The Extent of Genetic Variation in Human Genes

Eskender McCoy, Eran Elhaik, and Aravinda Chakravarti
McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

Introduction

The publication of the draft Human Genome in 2001 initiated an era of better understanding of the role of genes and their variants in complex diseases. In the following years, sequencing technology for finding genomic variants, such as single nucleotide polymorphisms (SNPs), have greatly advanced. Sequencing technology has enabled us to create a comprehensive catalog of coding SNPs (cSNPs) that may affect the final protein product. Where GWAS (Genome Wide Association Study) studies have genotyped random SNPs in patients and tested for association with a certain disease, limiting the study to cSNPs focuses the search to specific genes.

We propose a computational method to find genes that are involved in complex disorders using data from GWAS studies. To demonstrate the usefulness of our method we used (The International HapMap Consortium, 2007, Nature, 449, 851-861) GWAS data from a study of blood pressure. In this study, the authors attempted to identify variants associated with blood pressure using HapMap (phase 2) SNPs, most of which were non-coding.

We successfully identified cSNPs associated with hypertension that were not reported before. The candidate genes harboring those cSNPs should be studied in further detail.

Method

The cSNPs catalogue, was obtained from the NCBI website (http://www.ncbi.nlm.nih.gov/projects/SNP) and processed by the Galaxy browser (http://main.g2.bx.psu.edu/). We used NCBI tables: Snp130CodingDbSnp, SnpFunctionCode, and SNPAlleleFreq.

One caveat in our study is that minor allele frequency data obtained from NCBI was not calculated over a large number of samples or populations and therefore may not represent the worldwide allele frequency.

Functional cSNPs were further classified into: synonymous, nonsense, missense, and frameshift. Because only a few hundred cSNPs caused frameshift changes they were removed from the analysis.

We next calculated the LD (Linkage Disequilibrium) between cSNPS and neighboring SNPs (200 Kb upstream and downstream) using HapMap CEU genotypes and selected pairs that exhibit high association (r^2 >=0.8).

We examined the p-values of SNPs from the GWAS study that were strongly associated with cSNPs using a QQ-plot.

cSNPs that were associated with their neighboring SNP and had a significant *p*-value are likely to be associated with hypertension.

(R script was used for all our data analyses.)

Terms used

Genetic Variation: The nucleotide diversity between humans is about 0.1%(Halushka et al, 1999, Nature, 22, 239-247). SNPs along our genome are responsible for most of the variation between any two individuals. Although most gene variants are neutral, some may cause functional changes in proteins resulting in phenotypic differences between Individuals.

HapMap: The international HapMap project attempts to create a haplotype map of the human genome. By mapping all SNPs scientists hope to describe the common patterns of human variation.

Hypertension: High blood pressure (HBP) or hypertension means high pressure (tension) in the arteries: arteries being the vessels that carry blood from the heart to all the tissues and organs of the body.

Data Analysis

Functional category	Total cSNPs in functional category (EXP)	Total cSNPs in functional category (OBS)	Observed data compared with Expected (%)
synonymous	42,944	72,914	70%
nonsense	7,371	3,090	-41%
missense	125,812	98,993	-11%

Table 1: Expected (EXP) number of cSNPs was calculated using the codon table and assuming all possible single nucleotide changes for each codon.

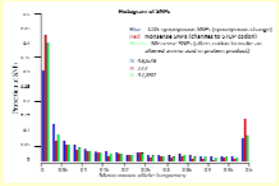

Figure 1 : Minor allele frequency of cSNPs from dbSNP dataset.

The observed frequency of cSNPs in each functional category deviates largely from their expected frequency.

The frequency of synonymous cSNPs was higher than expected because these cSNPs do not affect the gene product and thus do not affect the fitness of the individuals carrying them.

There are fewer nonsense- (-42%) and missense- (-11%) cSNPs than expected because these changes might be deleterious if they occurred in highly conserved proteins. Such deleterious changes would be selected against and be removed from the population over time.

Overall, there were fewer cSNPs than expected. It was estimated that $1.6x10^{-30}$% of the total nucleotides in the exon regions would have SNPs, but only $0.8x10^{-30}$% were actually there. This is most likely due to the fact that cSNPs that are negatively selected against are unobserved.

Frequencies

Because synonymous cSNPs do not change the protein structure and are neither selected for nor against, thus in figure 1 they can be considered as a control for the distribution of nonsense and missense cSNP frequency.

The majority of synonymous cSNPs are of low allele frequency (0-0.1) and their prevalence drops after 0.1.

Comparing the distribution of nonsense and missense cSNP distributions to the distribution of synonymous cSNPs shows a large difference primarily in the lowest and highest bins. As expected, the shape of the distribution differs because the vast majority of the nonsense and missense mutations are deleterious or lethal. Surprisingly, the proportion of nonsense cSNPs with allele frequency of 0.5 is higher than that of both synonymous and missense SNPs, suggesting that these changes may be beneficial.

Identifying cSNPs associated with hypertension

Figure 2a

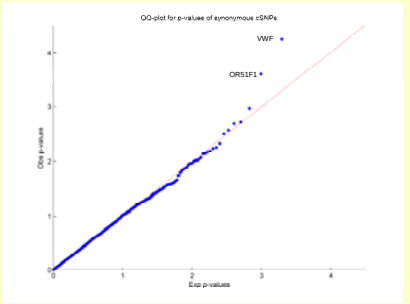

Figure 2b

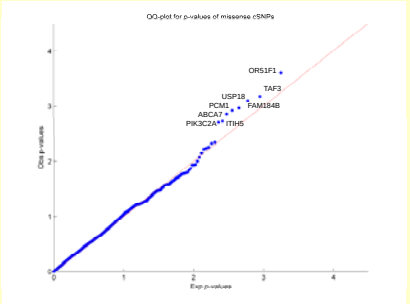

Figure 2a-b: QQ-plots for the p-values of hypertension SNPs. SNPs with the p-values that deviate from a straight are considered to be associated with hypertension. Genes harboring the cSNPs associated with those SNPs are marked.

The hypertension SNPs of interest were identified by plotting their *p*-values against a uniform distribution, so that SNPs unrelated to hypertension would fall on the straight line and those associated with hypertension deviate from it. Figures 2a and 2b mark the SNPs deviating from the straight line that are considered to be associated with hypertension. The cSNPs associated with those SNPs and the genes harboring them are likely candidates to be related with hypertension.

Conclusions

We demonstrated the usefulness of our computational methods to find coding SNPs and genes related with a complex disorder by analyzing GWAS data from a hypertension study. We identified several cSNPs related to hypertension and candidate genes. The exact role of these genes in hypertension should be further studied.

Acknowledgments

I would like to use this space to thank Vick Schneider and everyone else who runs the Center Scholars Program for giving me the another chance to work in a lab this year.