

SEARCH FOR MULTI-SNP DISEASE ASSOCIATION

*D. Brinza, A. Pereygin, M. Brinton, A. Zelikovsky**

Georgia State University, Atlanta, GA, USA

*Corresponding author; e-mail: alexz@cs.gsu.edu

Keywords: algorithm, disease association, SNP, genotypes

Summary

Motivation: Recent improvements in the accessibility of high-throughput genotyping have brought a great deal of attention to association studies for common complex diseases. The two-loci analysis was recently developed (Marchini et al, 2005) but multi-loci analyses are expected to find even stronger disease associations.

Results: A novel combinatorial method for finding disease-associated multi-SNP combinations was developed. Multi-SNP combinations significantly associating with diseases were found. For Crohn's disease data (Daly, et al., 2001), a few associated multi-SNP combinations with multiple-testing-adjusted to $p < 0.05$ were found, while no single SNP or pair of SNPs showed significant association. For a dataset for an autoimmune disorder (Ueda, et al., 2003), a few previously unknown associated multi-SNP combinations were found. For tick-borne encephalitis virus-induced disease, a multi-SNP combination within a group of genes showing a high degree of linkage disequilibrium significantly associated with the severity of the disease was found.

Availability: <http://alla.cs.gsu.edu/~software/DACS>

Introduction

Analysis of variation in suspected genes in disease and nondisease individuals is aimed at identifying SNPs with considerably higher frequencies among the disease individuals than among the nondisease individuals. Successful (as well as unsuccessful) searches for SNPs with statistically significant associations have recently been reported. Although common diseases can be caused by combinations of several unlinked gene variations, most searches are done on a SNP-by-SNP basis. In this paper, we address the computational challenge of searching for such multi-gene causal combinations.

False-discovery rates are usually very high for genome-wide searches. Although only statistically significant SNPs (with a $p < 0.05$ for frequency distribution) are reported, frequently these findings are not reproducible because the computed p-values are not adjusted for multiple testing. The standard Bonferroni adjustment is overly pessimistic; therefore, we adjusted for multiple testing by using a more accurate randomization method.

Formally, the computational problem is as follows. Given a population of disease and nondisease n genotypes or haplotypes with values of m SNPs, find all multi-SNP combinations with multiple-testing adjusted to $p < 0.05$ for the frequency distribution. We show that this problem is computationally feasible using the proposed novel searching techniques.

Methods and Algorithms

The search for disease-associated multi-SNP combinations, i.e., with the p-value of the frequency distribution below 0.05, among all possible combinations can be done by *Exhaustive Search* (ES) (e.g., Marchini et al, 2005). Since ES checks all 1-SNP, 2-SNP, ..., m -SNP combinations, its runtime is $O(n3^m)$ making it unfeasible even for small numbers of SNPs m . Further we searched only among multi-SNP combinations with $k < 3$ SNPs. We refer to k as the *search level* of the exhaustive search. In order to reduce the runtime of the exhaustive search, we propose to decrease the size of the input data set by extracting informative SNPs (*indexing SNPs*) from which one can reconstruct all other SNPs. In our experiments, we used a multiple linear regression based tagging method (He and Zelikovsky, 2006). The tradeoff between the number of chosen indexing SNPs and quality of reconstruction requires choosing the maximum number of index SNPs that can be handled by ES in a reasonable computational time. ES on indexing SNPs will be further referred as *Indexing Exhaustive Search* (IES).

Another way to fight extensive computations is to apply a faster search. Our new search method can find disease-associated multi-SNP combinations consisting of large numbers of SNPs and small search levels k . The new method is based on the notion of closure. Let C be a multi-SNP combination and let $snp(C)$ be the subset of SNPs with their values defining C . All individuals containing $snp(C)$ are partitioned into two subsets: $dis(C)$ consisting of individuals with disease and $nondis(C)$ consisting of individuals without disease. We search for C 's with larger disease frequency $dis(C)$ and lower nondisease frequency $nondis(C)$. Sometimes, the size of $nondis(C)$ can be decreased while keeping $dis(C)$ unchanged using the following closure operator. The set $dis(C)$ can have more SNP values in common than in $snp(C)$. Closure of C is a multi-SNP combination C' with $snp(C')$ equal to $snp(C)$ extended with such SNPs. Obviously, $dis(C') = dis(C)$ while $nondis(C') \subseteq nondis(C)$. The proposed *Combinatorial Search* (CS) finds the disease-associated multi-SNP combinations among closures of all j -SNP combinations ($j=1..k$, $k < m$). The corresponding *search level* is the number of SNPs k in multi-SNP combinations for which a closure is found. Because of the disease-closure, the same level of searching using a combinatorial search finds better associations than an exhaustive search. CS on indexing SNPs will be further referred as *Indexing Combinatorial Search* (ICS).

Implementations and Results

In our implementation, we avoid checking of those multi-SNP combinations which can not lead to statistically significant ones. Additionally, we never recheck any combinations after disease-closure that will form already checked combinations. The resulting implementation is several times faster than the exhaustive search.

Results from four methods used to search disease-associated multi-SNP combinations are reported for the following datasets. The first dataset was derived from human Chromosome 5q31, which may contain a genetic variant responsible for Crohn's disease, by genotyping 103 SNPs for 144 disease and 243 nondisease individuals (Daly. et al., 2001). The second dataset consisted of 108 SNPs sequenced from 330 kb of human DNA containing the genes, CD28, CTLA4 and ICONS, that were previously shown to be related to an autoimmune

disorder, from 384 disease and 652 nondisease individuals (Ueda. et al., 2003). The tick-borne encephalitis virus-induced dataset consists of 41 SNPs genotyped from DNA of 21 patients with severe tick-borne encephalitis virus-induced disease and 54 patients with mild disease. The missing genotypes were inferred using 2SNP software (Brinza and Zelikovsky, 2006).

The four methods compared are *Exhaustive Search* (ES), *Indexed Exhaustive Search* (IES(N))-ES on the indexed datasets obtained by extracting N indexed SNPs, *Combinatorial Search* (CS), *Indexed Combinatorial Search* [ICS(N)]-CS on the indexed datasets. Each method has been applied only to the search levels of 1 and 2. All experiments were run on a Processor Pentium 4 3.2Ghz, RAM 2Gb, OS Linux. The runtime is given in the last column of Table 1.

Table 1. 4 Methods for searching disease associated multi-SNPs combinations.

Search level	Search method	MT-unadjusted p corresp. to adjusted p=0.05	SNP combination with minimum p-value			# of SNP combin. with MT-adjusted p<0.05	Run-time msec
			expos. freq.	unexp. freq.	unadjust. p-value		
Dataset of Crohn's disease (Daly.et al., 2001)							
1	ES	1.6×10^{-3}	0.31	0.16	1.8×10^{-3}	0	900
	IES(30)	3.9×10^{-3}	0.30	0.16	4.7×10^{-3}	0	500
	CS	5.1×10^{-5}	0.30	0.11	2.0×10^{-5}	2	1000
	ICS(30)	2.2×10^{-3}	0.30	0.14	4.6×10^{-4}	1	600
2	ES	1.9×10^{-5}	0.30	0.13	3.1×10^{-4}	0	1500 0
	IES(30)	1.0×10^{-4}	0.31	0.14	4.4×10^{-4}	0	1000
	CS	1.5×10^{-6}	0.17	0.02	6.5×10^{-7}	2	7000
	ICS(30)	5.0×10^{-5}	0.17	0.04	3.7×10^{-5}	1	400
Dataset of autoimmune disorder (Ueda.et al., 2003)							
1	ES	1.3×10^{-3}	0.43	0.28	1.1×10^{-4}	2	1000
	IES(30)	3.1×10^{-3}	0.43	0.28	1.1×10^{-4}	4	600
	CS	1.8×10^{-4}	0.43	0.28	9.2×10^{-5}	2	1100
	ICS(30)	1.6×10^{-3}	0.43	0.28	1.1×10^{-4}	4	600
2	ES	2.7×10^{-6}	0.25	0.12	1.5×10^{-6}	2	3000 0
	IES(30)	8.0×10^{-5}	0.25	0.12	1.5×10^{-6}	9	3000
	CS	1.1×10^{-6}	0.16	0.06	8.5×10^{-7}	3	2000 0
	ICS(30)	4.7×10^{-5}	0.25	0.12	1.1×10^{-6}	10	1000
Dataset of tick-borne encephalitis virus-induced disease							
1	ES	6.1×10^{-3}	0.33	0.07	1.5×10^{-2}	0	80
	IES(20)	9.4×10^{-3}	0.33	0.07	1.5×10^{-2}	0	30
	CS	4.8×10^{-4}	0.33	0	1.3×10^{-4}	1	84
	ICS(20)	8.1×10^{-4}	0.33	0.02	8.1×10^{-4}	1	35
2	ES	2.5×10^{-4}	0.29	0	4.8×10^{-4}	0	820
	IES(20)	1.3×10^{-4}	0.29	0	4.8×10^{-4}	0	100
	CS	4.3×10^{-5}	0.33	0	1.3×10^{-4}	0	600
	ICS(20)	1.3×10^{-4}	0.29	0	4.8×10^{-4}	0	76

The relative qualities of the searching methods are compared using the number of statistically significant multi-SNP combinations found (Table 1, column 7). The statistical significance was adjusted to multiple testing and the adjusted 0.05 threshold is shown (third column of Table 1). In the 4th, 5th and 6th columns, we give the frequencies of the best multi-SNP combination among disease and nondisease populations and the unadjusted p-value, respectively.

Discussion

Comparing indexed counterparts with ES and CS shows that indexing is quite successful. Indeed, the indexed searches found the same multi-SNP combinations as the non-indexed searches for the second and third data sets but were much faster and the multiple-testing adjusted 0.05-threshold was higher and easier to meet.

Comparing the CS with the ES counterparts is advantageous to the former. Indeed, for the Crohn's disease data (Daly.et al., 2001), the ES on the first and second search levels is unsuccessful while the CS finds several statistically significant multi-SNP combinations. Similarly, for the tick-borne encephalitis virus-induced disease data, the CS and ICS(20) found a significant association on the first level while no association was found by the ES or IES(20). For the autoimmune disorder data (Ueda.et al., 2003), the CS found many more statistically significant multi-SNP combinations than the ES.

In addition, we have developed a new disease susceptibility prediction (DSP) method based on CS. In a leave-one-out test for the tick-borne encephalitis data, the accuracy of DSP is 90% which proves that the data contain a well-defined border between severe and mild forms. The accuracy of DSP is around 85% for the other two datasets which is significantly higher than the accuracy of previously known methods. These results show that the combinatorial disease-association search is more powerful than the existing methods when applied to disease susceptibility prediction.

We conclude that the proposed indexing approach and the combinatorial search method are very promising techniques for searching for statistically significant diseases-associated multi-SNP combinations and disease susceptibility prediction.

Acknowledgements

DB was supported by GSU Molecular Basis of Disease Fellowship, AZ was supported by NIH Award 1 P20 GM065762-01A1 and US CRDF Award MOM2-3049-CS-03, and AP was supported by CDC grant R01 CI000216.

References

- Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) High resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229-232.
- Brinza, D. and Zelikovsky, A. (2006) 2SNP: Scalable Phasing Based on 2-SNP Haplotypes, *Bioinformatics*, **22(3)**, 371--373.

Ueda, H., Howson, J.M.M., et al. (2003) Association of the T Cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease, *Nature*, **423**, 506-511.

He, J. and Zelikovsky, A. (2006) Tag SNP Selection Based on Multivariate Linear Regression, *Proc. of Intl Conf on Computational Science (ICCS 2006)* (to appear).

Marchini, J., Donnelly, P. and Cardon, L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases, *Nature Genetics*, **37**, 413-417.