BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 9 2012, pages 1280–1281
doi:10.1093/bioinformatics/bts124

Genetics and population analysis

Advance Access publication March 25, 2012

CpGassoc: an R function for analysis of DNA methylation microarray data

Richard T. Barfield[1,*], Varun Kilaru[2], Alicia K. Smith[2] and Karen N. Conneely[1,3]

[1]Department of Bioinformatics and Biostatistics, School of Public Health, [2]Department of Psychiatry & Behavioral Sciences and [3]Department of Human Genetics, School of Medicine, Emory University at Atlanta, GA, USA 30322

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: With the increasing availability of high-density methylation microarrays, there has been growing interest in analysis of DNA methylation data. We have developed CpGassoc, an R package that can efficiently perform the statistical analysis needed for increasingly large methylation datasets. CpGassoc is a modular, expandable package with functions to perform rapid analyses of DNA methylation data via fixed or mixed effects models, to perform basic quality control, to carry out permutation tests, and to display results via an array of publication-quality plots.

Availability and implementation: CpGassoc is implemented in R and is freely available at http://genetics.emory.edu/conneely; we are in the process of submitting it to CRAN.

Contact: rtbarfi@emory.edu

Received on December 20, 2011; revised on March 5, 2012; accepted on March 08, 2012

DNA methylation, the addition of a methyl group to a cytosine base followed by a guanine (CpG), is the most widely studied epigenetic modification. DNA methylation has a critical role in gene regulation, as methylation of the promoter region correlates with lower levels of gene expression (Bell *et al*., 2011), and recent studies have associated methylation patterns with complex traits (Breitling *et al*., 2011; Fackler *et al*., 2011; Javierre *et al*., 2010). The availability of high-density commercial methylation arrays has made it possible to perform genome-wide methylation analysis with increasingly dense coverage. For example, the Infinium HumanMethylation450 array includes >485 000 CpG sites, and this number can be expected to increase further over the coming years, allowing for larger and more expansive studies (Bibikova *et al.*, 2011; Illumina, 2011). However, despite technological advances in generating this data, the development of analytical software has lagged behind, particularly with respect to association analysis between methylation and complex traits. The popular R package methylumi performs quality control and data normalization, but not association analysis (Davis, 2012). Other packages such as Illumina's GenomeStudio, the R package minfi (Hansen and Aryee, 2011), and Significance Analysis of Microarrays (Tusher *et al*., 2001) can perform tests of association, but do not allow for additional covariates.

To provide a more flexible analysis tool, we created the R package CpGassoc to perform efficient analyses of quantitative methylation data. CpGassoc has two main analysis functions: cpg.assoc and cpg.perm. cpg.assoc uses fixed or mixed effects models to examine the relationship between a phenotype of interest and methylation of individual CpG sites across the genome. As input, cpg.assoc accepts a matrix or data frame of β-values (analogous to the proportion of DNA methylated). The program models the outcome as either β-values or the logit transformation of the β-values $\log(\beta/(1 - \beta))$, which can help stabilize the variance (Du *et al*., 2010). As predictors, users may specify a categorical or continuous phenotype of interest, an unspecified number of continuous or categorical covariates, and a fixed or random effect to control for batch or chip effects.

cpg.assoc outputs an R object of class 'cpg' that includes all relevant model information and association statistics for every CpG site tested, including effect sizes and standard errors, t or F-statistics, and both unadjusted and multiple-testing-adjusted P-values. To account for multiple testing, cpg.assoc can assess significance using (i) the Holm method—a step-down Bonferroni procedure (Holm, 1979), (ii) a false discovery rate (FDR) procedure or (iii) permutation testing using cpg.perm. Users can select from several FDR methods, including Storey's q-value method (Dabney *et al*, 2012; Storey, 2002) and all FDR methods available in the R function p.adjust; the Benjamini–Hochberg method (Benjamini, 2001) is used by default. cpg.perm can be used to obtain empirical P-values through permutation testing. cpg.perm performs a user-specified number of permutations in which the phenotype of interest is randomly re-assigned, and the data re-analyzed via cpg.assoc. cpg.perm outputs an R object of class 'cpg.perm' which is similar to the class 'cpg' but also includes a matrix containing information on each permutation and a vector containing three empirical P-values based on (i) the minimum observed P-value, (ii) the number of Holm-significant CpG sites and (iii) the number of FDR-significant sites.

CpGassoc is a modular and expandable package that currently includes 13 different functions with detailed help pages. The quality control function cpg.qc returns a matrix that can be read by cpg.assoc or cpg.perm, and R objects returned from cpg.assoc and cpg.perm can be passed to other functions within CpGassoc to easily generate publication-quality Q–Q plots, scatterplots or boxplots, and Manhattan plots (Fig. 1). In addition to standard Q–Q plots, CpGassoc can also produce a specialized Q–Q plot (Fig. 1A) where the expected quantiles of P-values or t-statistics are based on the number of sites for which the null hypothesis is maintained according to the Holm procedure. This produces a plot with confidence intervals based on the estimated null distribution of ordered P-values (or t-statistics); these confidence intervals will correspond directly to Holm significance. If a permutation test was

*To whom correspondence should be addressed.

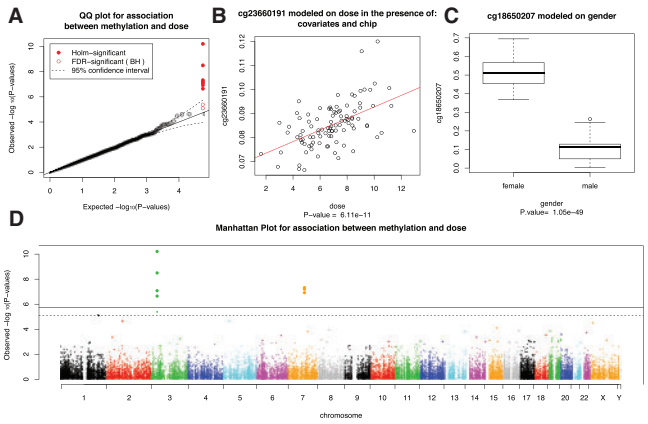

Fig. 1. Plots from CpGassoc: (**A**) Q–Q plot, (**B**) scatterplot, (**C**) boxplot and (**D**) Manhattan plot.

Table 1. Timing for analyses performed with various data sizes

Data size	Fixed effects model with no missing data	Fixed effects model with 5.9% missing data	Mixed effects model with 5.9% missing data
27 k, $n = 200$	2	10	971
27 k, $n = 1000$	14	73	1409
450 k, $n = 200$	37	153	15 744
450 k, $n = 1000$	658	1621	22 993

Times in seconds, based on Intel Xeon 2.8 GHz, with 12 GB RAM. Model includes one covariate and either fixed or random chip effects.

performed, empirical confidence intervals will be used. These plots allow visualization of the degree to which significant CpG sites deviate from the expected null distribution; standard Q–Q plots are also available for assessment of genomic inflation.

As noted above, cpg.assoc allows users to model chip or batch effects via either fixed or random effects. cpg.assoc can perform fixed effects analysis extremely quickly due to our algorithm for partitioning the CpG sites based on the presence of missing data. For CpG sites with no missing data, linear regression is performed via a computationally efficient matrix multiplication that allows multiple sites to be analyzed in one step. In contrast, sites with missing data are analyzed site-by-site via a loop, and the results of these partitioned analyses are then combined. To ensure that the analysis fits within the bounds of available memory, large datasets may be partitioned further; cpg.assoc determines the optimal number of CpG sites to be included in each partition based on the size of the data and the memory available in the environment. Mixed effects analyses that include random effects will take longer because they are computationally more intensive and must be performed site-by-site; for these analyses, the R package nlme is used (Pinheiro *et al.*, 2012). Timings for several analyses can be seen in Table 1; note that even for analyses of extremely large datasets fixed effects analyses can be performed very quickly.

In conclusion, we have created an effective and flexible tool for analysis of DNA methylation data. In addition to performing the analysis very quickly, CpGassoc has a variety of additional functions to assist researchers in their analyses. With the open-source nature of R, users may edit and customize CpGassoc according to their needs. The package is designed to be modular, with more functions easily integrated so that CpGassoc can continue to grow with the advancing field. Future modules we hope to incorporate include functions based on our ongoing research on normalization methods and adjustment for population stratification and heterogeneous cell types. CpGassoc is freely available at http: //genetics.emory.edu/conneely and we are in the process of submitting it to CRAN.

ACKNOWLEDGEMENTS

We are grateful to B.Barwick, D.Cutler, M.Epstein and Y.Sun for helpful discussions.

Conflict of Interest: none declared.

REFERENCES

Bell,J.T. *et al.* (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, **12**, R10.

Benjamini,Y. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*, **4**, 1165–1188

Bibikova,M. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

Breitling,L.P. *et al.* (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am. J. Hum. Genet.*, **88**, 450–457.

Dabney,A. *et al.* (2012) *qvalue: Q-value Estimation for False Discovery Rate Control*. R package version 1.28.0, R Foundation for Statistical Computing, Vienna.

Davis,S. *et al.* (2012) *Methylumi: Handle Illumina Data*. R package version 2.0.13, R Foundation for Statistical Computing, Vienna.

Du,P. *et al.* (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Fackler,M.J. *et al.* (2011) Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. *Cancer Res.*, **71**, 6195–6207.

Hansen,K.D. and Aryee,M. (2011) *minfi: Analyze Illumina's 450k Methylation Arrays*. R package version 1.0.0, R Foundation for Statistical Computing, Vienna.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

Illumina (2011) *Infinium HumanMethylation450 BeadChip*. San Diego.

Javierre,B.M. *et al.* (2010) Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.*, **20**, 170–179.

Pinheiro,J. *et al.* (2012) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-103, R Foundation for Statistical Computing, Vienna.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. Roy. Stat. Soc. B.*, **64**, 479–498.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.