To whom correspondence should be addressed.

Daniel J. Kliebenstein

Department of Plant Sciences

University of California, Davis

One Shields Avenue

Davis, CA 95616

USA

Phone: 530-754-7775

E-mail: kliebenstein@ucdavis.edu

Journal research area: Systems Biology, Molecular Biology, and Gene Regulation

Or Genetics, Genomics and Molecular Evolution
Title: Systems Biology uncovers the foundation of natural genetic diversity.

Daniel J. Kliebenstein

Department of Plant Sciences

University of California, Davis

One Shields Avenue

Davis, CA 95616

USA
Financial source (if any) and the experiment station or institution paper number;

The presented work was supported by a National Science Foundation Grant DBI 0820580 to DJK.

Present address(es) of authors if different from heading;

* To whom correspondence should be addressed.

Daniel J. Kliebenstein

Department of Plant Sciences

University of California, Davis

One Shields Avenue

Davis, CA 95616

USA

Phone: 530-754-7775

E-mail: kliebenstein@ucdavis.edu

Fax: 530-752-9659
Abstract

The vast majority of measurable phenotypes within species are not fixed and populations contain significant levels of natural genetic variation among individuals affecting phenotypes from development to metabolism to abiotic resistance. All of which are of interest to both basic and applied biologists from a myriad of fields. Despite the ubiquity of this variation there is very little known about the molecular underpinnings of natural genetic variation or the forces behind its maintenance or generation. Recent advances in both genomics and systems biology are beginning to allow some of the first direct empirical tests of a suite of parameters that while being a foundation of natural variation were largely left to the theoreticians. These include the following basic questions: what is Pleiotropy?; how many genes control a given quantitative trait?; where is the heritability? and how is conditional genetic variation generated. This review highlights progress made towards addressing these questions via the use of systems biological inquiries into natural variation.
Introduction

The vast majority of measurable phenotypes within species are not fixed and populations contain significant levels of natural genetic variation among individuals for traits ranging from development to metabolism to abiotic resistance. While the presence of this diversity is widely acknowledged, much less is known about the consequences of natural genetic variation or the forces behind its maintenance or generation which are fundamental to evolution and ecology. In addition, a wide range of applied fields are finding a need to better understand the systematic basis of natural variation, including efforts to improve crop yields via breeding and attempts towards individualized medicine for humans. These efforts are complicated because this diversity is typically polygenic and can involve complex interactions with numerous factors including, but not limited to, the environment, development, epistatic interactions between genes and potential higher-order interaction among these factors (Falconer and Mackay, 1996; Lynch and Walsh, 1998; Wentzell and Kliebenstein, 2008). While most studies on natural variation to-date have queried one or two phenotypes, there is an obvious need for more intricate, genomic and broad phenotypic analyses of natural variation to address the fundamentals of quantitative genetics.

Systems Biology is a newly defined and emerging field that attempts to conduct massively parallel experiments that differentially perturb a system to identify the properties of the system being tested (Ideker et al., 2001; Kitano, 2002). This has led to the systematic identification of networks controlling gene expression within bacteria, protein interactions in numerous species and metabolic networks (Fiehn, 2002; Rosenfeld et al., 2002; Covert et al., 2004; Martins et al., 2004). The utility of systems biology has gone beyond the ability to
describe individual phenotypes to begin describing the shape of the mechanistic interactions or rules that define how the networks can be shaped (Albert, 2005). Systems Biology has also begun to make great headway in helping to better understand and describe plant biology (Brady and Provart, 2009). This includes the development of key approaches to combine metabolomic and transcript data to identify novel enzymatic and regulatory genes (Hirai et al., 2005; Hansen et al., 2007). This approach has further been extended to establish links between gene expression and root development and between development and abiotic stress within *Arabidopsis thaliana* (Brady et al., 2007; Dinneny et al., 2008). As such, the concept and implementation of systems biology is showing great promise to establish both individual gene linkages and more theoretical rules guiding these linkages.

Numerous systems biology approaches are now being applied to the study of natural genetic variation including transcriptomics (Keurentjes et al., 2008; Kliebenstein, 2009), metabolomics (Keurentjes, 2009; Kliebenstein, 2009), proteomics (Stylianou et al., 2008), compilation of large physiological datasets (Keurentjes et al., 2008), network analysis tools (Hansen et al., 2008; Kliebenstein, 2009) and whole genome sequencing (Borevitz and Chory, 2004; Nordborg and Weigel, 2008). Each of these approaches has a myriad of strengths and aid in the rapid identification of the genes underlying phenotypes ranging from flowering time to biotic resistance mechanisms. However, their potential uses in understanding the genetic underpinnings of specific biological phenotypes have been reviewed extensively in the past several years as described in the preceding citations. Given the depth of published reviews on the use of systems biology to identify causal genes in natural genetic variation, I will instead use this review to focus on how these genomics approaches are allowing researchers to begin developing
a systems view of how the genetic architecture of natural variation is shaped and what evolutionary events are shaping this variation.

At its essence, the application of systems biology to the analysis of natural genetic variation is allowing the research community to begin directly testing the long standing and fundamental yet little understood underpinnings of quantitative genetics that had previously been yielded to the theoreticians. This combination also benefits Systems Biology experiments as natural genotypes typically differ at hundreds to thousands of genes providing access to dramatically more than a single perturbation at a time. This combination provides access to experimental tests of several theoretical underpinnings of natural genetic diversity that could be summed up with the following basic questions: What is Pleiotropy?, How many genes control a given quantitative trait? Where is the heritability? and How is conditional genetic variation generated? This list of questions is by no means exhaustive, for instance epistasis is a major factor in quantitative genetics but the use of systems biology to study epistasis has been dealt with in other reviews (Carlborg and Haley, 2004; Phillips, 2008; Kliebenstein, 2009). Each of the stated questions is beginning to be illuminated by recent systems biology interrogations of natural variation and will be a separate focus of this review.

**What is Pleiotropy?** One of the first lectures in any introductory genetics class attempts to define the meaning of pleiotropy. While the general definition is simple enough, “when a gene influences multiple traits/phenotypes,” the mechanistic basis generating pleiotropy, or even the frequency to which it occurs are not frequently investigated nor understood. Often in molecular or Mendelian genetics when a mutation in a single gene is observed to have a predominant impact on one phenotype and smaller affects on other phenotypes, this is disregarded as being a
secondary or indirect consequence and not an inherent part of the biology of the gene. Further, it is assumed that these secondary consequences are due to the presence of the gene within a network that responds both locally (direct effect) and distally (secondary effect) to the perturbation in the given gene. However, it is not known if these “secondary” linkages are random consequences of network architecture or if these distal network connections are present to generate pleiotropic linkages and that pleiotropy may be a direct part of the gene’s biology within the plant in the natural setting.

In quantitative genetic systems where most gene polymorphisms have moderate to small effects, it becomes more difficult to discern local from distal effects by using phenotypic magnitude. However, pleiotropy takes on added importance because the presence of pleiotropy means that selection upon one phenotype can have unexpected consequences on other phenotypes if the phenotypes can be effected by the same gene. Systems biology allows for a unique survey into pleiotropy as it allows for a dramatic expansion in the number of phenotypes that can be measured from the typical study of one to a dozen phenotypes to expression studies querying thousands of genes across natural populations. Within these studies, evidence for high levels of pleiotropy for natural genetic polymorphisms arises from the presence of hotspots when measuring quantitative variation in either transcripts or metabolites (Keurentjes et al., 2008; Keurentjes, 2009; Kliebenstein, 2009; Kliebenstein, 2009). These hotspots are regions of the genome where a genetic polymorphism is linked to phenotypic diversity in thousands of transcripts or dozens of metabolites, vastly more than would be expected by chance. Often, this pleiotropy is used to argue for the presence of a polymorphic regulatory gene in this hotspot. Several transcript hotspots have been cloned and shown to be regulatory genes that have local/direct effects on dozens of genes representing a local network and this effect then
permeates to distal/indirect effects on hundreds of additional genes (Keurentjes et al., 2007; Sønderby et al., 2007; Wentzell et al., 2007). In one case, the regulatory gene encoded a enzyme within a secondary metabolite pathway that altered both transcript and metabolite abundance for the pathway expanding the definition of ‘regulatory’ (Wentzell et al., 2007). While the terms distal or secondary are often used to imply a lack of biological linkage, investigations into these distal genes often suggest modifying biological linkages to the directly affected genes. However, the frequency by which distal/secondary genes contain a biological link to the primary phenotype remains to be tested in both natural and lab-induced variation studies. Of particular interest will be studies that compare the pleiotropic impacts of genetic perturbation to the given genes’ network architecture to test if the pleiotropic consequences could have been predicted.

While these systems analyses of natural variation have shown pleiotropy, it is typically amongst related molecular phenotypes which raises the question of “what is a distinct phenotype?” Should we consider transcript abundance for 100 genes controlled by a single transcription factor, to be 100 different phenotypes or 1 phenotype measured 100 different ways? To alleviate this issue, Keurentjes and colleagues took this approach a step further to test if there was a link between hotspots across different molecular phenotypes (transcripts, metabolites and proteins) as well as among a large compilation of physiological phenotypes (Fu et al., 2009). The ability to combine molecular and physiological phenotypes was greatly aided by the author’s use of a homozygous Arabidopsis thaliana population that had been phenotyped for different physiological traits by a large number of labs (Fu et al., 2009). In this study, the authors were able to show that there were a discrete number of hotspots that controlled transcript, metabolic, protein and physiological traits. Thus, they identified naturally polymorphic genetic loci that altered phenotypes at every measurable level within this population, the very definition
of pleiotropy. Interestingly, since these pleiotropic loci are naturally polymorphic, they are exposed to selection. It will require the identification of these underlying highly pleiotropic genes to determine if they exhibit patterns of selection that differ from less pleiotropic, yet also naturally variable, genes. It will be interesting to see more studies integrating molecular and phenotypic datasets in the future to uncover how frequently these highly pleiotropic loci arise.

How many genes? While it may seem surprising, there is little empirical data describing how many genes control natural variation for a phenotype within a species. In the absence of hard empirical data, there are two theoretical models that attempt to describe how many genes may control a phenotype. The oldest is the infinitesimal model where it is assumed that a quantitative trait is determined by an infinite number of loci with infinitesimal effects and without epistasis (Fisher, 1918). This has been a useful statistical model to allow the modeling of quantitative traits and genetic variation in a myriad of simulations but has always evoked unease in biologists with the word infinity. In contrast, observations on quantitative trait loci mapping in a number of species and populations has identified frequent genes of large effect, typically a defined number of genes and a range of effects, which has led to a counter model called the finite-locus model (Thompson and Skolnick, 1977; Cannings et al., 1978). However, this model is typically derived from data involving a structured population containing typically two genotypes from a species and as such it is not known how this model translates to an entire species. While both models provide useful approximations for simulation studies it is not fully understood which or if either model best describes biological reality.

Recent work querying large phenotypic datasets identified the presence of a few highly pleiotropic loci with moderate to large effects within a given population (Fu et al., 2009). This
observation was used to indicate that the finite model is more correct and that quantitative genetics may involve a few moderate loci (Fu et al., 2009). However, the presence of moderate effect loci diminishes the ability to identify other genetic determinants and as such there is a strong potential for false negative results in these studies (Beavis, 1994). This indicates that it is possible that mapping in a structured population is akin to an onion, if you remove one layer of moderate effect loci you may simply identify more loci that show a similar distribution of effects to the first layer. To directly test this potential with sufficient genotype-phenotype detection power, massive structured populations on the scale of thousands of individual genotypes generated from a biparental mating will be required.

As a first step to answering the question of how many genes control a naturally variable phenotype, a Nested Association Mapping (NAM) structure has been developed in maize (McMullen et al., 2009). In this structure a set of homozygous parents are crossed individually to a single concurrent parent to generate a set of moderate structured populations. These are then combined to create a single large population for phenotypic and genotype-phenotype analysis. The NAM population currently has nearly 5,000 lines from 25 different parent crosses. Analysis of a single phenotype, flowering time, within this population identified 333 significant loci controlling this phenotype, the vast majority with small to moderate effects (Buckler et al., 2009). Given the criteria used in selecting the parental germplasm, 333 genetic loci likely represents a minimal estimate for the number of loci controlling flowering time in Maize. This begins to support the concept of the infinitesimal (or at least a near-infinitesimal) model by suggesting a large number of loci. Further support for this concept was provided by an apparent lack of detectable epistasis in this NAM structure (Buckler et al., 2009; McMullen et al., 2009). However, other species frequently show epistasis for flowering time (Caicedo et al., 2004). It
remains to be seen if this is due to differences in the breeding systems between the species; i.e. Maize being outcrossing and Arabidopsis/rice being inbreeding; or if the disconnect results from wild versus domesticated species. Parallel studies of large populations from multiple species that differ in breeding design and domestication for a common broad set of phenotypes could begin to allow for direct empirical tests of how many genes control phenotypes within species and how the life history of the species impacts these estimates.

Where is the heritability? One common difficulty with natural variation studies is that the final model obtained rarely describes the full variation in the phenotypes. This gap in knowledge is often described as missing heritability. While this missing heritability may be due to missing loci, it may also be due to environmental interactions or untested epistatic interactions (Mackay, 2001). Systems analysis of natural variation within plants is beginning to identify another potential source for missing heritability: natural variation in gene methylation patterns (Richards, 2008). Studies have identified heritable natural variation for gene methylation in both Arabidopsis and rice (Takata et al., 2005; Vaughn et al., 2007; Woo and Richards, 2008). This has recently been extended to whole genome surveys showing extensive variation in gene methylation patterns and that methylation levels correlated with differential gene expression between Arabidopsis accessions (Zhang et al., 2008). Interestingly, the direction of this correlation depended upon the part of the gene being surveyed (Zhang et al., 2008). While there are a few instances where natural variation in methylation is linked to downstream phenotypes such as flowering time in Arabidopsis (Shindo et al., 2006), it remains to be seen how frequently epigenetic variation plays a role in natural phenotypic variation. In this review, epigenetic variation will be limited to that variation which could not have been predicted from DNA
sequence, e.g. Methylation, acetylation, etc. As such, a DNA polymorphism within a miRNA is not considered epigenetic for this reviews purposes.

One source of information addressing the potential role of epigenetic variation in phenotypic variation was the generation of structured populations that contained predominantly epigenetic variation (Johannes et al., 2009; Reinders et al., 2009). Both populations showed extensive heritable phenotypic variation. While some of this variation arose from genetic mutations caused by transposon movement, there was a significant fraction arising from heritable epigenetic variation. This raises the potential that each gene in a natural population or structured mapping population actually has two genotypes, one from the DNA sequence and the other from epigenetic modification(s). Interestingly, while both DNA and epigenetic genotypes are heritable, they are not equivalent. While the DNA genotype very rarely spontaneously mutates, studies have shown that the epigenetic genotype can convert at rates exceeding 20% per generation (Vaughn et al., 2007). Thus, the DNA genotype is not predictive of a gene’s epigenetic genotype. However, since most studies only measure and test the DNA genotype it is possible that the epigenetic genotype contains/encodes some of the heritability not explained by the DNA genotype. Thus, future studies querying the systems biology of epigenetic versus genetic control of natural variation will need to measure both the DNA and epigenetic genotype for a gene and use them concurrently to test for links with natural phenotypic variation.

The concept of an epigenetic genotype begins to raise a number of questions about how far the genotype analogy can be extended. For instance, two DNA SNPs are linked based upon the level of recombination which is correlated to physical distance in basepairs between the SNPs. However, it is not known if there is an equivalent concept of linkage between two methylation sites and their polymorphic state or alternatively, if two methylation sites behave
independently. It will be interesting to see if genetic concepts of linkage can be extended to methylation genotypes. This potential independence between DNA and epigenetic genotype for a gene even raises the potential for intragenic epistasis whereby the DNA and epigenetic genotypes for a gene interact to determine the phenotypic outcome. The simplest example would be where the DNA polymorphism toggles between a functional and non-functional allele and the epigenetic polymorphism toggles a silent and expressed gene. In this case, the plant would need the DNA to be in a functional state and the epigenetic state to be expressed before the gene was functional. This dual genotypic state at each gene will require new computational algorithms to fully interrogate.

**How is conditional genetic variation generated?** A major confounder of natural genetic variation studies is that the genetic loci are often conditioned upon the environment the organism is grown in (genotype × environment), the tissue wherein the phenotype is measured (genotype × development) and the age of the organism (genotype × ontogeny). These three topics, G×E, G×D and G×O, are fundamental to genetics studies yet often considered separately. However, it is possible to argue that all three topics are related by the simple supposition that most G× interactions are due to different regulatory networks and different structural genes interacting to control the generation of a phenotype under any given specific condition. This leads to two predominant hypotheses about what generates conditional genetics; either this variation lies in the regulatory networks controlling the genes directly causing a phenotype (i.e. enzymes or structural genes) or there is conditional genetic variation in the enzyme or structural genes. As both regulation and the fate of genome duplications are key to understanding biology the presence of G×E, G×D and G×O in natural genetic variation establishes this system as presenting
a unique opportunity to study the systems biology of the underlying networks and their evolution.

The regulatory hypothesis of G× argues that because genes are regulated in response to the environment, development or ontogeny any natural variation within these regulatory networks should be conditionally detected. Flowering time and light perception is one case where there is extensive G×E and the casual loci are regulatory proteins in the signaling network or even the photoreceptors themselves (Johanson et al., 2000; Caicedo et al., 2004; Filiault et al., 2008). The application of this finding to all phenotypes however may not be directly possible as, flowering time is a higher-order phenotype whereby a simple number (days to flower) incorporates data about broad swathes of plant physiology. Thus, it is formally possible that “simpler” phenotypes, ones likely to incorporate less diverse information such as secondary metabolism, may be different in their conditional genetics. However, in Arabidopsis glucosinolate secondary metabolism it was possible to link G×O variation to regulatory networks that controlled the differential expression across plant development of the two causal enzyme genes (Wentzell et al., 2008). Interestingly, further support for the regulatory hypothesis of G× came from a study of expression QTL in Barley whereby the authors were able to link potentially regulatory network variation to pathogen resistance QTL (Druka et al., 2008). Strikingly, the authors were able to show that they could measure variation in the regulatory network in the absence of the pathogen and that the G×E dependency of the resistance QTL was solely dependent upon the presence of the pathogen. This raises the possibility of querying network eQTL under control conditions in an attempt to directly predict G×E, G×D and G×O (Kliebenstein et al., 2006).
The other hypothesis for G× argues that conditionally detected polymorphisms lie within the genes directly causing a phenotype such as structural proteins and or enzymes rather than in the regulatory network. This hypothesis has often been supported by two lines of argument, the first is that regulatory networks are more critical to a plant's survival than individual downstream genes and as such more likely to be conserved. As shown above and in other studies there is ample evidence for natural genetic variation within regulatory networks and as such the network conservation argument does not appear valid (Keurentjes et al., 2007; Van Leeuwen et al., 2007). The stronger support for the structural gene explanation of G× comes about from the observation that structural and enzyme genes are often present in large gene families generated from whole genome duplications. The duplicated genes can then undergo expression sub-functionalization whereby one of the duplicated copies evolves to express under one condition (environment, tissue or ontogenic stage) and the other copy evolves to a different expression pattern over the same conditions (Freeling, 2009). As such, genetic variation in the genes would only be detected to influence the phenotype under the conditions when that specific gene is expressed leading to G×E, G×D and G×O.

The data supporting the concept of expression sub-functionalization between different tissues or environments is well established in a number of systems (Li et al., 2005; Casneuf et al., 2006; Kliebenstein, 2008; Zou et al., 2009). Additional studies are also beginning to show that sub-functionalization of gene copy expression can form a basis for the formation of QTLs at a genomic level in both Arabidopsis and Drosophila (Gu et al., 2004; Kliebenstein, 2008). More direct experimental support was found when tissue specific QTLs controlling phosphoglucomutase activity were linked to phosphoglucomutase paralogs with tissue specific expression patterns that mirrored the tissue specific QTL (Sergeeva et al., 2004). Additional
experimental support came from Arabidopsis glucosinolate metabolism whereby tissue specific QTLs for two different biochemical steps were shown to be due to developmental variation in expression patterns across the specific enzymes gene family (Lambrix et al., 2001; Li et al., 2008; Burow et al., 2009). While the current experimental support has largely been generated in G×D systems, genomics data has also suggested that G×E interactions for gene expression display a bias towards paralogous gene pairs (Landry et al., 2006). Intriguingly this duplicate gene concept of G×E, G×D or G×O QTLs suggests that it may be possible to rapidly identify candidate genes for these conditional QTLs by simply looking for whole genome duplication regions underlying multiple QTLs for a single phenotype. (Kliebenstein, 2009). Current experimental data supports a role for both regulatory and gene duplication in the generation of G×E, G×D or G×O QTLs and it will be interesting in the future once more causal loci are identified to begin testing what is the relative fraction of the two and if they differ based on biological differences between phenotypes.

Why answer the above questions? In the past decade, excitement within fields studying natural variation has partially shifted from the study of structured populations to the promise of genome wide association (GWA) mapping. GWA identifies association between phenotypes and genotypes using a sample of individuals from a species that are genotyped and phenotyped (Nordborg et al., 2002; Borevitz et al., 2007; Atwell et al., 2009). GWA with a large number of phenotypes was conducted within Arabidopsis thaliana and this showed that the number of genes controlling a phenotype may vary from trait to trait. However, one complexity in this analysis is determining how to conduct the statistics because the algorithms used make basic assumptions, such as normality of phenotypes, independence of traits and, typically, a finite
locus assumption. Knowing the shape of pleiotropic linkages between traits and the number of genes underlying this variation could allow these empirical estimates to be built into the GWA algorithms, potentially providing a dramatic increase in the accuracy of the observed genotype-to-phenotype linkages predicted. This would then allow for more rapid and correct gene identification, enhancing our ability to understand biology at a systematic level. Additionally, if epigenetic variation controls a significant fraction of heritability, then the individuals utilized will have to be genotyped at both the DNA and Epigenetic levels and the algorithms rewritten to incorporate this information. And finally, depending upon how conditional natural genetic variation is to the specific experimental design any association mapping experiment will only provide information about that specific condition. As such, there is likely to be a positive feedback loop of productivity linking systems biology analysis of structured and association mapping populations. In this loop, the structured populations may provide more solid empirical data to address the above fundamental questions and this can then be used to improve GWA’s ability to find the actual causal genes for natural phenotypic variation.

**Future Directions:** This complementary paring of systems biology and natural variation analysis will continue far into the future. It is likely, however, that it will bifurcate; however, with one path improving techniques that enable systems biology to more quickly identify the causal genes underpinning natural phenotypic variation. The second path will likely continue to develop bigger, faster, more powerful datasets to better enable us to measure the network underpinnings of natural variation and quantitative genetics. While these may appear separate paths, they actually will uncover significant observations that will benefit progress along both pathways,
each with the ultimate goal of being able to definitively state which genes control natural variation in a phenotype, how they control the variation and why the variation exists.

Acknowledgements: I would like to apologize to those authors whose citations I was unable to include because of size constraints.

References:


