BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 8 2013, pages 1086–1088
doi:10.1093/bioinformatics/btt081

Genetics and population analysis

Advance Access publication February 17, 2013

PleioGRiP: genetic risk prediction with pleiotropy

Stephen W. Hartley[1,*] and Paola Sebastiani[2]

[1]National Institutes of Health/National Human Genome Research Institute, 5625 Fishers Lane, Suite 5 N-01, Rockville, MD 20850, USA and [2]Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston MA 02118, USA

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Although several studies have used Bayesian classifiers for risk prediction using genome-wide single nucleotide polymorphism (SNP) datasets, no software can efficiently perform these analyses on massive genetic datasets and can accommodate multiple traits.

Results: We describe the program PleioGRiP that performs a genome-wide Bayesian model search to identify SNPs associated with a discrete phenotype and uses SNPs ranked by Bayes factor to produce nested Bayesian classifiers. These classifiers can be used for genetic risk prediction, either selecting the classifier with optimal number of features or using an ensemble of classifiers. In addition, PleioGRiP implements an extension to the Bayesian search and classification and can search for pleiotropic relationships in which SNPs are simultaneosly associated with two or more distinct phenotypes. These relationships can be used to generate connected Bayesian classifiers to predict the phenotype of interest either using genetic data alone or in combination with the secondary phenotype(s).

Availability: PleioGRiP is implemented in Java, and it is available from http://hdl.handle.net/2144/4367.

Contact: stephen.hartley@nih.gov

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

Received on November 16, 2012; revised on January 11, 2013; accepted on February 11, 2013

1 INTRODUCTION

A Naive Bayesian classifier (NBC) is a simple, efficient and robust tool that can be used to capture the complex genetic basis of a multigenic trait and can predict a subject phenotype based on the posterior probability of the trait given the individuals genetic profile (Okser *et al*., 2010; Sebastiani *et al*., 2012a). Such classifier can be built using a large number of genetic variants, each with a relatively small contribution to the total genetic risk. However, the use of NBCs and similar Bayesian methods are relatively uncommon in genetics, and although general-use software does exist that is capable of performing Bayesian modeling and prediction (Thomas *et al*., 1992), many of these software utilities are not well-suited to the high-dimensionality of genome-wide data (Scutari, 2010). We have developed a stable, easy-to-use software package capable of performing these analyses efficiently and accurately using exact Bayesian computations.

*To whom correspondence should be addressed.

In addition, although methods are available to identify associations between genetic loci and individual phenotypes, few methods can detect pleiotropic associations in which loci are simultaneously associated with multiple phenotypes and use the pleiotropic associations to improve genetic risk prediction. In Hartley *et al*. (2012), we showed that pleiotropic associations can be modelled via the construction of simple Bayesian networks, and that these models can be applied to produce single or ensembles of Bayesian classifiers that leverage pleiotropy to improve genetic risk prediction. Testing with simulated and real data has shown that these models may improve genetic risk prediction under certain circumstances. Our software package, PleioGRiP is capable of both 'Pleiotropic Genetic Risk Prediction', as well as prediction using NBCs.

2 FUNCTIONALITY AND CAPABILITIES

PleioGRiP takes as input a genome wide association study (GWAS) dataset with multiple phenotypes $T_1, T_2, \ldots$, identifies relationships between the single nucleotide polymorphisms (SNPs) $S_1, S_2, \ldots$ and the phenotypes and uses these relationships to generate classifiers and ensembles of classifiers that can predict the phenotypes, conditional on the genetic profiles, or one phenotype conditional on the genetic profile and the other phenotypes. The details of the method are described in Hartley *et al*. (2012) and include two main phases: (i) discovery of SNPs that could be used for prediction using a Bayesian-model-based approach and (ii) selection of a final set of the most predictive SNPs using cross-validation, or generation of ensemble of classifiers.

2.1 Model building

To build standard NBCs with a single phenotype, PleioGRiP first ranks the SNPs by the strength of evidence that each SNP is associated with the phenotype. Bayes factors are calculated under four different modes of inheritance—genotypic, allelic, dominant and recessive—and the highest Bayes factor is used for the ranking. Nested classifiers $\Sigma_1, \ldots, \Sigma_R$ are then constructed such that Σ_1 uses only the top ranked SNP, Σ_2 uses the top two SNPs and so forth. The predictive accuracy of the classifiers in the list is evaluated using cross-validation to select the optimal number of SNPs.

If two or more phenotypes are provided, then PleioGRiP can search for pleiotropic relationships between SNPs and multiple phenotypes. In this case, for each SNP, Bayes factors of the association between each SNP and each phenotype are calculated for each mode of inheritance, and the highest Bayes factor for each SNP and phenotype is selected. All SNPs with highest

Bayes factor that do not pass a user-assigned significance threshold (default threshold = e) are dropped. For each SNP that passes this threshold and hence is associated with at least one phenotype, the simultaneous association of the SNP with additional phenotypes is scored by the Bayes factor and accepted if the Bayes factor increases. SNPs are ranked by the Bayes factor of the selected associations, and nested classifiers $\Sigma_1, \ldots, \Sigma_R$ are constructed such that Σ_1 uses only the top ranked SNP, Σ_2 uses the top two SNPs and so forth.

2.2 Prediction

Once nested classifiers have been constructed, PleioGRiP can carry out genetic risk prediction using a variety of different classification rules. If only one phenotype is being used, prediction is carried out with either a single classifier Σ_r, where the optimal number r of SNPs is selected using cross-validation, or an ensemble of the first r classifiers. The single classifier prediction uses the posterior probability of the phenotype, given the genetic profile, that is computed using Bayes' theorem as described for example in Hartley *et al.*, 2012; Sebastiani *et al.*, 2012b. The ensemble of classifiers predicts based on the average of posterior probabilities of the phenotype computed from the classifiers $\Sigma_1, \ldots, \Sigma_r$. The ensembles of classifiers prediction method has been shown to provide improved classification in certain situations. In particular, it appears to be robust to inclusion of false positive associations (Hartley *et al.*, 2012).

If PleioGRiP is assigned to search for pleiotropic associations between SNPs and two or more phenotypes, then additional prediction methods are available. The joint prediction of all the phenotypes uses the joint posterior probability of the phenotypes, given the genotype data that are computed by Bayes' theorem, using the formula:

$$p(T_1 \& T_2 | S_1, \ldots, S_r) \propto p(T_1 \& T_2) p(S_1, \ldots, S_r | T_1 \& T_2) = p(T_1 \& T_2) \prod_j p(S_j | T_1) \prod_k p(S_k | T_2) \prod_l p(S_l | T_1 \& T_2)$$

This formula leverages the conditional independences embedded in NBCs to factorize the probability $p(S_1, \ldots, S_r \mid T_1 \& T_2)$ [See Hartley *et al.* (2012) for details]. Conditional prediction predicts the phenotype of interest for a given subject using the conditional probability of the subject's phenotype, given the genotype and the other secondary phenotype(s). Marginal prediction predicts an assigned phenotype of interest for a given subject using the posterior probability of the phenotype, given only the subject's genotype data. Both conditional and marginal posterior probabilities can be derived from the joint probability. Prediction rules can be generated for these types of prediction based on either single classifiers or ensembles of classifiers.

2.3 Cross-validation

Cross-validation can be used to: (i) determine the optimal number of SNPs to use in the final classifier; (ii) estimate various accuracy metrics of the classifiers; and/or (iii) select alternative classification thresholds. Cross-validation can be carried out using any user-defined number of folds, up to and including leave-one-out cross-validation. For each subject in the discovery set, the cross-validation risks are calculated using each applicable prediction method. Using these data, the cross-validation specificity, sensitivity and area under the ROC curve (AUC) can be calculated for each nested classifier and each prediction rule. The AUC can be used to estimate the number of SNPs to include in the final classifier to maximize the efficacy of prediction. In addition, the cross-validation data can be used to determine the prediction threshold that optimizes the Youdens J statistic for a given classifier and prediction method. In some cases, particularly when the phenotypes are unbalanced, this can substantially improve prediction. The utility is designed to be fast and can carry out these cross-validation functions on 4000-subject, 500 000 SNP datasets in minutes (for 10-fold cross-validation) on a PC with 64 bit Operating System, 8 GB RAM, Intel 5, 2 2.67 GHz.

2.4 Input and output files

PleioGRiP requires as input files a *ped* file with genotype data, a *map* file with SNP location and a *phenotype file* with the phenotypic data. Both binary and text-based PLINK files are accepted as format. The output is a set of tab-delimited text files, including a model summary file and various prediction summary files. The model summary file lists the top SNPs in descending order of significance, along with the log-Bayes factor, the selected single-SNP model and mode of inheritance and the full contingency table for that model. The prediction summary files include, for each nested classifier and each type of prediction: accuracy, specificity and sensitivity estimates along with Fisher exact tests of differential prediction, both with and without prediction threshold selection AUC's estimates and 95% confidence intervals. The utility also exports the raw classification statistics for each classifier. R scripts included in the software package can then perform classification using any of these classifiers, and using any user-supplied prediction threshold.

3 EXAMPLE

The program is distributed with an example dataset of 2000 subjects, 7500 SNPs and 75 causal SNPs. To run the program:

```
java -Xmx8g -jar PleioGRip.jar
-ped simData.bed -map simData.bim -fam simData.fam
-pheno simData.pheno -read plinkBed -model find_add SearchBC
-pleio_parentList 'DISEASE_A—DISEASE_B'
-optimize run10FoldCrossValidationOptimization
-mdlSubjListFile list_disc.txt -repSubjListFile list_rep.txt
-twoColSubjLists -logDir./logs/-classify replication Classification_plio
```

where *.bed, .bim, .fam, .pheno* are input files with genotypes and phenotype lists. The software release includes documentation with descriptions of all the options.

4 CONCLUSION

PleioGRiP is stable, computationally efficient and offers a flexible new tool to GWAS investigators interested in Bayesian genetic risk prediction analyses, either with or without pleiotropy.

Funding: NIH/NHLBI R21HL114237 and NIH/NIA U19AG023122.

Conflict of Interest: This article was prepared while Stephen Hartley, Ph.D. was employed at Boston University. The opinions expressed in this article are the author's own and do not reflect the views of the National Institutes of Health, or the United States government.

REFERENCES

Hartley,S.W. *et al.* (2012) Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction. *Front. Genet.*, **3**, 176.

Okser,S. *et al.* (2010) Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young finns study. *PLoS Genet.*, **6**, e1001146.

Scutari,M. (2010) Learning Bayesian networks with the bnlearn R package. *J. Stat. Softw.*, **35**, 1–22.

Sebastiani,P. *et al.* (2012a) Genetic signatures of exceptional longevity in humans. *PLoS One*, **7**, e29848.

Sebastiani,P. *et al.* (2012b) Nave Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: Not so different after all! *Front. Genet.*, **3**, 26.

Thomas,A. *et al.* (1992) Bugs: a program to perform Bayesian inference using Gibbs Sampling. In: Bernardo,J. *et al.* (ed.) *Bayesian Statistics 4*. Oxford, UK, Oxford University Press, pp. 837–842.