BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 20 2014, pages 2971–2972
doi:10.1093/bioinformatics/btu434

Genome analysis
Advance Access publication July 9, 2014

repfdr: a tool for replicability analysis for genome-wide association studies

Ruth Heller*, Shay Yaacoby and Daniel Yekutieli
Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv 6997801, Israel

Associate Editor: John Hancock

ABSTRACT

Motivation: Identification of single nucleotide polymorphisms that are associated with a phenotype in more than one study is of great scientific interest in the genome-wide association studies (GWAS) research. The empirical Bayes approach for discovering whether results have been replicated across studies was shown to be a reliable method, and close to optimal in terms of power.

Results: The R package *repfdr* provides a flexible implementation of the empirical Bayes approach for replicability analysis and meta-analysis, to be used when several studies examine the same set of null hypotheses. The usefulness of the package for the GWAS community is discussed.

Availability and implementation: The R package *repfdr* can be downloaded from CRAN.

Contact: ruheller@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 9, 2014; revised on June 22, 2014; accepted on July 3, 2014

1 INTRODUCTION

Replicating findings in genome-wide association studies (GWAS) research is important for two main reasons. First, GWAS are observational studies, and therefore there is always a danger that bias may explain away the discoveries. If a finding is replicated in more than one study, it is more convincingly not due to bias. Second, finding the associations that are common in more than one population is scientifically interesting.

A replicability analysis is primarily useful in the situation where a GWAS meta-analysis is used at a primary (or discovery) stage to suggest single nucleotide polymorphisms (SNPs) to test in additional follow-up studies. The SNPs that exhibit a strong evidence for replicated associations across the studies in the primary stage, need not be followed up, as the replicability of association has already been established. For example, in Voight *et al*. (2010), the associations of SNPs in gene TCF7L2 with type 2 diabetes in the studies considered in the primary stage produced a strong evidence of replicability, even when faced with the many SNPs examined at this stage, and therefore there was no need to follow-up on this gene.

The meta-analysis in the primary stage may identify SNPs that have not been discovered in the replicability analysis of the primary stage studies. This does not imply the absence of true replicated association, as it could be the result of small effect size and/or limited sample size (i.e. lack of statistical power). On the other hand, it could also be because of the effect not being replicated, as a discovered association in a meta-analysis may be entirely driven by a single study: if a *P*-value is close to zero in one study, the association will be significant in a meta-analysis even if there is no association in all the other studies. Using a follow-up study (or replication stage), the replicability of association may be verified. See Bogomolov and Heller (2013) for methods to establish replicability from follow-up studies.

Heller and Yekutieli (2014) present an empirical Bayes approach to replicability analysis. They estimate the probability that the association was replicated in the same direction in at least two studies, and declare as having replicated associations the SNPs with estimated probability small enough so that the estimated Bayes FDR is at most q (e.g. $q = 0.05$). For each SNP, the null state is that the association is replicated in the same direction in at most one study. More generally, Heller and Yekutieli (2014) show how to test different null states according to the interest of the user, where meta-analysis corresponds to the null state of no association in all the studies. They verify the optimality and FDR control of this approach in extensive simulations and demonstrate its usefulness and power in detecting SNPs that are associated with type 2 diabetes in several studies. Here, we present the R package *repfdr* that provides a flexible and efficient implementation of the method in Heller and Yekutieli (2014).

2 METHODS

Suppose n GWAS studies examine the same phenotype. The input data for analysis are the n vectors of one-sided *P*-values, one vector per study. Following Efron (2010), the *P*-values are transformed into z-scores (using the inverse standard normal cumulative distribution). The three main analysis steps are—(i) in each study, bin the z-scores and estimate the non-null probabilities for each bin; (ii) estimate the probabilities of association status by an expectation-maximization (EM) algorithm; (iii) estimate the Bayes FDR and select which hypotheses have replicated findings or meta-analysis findings, depending on whether the aim is replicability analysis or meta-analysis. Step (i) can be done using our function *ztobins* that calls the R package *locfdr*, or by any other software the user chooses. Our main function *repfdr* performs steps (ii) and (iii), requiring as input the binned z-scores, and the null and estimated non-null probabilities for each bin in each study. Our function *ldr* reports the estimated posterior probabilities for SNPs of interest. Details about the computation of the estimated parameters can be found in Heller and Yekutieli (2014).

*To whom correspondence should be addressed.

In step (ii) the number of parameters estimated by the EM is exponential in n: if the association can be either null, or positive, or negative, then the EM estimates 3^n parameters; if the association can be either null or non-null (e.g. only in one direction), then the EM estimates 2^n parameters. On a computer with 12 GB memory, it is possible to estimate these parameters for $n = 5$ studies using 3×10^6 SNPs, or for $n = 6$ studies using 10^6 SNPs, but there is not enough memory for $n = 6$ studies using 3×10^6 SNPs. Moreover, the greater the number of parameters, the larger the amount of data (SNPs) necessary for reliably estimating the parameters. If a large number of studies are available, we therefore recommend first grouping the studies into fairly homogeneous clusters of studies, then computing the meta-analysis P-value for each cluster of studies and finally applying our software to discover the replicated findings across clusters of studies.

The implementation of *repfdr* (written in R and C) allows it to take advantage of parallel processing, improving its efficiency. By default, the software automatically detects the number of available processing threads.

2.1 Example

We give a data example which is a simulation of three GWAS from the simulator HAPGEN2 (Su *et al.*, 2011), thus emulating real GWAS data. Each study is summarized by the z-scores for the test of association of an SNP with a binary outcome in the same 249 024 SNPs. Our specific question is—which SNPs show replicated association with the phenotype? i.e. for which SNPs the association with phenotype is present in more than one of the studies? In the package, it is shown how to summarize the findings and conclude about this question.

We have 3^3 possible vectors of association status for each SNP. When we test for no association, the null set is just the zero vector. We get 239 SNP discoveries at the Bayes FDR of 0.05. For replicability analysis, we have 13 states in our null set, and we discover 119 SNPs at the Bayes FDR of 0.05. Supplementary Figure S1 shows a Manhattan plot of the negative logarithm of the estimated Bayes FDRs for replicability (top) and for association (bottom) versus the genomic coordinates.

Noting that the replicability analysis findings are typically a subset of the findings from an analysis to detect associations, SNPs may be discovered only in the latter analysis for two possible reasons: either because of insufficient power for establishing replicability, or because of the association being present only in a single study. Among these SNPs, identification of those with replicated association may be done using a follow-up study that examines only these SNPs (239-119 = 120 SNPs in this example) in an independent study [e.g. using the methods in Bogomolov and Heller (2013), implemented in the Web application http://www.math.tau.ac.il/~ruheller/App.html].

3 CONCLUSION

In large problems where it is possible to estimate well the unknown parameters, the empirical Bayes replicability analysis is useful. In particular, this is so for GWAS, where discovering SNPs with replicated associations is of great scientific interest. Specifically, the replicability analysis using the P-values of several published GWAS that examine the same complex disease can shed light on the genetic architecture of the disease by identifying the SNPs that have replicated associations across studies as well as the SNPs associated with the disease that show no evidence (or inconsistent evidence) of replicability of associations. Our package provides a way of performing this analysis, and the theoretical justifications are given in Heller and Yekutieli (2014). The proposed approach is a general approach for assessing replicability in several studies when each study examines the same hypotheses. Therefore, it can be used for applications other than GWAS, as long as the marginal and non-null densities can still be reasonably well approximated for each study, and the dependency of the test statistics within each study is local.

Funding: The work of Ruth Heller and Shay Yaacoby was supported by grant no. 2012896 from the Israel Science Foundation (ISF).

Conflict of Interest: none declared.

REFERENCES

Bogomolov,M. and Heller,R. (2013) Discovering findings that replicate from a primary study of high dimension to a follow-up study. *J. Am. Stat. Assoc.*, **108**, 1480–1492.

Efron,B. (2010) *Large-Scale Inference*. Cambridge University Press, Cambridge, MR2724758.

Heller,R. and Yekutieli,D. (2014) Replicability analysis for genome-wide association studies. *Ann. Appl. Stat.*, **8**, 481–498.

Su,Z. *et al.* (2011) Hapgen2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.

Voight,B.F. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579–589.