Abstract
Genetic variation in a population creates an impressive spectrum of phenotypic diversity, particularly when changes in diet or the environment are imposed on the population. Genome-wide association studies have become a powerful tool for linking sequence variants with overlying systems level phenotypes, but they do not provide insight into the mechanisms through which genetic variation drives phenotypic variation. Systems genetics is an emerging discipline that provides a means to fill this knowledge gap by assembling the hierarchy of interactions among genes, proteins, and other intermediate phenotypes that manifest as phenotypic variation. When applied to nutrition, systems genetics enables the identification of pathways through which nutrients influence health and the determination of the mechanisms that cause individuals to differ in their response to diet. J. Nutr. 141: 515–519, 2011.

Fundamentals of systems genetics
Systems biology represents a paradigm shift in the study of health and disease that is distinct from, but complementary to, reductionist approaches that predominated in the past few decades as individual genes were cloned, sequenced, and characterized. Systems biology is anchored in the concept that traits are emergent properties of complex interactions among genes, proteins, cells, and tissues. Just as Galen and other pioneers of medical research attempted to understand the body based on physical interconnections between its components, systems biology constructs a systems level view of the organism by assembling the interconnections between traits measured at multiple levels of complexity and scale using a variety of statistical and computational tools (1).

Systems genetics is a specialized version of systems biology that can most succinctly be described as systems biology of populations. Populations, whether “natural” or laboratory derived, consist of individuals in which genetic variants at multiple loci across the genome create parallel variation in overlying phenotypes, akin to a multifactorial experimental model (2). Phenotypic variation can be viewed as a hierarchy of interactions among sets of genes, proteins, metabolites, cells, and tissues anchored to the underlying genetic variation (Fig. 1). The goal of systems genetics is to capture these interactions, while simultaneously identifying genetic variants that alter the overlying phenotypes of interest. Systems genetics thus uncovers the architecture of traits, providing mechanistic insight into the pathways that lead to phenotypic variation.

Practice of systems genetics
Fundamentally, systems genetics is practiced by phenotyping a population at multiple levels of scale, from molecular through higher order traits, to produce the building blocks with which trait architecture can be assembled. Molecular phenotyping typically includes microarrays, which support the identification of gene-phenotype networks and mapping of variants that regulate networks using approaches such as expression quantitative trait loci (QTL) (3). Considering the large number of traits profiled in a single microarray hybridization combined with the many other endpoints measured, all of which are performed across a population of individuals, data density is an inherent feature of systems genetics studies. Therefore, it is a discipline that is interdisciplinary by nature, requiring extensive collaborations among biologists, statisticians, and computational biologists.

Relationships among traits are extracted using a variety of computational approaches, all of which begin with some measure of pairwise correlation between traits (4). Typically, the process of assembling phenotypic networks begins at the level of
gene expression, where sets of transcripts with similar patterns of expression across the population are extracted from large scale gene expression data. The rationale behind coexpression networks is that genes encoding proteins that function in the same pathway will display coordinated expression across the population to the extent that they are regulated at the level of mRNA abundance (5). Progressing from the very large correlation matrix created from microarray data to identification of smaller sets of coexpressed genes requires some level of thresholding, i.e. selecting a correlation value above which relationships are considered meaningful (6). After thresholding, a variety of methods are used to identify putative coexpression networks and link them to higher order physiological traits. Graph algorithms, which represent transcripts as nodes and the correlations between transcripts as edges, are widely used to represent the interactions between genes after thresholding (7). Graphs can be weighted, in which edges retain information about the magnitude of correlation between transcripts, or unweighted, with all edges treated equally. Weighted gene coexpression network analysis (WGCNA) builds graphs that take into account the correlation value and weight the resultant network graphs in favor of high correlations (8). The resultant modules of interconnected genes are dense subgraphs in which many but not all nodes are interconnected. Clique extraction is an alternative approach that extracts perfectly interconnected sets of transcripts from the larger unweighted graph (9). Clique algorithms return sets of transcripts that are often smaller than modules from WGCNA but are also not disjoint, allowing transcripts to be assigned to multiple sets of interacting partners. Whether cliques, modules, or some other means to extract putative gene networks is used, functional annotation based on Gene Ontology (GO) enrichment, pathway mapping, and literature mining are used to identify dense subgraphs that are enriched in functionally related transcripts. Web-based tools such as the Database for Annotation, Visualization and Integrated Discovery are valuable resources for this step (10). In some cases, the relationship between functional enrichment is obvious (e.g. GO enrichment for cell cycle genes in a cancer study), whereas in others functional enrichment may highlight pathways of interest that, a priori, would not have been linked to the systems level traits under study (11). In addition to building the systems trait architecture, modules or cliques are informative in and of themselves. For example, putative functions of un-annotated genes can be inferred from the genes with which they show correlated expression, based on the concept of “guilt by association” (12).

Correlation-based methods are also used to associate modules with overlying higher order traits. If expression of a set of genes drives trait variation, it is reasonable to expect that the transcripts themselves would show correlation with the trait. Transcripts meeting this criterion are referred to as quantitative trait transcripts (QTT) and are valuable in prioritizing modules for further study (13). Jumbo-Luciano et al. (14) implemented QTT in a systems genetics study of body mass and composition traits in a panel of Drosophila melanogaster inbred lines. QTT for body weight, glycogen, glycerol, and TG content were assembled into modules using WGCNA based on intercorrelations among genes. This approach identified a number of significant modules for each trait, some of which were enriched in expected GO functions whereas others represented novel associations between gene networks and traits. Plaiser et al. (15) took a different approach using WGCNA to identify 2 coexpression modules consisting of hundreds of genes associated with genetic variation in plasma TG levels. Gene sets were mapped to phenotypes by first using principal component analysis to collapse modules into eigengenes that represented the most tightly correlated genes within a module. The relationships between eigengenes, rather than individual transcripts, and systems level phenotypes were evaluated using regression modeling. Using this approach, 2 eigengenes representing expression levels of several hundred genes explained 30% of variation in plasma TG level. Classical QTL mapping can also aid in linking gene sets to overlying phenotypes. Transcripts can be treated as quantitative traits to identify eQTL for each transcript within a module; overlap between the position of eQTL and the QTL for overlying systems traits would suggest a potential causal relationship between genes within the module and the trait (3). QTL mapping can also be used to infer causality between gene networks and higher order traits. For example, Chesler et al., applied a clique-based algorithm to brain expression data from a panel of BXD (C57BL/6J × DBA/2J) recombinant inbred (RI) strains to identify a set of 193 transcripts that were tightly intercorrelated with each other and with midbrain iron levels and several measures of locomotor activity (16). Multilocus QTL mapping determined that expression levels of many of the genes within this network were linked to a common genomic location, suggesting potential causal flow from genetic variation at this locus through the set of interaction transcripts to the higher order brain and behavior phenotypes. In practice, module-phenotype associations often are based on convergent evidence from combinations of approaches. For example, we recently identified acid phosphatase 1 (Acph1) as an important determinant of genetic variation in T lymphocyte subpopulations in the BXD strain panel based on its correlation with the overlying phenotypes, its position within the QTL for the trait (CD4+:CD8+ ratio in peripheral blood), and its membership in a large clique highly enriched in functions related to cell cycle (17).

Network-based approaches that are central to systems genetics are also ideal for determining mechanisms through which environmental variables such as nutrients affect the system across a population. The general approach described above is extensible to differential analyses through which networks and relationships that exist only under a specific condition can be identified and the “differential concept” can be applied at multiple levels of the network (18). For example, we used the concepts of differential correlation and differential
network analyses to identify sets of genes and key hubs within those networks that were specifically responsive to an environmental stressor (low-dose ionizing radiation exposure) (18). Differential correlation highlighted a network of radiation-specific interactions that were centered upon the hub gene topoisomerase 3a (Top3a), which is important for genomic stability and regulation of cell cycle checkpoint control after radiation exposure (19). The same concepts can be used to assemble nutrient-specific interaction networks or to compare and contrast the architecture through which nutrients differentially affect health in males and females or between different ethnic groups within a population.

Population-based models for systems genetics

Population-based models are central to systems genetics. The core criterion is a set of individuals or inbred lines in which each individual or line consists of its own unique set of allele combinations. Based on this criterion, a traditional F2 mapping population is sufficient. However, the limitation in using such a model lies in the need for extensive phenotyping and the fact that the same set of unique individuals cannot be recreated. Genetic reference populations overcome this limitation, because they consist of populations of individuals that can be reproduced indefinitely due to the fact that strains comprising the population are inbred. Because they are genetically stable, aside from random genetic drift over time, data from multiple studies with the same population can be integrated as if they were collected from a single set of individual mice. A number of mouse genetic reference populations exist, including the set of standard inbred strains central to the Mouse Phenome Database and RI strain sets created by intermixing 2 parental genomes and then fixing recombinations by inbreeding (20). Inbred mouse strains provide a spectrum of diversity for most phenotypes. However, many of the strains share overlapping breeding history and at the genomic level are identical by descent across a large percentage of the genome (21). As a result, these loci are effectively blind spots in terms of genetic mapping (22).

RI strain panels are derived from 2 parental genomes, but the random nature of meiotic recombination creates the opportunity for new allelic combinations to occur prior to fixation by inbreeding. Each individual strain represents a genetic mosaic of the 2 original parental genomes. However, if the 2 parental strains shared large regions of identical by descent loci, these regions will remain blind spots in the resulting inbred strains (22). The largest existing RI panel was created from the C3BL/6J and DBA/2J genomes and consists of 81 BXD strains that have been used to study a wide range of phenotypes (23). Use of the BXD strain set, as well as other RI strain panels, for systems genetics is facilitated by GeneNetwork. GeneNetwork is a Web-based resource that houses both genotype and phenotype data for a number of genetic reference populations and a suite of interconnected analysis tools that support QTL mapping and trait correlations across multiple studies (24).

All of the existing RI strain panels in rodents were created from 2 parental genomes and consist of at most 81 strains; both the limited set of input alleles and the numbers of “individuals” within the panels limit the power to make significant genetic associations, particularly for small effect allele combinations. The Collaborative Cross (CC) was first proposed at the Edinburgh meeting of the International Mouse Genome Conference in October of 2001 as a genetic reference population that would overcome both the genetic and statistical limitations of existing RI strain panels. The CC was conceived as a population comprised of a much larger number (1000 or more) of RI strains created from 8 rather than 2 parental genomes (25). The 8 parental strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ, CAST/EiJ, PWK/Ph, and WSB/EiJ) were selected based on predicted genetic diversity and included 3 wild-derived strains (CAST/EiJ, PWK/Ph, and WSB/EiJ) as well as strains prone to type I (NOD/ShiLtJ) and type II (NZO/HILtJ) diabetes. These 8 strains capture at least 90% of the known allelic diversity across the mouse genome (22).

Breeding of the CC in the US began at Oak Ridge National Laboratory and internationally at Tel Aviv University and the Western Australia Institute for Medical Research (26–28). The U.S. population has since been transferred to the University of North Carolina, Chapel Hill, where inbreeding will be completed and lines will be maintained and distributed for use by the systems genetics community. Phenotypic diversity in the CC promises to be broad, as expected. For example, interim generations of incipient CC lines were phenotyped for adiposity and other traits as inbreeding progressed at Oak Ridge National Laboratory. Adiposity in males and females from 59 CC lines ranged by >18-fold in males and 44-fold in females after at least 7 generations of inbreeding (Fig. 2).

Because genetic reference populations are stable over time and space, they enable the true potential of systems genetics in that data from any studies using the same population can be integrated, maximizing the discovery nature of the discipline. Lynch et al. (17) used the BXD RI panel to identify mechanisms of genetic sensitivity to environmental stressors. By integrating peripheral blood immunophenoype data from our studies with outcomes of Chlamydia psittaci infection produced by Miyairi et al. (29) using the same BXD strains, we uncovered a significant predictive relationship between the ratio of T and B lymphocytes in healthy individuals and pathogen load in infected individuals. In the absence of a priori knowledge,
such relationships would likely go undiscovered in the absence of a systems genetics framework. The same rationale is equally capable of uncovering mechanistic relationships between traits relevant to nutrition, such as the overlapping networks of interactions that make individuals susceptible to both diet-induced obesity and insulin resistance. Obesity is, in fact, a compelling model for the use of systems genetic due to the now established role of adipose tissue as much more than an energy storage depot and the recognition that changes in adipose mass affect many other parts of the system. Obesity and metabolic syndrome are arguably the most well-established pair of comorbid conditions related to nutrition. However, despite this well-established relationship, it is estimated that up to 25% of obese individuals are metabolically healthy (30). Conversely, there exist groups of individuals who are metabolically obese despite a lean phenotype (31). Uncoupling these conditions at the genetic level while defining the pathways that lead to each disorder could both identify new therapeutic targets and improve clinical understanding of personalized risk for patients presenting with obesity. More broadly, systems genetic would provide a means to uncover the genetic and functional correlations through which obesity increases the risk for a range of disparate disorders such as asthma, psoriasis, periodontal disease, and a number of cancers.

**Systems genetics for nutrition research**

Systems genetics is an ideal framework for nutrition research. It is widely accepted that individuals differ significantly in their metabolism of nutrients. Ferrara et al. (32) provided a compelling example of using systems genetics to uncover the mechanisms for differences in metabolism in an F2 cross between diabetes-resistant and susceptible inbred strains of mice. Quantitative metabolomic and microarray data from liver were integrated with QTL mapping to construct a causal network that mediated heritable differences in glutamate metabolism. Importantly, these authors also validated the relationship of their network to the linked metabolic traits by demonstrating that glutamate availability altered expression levels of genes in the network. Ultimately, this level of understanding across a population would support implementation of personalized nutrition. This work also raises the important issue of how to validate that networks uncovered by systems genetics are causal in development of phenotypes. In the case of Ferrara et al. (32), the authors manipulated glutamate levels in an ex vivo hepatocyte culture model to show that genes identified in the network were responsive to glutamate levels. In some cases, genetic manipulation of individual genes, such as hub genes central to networks, may be sufficient to alter the associated phenotype. Schadt et al. and Yang et al. (33,34) successfully used transgenic and knockout mouse models to 8 genes (of 9 tested) predicted to be causal for obesity in an F2 cross between C57BL/6J and DBA/2J strains of mouse models to 8 genes (of 9 tested) predicted to be causal for obesity in an F2 cross between C57BL/6J and DBA/2J strains of mouse. However, functional redundancy between related genes may dampen the results of manipulating any single gene. Further, most higher order traits are by nature complex, with many loci contributing modest effects to the resultant phenotype. The CC should provide additional means to test systems genetics predictions because of its size (~1000 lines or more) (25). Causal models can be developed in a set of randomly chosen strains and validated in an independent set of strains.

**Summary**

In summary, systems genetics is an emerging discipline that is orthogonal but complementary to traditional methods that emphasize the function of single genes or proteins. While it supports hypothesis testing, it also enables discovery of previously unsuspected relationships among genes, proteins, metabolites, biochemical pathways, and overlying traits relative to health and disease. As genetic reference populations become more widely used, novel relationships will continue to be discovered, prompting new hypotheses and alternative models. The CC, in particular, should accelerate both the rate of such developments and the adoption of systems genetics by the nutrition community.

**Acknowledgment**

The sole author had responsibility for all parts of the manuscript.

**Literature Cited**

1. Siegel RE. Galen’s system of physiology and medicine: an analysis of his doctrines and observations on blood flow, respiration, humors and internal diseases. S. Karger; 1968.

**Downloaded from** https://academic.oup.com/jn/article-pdf/141/3/515/4743605 by guest on 24 November 2018


