Development of bioinformatic tools and application of novel statistical methods in genome-wide analysis

van der Most, Peter Johannes

Chapter 4

lodGWAS: a software package for genome-wide association analysis of biomarkers with a limit of detection

Ahmad Vaez, Peter J. van der Most, Bram P. Prins, Harold Snieder, Edwin van den Heuvel, Behrooz Z. Alizadeh, Ilja M. Nolte

Bioinformatics, 2016, 32 (10): 1552-1554

Abstract

Summary: Genome-wide association study (GWAS) of a biomarker is complicated when the assay procedure of the biomarker is restricted by a Limit of Detection (LOD). Those observations falling outside the LOD cannot be simply discarded, but should be included into the analysis by applying an appropriate statistical method. However, the problem of LOD in GWAS analysis of such biomarkers is usually overlooked. 'lodGWAS' is a flexible, easy-to-use R package that provides a simple and elegant way for GWAS analysis of such biomarkers while simultaneously accommodating the problem of LOD by applying a parametric survival analysis method.

The package is available at:
https://cran.r-project.org/web/packages/lodGWAS

Supplementary information can be found at Bioinformatics Online:
https://academic.oup.com/bioinformatics

Introduction

Genome-wide association study (GWAS) of a biomarker such as immunoproteins (e.g. C-reactive protein), cytokines (e.g. interleukins) and hormones (e.g. testosterone) is complicated if the detection range of the biomarker is restricted by the assay procedure. So far GWAS analyses of such biomarkers inadequately dealt with the problem of the limit of detection. As a consequence, the identified associations might be biased. The limit of detection (LOD) is the smallest or largest concentration of a biomarker that can be reliably measured by the analytical procedure. Those measurements falling outside the LOD, so-called non-detects (NDs), cannot be considered as missing values since they do provide information about the distribution of the biomarker. Simple exclusion of NDs from the statistical analysis will therefore yield incorrect estimates. A number of statistical approaches have been proposed to deal with the problem of LOD while still including those NDs in the statistical analysis, such as: analyzing detect/non-detect dichotomies, substituting NDs with a constant smaller than or equal to the lower LOD, or substituting NDs with a random value between zero and lower LOD[1]. However these methods insufficiently account for the available information provided by NDs, particularly when NDs comprise a large proportion of the data.

Recently Dinse and colleagues proposed to apply survival analysis to overcome the problem of LOD[2]. They stated that NDs can be viewed as censored data, as they are known to fall within a certain interval. More details on this can be found in the Supplementary data. This approach could also be applied to GWAS data. However, currently available software packages for GWAS analysis are not flexible enough to properly account for NDs. We developed 'lodGWAS', a flexible, easy-to-use software package that is capable of performing GWAS analysis of biomarkers while accommodating the problem of LOD by applying survival analysis in which NDs are treated as censored data.

Approach

Implementation

lodGWAS is built as a package for R[3]. The R platform was chosen as it is operating system-independent, commonly used, open source, and can handle large datasets. lodGWAS depends upon the "survival" R package[4]. It appropriately treats NDs as censored data, and performs a genome-wide parametric survival analysis by including both 'measured' and 'censored' values. In this way, it allows full use of the available data.

Usage

The lodGWAS package provides two functions: the first enables the user to check the quality of the input data, and the second to perform a genome-wide censored survival analysis.

GWAS analysis is prone to errors in the input files. Therefore we emphasize the need for a thorough quality control of the input files, and in particular, the phenotype file. The function 'lod_QC' checks if the phenotype values within or outside LOD are coded correctly, provides a number of descriptive statistics about the phenotype coding, and generates warning messages on suspected problems. The main function is ʻlod_GWAS', which enables the user to perform a GWAS using survival analysis for censored data.

Results

Features

lodGWAS provides a number of options allowing for a flexible and user-customized GWAS analysis. It can handle NDs resulting from both lower and upper LODs (left- and right-censored data, respectively), as well as multiple lower and/or upper LOD levels (as in the case of multiple assay types). Additionally, the user can adjust the analysis for covariates, and specify the assumed distribution of the phenotype.

The genomic data, either genotyped and/or imputed, should be provided in dosage format, which contains more information than the best-guess genotype format. lodGWAS is capable of handling all three commonly used dosage formats, as it will automatically recognize whether there is one (dosage), two (two-probabilities) or three (three-probabilities) columns of data per individual.

The output data will be optionally saved in a compressed or uncompressed file. lodGWAS supports a variety of output file formats, ensuring compatibility with different downstream software packages, like we implemented in QCGWAS[5].

TABLE 1 | Four different GWAS analyses on serum levels of CRP in the LifeLines cohort study with and without assuming a lower LOD.

Software	Statistical Method	Complete data	Censored data (39%)
PLINK	Linear regression	#1: all samples	#3: only samples ≥ LOD
lodGWAS	Survival analysis	#2: all samples	#4: both < LOD & ≥ LOD

Example application in real human samples

As an example application of lodGWAS, we performed four GWAS analyses on serum levels of C-reactive protein (CRP) in the LifeLines cohort study[6]. CRP levels as well as GWAS data were available for 12,838 healthy individuals (see supplementary information for more details). Measurements of CRP were not restricted by lower or upper LOD, yielding a complete dataset. We analyzed the data with both PLINK[7], considered as *gold standard*, and lodGWAS. To demonstrate the robustness of the results of lodGWAS, we also performed a second set of analyses using an arbitrary lower LOD of 1.0 mg/L, which is equal to the LOD of some older CRP assays. In this way, CRP values less than this assumed LOD, i.e. 39% of the whole sample size, were discarded by PLINK, and were considered as left-censored by lodGWAS (Table 1). Figure 1 and supplementary Figure S1 show that GWAS results of analyses by PLINK and lodGWAS on the complete dataset were identical. The PLINK analysis that discarded 39% of data below LOD yielded biased results that only modestly correlated with the gold standard (Pearson's correlation of estimated effect sizes: r_β=0.58; Pearson's correlation of $-\log_{10}$(p-value): r_p=0.32). Moreover, the highly significant peaks that were observed on chromosomes 1, 12, and 19 completely disappeared in this analysis (Figure S1). The results of the lodGWAS analysis accounting for 39% NDs, however, were unbiased and showed high correlation to the gold standard (r_β=0.96; r_p=0.93).

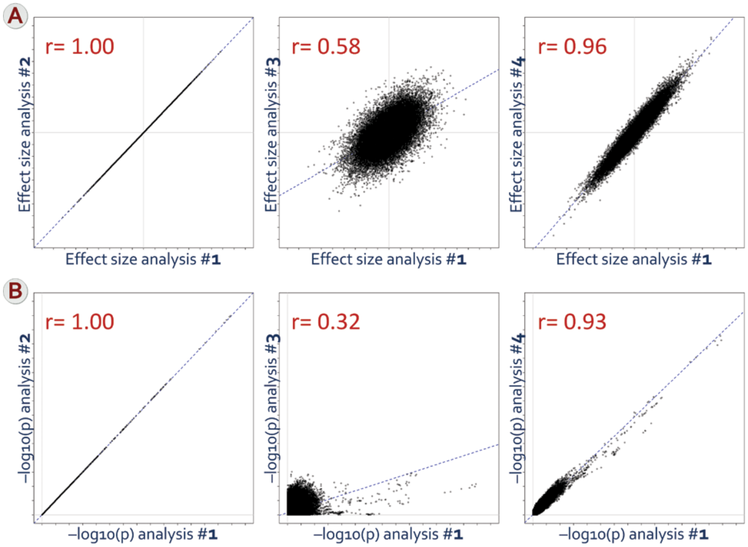

FIGURE 1 | The correlation scatter plots of the estimated (A) effect sizes, and (B) log transformed p-values of analyses #2, #3, and #4 versus analysis #1 (as described in Table 1).

Real applications

lodGWAS has already been applied by a number of cohorts as collaborators in a large-scale meta-GWAS project on serum levels of CRP within the context of the inflammation working group of the CHARGE consortium (http://www.chargeconsortium.com) (manuscript in preparation). For each of these cohorts the effect sizes of the individual cohort's GWAS were comparable to those of the meta-analysis (data not shown).

Conclusion

In this study we show the importance of applying an appropriate statistical method for GWAS analysis of biomarkers whose measurements are constrained by limits of detection. With lodGWAS we provide a flexible, easy-to-use R package for a simple and elegant way to perform GWAS analysis of such biomarkers while accommodating NDs.

Conflict of Interest: none declared.

References

1. Uh H, Hartgers FC, Yazdanbakhsh M, Houwing-Duistermaat JJ: Evaluation of regression methods when immunological measurements are constrained by detection limits. *Bmc Immunology* 2008; **9:** 59.
2. Dinse GE, Jusko TA, Ho LA *et al:* Accommodating Measurements Below a Limit of Detection: A Novel Application of Cox Regression. *Am J Epidemiol* 2014; **179:** 1018-1024.
3. R Core Team: R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing, 2014.
4. Therneau TM: Modeling Survival Data: Extending the Cox Model. New York, Springer, 2000.
5. van der Most PJ, Vaez A, Prins BP *et al:* QCGWAS: A flexible R package for automated quality control of genome-wide association results. *Bioinformatics* 2014; **30:** 1185-1186.
6. Scholtens S, Smidt N, Swertz MA *et al:* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol* 2015; **44:** 1172-1180.
7. Purcell S, Neale B, Todd-Brown K *et al:* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-575.