Epistasis and Its Implications for Personal Genetics

Jason H. Moore[1,2,3,4,5,*] and Scott M. Williams[6]

The widespread availability of high-throughput genotyping technology has opened the door to the era of personal genetics, which brings to consumers the promise of using genetic variations to predict individual susceptibility to common diseases. Despite easy access to commercial personal genetics services, our knowledge of the genetic architecture of common diseases is still very limited and has not yet fulfilled the promise of accurately predicting most people at risk. This is partly because of the complexity of the mapping relationship between genotype and phenotype that is a consequence of epistasis (gene-gene interaction) and other phenomena such as gene-environment interaction and locus heterogeneity. Unfortunately, these aspects of genetic architecture have not been addressed in most of the genetic association studies that provide the knowledge base for interpreting large-scale genetic association results. We provide here an introductory review of how epistasis can affect human health and disease and how it can be detected in population-based studies. We provide some thoughts on the implications of epistasis for personal genetics and some recommendations for improving personal genetics in light of this complexity.

Introduction

The discovery and characterization of *BRCA1* (MIM 113705) and *BRCA2* (MIM 600185) and their specific mutations as significant risk factors for familial breast cancer (MIM 114480) ushered in the era of commercial genetic testing.[1] There is no question that our ability to test for the presence of mutations in these two genes plays some role in understanding and preventing this form of cancer. However, much of familial breast cancer remains unexplained by *BRCA1* and *BRCA2*, and the elusive *BRCA3* has yet to be identified.[2] It is entirely possible that the remaining genetic risk factors for familial breast cancer are a combination of rare variants with intermediate penetrance and common variants, such as SNPs, that have low penetrance. However, the common disease-common variant (CDCV) model has thus far failed to uncover new variants that explain a large fraction of the genetic risk. The data, as reviewed by Ripperger et al., indicate that there is currently no evidence that genetic testing for variants of low penetrance is useful for predicting risk.[2] Therefore, before meaningful new genetic testing services can be offered, we must substantially improve our understanding of the genetic architecture of familial breast cancer, where genetic architecture is defined as (1) the set of genes and DNA sequence involved in the disease, (2) their variation in the population, and (3) their specific effects on the phenotype.[3] We argue here, based on the emerging data and analyses, that elucidating the genetic architecture of breast cancer and comparable diseases must focus on underlying complexity.

The current strategy for revealing genetic architecture is to carry out a genome-wide association study (GWAS) with a million or more SNPs or other variants that capture much of the common variation in the human genome by tagging blocks of variants that are in linkage disequilibrium.[4,5] This approach is based on the hypothesis that scanning the entire genome for single SNP associations in an unbiased or agnostic manner that ignores what we know about disease pathobiology will reveal much of the unexplained genetic architecture of a particular disease. The prevalent analytical strategy of searching for strong single SNP effects without regard to the rest of the genome or exposure was initially developed for diseases with few known etiologic factors. This approach has been applied universally to all GWAS analyses, producing deceptive results because of confounding, as occurs with smoking and lung cancer.[6]

Despite the promise of this technology and the time and financial resources already expended, the results have been generally underwhelming in terms of elucidating the genetic architecture of common complex disease and explaining a majority of the genetic risks. Consider, for example, the application of GWAS for identifying cancer-susceptibility genes. A recent review of these studies shows that a number of new susceptibility loci have been identified for several types of cancer, including breast, prostate, colorectal, lung, and skin.[7] The identification of new associations is certainly important. However, as Easton and Eeles note, the increase in risk for the susceptibility alleles at each of these loci is generally 1.3-fold or less.[7] For familial breast cancer, Easton et al. reported five significant replicated associations that were identified by GWAS in a three-stage study design.[8] Four of these variants were in known genes, and one was located in a hypothetical gene. Assuming a multiplicative model, these five loci combine to explain only 3.6% of the excess familial risk of breast cancer and, as suggested by Ripperger et al.,[2] were not deemed to be suitable for genetic testing as a result of their small effect sizes.[8] In a recent follow-up study with

[1]Computational Genetics Laboratory, Department of Genetics and Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH 03756, USA; [2]Department of Computer Science, University of New Hampshire, Durham, NH 03824-2619, USA; [3]Department of Computer Science, University of Vermont, Burlington, VT 05405, USA; [4]Translational Genomics Research Institute, Phoenix, AZ 85004, USA; [5]Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University, Providence, RI 02912, USA; [6]Center for Human Genetics Research and Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37232, USA

*Correspondence: jason.h.moore@dartmouth.edu

DOI 10.1016/j.ajhg.2009.08.006. ©2009 by The American Society of Human Genetics. All rights reserved.

two additional stages of testing and replication, two additional susceptibility loci were identified with odds ratios of 1.11 and 0.95, respectively, each accounting for much less than 1% of the familial risk of breast cancer.[9] When combined with the previous known genetic risk factors for familial breast cancer, the estimated fraction of risk explained is approximately 5.9%. This is in stark contrast to *BRCA1* and *BRCA2* mutations, which account for between 20% and 40% of familial breast cancer. Although the application of GWAS to familial breast cancer has generated new knowledge, it has not resulted in new genetic tests that can be used to predict and prevent familial breast cancer. These results are discouraging for more common diseases such as sporadic breast cancer and type 2 diabetes that are likely to have a much more complex genetic architecture. As Clark et al. predicted, our success with GWAS depends critically on the assumptions we make about disease complexity.[10]

The limits of GWAS, as revealed through the study of familial breast cancer, do not represent isolated examples. In fact, very few SNPs with odds ratios above 1.5 have been discovered and replicated for any common human disease, suggesting that their use in genetic testing will be limited. This limitation was pointed out in a recent study by Jakobsdottir et al. showing that SNPs identified by GWAS for a variety of diseases make poor classifiers of disease, thus calling into question their usefulness for risk assessment by genetic testing.[11] The same conclusions have been presented by Kraft et al.[12] Despite these cautions, commercial genotyping is currently being offered directly to the consumer at affordable prices, and, although there are appropriate disclaimers, it is obvious that the availability of cheap genetic testing is fostering the perception that the era of personal genetics is upon us. However, it is important to note that the ability to inexpensively measure one million or more SNPs in an individual's genome does not, in the absence of accurate genotype-phenotype maps, provide clinically relevant information in most situations. Nonetheless, several commercial direct-to-consumer genetic testing services are now available for less than $1000, and in some cases less than $500.[13]

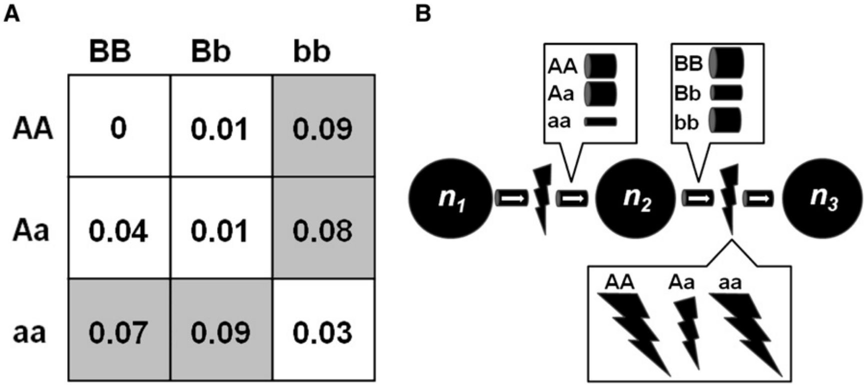
Barring future regulation, it appears as though personal genetics is here to stay. Given this reality, it is important to assess the impact that genetic architecture will have on the utility of the results being provided to the consumer. It is our working hypothesis that epistasis, or gene-gene interaction, plays an important role in the genetic architecture of common diseases and thus must be characterized if personal genetics is to have an impact on the health of the consumer. We provide here an introduction to epistasis and a theory for why it is ubiquitous in human biology. We then provide an overview of the analytical tools that are necessary to detect and characterize epistasis in genetic association studies. Finally, we provide a discussion of the implications of epistasis for personal genetics and then provide some recommendations for how to move forward under the assumption that the genetic architecture of common disease is extremely complex.

What Is Epistasis?

William Bateson, who first coined the term "genetics" (see historical account by Patrick Bateson[14]), also coined the word "epistasis" in the early 1900s to explain deviations from Mendelian inheritance.[15] The term "epistasis" literally means "standing upon," and Bateson used it to describe characters that were layered on top of other characters, thereby masking their expression. The *epi*static characters had to be removed before the underlying *hypo*static characters could be revealed. The commonly used definition of epistasis, an allele at one locus masking the expression of an allele at another locus, reflects this original definition. As reviewed recently by Tyler et al., eye color determination in *Drosophila* provides a classic example.[16] The genes *scarlet*, *brown*, and *white* play major roles in a simplified model of *Drosophila* eye pigmentation. Eye pigmentation in *Drosophila* requires the synthesis and deposition of both drosopterins, red pigments synthesized from GTP, and ommochromes, brown pigments synthesized from tryptophan. A mutation in *brown* prevents production of the bright red pigment, resulting in a fly with brown eyes, and a mutation in *scarlet* prevents production of the brown pigment, resulting in a fly with bright red eyes. In a fly with a mutation in the *white* gene, neither pigment can be produced, and the fly will have white eyes regardless of the genotype at the *brown* or *scarlet* loci. In this example, the *white* gene is epistatic to *brown* and *scarlet*: a mutant genotype at the *white* locus masks the genotypes at the other loci.

Since Bateson, there have been many different and evolving definitions of epistasis or gene-gene interaction.[16–29] For example, Fisher defined epistasis in a statistical manner as an explanation for deviation from additivity in a linear model.[30] This nonadditivity of genetic effects measured mathematically is different from Bateson's more biological definition of epistasis. We have previously made the distinction between Bateson's biological epistasis and Fisher's statistical epistasis.[29] This distinction is important to keep in mind when thinking about the genetic architecture of common human diseases because biological epistasis happens at the cellular level in an individual whereas statistical epistasis is a pattern of genotype-to-phenotype relationships that results from genetic variation in a human population. This distinction becomes important when attempting to draw a biological conclusion from a statistical model that describes a genetic association. Moore and Williams[29] and Phillips[22] have discussed the idea that more modern definitions of epistasis may be needed in light of our new knowledge about gene networks and biological systems. However, the classic definitions provided by Bateson and Fisher still provide a good starting point for thinking about gene-gene interactions.[15,30]

To illustrate the concept of statistical interaction, consider the following simple example of epistasis in the



	BB	Bb	bb
AA	0	0.01	0.09
Aa	0.04	0.01	0.08
aa	0.07	0.09	0.03

Figure 1. A Simple Biochemical Systems Model that Is Consistent with a Complex Genetic Model

(A) Penetrance function showing an exclusive OR (XOR) pattern of high-risk (shaded) and low-risk (unshaded) genotype combinations for two biallelic SNPs.

(B) A Petri net model of a biochemical system under the control of the two SNPs from the genetic model in (A). SNP_A controls the diameter of the "arc" or "pipe" carrying molecules of type 1, which are converted to molecules of type 2 at a constant rate governed by the first "transition" (lightning) on the left (wider pipes = larger diameter). SNP_A is pleiotropic and also controls the rate at which molecules of type 2 are converted to molecules of type 3 (wider lightning bolts = faster rate). SNP_B controls the arc or pipe carrying molecules of type 2 to the second transition, which converts them to molecules of type 3. When executed as part of a threshold model, the output of this system matches the distribution of high-risk and low-risk genotypes. This Petri net model demonstrates that a simple biochemical systems model can underlie a nonlinear genetic model.

form of a penetrance function. Penetrance is simply the probability (P) of disease (D) given a particular combination of genotypes (G) that was inherited (i.e., P[D|G]). Let us assume for two SNPs labeled *A* and *B* that genotypes *AA*, *aa*, *BB*, and *bb* have population frequencies of 0.25 whereas genotypes *Aa* and *Bb* have frequencies of 0.5. Let us also assume that individuals have a very high risk of disease if they inherit *Aa* or *Bb* but not both (i.e., the exclusive OR [XOR] logic function). What makes this model interesting is that disease risk is entirely dependent on the particular combination of genotypes inherited at more than one locus. The penetrances for each individual genotype in this model are all the same and are computed by summing the products of the genotype frequencies and penetrance values. Heritability can be calculated as outlined by Culverhouse et al.[31] Thus, in this model, there is no difference in disease risk for each single-locus genotype as specified by penetrance values. This model is labeled M170 by Li and Reich in their categorization of genetic models involving two SNPs and is an example of a pattern that is not separable by a simple linear function.[32] This model is a special case in which all of the heritability results from epistasis or nonlinear gene-gene interaction.

Although highly illustrative, the XOR model and others like it are often criticized for lack of biological plausibility. For example, the XOR model does not fit with Mendelian concepts of epistasis that are based on interactions between SNPs with recessive and dominant effects. There are two important points to keep in mind when embracing a complex view of genetic architecture: first, we do not yet know what a plausible epistasis model is, because we have yet to systematically evaluate nonlinear genetic models of human disease, and second, we have not yet begun to validate these models in experimental systems. Therefore, our knowledge of the diversity of genetic models underlying common diseases is in its infancy. However, we can begin to think about biological plausibility via computational thought experiments.[33] To this end, Moore and Hahn developed a computational system for discovering systems biology models that are consistent with epistasis models such as XOR.[34] Here, Petri nets were used as a discrete dynamic system modeling tool (see Moore and Hahn for details[34]). Figure 1A shows a variation on an XOR-based penetrance function in which *aa* or *bb* genotypes are high risk, but not in the presence of the nonhomologous homozygote. Figure 1B shows an example Petri net model that is consistent with the pattern of high-risk and low-risk cells in the XOR-based penetrance function. The importance of this result is that it shows how a simple discrete biochemical systems model can account for the nonlinear pattern observed in the XOR model. Although not an actual biological result, this thought experiment shows that a simple system can generate a complex genetic architecture that is not predicted by a one-SNP-at-a-time analytical approach. As reviewed by Moore and Williams, this represents a first step toward making the connection between biological and statistical epistasis.[29] Indeed, others have demonstrated biologically plausible models for transcriptional and biochemical networks that are consistent with the nonlinear XOR function.[35,36]

It is important to note that the data supporting epistasis in complex human diseases are emerging slowly. This is not surprising given that decisions to use models that do not incorporate complex interactions are based primarily on hypotheses of convenience and not on plausible biological phenomena that are inherently complex. Therefore, it is important to carefully consider biological plausibility in addition to analytical simplicity in designing analyses. The roles for experimental genetics and systems biology in constructing well-founded hypotheses of disease etiology are discussed below with this in mind.

Why Is Epistasis so Ubiquitous?

Moore[37] and Templeton,[24] for example, have argued that epistasis is likely to be a ubiquitous component of the genetic architecture of common human diseases. There are several reasons for this. First, as noted above, epistasis is not a new idea and remains a common phenomenon in the biological literature. Second, the ubiquity of biomolecular interactions in gene regulation and biochemical and metabolic systems suggests that the relationship between DNA sequence variations and biological endpoints is likely to involve interactions of multiple gene products. Third, positive results from studies of single polymorphisms typically do not replicate across independent samples. Fourth, and perhaps most importantly, epistasis is commonly found when properly investigated. These four reasons suggest that epistasis may be ubiquitous in human biology but do not provide an explanation for why. For that, we turn to evolutionary biology for a theory that may provide a compelling mechanism for epistasis.

Canalization is an idea introduced by Waddington to explain the buffering of phenotypes to genetic and environmental perturbations.[38] Evolutionary biologists have described canalization as stabilizing selection that ensures that systems evolve to a robust level.[39] In other words, evolution seeks to keep our blood pressure, glucose levels, and other important physiological and metabolic systems in a healthy range while ensuring that these measures are resistant to most genetic and environmental perturbations. Deviations from these healthy ranges are often categorized as diseases such as hypertension and diabetes. One manner in which evolution has succeeded in developing robust systems is by evolving redundant gene networks that are resistant to fluctuations, both genetic and environmental. This might explain why epistasis is so ubiquitous within the context of human disease. What we observe as disease might be the result of the accumulation of multiple mutations in different parts of a gene network that are needed to perturb a robust system from its evolved range. This might explain why most single variants explain very little of the risk for any given common disease. If this is true, it is essential to look for combinations of genetic variations in human populations as a way to capture the patterns of variation across networks that are needed to move individuals into unhealthy or disease phenotypes such as hypertension. In essence, evolution moves a population to a state where the vast majority of people are healthy, and this is often accomplished through complex networks that involve substantial epistasis. Epistasis as a robust gene network phenomenon has recently been discussed by Tyler et al.[16]

Assuming canalization has shaped human biology throughout history, one might ask why we see independent main effects in genetic association studies at all. Gibson suggests that human migration and recent bottlenecks might allow hidden or cryptic genetic variation to emerge as genetic risk factors.[39] Our recent evolutionary history may explain why genetic architecture is likely to be a mix of different types of genetic effects including epistasis, gene-environment interactions, and locus heterogeneity. Unfortunately, canalization is very difficult to determine experimentally. Nevertheless, it provides an important foundation to begin thinking about why the genetic architecture of common diseases is so complex. Gibson offers a few strategies for identifying the hallmarks of canalization.[39]

The Challenges of Detecting Statistical Epistasis in Genetic Association Studies

As discussed above, one of the early definitions of epistasis was deviation from additivity in a linear model.[30] The linear model plays an important role in modern genetic epidemiology because it has a solid theoretical foundation, is easy to implement with a wide range of different software packages, and is easy to interpret. Despite these good reasons to use linear models,[27,28] they do have limitations for explaining genetic models of disease because they have limited ability to detect nonlinear patterns of interaction.[40] The first problem is that modeling interactions requires looking at combinations of variables. Considering multiple variables simultaneously is challenging because the available data get spread thinly across multiple combinations of genotypes. Estimation of parameters in a linear model can be very problematic when the data are sparse. The second problem is that linear models are often implemented such that interaction effects are only considered after significant independent main effects are identified. This certainly makes model fitting easier, but it assumes that the most important predictors will have main effects. For example, the focused interaction testing framework (FITF) approach of Millstein et al. provides a powerful logistic regression approach to detecting interactions but conditions on main effects.[41] Furthermore, it is well documented that linear models have greater power to detect main effects than interactions.[42–44] Therefore, in using linear models, we are constrained not by biological reality but by statistical tools that were not necessarily developed to test realistic biological models. As a field, genetic epidemiology has preferred Fisher's definition of epistasis to Bateson's, and this has led to analytical approaches that significantly hurt our ability to model real genetic architecture. In fact, the historical sidetracking of Bateson's biological epistasis for Fisher's statistical definition, which he called "epistacy," has been noted.[20] The limitations of the linear model and other parametric statistical approaches have motivated the development of computational approaches such as those from machine learning and data mining that make fewer assumptions about the functional form of the model and the effects being modeled.[45–47] Several recent reviews highlight the need for new methods[48] and discuss and compare different strategies for detecting statistical epistasis.[28,49] The methods reviewed by Cordell[28] include novel approaches such as combinatorial partitioning[50,51] and logic regression[52,53] and machine learning approaches such as random forests,[54,55] for example. We briefly review one of these novel methods, multifactor dimensionality reduction (MDR), in the next section.

In addition to the challenge of modeling nonlinear interactions, GWAS introduces important computational challenges. The detection of epistasis in the absence of significant main effects requires combinations of SNPs to be systematically evaluated. As summarized by Moore and Ritchie[56] and Moore,[57] combinatorial assessment of SNPs in a GWAS is not computationally feasible beyond exploring two-way and three-way combinations. As we will briefly describe in the next section, addressing this problem will require using prior statistical and biological knowledge, because there are not enough computers in the world for a brute-force approach.

Finally, perhaps the most important challenge we face in detecting and characterizing epistasis is interpretation. As discussed above, going from a population-level statistical summary of gene-gene interactions to inferences about the biological interactions occurring at the cellular level is a significant and difficult leap. Conversely, translating our knowledge of gene networks and cellular function at the individual level to predictions about public health is equally difficult. As discussed by Moore and Williams, systems biology holds the promise to help us traverse this conceptual and practical divide.[29]

A Multifactor Dimensionality Reduction Approach to Modeling Statistical Epistasis

Thornton-Wells et al. have suggested that we need an analytical retooling to address the etiological complexity of common human disease.[48] As such, several novel approaches have been developed that are designed specifically to tackle complex problems such as modeling epistasis. As reviewed recently by Cordell,[28] multifactor dimensionality reduction (MDR) has emerged as one important new method for detecting and characterizing patterns of statistical epistasis in genetic association studies that complements the linear modeling paradigm. MDR was developed as a nonparametric (i.e., no parameters are estimated) and genetic model-free (i.e., no genetic model is assumed) data mining and machine learning strategy for identifying combinations of discrete genetic and environmental factors that are predictive of a discrete clinical endpoint.[57–63] Unlike most other methods, MDR was designed to detect interactions in the absence of detectable main effects and thus complements statistical approaches, such as logistic regression, and machine learning methods, such as random forests and neural networks. At the heart of the MDR approach is a feature or attribute construction algorithm that creates a new variable or attribute by pooling genotypes from multiple SNPs. The general process of defining a new attribute as a function of two or more other attributes is referred to as constructive induction, or attribute construction, and was first described by Michalski.[64] Constructive induction, using the MDR kernel, is accomplished in the following manner. Given a threshold *T*, a multilocus genotype combination is considered high risk if the ratio of cases (subjects with disease) to controls (healthy subjects) exceeds *T*; otherwise, it is considered low risk. Genotype combinations considered to be high risk are labeled G_1, whereas those considered low risk are labeled G_0. This process constructs a new one-dimensional attribute with values of G_0 and G_1. It is this new single variable that is assessed, via any classification method. The MDR method is based on the idea that changing the representation space of the data will make it easier for methods such as logistic regression, classification trees, or a naive Bayes classifier to detect attribute dependencies. As such, MDR significantly complements other classification methods such as those reviewed by Hastie et al.[46] This method has been confirmed in numerous simulation studies, and a user-friendly open-source MDR software package written in Java is freely available.

Since its initial description by Ritchie et al.,[58] many modifications and extensions to MDR have been proposed. These include, for example, entropy-based interpretation methods,[63] the use of odds ratios,[65] log-linear methods,[66] generalized linear models,[67] methods for imbalanced data,[68] permutation testing methods,[69] methods for missing data,[70] and different evaluation metrics.[71–73] The MDR approach has also been successfully applied to a wide range of different genetic association studies. For example, Andrew et al.[74] used MDR to model the relationship between polymorphisms in DNA repair enzyme genes and susceptibility to bladder cancer (MIM 109800). A highly significant nonadditive interaction was found between two SNPs in the xeroderma pigmentosum group D (*XPD*) gene (MIM 278730) that was a better predictor of bladder cancer than smoking. These results were later replicated in independent studies from a consortium.[75]

As discussed above, the biggest challenge to implementing methods such as MDR in GWAS is the combinatorial explosion of SNP interactions. The focus of many current MDR studies is on scaling this approach to GWAS data. Other than faster computer hardware for MDR[76] or parallel implementations,[77] there are two general strategies that are being pursued,[57,78] each of which has advantages and disadvantages.[79] The first is a filter approach that preselects SNPs that are likely to interact prior to MDR analysis[63,80]. Machine learning methods based on the ReliefF algorithm[81] look promising as statistical filter approaches for GWAS.[82–84] The second approach is the use of stochastic search algorithms, such as those reviewed by Michalewicz and Fogel,[85] to guide an MDR analysis. Methods based on evolutionary computing algorithms that perform parallel stochastic searches across MDR models have been extensively explored.[57,78,86–91] The key to the success of these algorithms is the availability of either statistical or biological knowledge that can be used to prioritize certain SNPs in the search process.[78,91] Otherwise, the algorithm is searching for a genetic needle in an effectively infinite genomic haystack. For example, Moore and White[86,87] showed how preprocessed ReliefF scores can be used to provide good building blocks for an evolutionary computing algorithm, and Pattin et al.[92] reviewed a role for protein-protein interaction databases as a source of biological knowledge that could be used in the same manner.

Implications of Epistasis for Personal Genetics

The current personal genetics paradigm that is being marketed directly to the consumer is built on the results of genetic association studies that ignore the complexity of the genotype-to-phenotype mapping relationship that results from epistasis and other phenomena. Indeed, it is now apparent that single SNPs typically have very small effects on risk and are not useful for predicting risk.[11,12] This presents a significant problem for those hoping to capitalize on SNPs or other genetic variations as useful markers of health and disease. We propose here that the full utility of personal genetics will not be realized until the full complexity of genetic architecture is embraced rather than ignored. For this to become a reality, those conducting genetic association studies will need to rigorously test hypotheses about epistasis, gene-environment interaction, locus heterogeneity, etc. via analytical methods that are powered to detect these phenomena. These results need to be reported for every study, in addition to the standard analyses reporting independent SNP effects. This will of course not be easy for some of the reasons outlined above, but it is absolutely necessary if we expect to bring genetics to the consumer in a meaningful manner.

To illustrate this point, all one needs to do is consider a well-characterized family history, which remains the most powerful predictor we have about risk for common diseases. As Kardia et al. discuss, a carefully recorded family history of coronary heart disease (CHD) is a powerful indicator of future risk even after adjustment for the effects of traditional risk factors such as age, smoking, and body mass index.[93] Family history captures large amounts of genomic and environmental sharing among relatives, and it implicitly incorporates nonlinear aspects. Therefore, family history provides genome-wide and not gene-specific risk, thereby enabling a better model of risk. Given these facts, genetic data analyses using a linear model will never approach the simplicity and cost effectiveness of using family history to identify individuals at risk for CHD. This is particularly true if we assume that epistasis and gene-environment interactions play an important role in disease susceptibility, as is expected for CHD[94] and its risk factors.[95]

Consider type 2 diabetes, or T2D (MIM 125853), as an example. As with CHD, family history of T2D is a strong predictor of risk. However, as noted by Williams et al., the results of GWAS analysis of T2D have been mixed, and it does not appear that the known genetic risk factors yet approach the predictive power of family history.[96] To illustrate this point, the authors present in Figure 2 their consumer genetic testing results for T2D from one of the available services, 23andMe. Note that none of the polymorphisms reported have genotypes with maximum relative risk levels above 2. In fact, as noted earlier, most have a maximum relative risk level of less than 1.5. Overall, this combination results in very small increases or decreases in risk, as is seen for the results for each author. The progression from single SNPs as risk factors to combinations of SNPs acting epistatically to the entire genome as a risk factor was recently discussed by Moore, who proposed that our individual "genometype" may ultimately prove the most useful strategy for personal genetics.[97] This is consistent for the power of family history and, if true, suggests that we need to fundamentally change our approach to genetic association analysis if the results are to be useful for personal genetics and other endeavors in human genetics.

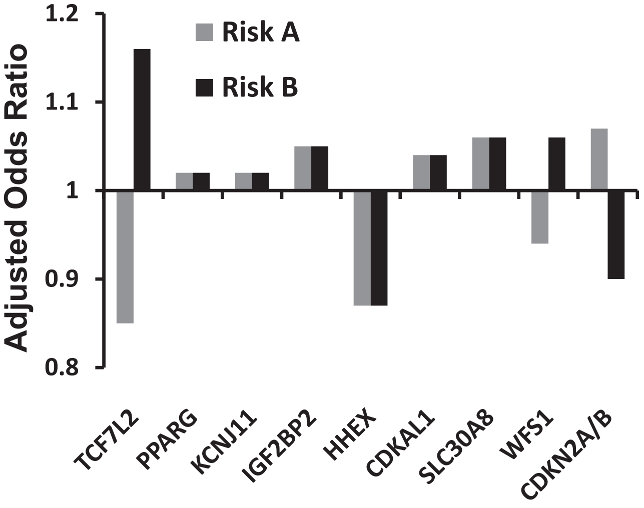


Figure 2. Panel of Genetic Markers for Type 2 Diabetes Provided by 23andMe for Each of the Authors
The profile of author A (gray bars) is associated with an overall slight decrease in risk under a multiplicative model, whereas the profile of author B (black bars) is associated with a slightly increased risk. Note that all adjusted odds ratios for individual genotypes are between 0.8 and 1.2.

Recommendations for Personal Genetics

We have presented here an introduction to epistasis, a theory for why epistasis is so common in human biology, a summary of the challenges in detecting and characterizing epistasis in genetic association studies, a summary of the MDR method for modeling epistasis, and a summary of the implications of epistasis for personal genetics. Below, we present five recommendations for what we can do to improve the usefulness of genetic association results for the providers and the consumers of personal genetics. We believe this should be a research priority given the central motivation in human genetics to deliver new knowledge that improves human health. Personal genetics may play an important role in the delivery of healthcare benefits by initially fostering a patient-physician dialog about results and perhaps even later by providing personalized genetics that will tailor treatment and prevention strategies to individual patients.

First, we need to greatly improve our knowledge of biological and statistical epistasis and its role in human health and disease. We know very little about the role of epistasis in human biology and public health because the focus for so long has been on the effects of single genes and single genetic variants in biological and clinical endpoints. Given the ubiquity of complexity in genetic architecture, with epistasis as a central component, we need to rephrase our research questions with this in mind. Instead of asking which SNP is associated with disease, we should be asking which combination of SNPs is associated with disease. Rephrasing the question in this manner necessitates a redefinition of the null hypotheses that needs to be tested via statistical and computational methods. The current status quo is to test the universal null hypothesis of no association via only linear statistical methods. Rejection of this null hypothesis allows an investigator to draw inferences about independent genetic effects but not nonlinear epistatic effects. Given that complexity, including epistasis, is likely a substantial component of biological reality, we propose the following set of plausible hypotheses for retooling our analytical approach to this problem. First, we recommend as a starting point to test the null hypothesis that the associations in the data are only linear and additive. A null hypothesis of linearity is consistent with the hypothesis testing in the status quo. For example, one could test the linear null hypothesis via methods such as MDR that were designed to model nonlinear interactions. Once the linear null hypothesis has been tested via proper nonlinear methods, the logical next step is to test the universal null hypothesis of no association via linear statistical methods that model the independent and additive effects. Rejection of the universal null, in addition to the linear null, provides a set of results generated in a systematic manner that embraces complexity. These results can then be interpreted biologically via experimental methods or can be interpreted statistically via approaches such as parsimony or information theory. At present, this proposal is more of a philosophical exercise than a practical one because fast and powerful analytical tools and software for detecting epistasis on a genome-wide scale are not yet available. Additionally, it is important to design these analyses to compensate for multiple testing issues that exist as part of evaluating interactions. Despite some of these current practical limitations, the biological evidence for epistasis is compelling enough to suggest that there are very few loci in the genome that will have universal effects on disease risk on the order of that seen, for example, for the apolipoprotein E (*APOE*) gene (MIM 107741) and Alzheimer's disease (MIM 104300). As such, we need to design our analyses accordingly. Genetic epidemiology has shaken the trees and picked the low-hanging fruit. It is time to climb the branches to identify the hard-to-reach fruit and to search for low-hanging fruit from a different perspective (see Figure 3).

With the above in mind, it is important to note that the current one-SNP-at-a-time approach to GWAS analysis that puts so much weight on replication[98] can actually provide important clues about the complexity of the underlying genetic architecture. For example, a recent study by Greene et al. demonstrated that failure to replicate a genetic association in a second independent sample can be an indication that single SNPs contribute to disease susceptibility through nonlinear interactions with one or more other SNPs.[99] Greene et al. showed that the power to replicate a SNP with a significant main effect can drop from more than 80% to less than 20% with a change in allele frequency at a second interacting SNP of less than 0.1.[99] Such small changes in allele frequency are often observed even when the replication sample is taken from the same population. This study recommended that SNPs that fail to replicate be followed up with epistasis analysis to check for interaction. This of course introduces epistasis analysis as an afterthought to a main-effects analysis, an approach that we now argue is inadequate. This study also raises the question of how much weight we should put on statistical replication under the assumption of complexity.

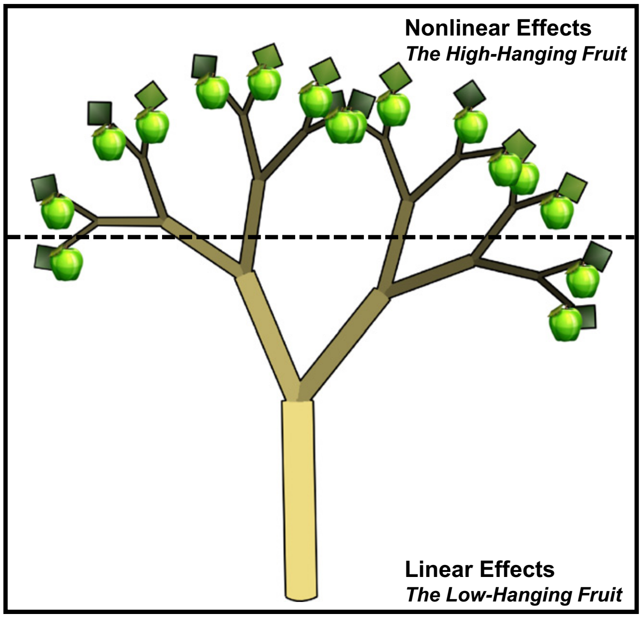


Figure 3. Low-Hanging and High-Hanging Genetic Fruit
Under the assumption that common diseases have a complex genetic architecture, we expect there to be few SNPs with moderate to large independent and additive main effects on disease susceptibility (i.e., low-hanging fruit). Rather, most SNPs of interest will be nestled in the branches and will only be found by embracing the complexity of the genotype-to-phenotype mapping relationship that is likely to be characterized by nonlinear gene-gene interactions and other phenomena such as gene-environment interaction and locus heterogeneity.

The believability of a statistical result relies more on the biological interpretation and experimental evidence than it does on the actual statistical finding.[99] Indeed, Bush et al.,[100] Holmans et al.,[101] and Saccone et al.[102] have shown that using biological knowledge to guide genetic association studies may provide more meaningful results. Yu et al. provide a hypothesis-testing framework for combining multiple SNPs from the same gene or from multiple genes in a pathway-based manner.[103] Askland et al. recently showed that patterns of SNPs in biological pathways are more likely to replicate than individual SNPs in GWAS.[104] Wilke et al. have suggested that we should not even begin to analyze a GWAS study until we have exhaustively studied each candidate gene and each pathway; only then will we have the appropriate knowledge base to make sense of GWAS results.[105] As Moore noted, there is a major shift in the field of genetic epidemiology away from the purely statistical approaches to more bioinformatic approaches that consider knowledge about gene function, gene networks, and biochemical pathways.[97] 2009 perhaps marks the turning point toward more of a systems approach that recognizes the role of epistasis and other complexities in genetic architecture.

Second, we need powerful analytical tools that are designed to address the complexity of genetic architecture resulting from epistasis and other phenomena. There is an important role for biostatisticians, bioinformaticists, and other analytically trained scientists in developing the next generation of statistical and computational tools that will embrace and directly confront the complexity that confounds current genetic association studies. The MDR algorithm that we briefly summarized here is a start, but it is only one example of the types of tools needed. Such tools are likely to come from the machine learning and data mining communities that are actively engaged in solving complex problems in other disciplines such as economics and engineering. In addition to powerful algorithms, we also need user-friendly software that can be used by geneticists and epidemiologists. As reviewed by Moore, these software packages need to be designed so that they are easy enough for a biologist to use but powerful and flexible enough for a statistician or computer scientist to use.[106] Moore suggests that the ideal analysis will be performed by the geneticist and the bioinformaticist jointly so that they can communicate information about the problem and the algorithms in real time.[106]

Third, we need better experimental methods for confirming statistical models of epistasis in animal models or in human cell culture. Interpreting genetic associations for common diseases such as type 1[107,108] and type 2[109,110] diabetes has not been easy, and it is clear that making inferences about etiology from any genetic model is a significant ongoing challenge.[111] Ultimately, we will need to rely on experimental biology to validate our genetic models. Although we are very good at perturbing single genes or pairs of genes, we are not very good at designing experiments to perturb complex systems. A call for multifactorial perturbation experiments has been made, but there has been little progress toward this end.[112] A step in the right direction is the Collaborative Cross initiative from the Complex Trait Consortium, which aims to develop a common reference panel of recombinant inbred mice that each have a mixture of genomes from several laboratory strains of mice and several wild strains.[113,114] This resource will provide a panel of mice for experimentation that more closely resembles the natural distribution of genetic diversity in humans than the widely used inbred laboratory strains do. Similar resources in other model systems such as *Drosophila* are also being developed.[115] Our ability to make full use of these panels and other resources to study epistasis will depend critically on our ability to perturb multiple genes simultaneously in these systems in a high-throughput manner.

Fourth, we need to remember the principles of classical genetics as we immerse ourselves in the excitement of cutting-edge genotyping technology that makes GWAS possible as well as in the emerging methods to rapidly sequence the entire genome. Indeed, Miller and Hollander[116] cautioned geneticists 15 years ago to not divert our attention elsewhere in light of “wondrous new molecular techniques.” This warning has been largely ignored. For example, pedigrees have been put aside in favor of large population-based case-control studies for GWAS. We predict there will be a return to pedigree-based studies and other methods consistent with classical genetics as it is realized that technology-centric approaches have significant shortcomings. The idea that pedigrees are still useful is supported by Culverhouse et al., who showed that purely epistatic models can give rise to increased allele sharing between affected siblings even in the absence of variation resulting from additivity or dominance.[31] Of course, as mentioned above, we still need to develop the analytical tools to model epistasis in pedigree-based studies. The integration of the pedigree disequilibrium test (PDT) of Martin et al.[117,118] with the MDR method described above has yielded a novel MDR-PDT approach to detecting interactions in general pedigrees.[119] This is of course only a start, but it is a step in the right direction. The only way to ensure that classical genetics is not forgotten in the genomics age is to make an effort to teach the classical concepts from the original literature in graduate school. This is often passed over in favor of recent literature on GWAS and other genomics methodologies. It is the blend of classical genetics with modern genetics that gives us the maximum ability to reveal the details of genetic architecture that are necessary for personal genetics.

Finally, we need to continue to integrate systems biology into human genetics in a meaningful manner. As Moore and Williams have discussed, one of the greatest contributions to our understanding of biological organisms was the merger of Darwin’s evolution of species by natural selection and Mendel’s principles of heredity.[29] This merger was referred to as “the modern synthesis” by Huxley[120] and others, and it paved the way for evolutionary and population genetics as we know them today. We are presently undergoing a “more modern” synthesis that merges multiple disciplines into what has been referred to as systems biology.[121] One goal of systems biology is to efficiently, accurately, and inexpensively measure most if not all of the biomolecules involved in one or more biochemical or physiological systems. Only after all of the relevant information is available will it be possible to mathematically model biomolecules with respect to interindividual phenotypic differences. We argue that the vast divide between biological and statistical epistasis will only be narrowed by our success in applying systems biology to genetics problems.[122] Our ability to measure information at multiple levels in the hierarchy between genes and disease will provide the basis for interpreting statistical models. Of course, more data is not the same as more knowledge, and our ability to translate systems biology into a deeper understanding of epistasis and human disease will depend critically on our analytical framework and the simplifying assumptions that we make.

Recognition of the complexity of genetic architecture and successful progress in terms of these five recommendations will provide the knowledge base that will be necessary for personal genetics to have the kind of impact on human health that we would all like to see. It is interesting to note that the focus on complexity in human genetics is not a new idea. More than 50 years ago, Snyder[123] suggested that "if human genetics is to progress along fresh pathways, the traditional atomistic approach must be supplemented by new methods which will provide information on multifactorial inheritance" and that "We must be able to analyze genetic variability without recourse to classical single-gene analyses." It is time for the human genetics community to embrace the complexity of human traits that was recognized by Snyder and others before many current human geneticists were born.

Acknowledgments

This work was supported by National Institutes of Health grants AI59694, LM009012, LM010098, and HD047447. The authors would like to thank two anonymous referees for their thoughtful suggestions.

Web Resources

The URLs for data presented herein are as follows:

23andMe, http://www.23andme.com
Multifactor Dimensionality Reduction (MDR) software, http://www.epistasis.org
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

References

1. Narod, S.A., and Foulkes, W.D. (2004). BRCA1 and BRCA2: 1994 and beyond. Nat. Rev. Cancer *4*, 665–676.
2. Ripperger, T., Gadzicki, D., Meindl, A., and Schlegelberger, B. (2009). Breast cancer susceptibility: Current knowledge and implications for genetic counselling. Eur. J. Hum. Genet. *17*, 722–731.
3. Weiss, K.M. (1995). Genetic Variation and Human Disease: Principles and Evolutionary Approaches (New York: Cambridge University Press).
4. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.
5. Wang, W.Y., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: Theoretical and practical concerns. Nat. Rev. Genet. *6*, 109–118.
6. Spitz, M.R., Amos, C.I., Dong, Q., Lin, J., and Wu, X. (2008). The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. J. Natl. Cancer Inst. *100*, 1552–1556.
7. Easton, D.F., and Eeles, R.A. (2008). Genome-wide association studies in cancer. Hum. Mol. Genet. *17*(*R2*), R109–R115.
8. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R., et al.; SEARCH collaborators; kConFab; AOCS Management Group (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature *447*, 1087–1093.
9. Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C.S., Humphreys, M.K., Platte, R., Morrison, J., Maranian, M., Pooley, K.A., Luben, R., et al.; SEARCH; GENICA Consortium; kConFab; Australian Ovarian Cancer Study Group (2009). Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat. Genet. *41*, 585–590.
10. Clark, A.G., Boerwinkle, E., Hixson, J., and Sing, C.F. (2005). Determinants of the success of whole-genome association testing. Genome Res. *15*, 1463–1467.
11. Jakobsdottir, J., Gorin, M.B., Conley, Y.P., Ferrell, R.E., and Weeks, D.E. (2009). Interpretation of genetic association studies: Markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet. *5*, e1000337.
12. Kraft, P., Wacholder, S., Cornelis, M.C., Hu, F.B., Hayes, R.B., Thomas, G., Hoover, R., Hunter, D.J., and Chanock, S. (2009). Beyond odds ratios—Communicating disease risk based on genetic profiles. Nat. Rev. Genet. *10*, 264–269.
13. Kaye, J. (2008). The regulation of direct-to-consumer genetic tests. Hum. Mol. Genet. *17*(*R2*), R180–R183.
14. Bateson, P. (2002). William Bateson: A biologist ahead of his time. J. Genet. *81*, 49–58.
15. Bateson, W. (1909). Mendel's Principles of Heredity (Cambridge: Cambridge University Press).
16. Tyler, A.L., Asselbergs, F.W., Williams, S.M., and Moore, J.H. (2009). Shadows of complexity: What biological networks reveal about epistasis and pleiotropy. Bioessays *31*, 220–227.
17. Snyder, L.H. (1935). The Principles of Heredity (Boston: Heath).
18. Hollander, W.F. (1955). Epistasis and hypostasis. J. Hered. *46*, 222–225.
19. Cheverud, J.M., and Routman, E.J. (1995). Epistasis and its contribution to genetic variance components. Genetics *139*, 1455–1461.
20. Miller, W.J. (1997). Dominance, codominance and epistasis. Braz. J. Genet. *20*, 663–665.
21. Phillips, P.C. (1998). The language of gene interaction. Genetics *149*, 1167–1171.
22. Phillips, P.C. (2008). Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet. *9*, 855–867.
23. Brodie, E.D. III. (2000). Why evolutionary genetics does not always add up. In Epistasis and the Evolutionary Process, J. Wolf, B. Brodie, III, and M. Wade, eds. (New York: Oxford University Press), pp. 3–19.
24. Templeton, A.R. (2000). Epistasis and complex traits. In Epistasis and the Evolutionary Process, J. Wolf, B. Brodie, III, and M. Wade, eds. (New York: Oxford University Press), pp. 41–57.
25. Wade, M.J. (2001). Epistasis, complex traits, and mapping genes. Genetica *112-113*, 59–69.

26. Wade, M.J., Winther, R.G., Agrawal, A.F., and Goodnight, C.J. (2001). Alternative definitions of epistasis: Dependence and interaction. Trends Ecol. Evol. *16*, 498–504.
27. Cordell, H.J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. *11*, 2463–2468.
28. Cordell, H.J. (2009). Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat. Rev. Genet. Published online May 12, 2009. 10.1038/nrg2579.
29. Moore, J.H., and Williams, S.M. (2005). Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. Bioessays *27*, 637–646.
30. Fisher, R.A. (1918). The correlations between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinburgh *52*, 399–433.
31. Culverhouse, R., Suarez, B.K., Lin, J., and Reich, T. (2002). A perspective on epistasis: Limits of models displaying no main effect. Am. J. Hum. Genet. *70*, 461–471.
32. Li, W., and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. Hum. Hered. *50*, 334–349.
33. Moore, J.H., Boczko, E.M., and Summar, M.L. (2005). Connecting the dots between genes, biochemistry, and disease susceptibility: Systems biology modeling in human genetics. Mol. Genet. Metab. *84*, 104–111.
34. Moore, J.H., and Hahn, L.W. (2004). Evaluation of a discrete dynamic systems approach for modeling the hierarchical relationship between genes, biochemistry, and disease susceptibility. Discrete Contin. Dyn. Syst. B *4*, 275–287.
35. Buchler, N.E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. Proc. Natl. Acad. Sci. USA *100*, 5136–5141.
36. Tagkopoulos, I., Liu, Y.C., and Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. Science *320*, 1313–1317.
37. Moore, J.H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum. Hered. *56*, 73–82.
38. Waddington, C.H. (1942). Canalization of development and the inheritance of acquired characters. Nature *150*, 563–565.
39. Gibson, G. (2009). Decanalization and the origin of complex disease. Nat. Rev. Genet. *10*, 134–140.
40. Moore, J.H., and Williams, S.M. (2002). New strategies for identifying gene-gene interactions in hypertension. Ann. Med. *34*, 88–95.
41. Millstein, J., Conti, D.V., Gilliland, F.D., and Gauderman, W.J. (2006). A testing framework for identifying susceptibility genes in the presence of epistasis. Am. J. Hum. Genet. *78*, 15–27.
42. Lewontin, R.C. (1974). Annotation: The analysis of variance and the analysis of causes. Am. J. Hum. Genet. *26*, 400–411.
43. Lewontin, R.C. (2006). Commentary: Statistical analysis or biological analysis as tools for understanding biological causes. Int. J. Epidemiol. *35*, 536–537.
44. Wahlsten, D. (1990). Insensitivity of the analysis of variance to heredity-environment interactions. Behav. Brain Sci. *13*, 109–161.
45. Mitchell, T. (1997). Machine Learning (New York: McGraw-Hill).
46. Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (New York: Springer-Verlag).
47. McKinney, B.A., Reif, D.M., Ritchie, M.D., and Moore, J.H. (2006). Machine learning for detecting gene-gene interactions: A review. Appl. Bioinformatics *5*, 77–88.
48. Thornton-Wells, T.A., Moore, J.H., and Haines, J.L. (2004). Genetics, statistics and human disease: Analytical retooling for complexity. Trends Genet. *20*, 640–647.
49. Motsinger, A.A., Ritchie, M.D., and Reif, D.M. (2007). Novel methods for detecting epistasis in pharmacogenomics studies. Pharmacogenomics *8*, 1229–1241.
50. Nelson, M.R., Kardia, S.L., Ferrell, R.E., and Sing, C.F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. *11*, 458–470.
51. Culverhouse, R., Klein, T., and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. Genet. Epidemiol. *27*, 141–152.
52. Kooperberg, C., Ruczinski, I., LeBlanc, M.L., and Hsu, L. (2001). Sequence analysis using logic regression. Genet. Epidemiol. *21* (*Suppl 1*), S626–S631.
53. Kooperberg, C., and Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. Genet. Epidemiol. *28*, 157–170.
54. Lunetta, K.L., Hayward, L.B., Segal, J., and Van Eerdewegh, P. (2004). Screening large-scale association study data: Exploiting interactions using random forests. BMC Genet. *5*, 32.
55. Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. Genet. Epidemiol. *28*, 171–182.
56. Moore, J.H., and Ritchie, M.D. (2004). STUDENTJAMA. The challenges of whole-genome approaches to common diseases. JAMA *291*, 1642–1643.
57. Moore, J.H. (2007). Genome-wide analysis of epistasis using multifactor dimensionality reduction: Feature selection and construction in the domain of human genetics. In Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data, X. Zhu and I. Davidson, eds. (Hershey, PA: IGI Global), pp. 17–30.
58. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., and Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. *69*, 138–147.
59. Ritchie, M.D., Hahn, L.W., and Moore, J.H. (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet. Epidemiol. *24*, 150–157.
60. Hahn, L.W., Ritchie, M.D., and Moore, J.H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics *19*, 376–382.
61. Hahn, L.W., and Moore, J.H. (2004). Ideal discrimination of discrete clinical endpoints using multilocus genotypes. In Silico Biol. *4*, 183–194.
62. Moore, J.H. (2004). Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev. Mol. Diagn. *4*, 795–803.

63. Moore, J.H., Gilbert, J.C., Tsai, C.T., Chiang, F.T., Holden, T., Barney, N., and White, B.C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J. Theor. Biol. *241*, 252–261.
64. Michalski, R.S. (1983). A theory and methodology of inductive learning. Artif. Intell. *20*, 111–161.
65. Chung, Y., Lee, S.Y., Elston, R.C., and Park, T. (2007). Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics *23*, 71–76.
66. Lee, S.Y., Chung, Y., Elston, R.C., Kim, Y., and Park, T. (2007). Log-linear model-based multifactor dimensionality reduction method to detect gene gene interactions. Bioinformatics *23*, 2589–2595.
67. Lou, X.Y., Chen, G.B., Yan, L., Ma, J.Z., Zhu, J., Elston, R.C., and Li, M.D. (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am. J. Hum. Genet. *80*, 1125–1137.
68. Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., and Moore, J.H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol. *31*, 306–315.
69. Pattin, K.A., White, B.C., Barney, N., Gui, J., Nelson, H.H., Kelsey, K.T., Andrew, A.S., Karagas, M.R., and Moore, J.H. (2009). A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. Genet. Epidemiol. *33*, 87–94.
70. Namkung, J., Elston, R.C., Yang, J.M., and Park, T. (2009). Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. Genet. Epidemiol. Published online February 24, 2009. 10.1002/gepi.20416.
71. Mei, H., Cuccaro, M.L., and Martin, E.R. (2007). Multifactor dimensionality reduction-phenomics: A novel method to capture genetic heterogeneity with use of phenotypic variables. Am. J. Hum. Genet. *81*, 1251–1261.
72. Bush, W.S., Edwards, T.L., Dudek, S.M., McKinney, B.A., and Ritchie, M.D. (2008). Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinformatics *9*, 238.
73. Namkung, J., Kim, K., Yi, S., Chung, W., Kwon, M.S., and Park, T. (2009). New evaluation measures for multifactor dimensionality reduction classifiers in gene-gene interaction analysis. Bioinformatics *25*, 338–345.
74. Andrew, A.S., Nelson, H.H., Kelsey, K.T., Moore, J.H., Meng, A.C., Casella, D.P., Tosteson, T.D., Schned, A.R., and Karagas, M.R. (2006). Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. Carcinogenesis *27*, 1030–1037.
75. International Consortium of Bladder Cancer. (2009). Polymorphisms in DNA repair genes, smoking, and bladder cancer risk: Findings from the International Consortium of Bladder Cancer. Cancer Res., in press.
76. Sinnott-Armstrong, N.A., Greene, C.S., Cancare, F., and Moore, J.H. (2009). Accelerating epistasis analysis in human genetics with consumer graphics hardware. BMC Res Notes *2*, 149.
77. Bush, W.S., Dudek, S.M., and Ritchie, M.D. (2006). Parallel multifactor dimensionality reduction: A tool for the large-scale analysis of gene-gene interactions. Bioinformatics *22*, 2173–2174.
78. Moore, J.H. (2009). Mining patterns of epistasis in human genetics. In Biological Data Mining, J.Y. Chen and S. Lonardi, eds. (New York: Chapman and Hall).
79. Greene, C.S., Kiralis, J., and Moore, J.H. (2009). Nature-inspired algorithms for the genetic analysis of epistasis in common human diseases: A theoretical assessment of wrapper vs. filter approaches. Proc. IEEE Cong. Evol. Comp. 800–807.
80. Wilke, R.A., Reif, D.M., and Moore, J.H. (2005). Combinatorial pharmacogenetics. Nat. Rev. Drug Discov. *4*, 911–918.
81. Robnik-Siknja, M., and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. *53*, 23–69.
82. McKinney, B.A., Reif, D.M., White, B.C., Crowe, J.E. Jr., and Moore, J.H. (2007). Evaporative cooling feature selection for genotypic data involving interactions. Bioinformatics *23*, 2113–2120.
83. Moore, J.H., and White, B.C. (2007). Tuning ReliefF for genome-wide genetic analysis. Lect. Notes Comput. Sci. *4447*, 166–175.
84. Greene, C.S., Penrod, N.M., Kiralis, J., and Moore, J.H. (2009). Spatially uniform reliefF (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Mining, in press.
85. Michalewicz, Z., and Fogel, D.B. (2004). How to Solve It: Modern Heuristics (New York: Springer).
86. Moore, J.H., and White, B.C. (2006). Exploiting expert knowledge in genetic programming for genome-wide genetic analysis. Lect. Notes Comput. Sci. *4193*, 969–977.
87. Moore, J.H., and White, B.C. (2007). Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In Genetic Programming Theory and Practice IV, R. Riolo, T. Soule, and B. Worzel, eds. (New York: Springer), pp. 11–28.
88. Greene, C.S., White, B.C., and Moore, J.H. (2008). Ant colony optimization for genome-wide genetic analysis. Lect. Notes Comput. Sci. *5217/2008*, 37–47.
89. Greene, C.S., Gilmore, J., Kiralis, J., Andrews, P.C., and Moore, J.H. (2009). Optimal use of expert knowledge in ant colony optimization for the analysis of epistasis in human disease. Lect. Notes Comput. Sci. *5483*, 92–103.
90. Greene, C.S., White, B.C., and Moore, J.H. (2009). Sensible initialization using expert knowledge for genome-wide analysis of epistasis using genetic programming. Proc. IEEE Cong. Evol. Comp. 1289–1296.
91. Greene, C.S., and Moore, J.H. (2009). Solving complex problems in human genetics using nature-inspired algorithms requires strategies which exploit domain-specific knowledge. In Nature-Inspired Informatics for Intelligent Applications and Knowledge Discovery, R. Chiong, ed. (Hershey, PA: IGI Global).
92. Pattin, K.A., and Moore, J.H. (2008). Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. Hum. Genet. *124*, 19–29.
93. Kardia, S.L., Modell, S.M., and Peyser, P.A. (2003). Family-centered approaches to understanding and preventing coronary heart disease. Am. J. Prev. Med. *24*, 143–151.
94. Sing, C.F., Stengård, J.H., and Kardia, S.L. (2003). Genes, environment, and cardiovascular disease. Arterioscler. Thromb. Vasc. Biol. *23*, 1190–1196.

95. Rea, T.J., Brown, C.M., and Sing, C.F. (2006). Complex adaptive system models and the genetic analysis of plasma HDL-cholesterol concentration. Perspect. Biol. Med. *49*, 490–503.

96. Williams, S.M., Canter, J.A., Crawford, D.C., Moore, J.H., Ritchie, M.D., and Haines, J.L. (2007). Problems with genome-wide association studies. Science *316*, 1840–1842.

97. Moore, J.H. (2009). From genotypes to genometypes: Putting the genome back in genome-wide association studies. Eur. J. Hum. Genet. Published online March 11, 2009. 10.1038/ejhg.2009.39.

98. Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., et al.; NCI-NHGRI Working Group on Replication in Association Studies (2007). Replicating genotype-phenotype associations. Nature *447*, 655–660.

99. Greene, C.S., Penrod, N.M., Williams, S.M., and Moore, J.H. (2009). Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS ONE *4*, e5639.

100. Bush, W.S., Dudek, S.M., and Ritchie, M.D. (2009). Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac. Symp. Biocomput. 368–379.

101. Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C., and Craddock, N.; Wellcome Trust Case-Control Consortium (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet. *85*, 13–24.

102. Saccone, S.F., Saccone, N.L., Swan, G.E., Madden, P.A., Goate, A.M., Rice, J.P., and Bierut, L.J. (2008). Systematic biological prioritization after a genome-wide association study: An application to nicotine dependence. Bioinformatics *24*, 1805–1811.

103. Yu, K., Li, Q., Bergen, A.W., Pfeiffer, R.M., Rosenberg, P.S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. Genet. Epidemiol. Published online March 30, 2009. 10.1002/gepi.20422.

104. Askland, K., Read, C., and Moore, J. (2009). Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum. Genet. *125*, 63–79.

105. Wilke, R.A., Mareedu, R.K., and Moore, J.H. (2008). The pathway less traveled: Moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. Curr. Pharmacogenomics Person Med. *6*, 150–159.

106. Moore, J.H. (2007). Bioinformatics. J. Cell. Physiol. *213*, 365–369.

107. Cordell, H.J., Todd, J.A., Bennett, S.T., Kawaguchi, Y., and Farrall, M. (1995). Two-locus maximum lod score analysis of a multifactorial trait: Joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. Am. J. Hum. Genet. *57*, 920–934.

108. Cordell, H.J., Todd, J.A., Hill, N.J., Lord, C.J., Lyons, P.A., Peterson, L.B., Wicker, L.S., and Clayton, D.G. (2001). Statistical modeling of interlocus interactions in a complex disease: Rejection of the multiplicative model of epistasis in type 1 diabetes. Genetics *158*, 357–367.

109. Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I., and Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat. Genet. *21*, 213–215.

110. Cox, N.J., Hayes, M.G., Roe, C.A., Tsuchiya, T., and Bell, G.I. (2004). Linkage of calpain 10 to type 2 diabetes: The biological rationale. Diabetes *53* (*Suppl 1*), S19–S25.

111. Page, G.P., George, V., Go, R.C., Page, P.Z., and Allison, D.B. (2003). "Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. Am. J. Hum. Genet. *73*, 711–719.

112. Jansen, R.C. (2003). Studying complex biological systems using multifactorial perturbation. Nat. Rev. Genet. *4*, 145–151.

113. Churchill, G.A., Airey, D.C., Allayee, H., Angel, J.M., Attie, A.D., Beatty, J., Beavis, W.D., Belknap, J.K., Bennett, B., Berrettini, W., et al.; Complex Trait Consortium (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat. Genet. *36*, 1133–1137.

114. Chesler, E.J., Miller, D.R., Branstetter, L.R., Galloway, L.D., Jackson, B.L., Philip, V.M., Voy, B.H., Culiat, C.T., Threadgill, D.W., Williams, R.W., et al. (2008). The Collaborative Cross at Oak Ridge National Laboratory: Developing a powerful resource for systems genetics. Mamm. Genome *19*, 382–389.

115. Ayroles, J.F., Carbone, M.A., Stone, E.A., Jordan, K.W., Lyman, R.F., Magwire, M.M., Rollmann, S.M., Duncan, L.H., Lawrence, F., Anholt, R.R., and Mackay, T.F. (2009). Systems genetics of complex traits in Drosophila melanogaster. Nat. Genet. *41*, 299–307.

116. Miller, W.J., and Hollander, W.F. (1995). Three neglected advances in classical genetics. BioScience *45*, 98–104.

117. Martin, E.R., Monks, S.A., Warren, L.L., and Kaplan, N.L. (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test. Am. J. Hum. Genet. *67*, 146–154.

118. Martin, E.R., Bass, M.P., Gilbert, J.R., Pericak-Vance, M.A., and Hauser, E.R. (2003). Genotype-based association test for general pedigrees: The genotype-PDT. Genet. Epidemiol. *25*, 203–213.

119. Martin, E.R., Ritchie, M.D., Hahn, L., Kang, S., and Moore, J.H. (2006). A novel method to identify gene-gene effects in nuclear families: The MDR-PDT. Genet. Epidemiol. *30*, 111–123.

120. Huxley, J. (1942). Evolution: The Modern Synthesis (London: Allen & Unwin).

121. Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. Annu. Rev. Genomics Hum. Genet. *2*, 343–372.

122. Moore, J.H. (2005). A global view of epistasis. Nat. Genet. *37*, 13–14.

123. Snyder, L.H. (1951). Old and new pathways in human genetics. Am. J. Hum. Genet. *3*, 1–16.