

SNP Analysis System for Detecting Complex Disease Associated Sites

Yoko Higashi¹, Hirotaka Higuchi¹, Takashi Kido², Hirohito Matsumine², Masanori Baba², Toshihiko Morimoto¹, and Masaaki Muramatsu^{2,3}

¹NTT Data Co. Ltd., Kayabacho Tower, 1-21-2, Shinkawa, Chuo-ku, Tokyo 104-0033, Japan.

²HuBit Genomix, Inc., JowareHanzomonm, 2-19, Hayabusa-cho, Chiyoda-ku, Tokyo 102-0092, Japan

³School of Biomedical Science, Tokyo Medical Dental University, 2-3-10, Kandasurugadai Chiyoda-ku, Tokyo 101-0062, Japan

higashiyu@nttdata.co.jp

Abstract

We developed a system that supports disease association studies to detect genes that may cause complex diseases. The main function of the system is to examine the possibility of each polymorphism being associated with a disease. Another important function is to perform linkage disequilibrium analysis and combine SNPs (Single Nucleotide Polymorphisms) together into LDblocks (Linkage-Disequilibrium-blocks) to improve statistical power for association study. Those analyses can be efficiently performed using an analysis pipeline of the new system with handy tools for eliminating the inadequate data and so on. Consequently, the number of SNPs the system can analyze is about 30 to 50 times higher than by the standard manual procedures per unit of time. The new system also has a sophisticated visualization tool. The main viewer displays the genomic structure and is linked to another main viewer showing the in-depth analysis result. These viewers let the user easily check and make an interpretation of the results. The new system should provide significant assistance for the genome research of complex diseases.

1. Introduction

Complex diseases, such as diabetic mellitus, hypertension and atherosclerosis, are thought to be caused by multiple genetic and environmental factors. While candidate causative genes of monogenic diseases, such as cystic fibrosis and Huntington's Disease, have been identified through parametric linkage analysis of the families of affected individuals [1], as for complex diseases for which the disease model parameters can't be set because of their low penetration rate, case-control association studies have been preferred. Although several candidate genes have been identified [2], more causative candidate genes are thought to remain unidentified. Inquiring studies of the causative polymorphisms will no doubt grow in number.

2. Method

2.1. Case-Control Association Study

A typical case-control association study is to examine the possibility of each polymorphism being associated with a disease by comparing the genotype frequency of cases with that of controls by the statistical tests, such as chi-square test and likelihood test [3]. Our system includes likelihood test algorithm, and regards a polymorphism whose p-value is under a threshold as strongly associated with a disease. Since the disease-associated genes are likely to be scattered over the whole genome, a genome-wide analysis needs to be performed. This, however, leads to the problem of multiple tests. For example, if the threshold is set 0.01, 1 % of them may become type I errors. Bonferroni's correction is usually applied to avoid this problem, but it is so strict that even the correct polymorphisms would be undetected. Therefore, it is the challenge to reduce the number of polymorphisms to be tested by combining several SNPs together into a meaningful unit.

2.2. Constructing haplotypes in Linkage-Disequilibrium-block (LDblock)

As for a meaningful unit, two points had to be called into account. (i) There are genomic regions inherited in clusters from generation to generation, which we call LDblocks (Linkage-Disequilibrium-blocks). It is in the LDblock that the disease associated polymorphisms reside and are inherited. (ii) Several polymorphisms in a gene are thought to contribute together to devastating effects on the protein products, e.g. producing abnormal quality and/or quantities of proteins. It seems useful to regard a gene as a unit. Thus our system automatically detect LDblocks in each gene. The genotype in a LDblock is referred to as diplotype, a pair of haplotypes (sequence on each gamete). Diplotypes of subjects are constructed by linkage disequilibrium analysis, for which there are three popular methods: the

parsimony method [4], the expectation and maximization (EM) algorithm [5], and the Bayesian statistical method [6][7]. Since the parsimony method seems less accurate than the others [8], and the Bayesian statistical algorithm takes far longer time than the others [8], our system includes an EM algorithm implemented in the software, LDSUPPORT [9]. Originally LDblocks had been manually detected using various linkage disequilibrium measures, such as Lod Score. We developed an LDblock-auto-detecting algorithm and implemented it in the system so that the LDblocks would not have to be manually detected.

2.3. Programming Language and Platform

The analysis pipeline of the system was written in C, Perl and Java. And the main visualization tool employs SVG. The system currently runs on Linux, but in theory, it should also be able to run on UNIX and Windows as compiled for each platform. We will make UNIX and Windows version in the future.

3. Results

The number of polymorphisms one can analyze using our system is about 30 to 50 times higher than by the standard manual procedures per unit of time. Although separate analysis software, such as PHASE [6] and EH [10], has already existed and statistical analysis system (SAS) can also be used, analysis of a massive number of polymorphisms using separate pieces of softwares is not feasible. Furthermore, there is no other software which automatically improves statistical power by detecting LDblocks and constructing haplotypes. Our visualization tool shows genomic structure accompanied with the analysis result (Figure 1). As real causative polymorphisms are likely to be located in the vicinity of the analyzed polymorphisms rather than themselves, it is important to examine the surrounding genes and polymorphisms.



Figure 1. The viewer for genome structure

4. Discussion

The software analysis package, HapScope, is equipped with an analysis pipeline and a sophisticated visualization tool [11]. Its main function is SNP discovery and haplotype construction in the candidate region. Unlike HapScope, the main function of our system is to examine the possibility of each polymorphism being associated with a disease. To maximize statistical power, our system automatically determines the proper sizes of LDblocks and the appropriate numbers of haplotypes. Since any statistical tests other than likelihood test currently used by our system are applicable, genotype and haplotype data files can be exported for other analyses, increasing the user's freedom to choose the analysis. Equipped with the informative viewers, our system should provide a significant assistance for the research of complex diseases.

5. References

[1]J.M.Rommens et al, "Identification of the cystic fibrosis gene: chromosome walking and jumping", *Science* **245**: 1059-1065 (1989)

[2]K.Ozaki et al, "Functional SNPs in the lymphotoxin- α gene that are associated with susceptibility to myocardial infarction", *Nature Genet.* 1047 (2002)

[3]D.Fallin et al., "Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease", *Genome Res* **11**: 143-151 (2001)

[4]A.G.Clark, "Inference of haplotypes from PCR-amplified samples of diploid populations", *Mol.Biol.Evol.* **7**: 111-122

[5]L.Excoffier and M.Slatkin, "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population", *Mol.Biol.Evol.* **12**: 921-927

[6]M.Stephens et al., "A New Statistical Method for Haplotype Reconstruction from Population Data", *Am.J.Hum.Genet.* **68**: 978-989 (2001)

[7]T.Niu et al., "Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms", *Am.J.Hum.Genet.* **70**: 157-169 (2002)

[8]C.F.Xu et al., "Effectiveness of computational methods in haplotype prediction", *Hum.Genet.* **110**: 148-156

[9]Y.Kitamura et al., "Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm", *Ann.Hum.Genet.* **66**: 183-193

[10]X.Xi and J.Ott, "Testing linkage disequilibrium between a disease gene and marker loci", *Am.J.Hum.Genet.* **53**: 1107 (1993)

[11]J.Zhang et al., "HapScope: a software system for automated and visual analysis of functionally annotated haplotypes", *Nucl.Acid.Res.* **30**: 5213-5221 (2002)