Systematic Analysis of Functional SNP with Type 2 Diabetes Mellitus (T2DM) in the Korean Population

Hyo-Jeong Ban  
hjban81@ngri.go.kr

Keun-Joon Park  
park-kj@ngri.go.kr

Ji-Hong Kim  
jihongkim@ngri.go.kr

Hyun-Woo Han  
hanhw@cdc.go.kr

Division of Bio-Medical Informatics, Center for Genome Science, National Institute of Health, Korea Center for Disease Control and Prevention, 194, Tongil-Lo, Eunpyung-Gu, Seoul 122-701, Republic of Korea

Keywords: single nucleotide polymorphism, support vector machine, bioinformatics, association study

1 Introduction

The relationship between complex disease and genetic sequence variation is complex. A Single Nucleotide Polymorphism (SNP) is a genomic DNA sequence variation at a single nucleotide base pair. Recently, many SNP association studies are identifying many SNP alleles related to various complex disorders. However, it is very difficult to understand the potential biological effect of SNP alleles. The purpose of this study is to find potential functional SNP alleles from human coding and non-coding regions of 97 Type 2 Diabetes Mellitus (T2DM) or non-insulin-dependent diabetes mellitus (NIDDM) related potential genes in the Korean population.

This research was performed as the association study, using the candidate gene approach, in which knowledge of disease pathophysiology is used to develop hypotheses about the association of SNP variants in a specific gene or group of genes with T2DM. At first, several subclasses of data set were constructed based on the clinical and epidemiological data analysis. The association between the case-control status and each individual SNP was estimated from the sub-data set, and we have tried to find the significant SNPs combination related to T2DM development using basic statistical analysis and SVM (Support Vector Machine) discrimination analysis. After these estimations of significant SNP search, we have analyzed the potential impact of each SNP on transcription regulation, protein function, subcellular localization of protein, microRNA target site, RNA splicing, etc.

2 Method and Results

420 SNPs spanning the potential T2DM related genes group were selected from previous genetic association study in a population-based cohort study by the Center for Genome Science, Korea Center for Disease Control and Prevention (KCDC). The previous association study also contains various clinical and epidemiological data for the cohort populations in Korea. This study has been investigated mainly the effects of obesity, hypertension, dyslipidemia, and insulin resistance related phenotypes in the T2DM case-control study of more than 900 Koreans. We performed some subsets of the case-control study by using the clinical and epidemiological data. Then, we evaluated the distribution of 420 SNPs in the potential T2DM related genes in each sub-data set of case-control study. In the present study, we also applied a Support Vector Machine (SVM) to find the important SNP or SNP sets within each case-control sub-data set. To validate the results of our SVM analysis, we performed 10-fold cross-validation test, and the important functional SNP sets were constructed by forward feature selection method [1].

Predicted meaningful SNP sets were investigated the putative biological effects of each SNP using various computational biology analysis tools and databases. We have considered all SNPs in the several different regions of promoter, exon, intron and UTR. Table 1 indicates the definitions of the analysis targets at each region of gene structure, and their potential biological effects. We have considered the change of
these effects between major and minor allele of SNP. Figure 1 shows the data flow diagram from the making of sub-data set to the result of SNP function prediction and ranking.

Table 1: Analysis targets of potential biological effect between major and minor allele of SNP

<table>
<thead>
<tr>
<th>Region of Gene Structure</th>
<th>Analysis Target</th>
<th>Potential Biological Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter</td>
<td>TF binding site</td>
<td>Transcription Factor binding ability</td>
</tr>
<tr>
<td>Exon</td>
<td>Domain and Motif</td>
<td>Conserved Domain and Motif within the coding-regions</td>
</tr>
<tr>
<td></td>
<td>ESE</td>
<td>Affinity of Exonic Splicing Enhancers (ESEs)</td>
</tr>
<tr>
<td>Intron</td>
<td>Splice Site</td>
<td>Splice Site at the exon-intron boundary regions</td>
</tr>
<tr>
<td></td>
<td>ISE</td>
<td>Affinity of Intronic Splicing Enhancers (ISEs)</td>
</tr>
<tr>
<td>UTR</td>
<td>miRNA target</td>
<td>miRNA related post-transcriptional regulation</td>
</tr>
</tbody>
</table>

Figure 1: Work flow from the clinical, epidemiological data and SNPs data to functional SNP analysis in silico.

To validate the results of putative effect of newly determined SNP is difficult, thus we independently have prepared the known disease related SNPs data set from the literature. Although this known example of data set has very limited knowledge, it is very important to find optimal condition of method from several different bioinformatics tools or to assign the useful ranking system.

3 Discussion

We have developed a new functional SNP analysis system in silico, and analyzed our genomic-clinical-epidemiological information combined case-control study data with T2DM in the Korean population. One of the future directions of the research is to incorporate more disease related genomic variation information in order to construct an analysis system of bio-medical information for personalized medicine.

Reference