Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw418
Advance Access Publication Date: 13 July 2016
Applications Note

OXFORD

Genetic and population analysis

LAMPLINK: detection of statistically significant SNP combinations from GWAS data

Aika Terada[1,2,3,*], Ryo Yamada[4], Koji Tsuda[2,3,5] and Jun Sese[3,6,*]

[1]PRESTO, Japan Science and Technology Agency, Saitama 332-0012, Japan, [2]Department of Computational Biology and Medical Science, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan, [3]Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan, [4]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, 606-8507 Japan, [5]Center for Materials Research by Information Integration, NIMS, Ibaraki, 305-0047 Japan and [6]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on February 22, 2016; revised on June 20, 2016; accepted on June 25, 2016

Abstract

Summary: One of the major issues in genome-wide association studies is to solve the missing heritability problem. While considering epistatic interactions among multiple SNPs may contribute to solving this problem, existing software cannot detect statistically significant high-order interactions. We propose software named LAMPLINK, which employs a cutting-edge method to enumerate statistically significant SNP combinations from genome-wide case–control data. LAMPLINK is implemented as a set of additional functions to PLINK, and hence existing procedures with PLINK can be applicable. Applied to the 1000 Genomes Project data, LAMPLINK detected a combination of five SNPs that are statistically significantly accumulated in the Japanese population.
Availability and Implementation: LAMPLINK is available at http://a-terada.github.io/lamplink/.
Contact: terada@cbms.k.u-tokyo.ac.jp or sese.jun@aist.go.jp
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Background

Genome-wide association studies (GWASs) have identified hundreds of loci associated with various complex human traits (Welter *et al.*, 2014). These studies conduct screening of individual single nucleotide polymorphisms (SNPs) using statistical tests to assess the association of each SNP with a phenotype. However, this procedure is known to cause the 'missing heritability', namely, a large proportion of heritability remains unexplained by loci identified (Maher, 2008), hence it is increasingly important to evaluate combinatorial effects of SNPs (Wei *et al.*, 2014).

Several types of software have been developed to detect interactions among SNPs related to a phenotype (Purcell *et al.*, 2007; Zhang and Liu, 2007; Calle *et al.*, 2010; Wan *et al.*, 2010; Kam-Thong *et al.*, 2012; Van Lishout *et al.*, 2013). However, few methods can simultaneously overcome two major problems. One is that statistical validity is not performed. Most methods that can enumerate higher-order interactions do not evaluate statistical significance of the results. The other is that a combination size is limited in practical application. Existing statistical techniques such as logistic regression and multifactor dimensionality reduction can be used to find combinatorial effects. When we investigate all combinatorial effects, these techniques have to be applied to all possible combinations, which is too computationally intensive. Both problems need to be overcome if high-order interaction analysis is to be successfully performed.

A recently proposed statistical method called Limitless Arity Multiple-testing Procedure (LAMP) (Terada *et al.*, 2013) provides a possibility of detecting statistically significant higher-order interactions. LAMP is a multiple testing procedure for listing statistically significant combinatorial effects by introducing a theoretical upper bound of family-wise error rate tighter than Bonferroni correction. Its application to GWAS analysis may uncover synergistic effects of SNPs associated with diseases.

We therefore developed LAMPLINK, a software that incorporates LAMP with a widely used GWAS analysis software PLINK (Purcell *et al.*, 2007). Applied to the 1000 Genomes Project data, it detected a combination of five SNPs accumulated in the Japanese population with statistical significance.

2 Methods and implementation

LAMPLINK is implemented by adding options for detecting statistically significant high-order interactions of SNPs to PLINK (version 1.07), allowing for use of all options and files in LAMPLINK. Figure 1 shows a typical analytical procedure for detecting SNP combinations using LAMPLINK. LAMPLINK performs a case–control analysis for GWAS data using Fisher's exact test or chi-squared test, and enumerates statistically significant combinations associated with a given phenotype. The additional options are shown in Supplementary Table S1, and the details of LAMPLINK are described in Supplementary Text. LAMPLINK runs with C and Python 2.7 on Linux.

2.1 Detection of statistically significant SNP combinations

The `--lamp` option with `--model-dom` (or `--model-rec`) can be used for enumerating statistically significant SNP combinations (Procedure 1 in Fig. 1a). The input and output filenames are specified with the `--file` (or `--bfile` for binary format) and `--out` options, respectively. When you set `--model-dom`, LAMPLINK detects statistically significant combinations of SNPs according to a dominant exclusive model, whereas `--model-rec` uses a recessive exclusive model. These two genetic models are defined in Supplementary Text.

LAMPLINK results are exported to files: '<lamp_out_filename>.lamp' and '<lamp_out_filename>.lamplink'. The former file reports all SNP combinations statistically significantly associated with the phenotype. The latter file reports detailed information about each SNP in a format similar to the result generated by PLINK for association analysis. All columns of the result files are listed in Supplementary Table S2.

2.2 Elimination of redundant SNP combinations

Procedure 1 may end up listing combinations of SNPs that are in the same linkage disequilibrium (LD) region, which may prevent understanding of SNP–phenotype associations. The `--lamp-ld-remove` option is useful to filter out uninformative combinations (Procedure 2 in Fig. 1a). Using this option eliminates SNP combinations whose members have r^2 higher than the user-specified threshold, on the assumption that they are located in the same LD region. If all r^2 scores computed for SNP pairs in each chromosome are higher than the threshold, the combination is removed.

3 Analysis of exome data

We applied LAMPLINK to human exome data provided by the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010), including 12 758 SNPs and 697 individuals from seven populations. We demonstrated a case–control study by regarding 105 Japanese individuals as cases and the remaining as controls. We detected combinations of SNPs accumulated in Japanese with statistical significance. All of the settings and commands for this demonstration are described in Supplementary Text. The experiments were run on a machine with an Intel Xeon E5-2680v2 processor at 2.6 GHz running Red Hat Enterprise Linux 6.4.

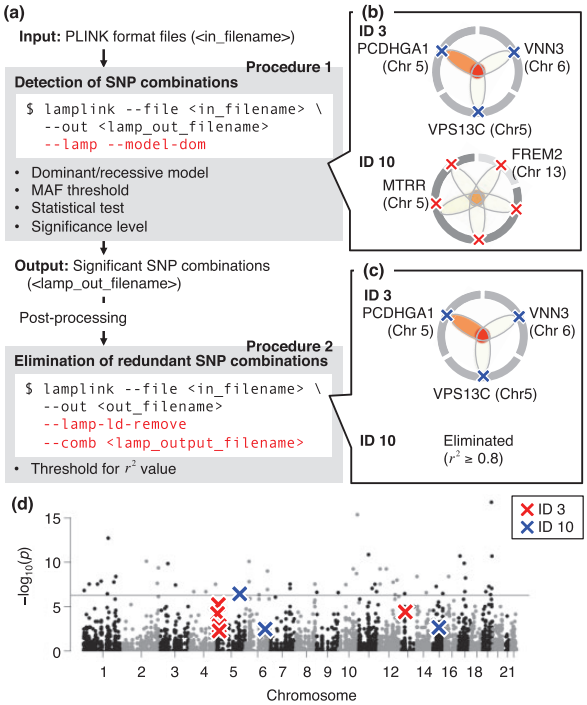

Fig. 1. Overview of LAMPLINK. (**a**) Workflow to detect statistically significant SNP combinations. (**b**) Two significant combinations including three and five SNPs (IDs 3 and 10 in Supplementary Table S3) detected in Procedure 1. Each petal corresponds to the *P*-value of a single SNP, and the central circle represents that of the SNP combination. Color shows the adjusted *P*-values. The *P*-value of the combination was smaller than the *P*-values of any single SNP, suggesting the existence of an epistatic effect among the three SNPs. (**c**) Detected combination after Procedure 2. ID 10 has been eliminated because it includes pairs of SNPs whose $r^2 \geq 0.8$. (**d**) Manhattan plot of *P*-values from the test of the association between the Japanese population and other populations. Crosses represent significant SNP combinations in (b). The horizontal line indicates the adjusted significance level (5.49e-07)

We compared the time performance of LAMPLINK with the `--epistatic` option in PLINK, which exhaustively analyzes the relationship of pairs of SNPs to a phenotype. The calculation time of LAMPLINK was 21.281 s, whereas PLINK required over 150 min to investigate all pairs of SNPs, showing that LAMPLINK has the ability to identify combinatorial effects of SNPs within a short time despite investigating all possible combinations of SNPs. A detailed time performance analysis of LAMPLINK is provided in Supplementary Text.

Procedure 1 detected 106 statistically significant SNP combinations, including 10 SNP combinations that consisted of three or more SNPs (Supplementary Table S3). These combinations could not be detected by PLINK.

Figure 1(b) illustrates two statistically significant combinations (IDs 3 and 10 in Supplementary Table S3). ID 3 consisted of three SNPs located within the genes *PCDHGA1*, *VPS13C* and *VNN3*. These SNPs are located within different genes on different chromosomes (Fig. 1d). ID 10 consisted of five SNPs. Four SNPs are located within the same gene *MTRR*, and hence this combination is eliminated in Procedure 2 (Fig. 1c). We discuss these results in detail in Supplementary Text.

These two results show that LAMPLINK has the ability to detect statistically significant SNP combinations from genome-wide case–control data. By replacing the phenotype with a disease, it might be

possible to identify causal mutations of complex diseases. LAMPLINK is the first implementation that can detect statistically sound high-order interactions from tens of thousands of markers. Hence, LAMPLINK may contribute to the identification of combinatorial effects from multiple markers by re-analysis of existing GWAS datasets.

4 Future work

LAMPLINK currently supports two genetic models (dominant and recessive exclusive models), but it cannot handle the combination of recessive and dominant models (known as the jointly recessive-dominant model for two loci; Li and Reich, 2000) due to a theoretical limitation in LAMP. Future work includes supporting the jointly recessive-dominant model as well as the threshold (Greenberg, 1981) and additive models (Neuman and Rice, 1992), which may help solving the problem of missing heritability.

We also plan to incorporate other statistical models into LAMPLINK for analyzing various types of data. For example, statistical assessment using the Mann–Whitney *U*-test or a regression model is useful to analyze numerical traits data. Additionally, LAMP has been developed to avoid spurious results caused by a confounding variable (e.g. age or gender of patients) (Terada *et al.*, 2016). Incorporating these methods will greatly improve the versatility of our software.

Acknowledgement

We thank Assist Prof. David duVerle and Dr. Raissa Relator for insightful comments and suggestions. Supercomputing resources were provided by NIG (ROIS).

Funding

This work was supported by Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Agency (JST) to A.T.; Core Research for Evolutional Science and Technology (CREST), JST to R.Y., K.T. and J.S.; and KAKENHI [14469361 to R.Y., 15H05711 to K.T. and 15H01717, 15H05713 and 24240044 to J.S.].

Conflict of Interest: none declared.

References

Calle,M.L. *et al.* (2010) mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. *Bioinformatics*, **26**, 2198–2199.

Greenberg,D.A. (1981) A simple method for testing two-locus models of inheritance. *Am. J. Hum. Genet.*, **33**, 519–530.

Kam-Thong,T. *et al.* (2012) GLIDE: GPU-based linear regression for detection of epistasis. *Hum. Hered.*, **73**, 220–236.

Li,W. and Reich,J. (2000) A complete enumeration and classification of two-locus disease models. *Hum. Hered.*, **50**, 334–349.

Maher,B. (2008) Personal genomes: the case of the missing heritability. *Nature*, **456**, 18–21.

Neuman,R.J. and Rice,J.P. (1992) Two-locus models of disease. *Genet. Epidemiol.*, **9**, 347–365.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Terada,A. *et al.* (2013) Statistical significance of combinatorial regulations. *Proc. Natl. Acad. Sci. USA*, **110**, 12996–13001.

Terada,A. *et al.* (2016) Significant pattern mining with confounding variables. In *Proceedings of PAKDD 2016*, pp 277–289.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Van Lishout,F. *et al.* (2013) An efficient algorithm to perform multiple testing in epistasis screening. *BMC Bioinformatics*, **14**, 138.

Wan,X. *et al.* (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.

Wei,W.H. *et al.* (2014) Detecting epistasis in human complex traits. *Nat. Rev. Genet.*, **15**, 722–733.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Zhang,Y. and Liu,J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.*, **39**, 1167–1173.