BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 2 2015, pages 259–261
doi:10.1093/bioinformatics/btu621

Genome analysis

Advance Access publication September 25, 2014

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

Thomas W. Winkler[1,*], Zoltan Kutalik[2,3,4], Mathias Gorski[1], Claudio Lottaz[5], Florian Kronenberg[6] and Iris M. Heid[1,*]

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

Associate Editor: Janet Kelso

**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 GWAMA 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-wide 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.

*To whom correspondence should be addressed.

*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 scatterplots 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.

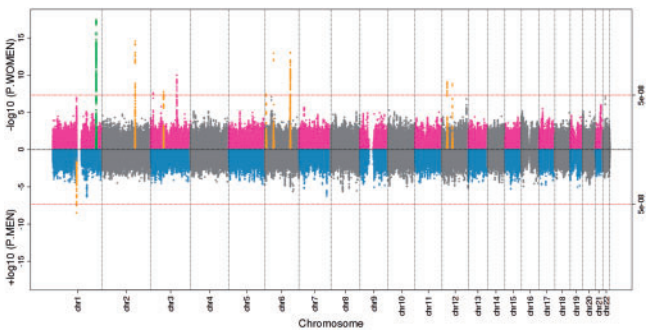


**Fig. 1.** Miami-Plot showing publically available women- and men-specific association *P*-values for waist–hip ratio adjusted for BMI (Randall *et al.*, 2013). Known main effect loci are highlighted in green (Lindgren *et al.*, 2009) or orange (Heid *et al.*, 2010). Based on the locus coloring, one can see two novel (women-specific) loci on chromosome 3

## 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

Aschard,H. *et al.* (2010) Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Hum. Hered.*, **70**, 292–300.

Behrens,G. *et al.* (2011) To stratify or not to stratify: power considerations for population-based genome-wide association studies of quantitative traits. *Genet. Epidemiol.*, **35**, 867–879.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, **57**, 289–300.

Cochran,W.G. (1954) The combination of estimates from different experiments. *Biometrics*, **10**, 101–129.

Cox,D.R. and Hinkley,D.V. (1979) *Theoretical Statistics*. Chapman and Hall, Halsted Press, London, New York.

Heid,I.M. *et al.* (2010) Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.*, **42**, 949–960.

Hutter,C.M. *et al.* (2013) Gene-environment interactions in cancer epidemiology: a national cancer institute think tank report. *Genet. Epidemiol.*, **37**, 643–657.

Johnson,R.C. *et al.* (2010) Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genom.*, **11**, 724.

Lindgren,C.M. *et al.* (2009) Genome-wide association scan meta-analysis identif three Loci influencing adiposity and fat distribution. *PLoS Genet*., **5**, e1000508.

Magi,R. *et al.* (2010) Meta-analysis of sex-specific genome-wide association studies. *Genet. Epidemiol*., **34**, 846–853.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Randall,J.C. *et al.* (2013) Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet*., **9**, e1003500.

Stouffer,S.A. (1949) *The American Soldier: Adjustment During Army Life*. Vol. 1, Princeton University Press, Princeton, NJ.

Visscher,P.M. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet*., **90**, 7–24.