Chapter XIV
Prioritizing Disease Genes and Understanding Disease Pathways

Xiaoyue Zhao
Bionovo Inc., USA

Lilia M. Iakoucheva
Rockefeller University, USA

Michael Q. Zhang
Cold Spring Harbor Laboratory, USA

ABSTRACT

Genetic factors play a major role in the etiology of many human diseases. Genome-wide experimental methods produce an increasing number of genes associated with such diseases. This chapter introduces data sources, bioinformatics tools, and computational methods for prioritizing disease candidate genes and identifying disease pathways. The main strategy is to examine the similarity among the candidate genes and known disease genes at the functional level. The authors review different similarity measures and prevailing methods for integrating results from different functional aspects. They hope this chapter will help advocate many useful resources that the researchers can use to investigate diseases of their interest.

INTRODUCTION

Genetic factors play a major role in the etiology of many diseases, including cancers and neurological disorders. Identifying genes that confer increased risk to the disease, and elucidating cellular and molecular processes in which these genes participate are very important problems in biomedical research. Genome-wide experimental methods, such as linkage, association (Botstein & Risch, 2003) and recently copy number variation (CNV) studies (McCarroll & Altshuler, 2007; Sebat, 2007), are all aimed at narrowing down genomic regions containing candidate disease genes. However, due to the linkage disequilibrium and the limited resolution of genome-wide

technologies, the disease-associated regions could contain hundreds of candidate genes. The list of genes produced from such studies is constantly growing. The traditional one-gene-at-a-time approach is a time-consuming and expensive step to validate the disease-causing genes using experimental methods. Therefore it is of great importance and also a challenging task to use computational methods to prioritize disease gene candidates. Computational methods could greatly speed up the efforts directed towards elucidating disease mechanisms and ultimately translating genetic findings into effective prevention, diagnosis and treatment.

The recent availability of a large variety of genomic data and modern high-throughput technologies provide unique opportunities and complementary powerful resources for this purpose. Although disease-gene relationships are not simple (such as different diseases may be caused by mutations in the same gene, and the same disease may be caused by mutations in different genes), disease genes usually share at least some common characteristics including sequence features, expression patterns, involvement in the same protein-protein interaction sub-network, common gene ontology annotations, shared pathways and others (Goh et al., 2007; Oti & Brunner, 2007). For example, it was shown that genes involved in the same disease share up to 80% of their annotations in the GO and InterPro databases (Mulder et al., 2007). The similarity among disease genes is not restricted to the sequences and annotations; the similarity in their functions could also be noted. This leads to the main strategy in prioritizing disease genes, that is, to examine the similarity among candidate genes and known disease genes at the functional level (Han, 2008).

This functional approach is starting to yield new insights into the underlying biology of more and more diseases. For example, Mootha et al. integrated gene co-expression data and tandem mass spectra proteomics data to pinpoint a single candidate gene in a physical map of candidate disease loci for a familial cytochrome c oxidase deficiency LSFC (Mootha et al., 2003). In another study, the PPI network for Huntington's disease (HD) was generated to identify many new interactions and to pinpoint several novel proteins which may be involved in HD (Goehler et al., 2004). Further insights into neuronal toxicity in HD were revealed by another recently developed huntingtin-centered network (Kaltenbach et al., 2007). In the latter study, the authors identified a large number of new proteins that bind to normal and mutant forms of the huntingtin protein using two complementary approaches, high-throughput yeast two-hybrid screening and affinity pull down followed by mass spectrometry. A recent study by Lim et al. found a high degree of connectivity between different ataxia-causing proteins by constructing an ataxia-centered protein interaction network (Lim et al., 2006). Genes associated with breast cancer and a potential link between breast cancer susceptibility and centrosome dysfunction were identified via a network modeling strategy employed by Pujana et al. (Pujana et al., 2007). In the above study, known breast cancer susceptibility genes BRCA1 and BRCA2 were used as baits to generate hundreds of potential functional associations by combining gene expression with functional genomic and proteomic data. In our recent study of autism spectrum disorders (ASD), we tested whether newly identified ASD candidate loci are enriched in genes that are functionally related in transcriptional networks, protein-protein interaction networks, pathways and biological processes (manuscript in preparation) . We found evidence for increased connectivity between autism copy number variant genes and a set of known ASD genes, allowing us to prioritize ASD candidate genes, biological processes and cellular pathways for further studies. Taken together, all these examples suggest that the combination of experimental and computational approaches at a systems biology level could provide important insights into mechanisms of various diseases.

In this chapter, we describe a variety of useful data sources and bioinformatics tools and methods that can help prioritize disease genes and identify disease pathways. A selection of tools and resources is given in Table 1. We present the principles underlying different methods to help understand their advantages and disadvantages, and emphasize the importance of integrating results from different functional aspects. We use a few detailed studies to illustrate the power of this approach.

RESOURCES AND METHODS

When provided with a diverse list of genes associated with a particular disease, there are two frequently asked questions: what genes in this list are functionally related to each other? And what genes are functionally related to the reference genes that have already been implicated in the same disease or diseases with similar phenotypes?