Published in final edited form as:
Curr Protoc Hum Genet. 2011 July ; 0 1: Unit1.14. doi:10.1002/0471142905.hg0114s70.

Analysis of Gene-Gene Interactions

Diane Gilbert-Diamond[1] and **Jason H. Moore**[1,2,3,4]

[1]Computational Genetics Laboratory, Departments of Genetics and Community and Family Medicine, Dartmouth Medical School, Lebanon, New Hampshire

[2]Department of Computer Science, University of New Hampshire, Durham, New Hampshire

[3]Department of Computer Science, University of Vermont

[4]Institute for Quantitative Biomedical Sciences, Dartmouth College

Abstract

The goal of this unit is to introduce gene-gene interactions (epistasis) as a significant complicating factor in the search for disease susceptibility genes. This unit begins with an overview of gene-gene interactions and why they are likely to be common. Then, it reviews several statistical and computational methods for detecting and characterizing genes with effects that are dependent on other genes. The focus of this unit is genetic association studies of discrete and quantitative traits because most of the methods for detecting gene-gene interactions have been developed specifically for these study designs.

Keywords

epistasis; genetics; statistics; bioinformatics

INTRODUCTION

One goal of human genetics is to identify genes with specific DNA sequence variations that increase or decrease susceptibility to disease. Success in this endeavor will depend largely on the genetic architecture of the disease, which can be defined as the (1) number of genes that impact disease susceptibility, (2) distribution of alleles and genotypes at those genes, and (3) manner in which the alleles and genotype impact disease susceptibility (Weiss, 1993). It is anticipated that the genetic architecture of common diseases that represent the bulk of the public health burden is likely to be very complex (e.g., Moore, 2003; Sing et al.,

INTERNET RESOURCES

http://dceg.cancer.gov/tools/design/POWER

The Power program for estimation of sample size and power for two-locus interactions in both cohort and case-control studies.

http://hydra.usc.edu/fitf

The focused interaction testing framework (FITF) for detecting gene-gene interactions using logistic regression.

http://hydra.usc.edu/gxe

The Quanto program for estimation of sample size and power in matched case-control, case-sibling, case-parent, and case-only designs.

http://www.epistasis.org

Computational Genetics Laboratory at Dartmouth Medical School, where open-source software for MDR can be obtained.

2003; Thornton-Wells et al., 2004; Eichler et al., 2010), i.e., there are likely to be many susceptibility genes, each with combinations of rare and common alleles and genotypes, that impact disease susceptibility primarily through nonlinear interactions with genetic and environmental factors.

There are many phenomena such as phenocopy and locus heterogeneity that contribute to the complexity of the mapping between genotype and phenotype (Thornton-Wells et al., 2004). The goal of this unit is to introduce the concept of gene-gene interactions (epistasis) as a significant complicating factor in the search for disease susceptibility genes. This unit begins with an overview of gene-gene interactions and why they are likely to be common; then it reviews several statistical and computational methods for detecting and characterizing genes with effects that are dependent on other genes. The focus of this unit is genetic association studies of discrete and quantitative traits (see Weiss, 1993), since most of the methods for detecting gene-gene interactions have been developed specifically for these study designs.

WHAT ARE GENE-GENE INTERACTIONS?

The concept of epistasis or gene-gene interaction is not new. In fact, the idea has been around for at least 100 years and was recognized by early geneticists as an explanation for deviations from simple Mendelian ratios. William Bateson (1909) has been credited by Hollander (1955) and more recently by Phillips (1998, 2008) as the first to use the term epistasis, which literally translated means "resting upon." A commonly used textbook definition of epistasis is one gene masking the effects of another gene (e.g., Neel and Schull, 1954; Griffiths et al., 2008). A classic example of epistasis comes from studies of the shape of seed capsules from crosses of a weedy plant called shepard's purse (*Bursa bursa-pastoris*) by Shull (1914). In this study, crosses from doubly heterozygous plants yielded Mendelian ratios of fifteen triangular capsules to one oval capsule. It is generally believed that there are two pathways with dominant loci that lead to the triangular shape. It is only when both pathways are blocked by recessive alleles that the oval-shaped seed capsule is produced. This is an example of a recessive-by-recessive interaction since having two recessive genotypes leads to a different phenotype than with only one from either locus.

The shepard's purse example from Shull (1914) is an example of biological epistasis, i.e., the gene-gene interaction has a biological basis. This is exactly what Bateson (1909) had in mind when he coined the term. This is in contrast to the concept of statistical epistasis or epistacy that was first used by Fisher (1918) to describe deviations from additivity in a linear statistical model. Making biological inferences about epistasis from statistical models can be difficult (Cordell, 2002; Moore, 2005; Moore and Williams, 2005), although there are some approaches that take steps towards doing so (e.g., Cheverud and Routman, 1995). Wade et al. (2001) present useful concepts of biological and statistical epistasis from an alternative, evolutionary biology perspective.

A simple example of statistical epistasis in the form of penetrance functions is presented in Table 1.14.1. Penetrance is simply the probability (P) of disease (D) given a particular combination of genotypes (G) that was inherited, i.e., $P[D|G]$. The model illustrated in Table

1.14.1 is an extreme example of epistasis between two single nucleotide polymorphisms (SNPs) A and B. Let's assume that AA, aa, BB, and bb have population frequencies of 0.25, while genotypes Aa and Bb have frequencies of 0.5 (values in parentheses in Table 1.14.1). What makes this model interesting is that disease risk is entirely dependent on the particular *combination* of genotypes inherited. Individuals have a very high risk of disease if they inherit Aa or Bb but not both (i.e., the exclusive OR function). The penetrance for each individual genotype in this model is 0.05 and is computed by summing the products of the genotype frequencies and penetrance values. Thus, in this model there is no difference in disease risk for each single genotype as specified by the single-genotype penetrance values (all 0.05). This model is labeled M170 by Li and Reich (2000) in their categorization of genetic models involving two SNPs and is an example of a pattern that is not linearly separable. Heritability or the size of the genetic effect is a function of these penetrance values (e.g., Culverhouse et al., 2002). The model specified in Table 1.14.1 has a heritability of 0.053, which represents a relatively small genetic effect size. This model is a special case where all of the heritability is due to epistasis.

WHY ARE GENE-GENE INTERACTIONS LIKELY TO BE COMMON?

Moore (2003) outlines a working hypothesis stating that epistasis is a ubiquitous component of the genetic architecture of common human diseases. This working hypothesis is based on both historical and emerging research results.

First, the idea that epistasis is important is not new. As discussed above, the recognition that deviations from Mendelian ratios are due to interactions between genes has been around for nearly 100 years. This is important because it is an idea that has prevailed through time and is still recognized today.

Second, the ubiquity of biomolecular interactions in gene regulation and biochemical and metabolic systems suggests that the relationship between DNA sequence variations and clinical endpoints is likely to involve gene-gene interactions. This is perhaps the most important piece of evidence supporting the working hypothesis. For example, transcription of any given eukaryotic gene can be regulated by as many as 100 or more different proteins that act through protein-protein and protein-DNA interactions. It is likely that these biomolecular interactions are mediated by DNA sequence variations in the genes that encode the individual proteins.

Third, positive results from studies of single polymorphisms typically do not replicate across independent samples. This is true for both linkage and association studies. For example, Hirschhorn et al. (2002) reviewed more than 600 association studies for consistency of results. Of those in which the same polymorphism had been studied in three or more independent samples, there were only six results that were consistently replicated. While many of these conflicting reports arose from inadequately powered or designed studies, the majority of the conflicting results cannot be explained. Moore and Williams (2002) suggest that one reason studies of single polymorphisms typically do not replicate across independent samples is because gene-gene interactions are more important.

Fourth, gene-gene interactions are commonly found when properly investigated (see Templeton, 2000).

Why is epistasis so difficult to detect? What is the proper way to detect epistasis? These questions are addressed in the next several sections.

WHY ARE GENE-GENE INTERACTIONS DIFFICULT TO DETECT?

Epistasis is difficult to detect and characterize using traditional parametric statistical methods such as linear and logistic regression because of the sparseness of the data in high dimensions. That is, when interactions among multiple polymorphisms are considered, there are many multilocus genotype combinations that have very few or no data points. For example, with two SNPs that each has three genotypes, there are nine two-locus genotype combinations (e.g., Table 1.14.1). In the case of three SNPs, there are 27 three-locus genotype combinations. Thus, as each additional SNP is considered, the number of multilocus genotype combinations goes up exponentially. The result of this added dimensionality is that exponentially larger sample sizes are needed to have enough data to estimate the interaction effects. This phenomenon has been referred to as the curse of dimensionality (Bellman, 1961); for methods such as logistic regression, it can lead to parameter estimates that have very large standard errors, resulting in an increase in type I errors (see *APPENDIX 3M*) Concato et al., 1993; Peduzzi et al., 1996; Hosmer and Lemeshow, 2000).

In addition, detecting gene-gene interactions using traditional procedures for fitting regression models can be problematic, leading to an increase in type II errors and a decrease in power (see *APPENDIX 3M*). For example, forward selection (see Neter, 1990) is limited because interactions are only tested for those variables that have a statistically significant independent main effect. Those DNA sequence variations that have an interaction effect, but no or minimal main effect, will be missed. With backward elimination (see Neter, 1990), a complete model that includes all main effects and all interaction terms may require too many degrees of freedom. Stepwise procedures are more flexible than either forward selection or backward elimination, but can also suffer from requiring too many degrees of freedom. Detecting interactions among variables is a well-known challenge in statistics and data mining (Freitas, 2001).

METHODS FOR DETECTING GENE-GENE INTERACTIONS IN ASSOCIATION STUDIES OF DISCRETE TRAITS

Logistic Regression

Logistic regression is the workhorse of modern epidemiology. This approach is popular because it produces outputs in the range of 0 to 1 that can be used with a threshold to model discrete endpoints such as case-control status. Logistic regression models the probability of disease (p) as a linear function of independent variables (see Hosmer and Lemeshow, 2000; Kleinbaum and Klein, 2002). A logit transformation of p, $\ln[p/(1 - p)]$, is used to prevent p from taking on values less than zero or greater than one. By expressing the linear function in terms of exponentials, p can be modeled as $p = (e^{\alpha + \beta X})/(1 + e^{\alpha + \beta X})$, where e is the exponential, α and β are regression coefficients (i.e., parameters), and X is an independent variable. For a discrete independent variable such as a polymorphism, an odds ratio relating genotypes to probability of disease can be estimated from e^{α}. The independent main effects of two polymorphisms, A and B, can be modeled as $p = (e^{\alpha+\beta_1 A+\beta_2 B})/(1+e^{\alpha+\beta_1 A+\beta_2 B})$. The interaction between A and B can be modeled by adding a product term of the form β_3AB to the equation. A test of the null hypothesis of no interaction can be carried out by testing whether $\beta_3 = 0$. Rejection of this null hypothesis provides evidence for an interaction on a multiplicative scale.

The advantage of logistic regression is that interactions can be modeled relatively easily, the statistical theory is very well characterized, and the approach can be implemented on a standard desktop computer using a variety of freely and commercially available statistical packages. As described in the above section, an important disadvantage is that very large sample sizes are needed to accurately estimate the parameters in the model when there are many independent variables. Marchini et al. (2005) explored the role of logistic regression for detecting gene-gene interactions in the presence of independent effects on a genome-wide scale and found that models that included interactions were more powerful than traditional main effects models when the interaction effects were large relative to the main effects.

Several studies provide guidance on evaluating the power of a planned gene-gene interaction study using logistic regression. For example, the Power program of Garcia-Closas and Lubin and Gails (1999) allows estimation of sample size and power for two-locus interactions in both cohort and case-control studies. This program is available for free from http://dceg.cancer.gov/tools/design/POWER and is relatively easy to use. An additional program called Quanto is freely available from http://hydra.usc.edu/gxe for estimation of sample size and power in matched case-control, case-sibling, case-parent, and case-only designs. The software and methods are described in detail by Gauderman (2002).

Several alternatives to standard logistic regression for discrete clinical endpoints have also been developed. For example, Hoh et al. (2000) and Hoh and Ott (2001) have developed a combination of sequential and resampling methods for summing associations statistics to detect combined effects of multiple SNPs. This approach uses standard statistics in a novel way to detect multilocus effects. Application of this method to a coronary artery restenosis case-control data set yielded a highly significant interaction among seven SNPs from seven different genes (Zee et al., 2002). These associations would not have been identified using standard logistic regression analysis due to a lack of degrees of freedom for estimating all the interactions terms. The use of penalized logistic regression (Park and Hastie, 2008; Winham, 2011) shows some promise for overcoming some of these limitations.

The use of logic functions for defining new variables that can be included in a logistic regression analysis may also be useful (Kooperberg et al., 2001). Software for logic regression is freely available as a package for the logistic regression analysis model (R; see http://www.r-project.org). The focused interaction testing framework (FITF) of Millstein et al. (2006) provides a staged likelihood ratio-based approach to detecting interactions using logistic regression. The FITF software is freely available from http://hydra.usc.edu/fitf.

Multifactor Dimensionality Reduction

An alternative and complimentary method to logistic regression, reviewed by Cordell (2009), is multifactor dimensionality reduction (MDR). 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 strategy for identifying combinations of discrete genetic and environmental factors that are predictive of a discrete clinical endpoint (Ritchie et al., 2001, 2003; Hahn et al., 2003; Hahn and Moore, 2004; Moore, 2004, 2007; Moore et al., 2006; Velez et al., 2007). Unlike most other methods, MDR was designed to detect interactions in the absence of detectable main effects and thus complements approaches such as logistic regression.

At the heart of the MDR approach is a feature or attribute construction algorithm that creates a new variable or attribute by pooling, e.g., genotypes from multiple SNPS. The 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 developed by Michalski (1983). Constructive induction using the MDR kernel is accomplished in the following way. 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 while those considered low-risk are labeled G_0. This process constructs a new one-dimensional attribute with levels G_0 and G_1. It is this new single variable that is assessed using any classification method. The MDR method is based on the idea that changing the representation space of the data will make it easier for a classifier such as a decision tree or a naive Bayes learner to detect attribute dependencies (see Hastie et al., 2001). Open-source software in Java and C are freely available from http://www.epistasis.org.

Consider the simple example presented above and in Table 1.14.1. This penetrance function was used to simulate a data set with 200 cases (diseased subjects) and 200 controls (healthy subjects) for a total of 400 instances. All attributes in these data sets are categorical. The SNPs each have three levels (0, 1, 2) while the class has two levels (0, 1) that code controls and cases. Figure 1.14.1A illustrates the distribution of cases (left bars) and controls (right bars) for each of the three genotypes of SNP1 and SNP2. The dark-shaded cells have been labeled “high-risk” using a threshold of $T = 1$. The light-shaded cells have been labeled “low-risk.” Note that when considered individually, the ratio of cases to controls is close to 1 for each single genotype. Figure 1.14.1B illustrates the distribution of cases and controls when the two functional SNPs are considered jointly. Note the larger ratios that are consistent with the genetic model in Table 1.14.1. Also illustrated in Figure 1.14.1B is the distribution of cases and controls for the new single attribute constructed using MDR. This new single attribute captures much of the information from the interaction and could be assessed using logistic regression, for example.

Since its initial description by Ritchie et al. in 2001, numerous extensions and variations on the MDR method have been developed, including, for example, the incorporation of odds ratios (Chung et al., 2007) and Fisher's exact test (Gui et al., 2011) to increase model robustness, the implementation of generalized linear models (Lou et al., 2007) and other model based methods (Calle et al., 2008, 2010) that can accomodate discreet and continuous outcomes, entropy-based interpretation methods (Moore et al., 2006), permutation testing methods (Edwards et al., 2010; Greene et al., 2010a; Pattin et al., 2009), and different evaluation metrics (Bush et al., 2008; Mei et al., 2007; Namkung et al., 2009a). Methods have also been developed to handle imbalanced data (Valez et al., 2007), missing data (Namkung et al., 2009b), sparse or empty cells (Lee et al., 2007), covariate adjustment (Lou et al., 2007, Calle et al., 2008, Gui et al., 2011) and family data (Cattaert et al., 2010; Lou et al., 2008; Martin et al., 2006).The MDR method and its variations have been successfully applied to detecting gene-gene and gene-environment interactions for a wide variety of different common human diseases and clinical endpoints including, e.g., bladder cancer (Andrew et al., 2006, 2008; Chen et al. 2007; Huang et al. 2007), amytrophic lateral sclerosis (ALS) (Green et al., 2010b), and eczema (Mahachie et al., 2010). The MDR method has also been proposed for studies in pharmacogenetics and toxicogenetics (e.g., Wilke et al., 2005).

METHODS FOR DETECTING GENE-GENE INTERACTIONS IN ASSOCIATION STUDIES OF QUANTITATIVE TRAITS

Linear Regression

Linear regression is a popular choice for modeling quantitative traits because the models are easy to interpret and there is a well formulated mathematical theory underlying the method. Linear regression is a parametric statistical approach for modeling a continuous outcome variable (Y) as a linear function of discrete and/or continuous predictor variables (X_1, X_2, etc.). The linear model relating X to Y looks something like $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where β_0 is the intercept, β_1 and β_2 are the regression coefficients, and ε is the unexplained error in the model. In this model, the slope or regression of Y on X_1 is constant across the range of values for X_2. This means that the relationship between Y and X_1 is independent of X_2. Thus, the effects of the two predictor variables are purely additive. Deviations from additivity (i.e., interaction) can be measured by including a product term in the model as seen above for logistic regression. Here, the term $\beta_3 X_1 X_2$ would be added to account for any interaction. The presence of an interaction term in the linear model allows there to be a different regression relationship between Y and X_1 for each value of X_2. Thus, the null hypothesis of no interaction is equivalent to $\beta_3 = 0$. For genetic studies, it is customary to encode polymorphisms as dummy variables that specify certain types of genetic effects. Each polymorphism with N genotypes should be encoded by $N - 1$ dummy variables. A detailed description of linear regression methods is given by Neter et al. (1990). The use of linear regression to test for interactions is presented in detail by Aiken and West (1991). Details about power calculations for linear regression are provided by Cohen (1988).

As with logistic regression, advantages of linear regression are that interactions can be modeled relatively easily, the statistical theory is very well characterized, and the approach can be implemented on a standard desktop computer using a variety of freely and commercially available statistical packages. However, an important disadvantage is that very large sample sizes are needed to accurately estimate the parameters in the model when there are many independent variables. The limitations of linear regression approaches for detecting gene-gene interactions have been described by Wahlsten (1990). Some examples of using linear regression to detect gene-gene interactions include Hamon et al. (2004) and Asselbergs et al. (2007). Alternative methods for analyzing quantitative traits include MDR (Lou, 2007) and the combinatorial partitioning method, which will be discussed in the next section.

Combinatorial Partitioning Method

The combinatorial partitioning method (CPM) of Nelson et al. (2001) is one of the few alternatives to linear regression that have been developed. The CPM simultaneously considers multiple polymorphic loci to identify combinations of genotypes that are most strongly associated with variation in a quantitative trait. First, all possible multilocus genotypes are identified, and this multilocus genotype space is divided into partitions (groups that include one or more of the possible genotypes). The partitions are combined into sets in which every possible genotype in the multilocus genotype space is included in one, and only one, of the partitions of that set. Each possible set is then evaluated by two criterion: 1. if the proportion of explained variability in the quanitiative trait exceeds a predetermined threshold, using a method based on within- and between-partition variance (see Nelson, 2001 for details) 2. if the number of observations in each partition exceeds a pre-determined lower bound, e.g. 5, to ensure sufficient degrees of freedom for reliable within-partition estimates. Those sets that pass these criteria are then validated using multi-fold cross-validation (Stone, 1978). From the collection of validated sets, the most predictive sets are chosen to make inferences about the genotype-phenotype relationships using methods such as simple linear regression as described above. As with MDR, the partitioning of CPM serves to collapse the multiple dummy variables needed to encode multiple polymorphisms and their interactions intofewer variables (i.e., constructive induction), thereby reducing the dimensionality associated with modeling interactions..

When applied to modeling the relationship between eighteen diallelic loci from six cardiovascular disease susceptibility genes and interindividual variability in plasma triglycerides, Nelson et al. (2001) found nonadditive epistatic interactions between multiple loci. Although preliminary, these results suggest that CPM may be a valuable tool for the exploratory analysis of nonadditive gene-gene interactions. This is also the conclusion of Moore et al. (2002a,b), who applied CPM in an exploratory analysis of interactions among arterial thrombosis candidate genes. While CPM may provide a powerful alternative to linear regression, there are several important limitations, the most important of which is that the approach is very computationally intensive, since it must combinatorially sift through many genotype partitions. To address this limitation, Culverhouse et al. (2004) have developed the restricted partitioning method (RPM), which restricts the number of genotypic partitions evaluated. While CPM searches over all possible partitions, RPM only evaluates those partitions whose genotypes have statistically similar mean values of the quantitative trait. The reasoning is that partitions with large within-partition variance are unlikely to explain sufficient variability in the quanitiative trait. Interestingly, this method has been extended to case-control data and may complement the MDR method discussed earlier (Culverhouse, 2007, Hua, 2010).

DETECTING GENE-GENE INTERACTIONS ON A GENOME-WIDE SCALE

Biomedical sciences are undergoing an information explosion and an understanding implosion. That is, our ability to generate data is far outpacing our ability to interpret it. This is especially true in the domain of human genetics, where it is now technically and economically feasible to measure over a million SNPs across the human genome in each individual. An important goal in human genetics is to determine which of the multitude of genetic variants are useful for predicting who is at risk for common diseases. This genome-wide approach was expected to revolutionize the genetic analysis of common human diseases (Hirschhorn and Daly, 2005; Wang et al., 2005) and quickly replaced the traditional candidate-gene approach that focuses on several genes selected by their known or suspected function. The success of GWAS studies in identifying the genetic underpinnings of common diseases has been limited, however, and some of the unexplained heretability of common diseases may be explained by gene-gene interactions (Eichler et. al., 2010).

Moore and Ritchie (2004) have outlined three significant challenges that must be overcome to successfully identify nonadditive gene-gene interactions using a genome-wide approach. First, powerful data mining and machine learning methods will need to be developed to statistically model the relationship between combinations of DNA sequence variations and disease susceptibility. The MDR and CPM approaches were discussed above as alternatives to logistic and linear regression. A second challenge is the selection of genetic features or attributes that should be included for analysis. If interactions between genes explain most of the heritability of common diseases, then combinations of DNA sequence variations will need to be evaluated from a list of thousands of candidates. Filter and wrapper methods will play an important role because there are more combinations than can be exhaustively evaluated. A third challenge is the interpretation of gene-gene interaction models. Although a statistical model can be used to identify DNA sequence variations that confer risk for disease, this approach cannot be translated into specific prevention and treatment strategies without interpreting the results in the context of human biology. Making etiological inferences from computational models may be the most important and the most difficult challenge of all (Moore and Williams, 2005).

Combining the concept of nonadditive interaction described above with the challenge of variable selection yields what Goldberg (2002) calls a needle-in-a-haystack problem. That is, there may be a particular combination of SNPs that together with the right nonlinear function are a significant predictor of disease susceptibility. However, individually they may not look any different than thousands of other SNPs that are not involved in the disease process and are thus noisy. Under these models, the computational algorithm is truly looking for a genetic needle in a genomic haystack. A report from the International HapMap Consortium (Altshuler et al., 2005) suggests that approximately 300,000 carefully selected SNPs may be necessary to capture all of the relevant variation across the Caucasian human genome. Assuming this is true (it is probably a lower boundary), one would need to scan 4.5×10^{10} pairwise combinations of SNPs to find a genetic needle. The number of higher-order combinations is astronomical. Indeed, the current state of the art chips genotype over a million SNPs, leading to over 5.0×10^{11} pairwise combinations. What is the optimal approach to this problem?

Two approaches are generally used to select attributes for predictive models. The filter approach preprocesses the data by algorithmically or statistically assessing the quality or relevance of each variable and then using that information to select a subset for classification. The wrapper approach iteratively selects subsets of attributes for classification using either a deterministic or stochastic algorithm. The key difference between the two approaches is that the classifier plays no role in selecting which attributes to consider in the filter approach. As Freitas (2002) reviews, the advantage of the filter is speed, while the wrapper approach has the potential to do a better job classifying. Filter strategies, such as Relief (Kira and Rendell, 1992) and stochastic wrapper strategies such as genetic programming (Moore and White, 2006, 2007a,b) show promise for attribute selection (Moore, 2007). Recent extensions that have improved the power of Relief include Tuned ReliefF (TURF) (Moore et al., 2006), Spatially Uniform ReliefF (SURF) (Greene et al., 2009), and SURF* (Greene et al., 2010c). The Explicit Test of Interaction can also be used to pre-select SNPs that are likely to interact with other SNPs in relation to the outcome (Greene et al., 2010a).

The accumulated biological knowledge about the structure and function of various genes can also be used to prioritize which genetic variations to analyze and thus reduce the number of gene-gene interaction tests performed (Moore et. al, 2010). For example, expert knowledge about Gene Ontology (GO), chromosomal location, protein-protein interactions (Pattin and Moore, 2008, Emily et. al., 2009), and regulatory networks (Cowper-Sal Lari et. al., 2010) could all be used as biological filters. Some approaches to integrate expert knowledge into gene-gene interaction analyses include the Biofilter algorithm presented by Bush et. al. (2009) and the INTERSNP software package introduced by Herold et. al. (2009). Askland et. al. (2009) also demonstrated how the exploratory visual analysis (EVA) method can be used to select SNPs in specific pathways and GO groups. Additional work in this area is needed.

As sequencing technology advances, it is rapidly becoming feasible to sequence the entire genome of individuals for genetic analysis studies (Mardis, 2011). With whole genome sequences, the data available for analyses will grow dramatically and data mining and machine learning methods will become increasingly essential.

SUMMARY

This unit defines epistasis or gene-gene interaction and provides a rationale for why such interactions are likely to be common and why they are difficult to detect. Further, the unit summarizes several traditional and several new statistical and computational methods for detecting gene-gene interactions in association studies of both discrete and continuous traits. This brief introduction provides the foundation necessary to better understand the nature of gene-gene interactions and provides a starting point for deciding on an analytical approach for detecting interactions in epidemiological and genetic studies of common human diseases. Extending these methods to the analysis of genome-wide association data is a significant challenge that still needs to be addressed.

Acknowledgments

The authors are grateful to Davnah Urbach for her insightful suggestions for improving this manuscript. This work was financially supported by Grant R25CA134286 from the National Cancer Institute, NIH.

LITERATURE CITED

Aiken, LS.; West, SG. Multiple Regression: Testing and Interpreting Interactions. Thousand Oaks, Calif.: Sage Publications; 1991.

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]

Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, Tosteson TD, Schned AR, Karagas MR. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking, and bladder cancer susceptibility. Carcinogenesis. 2006; 27:1030–1037. [PubMed: 16311243]

Andrew AS, Karagas MR, Nelson HH, Guarrera S, Polidoro S, Gamberini S, Sacerdote C, Moore JH, Kelsey KT, Demidenko E, Vineis P, Matullo G. DNA repair polymorphisms modify bladder cancer risk: A multifactor analytic strategy. Hum Hered. 2008; 65:105–118. [PubMed: 17898541]

Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum. Genet. 2009; 125:63–79. [PubMed: 19052778]

Asselbergs FW, Williams SM, Hebert PR, Coffey CS, Hillege HL, Navis G, Vaughan DE, van Gilst WH, Moore JH. Epistatic effects of polymorphisms in genes from the renin-angiotensin, bradykinin, and fibrinolytic systems on plasma t-PA and PAI-1 levels. Genomics. 2007; 89:362–369. [PubMed: 17207964]

Bateson, W. Mendel’s Principles of Heredity. Cambridge: Cambridge University Press; 1909.

Bellman, R. Adaptive Control Processes. Princeton, N. J.: Princeton University Press; 1961.

Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinform. 2008; 9:238–255.

Bush WS, Dudek SM, Ritchie MD. Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac. Symp. Biocomput. 2009:368–379. [PubMed: 19209715]

Calle ML, Urrea V, Vellalta G, Malats N, Steen KV. Improving strategies for detecting genetic patterns of disease susceptibility in association studies. Stat. Med. 2008; 27:6532–6546. [PubMed: 18837071]

Calle ML, Urrea V, Malats N, Van Steen K. mbmdr: An R package for exploring gene–gene interactions associated with binary or quantitative traits. Bioinformatics. 2010; .26:2198–2199. [PubMed: 20595460]

Cattaert T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD, Edwards TL, Van Steen K. FAM-MDR: A flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. PLoS ONE. 2010; 5:e10304. [PubMed: 20421984]

Chen M, Kamat AM, Huang M, Grossman HB, Dinney CP, Lerner SP, Wu X, Gu J. High-order interactions among genetic polymorphisms in nucleotide excision repair pathway genes and smoking in modulating bladder cancer risk. Carcinogenesis. 2007; 28:2160–2165. [PubMed: 17728339]

Cheverud JM, Routman EJ. Epistasis and its contribution to genetic variance components. Genetics. 1995; 139:1455–1461. [PubMed: 7768453]

Chung Y, Lee SY, Elston RC, Park T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics. 2007; 23:71–76. [PubMed: 17092990]

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. Mahwah, N.J.: Lawrence Erlbaum Associates; 1988.

Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann. Intern. Med. 1993:118201–118210.

Cordell HJ. Epistasis: What it means, what it doesn’t mean, and statistical methods to detect it in humans. Hum Mol Genet. 2002; 11:2463–2468. [PubMed: 12351582]

Cordell HJ. Genome-wide association studies: Detecting gene–gene interactions that underlie human diseases. Nat. Rev. Genet. 2009; 10:392–404. [PubMed: 19434077]

Cowper-Sal Lari R, Cole MD, Karagas MR, Lupien M, Moore JH. Layers of epistasis: genome-wide regulatory networks and network approaches to genome-wide association studies. Wiley Interdiscip Rev Syst Biol Med. 2010 (In Press).

Culverhouse R. The use of the restricted partition method with case-control data. Hum. Hered. 2007; 63:93–100. [PubMed: 17283438]

Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: Limits of models displaying no main effect. Am. J. Hum. Genet. 2002; 70:461–471. [PubMed: 11791213]

Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. Genet. Epidemiol. 2004; 27:141–152. [PubMed: 15305330]

Edwards TL, Turner SD, Torstenson ES, Dudek SM, Martin ER, Ritchie MD. A general framework for formal tests of interaction after exhaustive search methods with applications to MDR and MDR-PDT. PLoS ONE. 2010; 5:e9363. [PubMed: 20186329]

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Review Genetics. 2010; 11:446–450.

Emily M, Mailund T, Hein J, Schauser L, Schierup MH. Using biological networks to search for interacting loci in genome-wide association studies. Eur. J. Hum. Genet. 2009; 17:1231–1240. [PubMed: 19277065]

Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. 1918; 52:399–433.

Freitas AA. Understanding the crucial role of attribute interaction in data mining. Artif. Intel. Rev. 2001; 16:177–199.

Freitas, AA. Data Mining and Knowledge Discovery with Evolutionary Algorithms. New York: Springer; 2002.

Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene-environmental interactions: Comments on different approaches. Am. J. Epidemiol. 1999; 149:689–693. [PubMed: 10206617]

Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am. J. Epidemiol. 2002; 155:478–484. [PubMed: 11867360]

Goldberg, DE. The Design of Innovation. Boston: Kluwer; 2002.

Greene CS, Penrod, N.M, Kiralis J, Moore JH. Spatially uniform reliefF (SURF) for computationally-efficient filtering of gene-gene interacitons. BioData Mining. 2009; 2:5–14. [PubMed: 19772641]

Greene CS, Himmelstein DS, Nelson HH, Kelsey KT, Williams SM, Andrew AS, Karagas MR, Moore JH. Enabling personal genomics with an explicit test of epistasis. Pac. Symp. Biocomput. 2010a: 327–336. [PubMed: 19908385]

Greene CS, Sinnott-Armstrong NA, Himmelstein DS, Park PJ, Moore JH, Harris BT. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. Bioinformatics. 2010b; 26:694–695. [PubMed: 20081222]

Greene CS, Himmelstein DS, Kiralis J, Moore JH. The Informative Extremes: Using Both Nearest and Farthest Individuals Can Improve Relief Alogorithms in the Domain of Human Genetics. EvoBIO 2010, LNCS. 2010c; 6023:182–193.

Griffiths, AJF.; Wessler, SR.; Lewontin, RC.; Carroll, SB. Introduction to Genetic Analysis. 9th ed. New York: W.H. Freeman & Co.; 2008. p. 243

Gui J, Andrew AS, Andrews P, Nelson HH, Kelsey KR, Karagas MR, Moore JH. A robust multifactor dimensionality reduction method for detecting gene–gene interactions with application to the genetic analysis of bladder cancer susceptibility. Ann. Hum. Genet. 2011; 75:20–28. [PubMed: 21091664]

Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003; 19:376–382. [PubMed: 12584123]

Hahn LW, Moore JH. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. In Silico Biology. 2004; 4:183–194. [PubMed: 15107022]

Hamon SC, Stengard JH, Clark AG, Salomaa V, Boerwinkle E, Sing CF. Evidence for nonadditive influence of single nucleotide polymorphisms within the apolipoprotein E gene. Ann. Hum. Genet. 2004; 68:521–535. [PubMed: 15598211]

Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. New York: Springer; 2001.

Becker T. INTERSNP: Genome-wide interaction analysis guided by a priori information. Bioinformatics. 2009; 25:3275–3281. [PubMed: 19837719]

Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics. 2005; 6:95–108.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet. Med. 2002; 4:45–61. [PubMed: 11882781]

Hoh J, Ott J. A train of thoughts on gene mapping. Theor. Popul. Biol. 2001; 60:149–153. [PubMed: 11855949]

Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J. Selecting SNPs in two-stage analysis of disease association data: A model-free approach. Ann. Hum. Genet. 2000; 64:413–417. [PubMed: 11281279]

Hollander WF. Epistasis and hypostasis. J. Hered. 1955; 46:222–225.

Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. New York: John Wiley & Sons; 2000.

Hua X, Zhang H, Zhang H, Yang Y, Kuk AYC. Testing multiple gene interactions by the ordered combinatorial partitioning method in case-control studies. Bioinformatics. 2010; 26:1871–1878. [PubMed: 20538724]

Huang M, Dinney CP, Lin X, Lin J, Grossman HB, Wu X. High-order interactions among genetic variants in DNA base excision repair pathway genes and smoking in bladder cancer susceptibility. Cancer Epidemiol. Biomarkers Prev. 2007; 16:84–91. [PubMed: 17220334]

Kira, K.; Rendell, L. Proc AAAI'92. San Jose, CA: 1992. The feature selection problem: Traditional methods and new algorithm.

Kleinbaum, DG.; Klein, M. Logistic Regression: A Self-Learning Text. New York: Springer-Verlag; 2002.

Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. Genet. Epidemiol. 2001; 21:S626–S631. [PubMed: 11793751]

Lee SY, Chung Y, Elston RC, Kim Y, Park T. Log-linear model-based multifactor dimensionality reduction method to detect gene-gene interactions. Bioinformatics. 2007; 23 2589-90255.

Li W, Reich J. A complete enumeration and classification of two-locus disease models. Hum Hered. 2000; 50:334–349. [PubMed: 10899752]

Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am. J. Hum. Genet. 2007; 80:1125–1137. [PubMed: 17503330]

Lubin JH, Gails MH. On power and sample size for studying features of the relative odds of disease. Am. J. Epidemiol. 1990; 131:552–566. [PubMed: 2301364]

Mahachie JM, Baurecht H, Rodríguez E, Naumann A, Wagenpfeil S, Klopp N, Mempel M, Novak N, Bieber T, Wichmann HE, Ring J, Illig T, Cattaert T, Van Steen K, Weidinger S. Analysis of the high affinity IgE receptor genes reveals epistatic effects of FCER1A variants on eczema risk. Allergy. 65:875–882. [PubMed: 20028371]

Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005; 37:413–417. [PubMed: 15793588]

Mardis E. A decade's perspective on DNA sequencing technology. Nature. 2011; 470:198–203. [PubMed: 21307932]

Martin ER, Hahn LW, Bass M, Ritchie MD, Moore JH. A combined multifactor dimensionality reduction and pedigree disequilibrium test (MDR-PDT) approach for detecting gene-gene interactions in pedigrees. Genet. Epidemiol. 2006; 30:111–123. [PubMed: 16374833]

Mei H, Cuccaro ML, Martin ER. Multifactor dimensionality reduction-phenomics: A novel method to capture genetic heterogeneity with use of phenotypic variables. Am. J. Hum. Genet. 2007; 81:1251–1261. [PubMed: 17999363]

Michalski RS. A theory and methodology of inductive learning. Artif. Intell. 1983; 20:111–161.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. Am. J. Hum. Genet. 2006; 78:15–27. [PubMed: 16385446]

Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum. Hered. 2003; 56:73–83. [PubMed: 14614241]

Moore JH. Computational analysis of gene-gene interactions in common human diseases using multifactor dimensionality reduction. Expert Rev. Mol. Diag. 2004; 4:795–803.

Moore JH. A global view of epistasis. Nat Genet. 2005; 37:13–14. [PubMed: 15624016]

Moore, JH. Genome-wide analysis of epistasis using multifactor dimensionality reduction: Feature selection and construction in the domain of human genetics. In: Zhu, X.; Davidson, I., editors. Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data. Hershey, Penn.: IGI Global; 2007. p. 17-30.

Moore JH, Ritchie MD. The challenges of whole-genome approaches to common diseases. JAMA. 2004; 291:1642–1643. [PubMed: 15069055]

Moore JH, White BC. Exploiting expert knowledge in genetic programming for genome-wide genetic analysis. Lect. Notes Comput. Sc. 2006; 4193:696–977.

Moore JH, White BC. Tuning ReliefF for genome-wide genetic analysis. Lect. Notes Comput. Sc. 2007a; 4447:166–175.

Moore, JH.; White, BC. Genome-wide genetic analysis using genetic programming. The critical need for expert knowledge. In: Riolo, R.; Soule, T.; Worzel, B., editors. Genetic Programming Theory and Practice IV. New York: Springer; 2007b. p. 11-28.

Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics. 2010; 26:445–455. [PubMed: 20053841]

Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. Ann. Med. 2002; 34:88–95. [PubMed: 12108579]

Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. Bioessays. 2005; 27:637–46. [PubMed: 15892116]

Moore JH, Lamb JM, Brown NJ, Vaughan DE. A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. Clin. Genet. 2002a; 62:74–79. [PubMed: 12123491]

Moore JH, Smolkin ME, Lamb JM, Brown NJ, Vaughan DE. The relationship between plasma t-PA and PAI-1 levels is dependent on epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms. Clin. Genet. 2002b; 62:53–59. [PubMed: 12123488]

Moore JH, Gilbert JC, Tsai C-T, Chiang FT, Holden W, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J. Theoretical Biol. 2006; 241:252–261.

Namkung J, Kim K, Yi S, Chung W, Kwon MS, Park T. New evaluation measures for multifactor dimensionality reduction classifiers in gene–gene interaction analysis. Bioinformatics. 2009a; 25:338–345. [PubMed: 19164302]

Namkung J, Elston RC, Yang JM, Park T. Identification of gene–gene interactions in the presence of missing data using the multifactor dimensionality reduction method. Genet. Epidemiol. 2009b; 33:646–656. [PubMed: 19241410]

Neel, JV.; Schull, WJ. Human Heredity. Chicago: University of Chicago Press; 1954.

Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. 2001; 11:458–470. [PubMed: 11230170]

Neter, J.; Wasserman, W.; Kutner, MH. Applied Linear Statistical Models. Chicago: Irwin; 1990.

Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics. 2008; 9:30–50. [PubMed: 17429103]

Pattin KA, Moore JH. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. Hum. Genet. 2008; 124:19–29. [PubMed: 18551320]

Pattin KA, White BC, Barney N, Gui J, Nelson HH, Kelsey KT, Andrew AS, Karagas MR, Moore JH. A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. Genet. Epidemiol. 2009a; 33:87–94. [PubMed: 18671250]

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 1996; 49:1373–1379. [PubMed: 8970487]

Phillips PC. The language of gene interaction. Genetics. 1998; 149:1167–1171. [PubMed: 9649511]

Phillips PC. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics. 2008; 9:855–867.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 2001; 69:138–147. [PubMed: 11404819]

Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet. Epidemiol. 2003; 24:150–157. [PubMed: 12548676]

Shull GH. Duplicate genes for capsule form in *Bursa bursa pastoris*. J. Ind. Abst. Vererb. 1914; 12:97–149.

Sing CF, Stengård JH, Kardia SL. Genes, environment, and cardiovascular disease. Arterioscler. Thromb. Vasc. Biol. 2003; 23:1190–1196. [PubMed: 12730090]

Stone M. Cross-validation: A review. Math. Operationsforsch. Statist. Ser. Statistics. 1978; 9:127–129.

Templeton, AR. Epistasis and complex traits. In: Wolf, J.; Brodie, B., III; Wade, M., editors. Epistasis and the Evolutionary Process. New York: Oxford University Press; 2000.

Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: Analytical retooling for complexity. Trends Genet. 2004; 20:640–647. [PubMed: 15522460]

Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol. 2007; 31:306–315. [PubMed: 17323372]

Wade MJ, Winther RG, Agrawal AF, Goodnight CJ. Alternative definitions of epistasis: Dependence and interaction. Trends Ecol. Evol. 2001; 16:498–504.

Wahlsten D. Insensitivity of the analysis of variance to heredity-environment interaction. Behav. Brain Sci. 1990; 13:109–120.

Wang WY, Barratt BJ, Clayton DG, Todd TA. Genome-wide association studies: Theoretical and practical concerns. Nature Rev. Genet. 2005; 6:109–118. [PubMed: 15716907]

Weiss, KM. Genetic Variation and Human Disease. Cambridge: Cambridge University Press; 1993.

Wilke RA, Reif DM, Moore JH. Combinatorial pharmacogenetics. Nat. Rev. Drug Discov. 2005; 4:911–918. [PubMed: 16264434]

Winham S, Wang C, Motsinger-Reif AA. A Comparison of Multifactor Dimensionality Reduction and L1-Penalized Regression to Idenity Gene-Gene Interactions in Genetic Association Studies. Statistical Applications in Genetics and Molecular Biology. 2011; 10 Iss 1, Art 4.

Zee RY, Hoh J, Cheng S, Reynolds R, Grow MA, Silbergleit A, Walker K, Steiner L, Zangenberg G, Fernandez-Ortiz A, Macaya C, Pintor E, Fernandez-Cruz A, Ott J, Lindpainter K. Multilocus interactions predict risk for post-PTCA restenosis: An approach to the genetic analysis of common complex disease. Pharmacogenomics J. 2002; 2:197–201. [PubMed: 12082592]

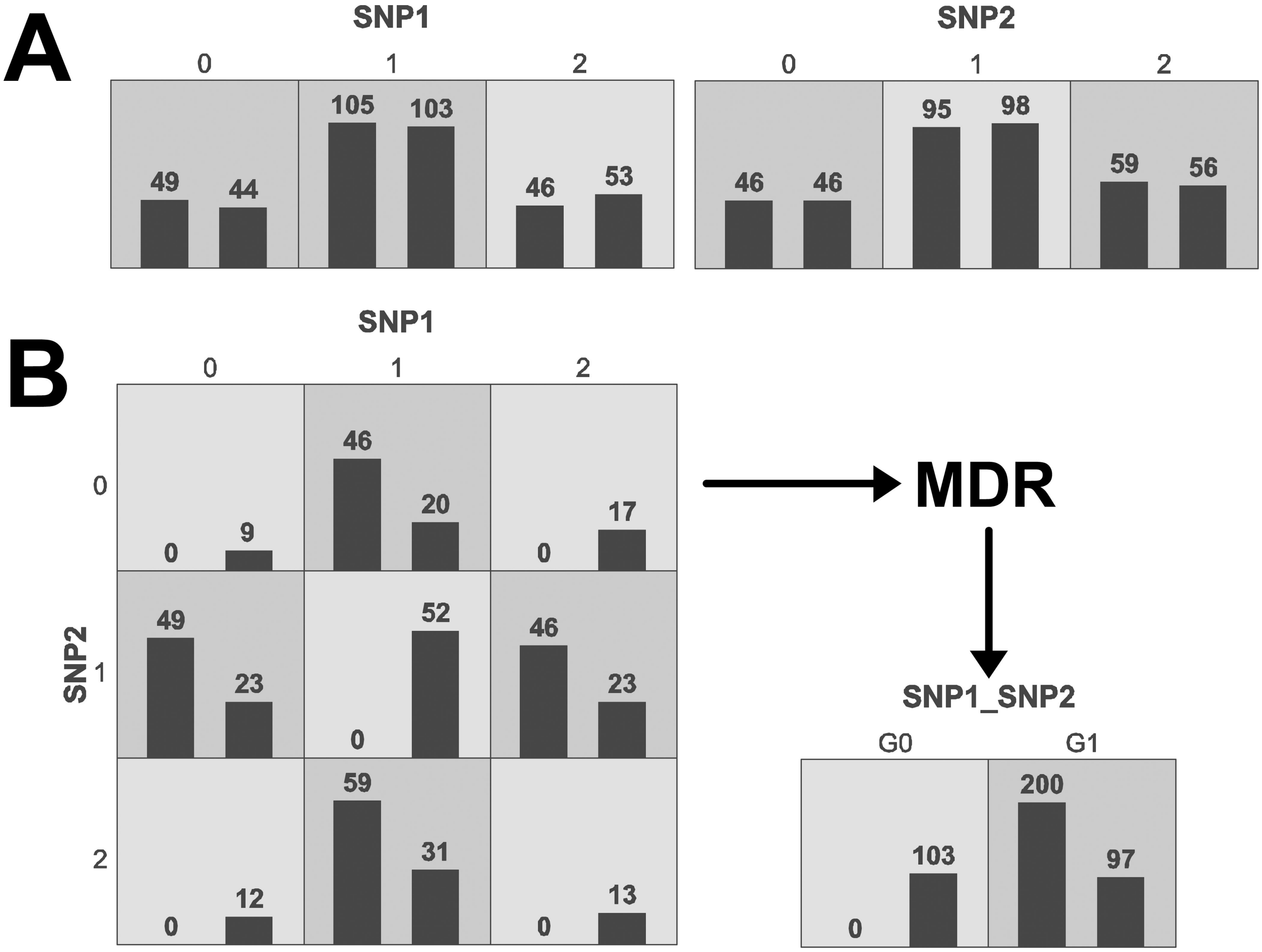

Figure 1.14.1.

(**A**) Distribution of cases (diseased subjects; left bars) and controls (healthy subjects; right bars) across three genotypes (0, 1, 2) for two simulated interacting single nucleotide polymorphisms (SNPs). Note that the ratio of cases to controls for these two SNPs is nearly identical. The dark shaded cells signify "high-risk" genotypes (empirically determined). (**B**) Distribution of cases and controls across nine two-locus genotype combinations. Note that considering the two SNPs jointly reveals larger case-control ratios. Also illustrated is the use of the attribute construction function (see Ritchie et al., 2001 for MDR method) that produces a single attribute (SNP1_SNP2 with two levels, G_0 and G_1) from the two SNPs.

Table 1.14.1

Penetrance Values for Combinations of Genotypes from Two SNPs Exhibiting Interactions in the Absence of Independent Main Effects

		Table penetrance		Margin penetrance
Genotype frequencies	AA (0.25)	Aa (0.50)	aa (0.25)	
BB (0.25)	0	0.1	0	0.05
Bb (0.50)	0.1	0	0.1	0.05
bb (0.25)	0	0.1	0	0.05
Margin penetrance	0.05	0.05	0.05	