A Tool for Detecting Complementary Single Nucleotide Polymorphism Pairs in Genome-Wide Association Studies for Epistasis Testing

GIZEM CAYLAK,1 OZNUR TASTAN,2 and A. ERCUMENT CICEK1,3

ABSTRACT

Detecting interacting loci pairs has been instrumental to understand disease etiology when single locus associations do not fully account for the underlying heritability. However, the number of loci to test is prohibitively large. Epistasis test prioritization algorithms rank likely epistatic single nucleotide polymorphism (SNP) pairs to limit the number of statistical tests. Potpourri detects epistatic SNP pairs by diversifying the selected SNPs’ genomic regions and investigating their co-occurrence patterns over the case cohort. It can also input and further prioritize SNPs in regulatory or coding regions. The program identifies and returns a list of prioritized SNP pairs for epistasis testing. This article describes how to use the program and the details of the input and output data.

Keywords: complementation, diversification, epistasis test prioritization, population cover, submodular optimization.

1. INTRODUCTION

Exhaustive identification of epistatic loci, even just pairs, is intractable for large genome-wide association studies (GWAS) due to large number of single nucleotide polymorphisms (SNPs). A common workaround is to minimize the number of tests by prioritizing the pairs of SNPs to be tested. Potpourri works on this principle of prioritizing epistasis tests to gain statistical power. The algorithm couples SNPs that are individually informative but are located at diverse genomic regions by selecting such pairs through maximizing a submodular function. A similar approach has been shown to be useful for feature selection for phenotype prediction (Yilmaz et al., 2019). Subsequently, Potpourri performs epistasis tests for SNP pairs that also cocover the case cohort. The algorithm can also further prioritize SNPs in coding and regulatory regions.

2. APPLICATION

Potpourri is implemented in MATLAB as a set of scripts. Underneath, it also uses the scripts from the related methods from the literature (Cowman and Koyutürk, 2017; Yilmaz et al., 2019). The only
dependency required for the software is the BOOST C++ library. A makefile is provided. Only installation step needed is typing “make” or running the build_mex.m script in the code directory.

The first input \((X)\) is a matrix wherein each row represents a sample, and each column represents the loci for which a variation is detected in the GWAS. The entries of \((X)\) represent the genotype of the sample in that locus with respect to number of minor alleles observed: homozygous major (0), heterozygous (1), and homozygous minor (2). The second input \((Y)\) is a binary vector representing the label of each sample. The third input \((W)\) represents the undirected SNP–SNP network as a sparse binary matrix. The input \((k)\) is the number of SNPs to be included in the diverse SNP set that constitutes the seeds for \(R_x \in R\). \(b\) sets the number of upstream and downstream neighbor SNPs to be included for each seed for \(R_x \in R\). The next input \((R)\) is a binary vector that represents whether an SNP is in a regulatory or coding region and \(\omega\) is the constant using which Potpourri increases the prize of such SNPs for making their selection more likely. \(\omega = 1\) indicates that this information is ignored. SNP_info is a matrix that includes the RS ID, chromosome number, and the genomic position for every considered SNP in \(X\). Finally, maxMarginal-Significance is an integer value that sets a \(p\)-value threshold to eliminate SNPs with high marginal effects as discussed in Cowman and Koyutürk (2017). It is between 1 and 6, the latter being stricter.

The program generates multiple result files with different extensions. The name for all output files is provided to the function through the outputFileName parameter. The program outputs two main result files along with some other temporary files: the summary file and the reciprocally epistatic SNP pairs list file. Summary file reports statistics about the run: (i) the number of SNPs input, (ii) the number of case samples input, (iii) maximum marginal significance threshold, (iv) the number of tests performed, and (v) the number of reciprocally epistatic pairs reported. reciprocalPairs.formatted file reports reciprocally epistatic SNP pairs. Here, each row represents a pair. First, the chi-square statistic for the pair is given. Then, the SNP IDs, chromosome numbers, and positions of each SNP are listed in this order.

3. CONCLUSION

Potpourri is a new epistasis test prioritization technique. In this study, we provide details of how to use its MATLAB-based implementation. The implementation requires \(~23\) minutes to process: \(~375k\) SNPs and selecting \(1000\) SNPs as seeds on an Intel(R) Xeon(R) E5-2650 v3 Ten-Core Haswell Processor (2.3 GHz 25 M 9.6 GT/s QPI) when using 251 GB RAM.

ACKNOWLEDGMENTS

We thank WTCCC and the study participants. We used SPADIS and LINDEN implementations in our study. We thank S. Yilmaz, T. Cowman, and M. Koyuturk. A.E.C. acknowledges the support of TUBA GEBIP award. O.T. and A.E.C. acknowledge the support of Bilim Akademisi BAGEP awards.

AVAILABILITY AND IMPLEMENTATION

The software is available at http://ciceklab.cs.bilkent.edu.tr/potpourri

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

This study was supported by TUBITAK through Career Grant no. 116E148 to A.E.C.
REFERENCES


Address correspondence to:
Dr. A. Ercument Cicek
Computer Engineering Department
Bilkent University
Cankaya, Ankara
06800, Turkey

E-mail: cicek@cs.bilkent.edu.tr