Assessing Collective Disease Association of Multiple Genomic Loci

Marzieh Ayati [1], Pamela L. Clark [2,3], Mark R. Chance[2] and Mehmet Koyutürk [1,2]

[1] Department of Electrical Engineering and Computer Science, [2] Center for Proteomics and Bioinformatics

[3] Systems Biology and Bioinformatics Program, Case Western Reserve University, Cleveland, OH 44106, USA

Abstract

In recent years, genome-wide association studies (GWAS) have successfully identified loci that harbor susceptible genetic variants for a large number of complex diseases. However, susceptible loci identified by GWAS so far generally account for a limited fraction of the genotypic variation in patient populations. Predictive models based on identified loci also have modest success in classifying phenotype (risk assessment) and therefore are of limited practical use. More recently, there has been considerable attention on identifying epistatic interactions; i.e., the improved association of pairs of loci with the disease as compared to the aggregation of the two individual loci. However, the large number of pairs to be tested for epistasis poses significant challenges, in terms of both computational (runtime) and statistical (multiple hypothesis testing) considerations. Here, we propose a new criterion, termed difference in allele distributions (DAD), for assessing the collective association of multiple genetic variants with a disease of interest. DAD is based on the comparison of the distribution of interested alleles for a set of genomic loci among case and control samples, using Kolmogorov-Smirnov statistics. This formulation of the coordination among multiple variants allows employment of efficient heuristic algorithms for the identification of sets of multiple loci that are collectively associated with disease. This formulation also enables application of permutation tests for empirical assessment of statistical significance, thereby directly correcting for multiple hypotheses. We test the proposed method on two independent data sets for two complex diseases, Psoriasis and Type 2 Diabetes (T2D), in terms of the statistical significance of identified associations and the performance of resulting risk assessment models. Our results show that, as compared to individual variants, multi-variant features provide better predictive performance in risk assessment and they are also more reproducible.

Introduction

Earlier GWAS focused on identifying individual loci associated with diseases using standard statistical tests comparing the distribution of genotypes or minor alleles in case and control populations. However, increasing empirical evidence from model organisms and human studies suggests that complex interactions among two or more loci contribute broadly to complex traits. Indeed, the individual locus associations identified in GWAS are often not reproducible. An increasing number of studies report the presence of statistically significant epistatic interactions in complex diseases. Nevertheless, identifying epistatic interactions remains a computationally and statistically challenging problem. The computational challenges are tackled to a certain extent by algorithms that prune out certain pairs of loci without explicitly testing them.

Methods

$$m(c,s) = \begin{cases} 2 & \text{if } g(c,s) \text{is Homozygous of minor allele} \\ 1 & \text{if } g(c,s) \text{is Heterozygous} \\ 0 & \text{otherwise} \end{cases}$$

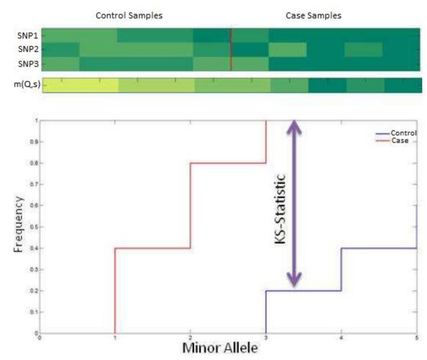
For a given set of SNP Q and set s of samples:

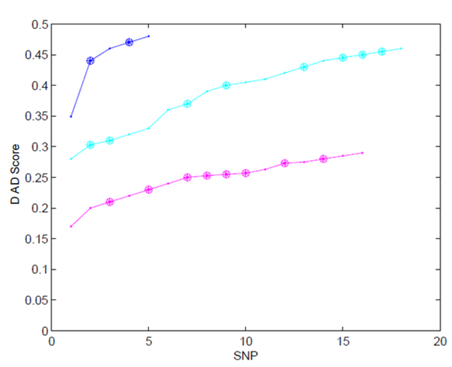
$$m(Q,s) = \sum_{c \in Q} m(c,s)$$

We define the score of subset as a Difference of Allele Distributions (DAD).

$$F_Q^{(\pi)}(k) = \frac{|\{s \in S : m(Q,s) \leq k \text{ and} f(s) = \pi\}|}{|\{s \in S : f(s) = \pi\}|}$$

$$K_Q = \sup_{0 \leq k \leq 2|Q|} |F_Q^{(1)}(k) - F_Q^{(0)}(k)|$$





Results

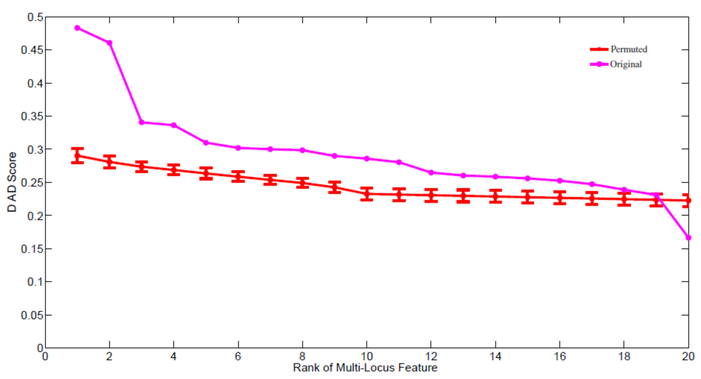
We use two GWAS datasets for two diseases T2D and PS

- Wellcome Trust Case-Control Consortium (WTCCC) [1,2]
- Database of Genotypes and Phenotypes (dbGaP) [3,4]

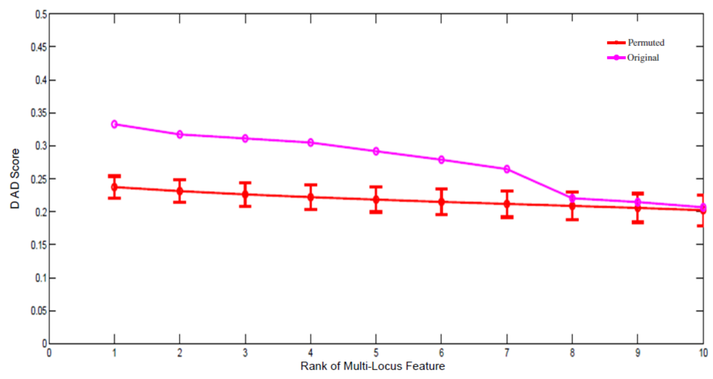
Statistical Significance of High Scoring Multi-Locus Sets:

We assess the statistical significance of each high-scoring locus set empirically, using a phenotype permutation test that preserves the distribution of genotypes within and between samples

Type 2 Diabetes



Psoriasis



Performance of Identified Feature in Risk Assessment:

A model for risk assessment is built using a Naive Bayes classifier.

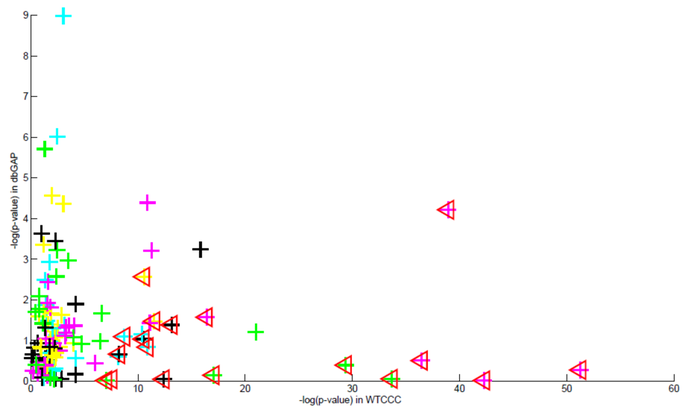
We assess the performance of the significant features in risk assessment

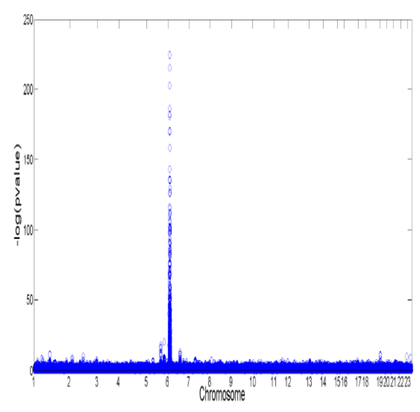
- Cross validation
- Cross-classification on an independent dataset

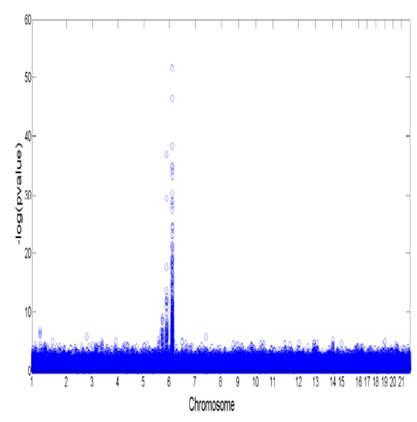
$$P(f(s) = \pi | m(Q_1,s), \ldots, m(Q_n,s)) = P(f(s) = \pi) \prod_{i=1}^{n} P(m(Q_i,s) | f(s) = \pi)$$

Feature Type	Number of Features in Final Model	Mean of Multi-Loci Feature Size	AUC Cross Validation	AUC Independent Data Set
Multi-locus	17	16	0.84(0.0001)	0.50
Single-locus	18	-	0.76(0.0008)	0.47

Feature type	Number of Features in Final Model	Mean of Multi-Loci Feature Size	AUC Cross Validation	AUC Independent Dataset
Multi-locus	7	18	0.81	0.62
Single-locus	4	-	0.75	0.65







Conclusion

The results show that the multiple-locus features may contain loci that are not individually significant on the discovery data, but are individually significant on the validation data, suggesting that the proposed multi-locus based model can extract relevant markers that are not readily captured by isolated analysis of individual loci. For this reason, multi-locus markers may serve as more reliable features for risk assessment. Moreover, the result suggests that the major allele also can be associated with the disease.

Acknowledgements

This work was supported by National Institutes of Health Grant R01-LM011247.

References

[1] W. T. C. C. Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447:661–678, 2007.

[2] Strange A, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet. 2010

[3] McCarty CA, Chisholm RL, et al. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011

[4] R. P. Nair, K. C. Duffin, and et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kB pathways. Nature genetics, 2009.