Chapter 38
Addressing the Challenges of Detecting Epistasis in Genome-Wide Association Studies of Common Human Diseases Using Biological Expert Knowledge

Kristine A. Pattin
Dartmouth Medical School, USA

Jason H. Moore
Dartmouth Medical School, USA

ABSTRACT

Recent technological developments in the field of genetics have given rise to an abundance of research tools, such as genome-wide genotyping, that allow researchers to conduct genome-wide association studies (GWAS) for detecting genetic variants that confer increased or decreased susceptibility to disease. However, discovering epistatic, or gene-gene, interactions in high dimensional datasets is a problem due to the computational complexity that results from the analysis of all possible combinations of single-nucleotide polymorphisms (SNPs). A recently explored approach to this problem employs biological expert knowledge, such as pathway or protein-protein interaction information, to guide an analysis by the selection or weighting of SNPs based on this knowledge. Narrowing the evaluation to gene combinations that have been shown to interact experimentally provides a biologically concise reason why those two genes may be detected together statistically. This chapter discusses the challenges of discovering epistatic interactions in GWAS and how biological expert knowledge can be used to facilitate genome-wide genetic studies.

DOI: 10.4018/978-1-4666-3604-0.ch038

INTRODUCTION

The fields of human genetics and genetic epidemiology have benefited greatly from the completion of the Human Genome Project in 2003 and the HapMap Project in 2005. The availability of dense maps of single-nucleotide polymorphisms (SNPs) along with high-throughput genotyping technologies has set the stage for routine genome-wide association studies (GWAS) that are expected to significantly improve our ability to identify susceptibility loci across the human genome. To be able to identify genetic variants that are associated with susceptibility to common-complex diseases is an important goal of the aforementioned fields, and the end goal of this endeavor is to utilize these genetic association results to develop better strategies for disease diagnosis, prevention, and treatment.

The GWAS is the current strategy for identifying and characterizing genetic predictors of disease and provides the capability to assess the role of one million or more SNPs in determining disease susceptibility (Hirschhorn & Daly, 2005; Wang et al., 2005). While great strides have been taken to optimize and establish the technical details of measuring a large representative set of SNPs in an accurate and efficient manner (Spencer et al., 2009), the analytical methods for determining which SNPs are important are in their infancy. These methods are based on assumptions such as each SNP having a large and independent effect on disease risk (Clark et al., 2005). It is recognized that most SNPs discovered have small effects on disease susceptibility making them less than ideal targets for medical research or genetic testing. One potential reason for this is that the current analytical framework follows the assumption that each associated SNP will have a detectible effect on disease risk that is independent of all the other variations in the genome as well as independent of the ecological context of each sampled human subject. While the one SNP at a time analytical approach is logical in the sense that it is time efficient and the results are easy to interpret, it is not comprehensive because it fails to acknowledge the complexity of the diseases at hand. If we assume a disease to have a complex genetic architecture, single SNP analyses may only reveal a small portion of the total genetic effects. It is evident that there needs to be an analytical retooling to address the complexity of common diseases (Thorton-Wells et al., 2004).

Common-complex diseases have a much more complex etiology that is due to phenomena such as epistasis (gene-gene interaction), plastic reaction norms (gene-environment interaction), and locus heterogeneity. Therefore, epistasis is a critical genetic component in determining disease susceptibility, where numerous points of genetic variation interact to influence risk. To be able to detect and characterize these interactions is pertinent to our understanding of the biological mechanisms underlying these diseases. However, there are many important challenges that need to be addressed if we wish to completely explore epistasis in a GWAS in order to gain a more coherent understanding of the genetic architecture of a complex trait and its interacting elements. The complexity of the genotype-to-phenotype mapping relationship for common diseases suggests that we are unlikely to identify important genetic variants until we acknowledge and address the many phenomena that create nonlinear patterns in genetic association data (Templeton, 2000; Moore, 2003; Sing et al., 2003; Thornton-Wells et al., 2004; Rea et al., 2006; Moore & Williams, 2009). Not only is there a need for statistical methods powerful enough to model the relationship between SNP interactions and disease susceptibility, but there is a technical challenge that needs to be addressed as well.

In order to detect and characterize epistasis in GWAS, there needs to be an effective way to statistically explore all possible combinations of SNPs, and while many methods have been developed to do so in smaller data sets, analysis of all SNP combinations in GWAS remains computationally daunting with all current existing methods. Re-