EasyStrata: evaluation and visualization of stratified genome-wide association meta-analysis data

Thomas W. Winkler1,*, Zoltan Kutalik2,3,4, Mathias Gorski1, Claudio Lottaz5, Florian Kronenberg6 and Iris M. Heid1,*

1Department of Genetic Epidemiology, University of Regensburg, D-93053 Regensburg, Germany, 2Department of Medical Genetics, University of Lausanne, CH-1005 Lausanne, Switzerland, 3Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), CH-1010 Lausanne, Switzerland, 4Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland, 5Department of Statistical Bioinformatics, Institute for Functional Genomics, University of Regensburg, D-93053 Regensburg, Germany and 6Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, A-6020 Innsbruck, Austria

ABSTRACT

Summary: The R package EasyStrata facilitates the evaluation and visualization of stratified genome-wide association meta-analyses (GWAMAs) results. It provides (i) statistical methods to test and account for between-strata difference as a means to tackle gene–strata interaction effects and (ii) extended graphical features tailored for stratiﬁed GWMA results. The software provides further features also suitable for general GWAMAs including functions to annotate, exclude or highlight speciﬁc loci in plots or to extract independent subsets of loci from genome-wide datasets. It is freely available and includes a user-friendly scripting interface that simpliﬁes data handling and allows for combining statistical and graphical functions in a ﬂexible fashion.

Availability: EasyStrata is available for free (under the GNU General Public License v3) from our Web site www.genepi-regensburg.de/easystrata and from the CRAN R package repository cran.r-project.org/web/packages/EasyStrata/.

Contact: thomas.winkler@ukr.de or iris.heid@ukr.de

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on April 16, 2014; revised on July 7, 2014; accepted on September 12, 2014

1 INTRODUCTION

Genome-wide association meta-analyses (GWAMAs), in which multiple study-speciﬁc genome-wide association (GWA) results are pooled, have resulted in a 10- to 20-fold increase in the number of known genes contributing to complex traits and diseases (Visscher et al., 2012). Meanwhile, a focus is also on gene–environment–interaction analyses (GEWIS) that are conducted to further characterize genetic main effects as well as to discover novel genetic associations that are only present under certain environmental conditions (gene–environment interaction, GxE). Although some GEWIS have already extended from single-study analyses to meta-analyses, so far, only few reported replicable GxE effects (Hutter et al., 2013). For categorical environmental factors E, GEWIS meta-analyses can be implemented as GWAMAs stratified by E, for example, a GWAMA stratified by sex or by smoking status. Stratified GWAMAs do not only improve power to detect stratum-sensitive genetic main effects (Behrens et al., 2011) but also allow for testing gene–strata (GxS) interaction and joint (main + interaction) effects (Aschard et al., 2010; Magi et al., 2010; Randall et al., 2013). Although a variety of methods for the analysis of stratified GWAMA results exist, the availability of software tools is limited (see Supplementary Table S1 for a comparison with other GWAS tools).

We have developed an R-package called EasyStrata, which allows the user to obtain statistical and graphical summaries for comparisons across strata and to investigate potential GxS effects. The software was developed within the GxE working groups of the GIANT (Genetic Investigation of ANthropometric Traits) consortium, and the functionality of the package is exempliﬁed on GWAMA results for anthropometric traits, which are publically available at www.broadinstitute.org/collaboration/giant. EasyStrata is applicable to stratified GWAMAs of continuous or dichotomous outcomes, and many of the functions are also applicable for ‘non-stratified’ GWAMAs (Supplementary Note).

2 IMPLEMENTATION

2.1 Features and functionality

The basis for the EasyStrata analyses are GWAMA results for each single-nucleotide polymorphism (SNP) genome-wide and by stratum (m strata) (Supplementary Fig. S1): the stratum-specific meta-analyzed beta estimates and standard errors [inverse-variance weighted meta-analysis (Cox and Hinkley, 1979)] or Z-scores and sample sizes [sample size–weighted Z-score–based meta-analysis (Stouffer, 1949)] as well as other information (e.g. stratum-speciﬁc association P-values). Examples for strata are men and women (m = 2), or older age group and younger (m = 2), a combination of these (m = 4), smoking status (non-smoker, previous smokers, current smokers, m = 3) or other categorical exposures.
2.1.1 Statistical functionality To evaluate stratified GWAMA results, we have implemented statistical approaches to estimate (i) the overall (i.e. strata-combined) effect by meta-analysis of the m strata results (Cox and Hinkley, 1979; Stouffer, 1949); (ii) the joint effect calculated from m strata results (Aschard et al., 2010); (iii) the difference between two strata results as a means to test for GxS effects (Randall et al., 2013); and (iv) the heterogeneity between m strata (Cochran, 1954) (see Supplementary Table S2 for a summary of implemented statistics). All tests are applicable for stratum-specific beta estimates and standard errors as well as for stratum-specific Z-scores and sample sizes. We also provide functions to correct the computed P-values for multiple testing by false discovery rate (Benjamini and Hochberg, 1995) or Bonferroni correction (Johnson et al., 2010) and functions to clump results into independent—in terms of physical distance or linkage disequilibrium (LD)—subsets of significant SNPs.

2.1.2 Graphing functionality To visualize stratified GWAMA results, we have implemented state-of-the-art graphical functions, such as Quantile-Quantile plot (QQ), scatterplot and Manhattan plot. More specifically, we provide graphical features that are tailored for between-strata comparison: EasyStrata allows for contrasting two Manhattan plots in so-called ‘Miami’ plots (Fig. 1), for displaying multiple QQ curves in a single graph (Supplementary Fig. S2), and for extending scatter-plots by further dimensions (Supplementary Fig. S3). The graphical functionality is complemented by other convenient features, such as highlighting specific regions in Manhattan or Miami plots (Fig. 1), excluding specific regions from QQ plots to focus on the potential of novel associations, omitting less significant SNPs to substantially improve plotting speed for large datasets, breaking up the scale of the y-axis to ensure proper presentation of extremely significant SNPs, or creating panels of plots to provide a quick overview on the singular studies of the GWAMA or on various traits (Supplementary Figs S4–S7).

2.2 Usage
Our open-source software is written in R and makes use of the ‘cairo’ and the ‘plotrix’ packages. Extracting independent loci using LD-based thresholds requires the software PLINK (Purcell et al., 2007). For stratified GWAMA results based on HapMap-imputed studies, we recommend to have at least 4 GB of random access memory (RAM) available (see Supplementary Table S3 for an evaluation of runtime and RAM allocation). EasyStrata is started by calling the function ‘EasyStrata’ with an ecf-file as parameter: EasyStrata(’/path2ecf/example.ecf’). The user-defined ecf-file is a text-file that provides a flexible scripting interface that allows for generating customized analysis pipelines (Supplementary Fig. S8). A number of template ecf-pipelines (e.g. those that were used to create the here presented figures) can be downloaded from our Web site.

3 CONCLUSIONS
With EasyStrata, we provide a user-friendly R package that facilitates evaluation or graphical presentation of stratified GWAMA results. We have developed this software as analysts the GIANT consortium that meta-analyzed more than a hundred studies to investigate the genetic underpinning of anthropometric traits and to identify potential GxE effects. For example, the functionality of our software has been used to evaluate sex-stratified GWAMAs for multiple anthropometric traits (Randall et al., 2013). The automated pipeline approach of EasyStrata can save time and minimize errors from manually extracting and merging the data. This software is highly useful for analysts to cope with the increased complexity of high-dimensional stratified GWAMAs data.

ACKNOWLEDGEMENTS
EasyStrata was developed and tested using data from the Genetic Investigation of ANthropometric Traits Consortium (GIANT, http://www.broadinstitute.org/collaboration/giant).

Funding: German Federal Ministry of Education and Research (BMBF 01ER1206); National Institutes of Health (NIH, R01-DK075787/01A1, CFDA 93 848); Swiss National Science Foundation (31003A-143914); Swiss Institute of Bioinformatics.

Conflict of interest: none declared.

REFERENCES


