LEP: A Statistical Method Integrating Individual-Level and Summary-Level Data of the Same Trait From Different Populations

Biomedical Informatics Insights
Volume 11: 1–3
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1178222619881624

Mingwei Dai[1], Jin Liu[2] and Can Yang[3]

[1]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China. [2]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore. [3]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China.

ABSTRACT: Statistical approaches for integrating multiple data sets in genome-wide association studies (GWASs) are increasingly important. Proper utilization of more relevant information is expected to improve statistical efficiency in the analysis. Among these approaches, LEP was proposed for joint analysis of individual-level data and summary-level data in the same population by leveraging pleiotropy. The key idea of LEP is to explore correlation of the association status among different data sets while accounting for the heterogeneity. In this commentary, we show that LEP is applicable to integrate individual-level data and summary-level data of the same trait from different populations, providing new insights into the genetic architecture of different populations.

KEYWORDS: Genome-wide association study, integrative analysis, polygenicity, pleiotropy, heterogeneity

RECEIVED: September 9, 2019. **ACCEPTED:** September 19, 2019.

TYPE: Commentary

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by grant numbers 11829101 and 11931014 from National Science Funding of China.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Can Yang, Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China. Email: macyang@ust.hk

COMMENT ON: Dai M, Wan X, Peng H, et al. Joint analysis of individual-level and summary-level GWAS data by leveraging pleiotropy. *Bioinformatics*. 2019;35(10):1729-1736. doi:10.1093/bioinformatics/bty870. PubMed PMID: 30307540. https://www.ncbi.nlm.nih.gov/pubmed/30307540.

The flourishing growth of genome-wide association studies (GWASs) has provided comprehensive understanding of genetic determinants of disease susceptibility,[1,2] shedding light on better prevention and treatment of diseases. The results from GWAS suggested the existence of "polygenicity" for complex diseases, which means that a complex disease is often affected by many variants with small effects. Due to polygenicity, limited sample size of a single GWAS often has a relatively low statistical power of association identification and poor predictive ability.

To this end, many methods have been proposed to effectively improve statistical efficiency by combining multiple data sets.[3,4] These methods might take different types of data as input; integrating different sources of data is often feasible by leveraging pleiotropy.[5,6] Recently, we have proposed a statistical method named LEP[7] to integrate the individual-level genotype data and summary statistics in GWASs. LEP and other statistical methods that integrate individual-level data and summary-level data are becoming increasingly important. This is because we often have limited individual-level data (usually a few thousands of samples at hand) but can get access to summary-level data through many public gateways. Working on limited samples with individual-level data may lead to great uncertainty on the estimation of genetic effects on a complex trait. Fortunately, genome-wide summary-level data bring additional information about genetic effects on the trait. LEP explores this kind of information in the joint analysis of individual-level data and summary-level data.

Originally, LEP was designed to integrate multiple traits of the same population by exploring pleiotropy among them. More specifically, pleiotropy means that a variant can affect multiple seemingly unrelated traits. LEP integrates the individual-level data and the summary-level data by modeling their pleiotropic relationship. By introducing γ_j and Γ_j to indicate whether the jth variant is associated with the trait for the individual-level data and the trait for the summary-level data, respectively, LEP characterizes the pleiotropic relationship between the trait for the individual-level data and the trait for the summary-level data through the following probabilistic model

$$\begin{aligned} u &:= Pr\left(\Gamma_j = 1 \mid \gamma_j = 1\right) \\ v &:= Pr\left(\Gamma_j = 0 \mid \gamma_j = 0\right) \end{aligned} \tag{1}$$

Comprehensive simulation studies and real-data analysis demonstrated the effectiveness of LEP by leveraging pleiotropy in the presence of heterogeneity among the individual-level and summary-level data.

For a given trait/disease, GWASs have been conducted in different populations. As a matter of fact, many GWASs have been conducted in the populations of European ancestry. Because the allele frequency and linkage disequilibrium (LD) pattern of samples from different populations can be quite different,[6,8,9] heterogeneity of genetic effects widely exists and the discoveries in 1 population could not be directly transferred to

Table 1. Information of the GWAS data for Crohn's disease from different populations.

GWAS	CASES	CONTROLS	NSNP	ANCESTRY	TYPE
WTCCC	2,005	3,004	308,950	England	Individual-level
Belgium	537	913	953,242	Belgium	Summary-level
Cedars-Sinai	925	2,882	953,242	USA	Summary-level
Early Onset	1,689	6,197	953,242	USA, Italy etc.	Summary-level
NIDDK	956	982	953,242	USA	Summary-level
German	479	1,145	953,242	German	Summary-level
Total	4,586	12,119			

Abbreviations: GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism; WTCCC, Welcome Trust Case Control Consortium.
After extracting overlapped SNPs of individual-level data (after quality control) and summary statistics, we had the individual-level data $\{\mathbf{X}\}_{N\times M}, \mathbf{y}_{N\times 1}$ and a *P*-value matrix $\mathbf{P} \in \mathbb{R}^{M\times K}$, where $N = 4,536$ is the number of samples, and $M = 248,409$ is the number of overlapped SNPs of individual-level data and summary-level data. The samples from the Cedars-Sinai Medical Center were divided into 2 studies (Cedar 1 and Cedar 2) and the samples from NIDDK were divided into the Jewish study (NiddkJ) and the non-Jewish study (NiddkNJ).

Table 2. Estimated parameters *u*, *v* for every single GWAS jointly analysis with WTCCC data.

	BELGE	CEDAR 2	EARLY ONSET	CEDAR 1	NIDDKJ	GERMAN	NIDDKNJ
$\hat{u}$	1	1	1	1	1	1	1
$\hat{v}$	0.713	0.5698	0.9704	0.8954	0.8953	0.9168	0.8834
Accuracy	63.85% ± 0.54%	63.50% ± 0.59%	66.30% ± 0.54%	64.26% ± 0.52%	63.54% ± 0.40%	64.08% ± 0.55%	64.09% ± 0.43%

Abbreviation: GWAS, genome-wide association studies.
Accuracy is calculated from 10 replications.

another population. The study of different approaches to deal with the heterogeneous genetic effects in different populations is gaining increasing attention. Although LEP was designed to explore pleiotropy among different traits, the essential idea of LEP is to make use of the correlation of association status of multiple GWASs while accounting for the heterogeneity. Clearly, the probabilistic model given in equation (1) can account for heterogeneity in the presence of either pleiotropy or correlated genetic effects of the same trait in different populations. The pair of parameters $\{u, v\}$ measures the extent to which the genetic determinants of disease risk are likely to be shared by or specific to populations.

As an illustrative example, we applied LEP to analyze GWAS data of Crohn's disease (CD) from several different populations. The individual-level data are from the Welcome Trust Case Control Consortium (WTCCC).[10] The summary-level data of CD are from the study by Franke et al,[11] composed of the *P*-values of 7 GWASs in total. These data sets are summarized in Table 1 (detailed information can be found in the study by Dai et al[12]). We first applied Bayesian variable selection regression[13] to the individual-level data and obtained accuracy of $63.2\% \pm 0.4\%$ (measured by the area under the curve [AUC]). Then, we applied LEP to incorporate summary-level data sets and the accuracy was improved, as shown in Table 2. The corresponding estimated parameters $\{u, v\}$ are also given in Table 2, indicating that LEP successfully accounts for heterogeneity.

In summary, LEP can effectively account for heterogeneity when integrating individual-level data and summary-level data from GWAS. As a result, not only can LEP be applied to leverage pleiotropy for analysis of multiple traits in the same population but also it can serve as an effective tool to analyze the same trait across different populations.

Author Contributions

MD performed the data analysis, CY conceived the idea of this study and MD, JL and CY wrote the manuscript.

ORCID iDs

Jin Liu https://orcid.org/0000-0002-5707-2078
Can Yang https://orcid.org/0000-0002-4407-3055

REFERENCES

1. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017;101:5-22.
2. Yang C, Wan X, Liu J, et al. Introduction to statistical methods for integrative data analysis in genome-wide association studies. In: Wong KC, ed. *Big Data Analytics in Genomics*. Basel, Switzerland: Springer; 2016:3-23.
3. Flannick J, Florez JC. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet*. 2016;17:535-549.

4. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*. 2010;42:1118-1125.
5. Stearns FW. One hundred years of pleiotropy: a retrospective. *Genetics*. 2010;186:767-773.
6. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. *Nat Rev Genet*. 2019;20: 567-581.
7. Dai M, Wan X, Peng H, et al. Joint analysis of individual-level and summary-level GWAS data by leveraging pleiotropy. *Bioinformatics*. 2019;35:1729-1736.
8. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51:584-591.
9. Gurdasani D, Barroso I, Zeggini E, et al. Genomics of disease risk in globally diverse populations. *Nat Rev Genet*. 2019;20:520-535.
10. Burton PR. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661-678.
11. Franke A, McGovern DPB, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*. 2010;42:1118-1125.
12. Dai M, Ming J, Cai M, et al. IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. *Bioinformatics*. 2017;33:2882-2889.
13. Carbonetto P, Stephens M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal*. 2012;7:73-108.