CHAPTER 10

Understanding the Functional Importance of Human Single Nucleotide Polymorphisms

Saurabh Asthana and Shamil Sunyaev*

Abstract

Single nucleotide polymorphisms (SNPs) are the major source of human genetic variation, and the functional subset of SNPs, predominantly in protein coding regions, contributes to phenotypic variation. However, much of the variation in coding regions may not produce any functional effects. There are two broad strategies for classifying polymorphism as functional or neutral: sequence-based methods predict functional significance based on conservation scores calculated from alignments of homologous gene sequences; structure-based methods map variations to known protein structures and predict likely effects based on properties of proteins. Several tools have been developed to classify polymorphism as functional or neutral based on these methods. It was shown that most of functional SNPs are evolutionarily deleterious. Though the utility of the tools has not yet been adequately demonstrated, they may have important applications in the area of medical genetics.

Introduction

The observation that organisms differ from each other extends back to the earliest human history. Aristotle developed taxonomic systems to categorize diverse populations into hierarchies, recognizing that the degree of differentiation between organisms corresponds to the degree of their separation by familial relationships. By now we have come to understand that inherited differences are transmitted via genetic material, and that the differences between individuals must ultimately translate into differences in their genetic sequence. One of the enduring puzzles of biology is understanding variation—what is it that makes sister different from sister? How do these changes in genetic material manifest as changes in outward appearance, behavior or biochemical makeup?

At its most basic level, genetic variation consists of simple changes in sequence—base-pair substitutions, insertions and deletions. What is commonly understood by the term "allele", i.e., two functionally divergent forms of the same gene, in the end might consist of only a single differing nucleotide base-pair. The vast majority (90%) of genetic variation in humans consists of single nucleotide polymorphisms (SNPs).[1] But all of this variation need not translate into observable phenotypic variation; most of it will be functionally neutral. The majority of SNPs occur in intergenic or intronic noncoding regions of the genome. Most noncoding SNPs are unlikely to have a functional impact; only a small minority is believed to have

*Corresponding Author: Shamil Sunyaev—Genetics Division, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Harvard Medical School New Research Building, 77 Ave. Pasteur, Boston, Massachusetts 02115, U.S.A . Email: ssunyaev@rics.bwh.harvard.edu.

Discovering Biomolecular Mechanisms with Computational Biology, edited by Frank Eisenhaber.

functional significance, predominantly due to the effect on gene expression. Of the fraction of SNPs that does occur in coding sequence approximately half are synonymous substitutions, which rarely produce an observable effect on phenotype. Even the remainder, which is guaranteed to result in amino acid variations, does not necessitate functionally divergent protein products.

Although some variants in noncoding regions could have phenotypic effects, the great majority of functional variation likely falls in coding regions. The structure of intergenic regions are also so poorly understood that they are largely impenetrable to analysis at the moment. For this reason, we limit our consideration of variation to nonsynonymous coding SNPs.

Identifying functional variation might be valuable in several contexts. First, we would expect the majority of functional variation to be detrimental (since beneficial variation is believed to be rare). Specifying functionally significant sequence divergence could therefore provide avenues to understanding disease susceptibility. Second, pinpointing the functional significance of nucleotide substitutions between species may shed light on the basic mechanisms of evolution, and reveal how genetic variation is translated into phenotypic variation.

There are a number of strategies we may follow to answer this question.

Comparative Sequence Analysis

Because purifying selection will eliminate variation at functionally important positions, as genes evolve and diverge functionally important positions will show greater conservation between species. Since selection operates exactly on the basis of phenotypic significance, conservation should be expected to be an excellent guide to functionality.

The availability of sequence information from hundreds of species allows the quick retrieval of many protein homologs of a gene of interest. A number of standard techniques exist for constructing multiple alignments of homologous sequences.

A very rude measure of conservation can be obtained simply by examining the degree of entropy at a particular position in a multiple sequence alignment of homologous protein sequences. Low entropy, i.e., high conservation, would suggest the position is important. This simple measure has been shown to be an effective discriminator for functionality.[2]

A more sophisticated measure based on multiple sequence alignments uses position-specific scoring matrices. A probability measure of the likelihood that a variant is permissible (profile score) may be generated for each amino acid at each position. This profile score accounts for phylogenetic history and amino acid frequency as well as conservation. Profile scores are employed in two tools that predict functional variation based on multiple sequence alignments, PolyPhen[3] and SIFT.[4]

Any set of homologous sequences is presumably descended from a common ancestral sequence. Accordingly, common sequence identity may be the result of common descent, making it necessary to segregate the effects of phylogeny from the effects of selection on conservation.

Most profile scores employ sequence-weighting techniques to discriminate between the effect of selection and simple phylogenetic proximity, so that closely related sequences are downweighted. In the ideal each sequence is weighted according to the information it adds to the alignment with regard to the effect of selection. Sequences that are phylogenetically more distant from the others in the alignment will provide the most information—if an amino acid has been conserved across a huge evolutionary distance it is highly likely to be important. Two basic strategies exist for determining evolutionary distance. One is to weight the sequences according to a reconstruction of their phylogenetic tree. The other is to weight sequences according to some metric based on sequence divergence, e.g., pairwise identity or from the degree of identity at aligned positions.[5] Most sequence-weighting techniques apply the same weight to the entire sequence, but some give position-specific weight.

A separate strategy for quantifying conservation is also based on phylogeny, counting the minimum number of amino acid substitutions from an ancestral sequence required to produce the pattern of variation in a multiple sequence alignment.[7]