Bioinformatics, 32(8), 2016, 1214–1216
doi: 10.1093/bioinformatics/btv741
Advance Access Publication Date: 18 December 2015
Applications Note

OXFORD

Genome analysis

traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals

Li Chen[1] and Zhaohui S. Qin[2,3]*

[1]Department of Mathematics and Computer Science, [2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322 USA and [3]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.

Associate editor: John Hancock

Received on 16 October 2015; revised on 27 November 2015; accepted on 12 December 2015

Abstract

Summary: Genome-wide association studies (GWASs) have successfully identified many sequence variants that are significantly associated with common diseases and traits. Tens of thousands of such trait-associated SNPs have already been cataloged, which we believe form a great resource for genomic research. Recent studies have demonstrated that the collection of trait-associated SNPs can be exploited to indicate whether a given genomic interval or intervals are likely to be functionally connected with certain phenotypes or diseases. Despite this importance, currently, there is no ready-to-use computational tool able to connect genomic intervals to phenotypes. Here, we present *traseR*, an easy-to-use R Bioconductor package that performs enrichment analyses of trait-associated SNPs in arbitrary genomic intervals with flexible options, including testing method, type of background and inclusion of SNPs in LD.

Availability and implementation: The *traseR* R package preloaded with up-to-date collection of trait-associated SNPs are freely available in Bioconductor

Contact: zhaohui.qin@emory.edu

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

1 Introduction

Genome-wide association studies (GWASs) have been conducted *en masse* in the past decade and have been tremendously successful in identifying sequence variants that are significantly associated with common diseases and traits (Stranger *et al.*, 2011). To this day, thousands of GWAS have been conducted and reported, across diverse spectrums of diseases as well as qualitative and quantitative phenotypes. Resources, such as association result browser (http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm) and NHGRI GWAS catalog (Welter *et al.*, 2014), have been established to catalog all the trait-associated variants. Currently, the association result browser contains 44 124 association results (checked on October 10, 2015), which corresponds to 30 553 (autosomes plus chromosome X) unique trait-associated single nucleotide polymorphisms (taSNPs), linking to 573 diseases or phenotypes. We believe such a catalog of taSNPs offers scientists a unique perspective to explore and annotate the functional potential of any given genomic intervals.

Maurano *et al.* showed that regulatory DNA marked by deoxyribonuclease I (DNase I) hypersensitive sites (DHSs) was enriched with noncoding GWAS SNPs (Maurano *et al.*, 2012). Recent studies from ENCODE and Roadmap Epigenome consortia systematically examined enrichment of taSNPs in ChIP-seq peaks of transcription factors and histone marks, and unveiled biologically interesting connections (Roadmap Epigenomics *et al.*, 2015; Schaub *et al.*, 2012).

These results indicate the utilities of conducting taSNP enrichment analyses in genomic intervals of interest. We believe that it will be a powerful tool for researchers to be able to query any given set of genomic intervals to see whether taSNPs are enriched in these particular

neighbourhoods of the genome and more importantly, which specific traits show significant enrichment. Typical genomic intervals include ChIP-seq peaks, differentially methylated regions and putative enhancers. In this way, we can build hypotheses linking these intervals to phenotypes. This is similar to the gene ontology (GO) (Ashburner *et al.*, 2000) term enrichment analysis or the gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005) except that GO terms or functional categories are replaced by traits, and a set of genes is replaced by a set of genomic intervals. We believe taSNPs can bring important functional insights to genomic regions, especially intergenic regions. Despite the great utilities, currently there is no off-the-shelf computational tool available to carry out this non-trivial test. To cater for this demand, we developed an R Bioconductor package named *traseR*, TRait-Associated SNP EnRichment analysis, which offers a turnkey solution for enrichment analysis of taSNPs.

2 Methods

traseR provides multiple options, including testing method, type of background and inclusion of SNPs in linkage disequilibrium (LD), to conduct statistical tests of taSNP enrichment for a given set of query genomic intervals. We here provide a brief description. More details about these methods can be found in the Supplementary Material.

2.1. Background SNPs

All SNPs from the CEU panel of the phase I 1000 Genomes with minor allele frequency (MAF) greater than 0.05 are used as background SNPs (6 571 512 SNPs genome-wide excluding those from the Y chromosome). These SNPs have a comparable MAF distribution to the taSNP collection.

2.2 Enrichment tests

2.2.1 Contingency table-based tests

The null hypothesis assumes that the proportion of taSNPs among all SNPs is the same both within and outside of the query genomic intervals. We classify all SNPs into either within/outside (query genomic intervals), or taSNPs/non-taSNPs, then construct a two-by-two contingency table and run a χ^2 test or Fisher's exact test to assess the enrichment level of the taSNPs. Alternatively, we classify every single base in the genome (except for chromosome Y) into either within/outside (query genomic intervals), or taSNPs/non-taSNPs, and conduct the test accordingly.

2.2.2 Binomial test

The null hypothesis states that the chance of observing a SNP being a taSNP is the same in query genomic intervals as in the whole genome (excluding chromosome Y). Therefore, under the null hypothesis, the number of observed taSNPs out of all SNPs in the query genomic intervals follows a binomial distribution with probability equal to the genome-wide proportion of all taSNPs among all SNPs. Alternatively, we can use all bases in the genome as the background and conduct the test accordingly.

2.2.3 Non-parametric statistical testing procedure

Because typical query genomic intervals only span a small fraction of the whole genome, and the genome-wide distribution of SNPs is not uniform, it is desirable to conduct a nonparametric test in which a set of randomly selected control intervals is compared to the query genomic intervals for taSNP enrichment, rather than imposing distribution assumptions. For each permutation, *traseR* generates a matching control interval of the same size and on the same chromosome as each query genomic interval. The process is repeated 10 000 times (or a number specified by the user) to obtain an empirical *P*-value.

Table 1. Top-ranked traits for peripheral T cell H3K4me1 peaks

Trait	*P* value	OR	#taSNP hits	#taSNP
All	2.7e-48	1.5	1846	30 553
Behcet syndrome	4.4e-23	6.3	59	274
Diabetes mellitus, type 1	1.7e-11	5.0	33	185
Lupus erythematosus	6.2e-09	3.9	32	223
Arthritis, rheumatoid	1.4e-07	5.1	20	112
Multiple sclerosis	1.6e-05	2.9	26	236
Autoimmune diseases	5.2e-05	15.9	6	15

2.3. Linkage disequilibrium

As an option, *traseR* allows users to expand the taSNP set to include all the SNPs that are in tight linkage disequilibrium (LD) ($r^2 > 0.8$) with any of the taSNPs. The extended taSNP set contains 78 247 unique SNPs. Inclusion of SNPs in LD with the taSNPs is preferable if there is a limited number of taSNPs associated with the traits of interest.

3 Results

We collect a compendium of up-to-date taSNPs from dbGaP and NHGRI. There are 30 553 unique taSNPs spanning 573 phenotypes, all of which have been preloaded into the *traseR* package. *traseR* takes in a bed format input file that contains the query genomic intervals, then performs a user-specified enrichment test on all taSNPs combined, as well as taSNPs associated with each of the 573 traits. In the output, *traseR* reports the overall enrichment level of all taSNPs in the query genomic intervals, followed by a ranked list of traits that show statistically significant enrichment. Accordingly, *P* values, FDR *q*-values and odds ratios are also reported for each trait.

We demonstrate *traseR*'s utilities by displaying a sample result (H3K4me1 peak regions in peripheral T cell) (Roadmap Epigenomics *et al.*, 2015) (Table 1). The top-ranked traits are all immune-related. Moreover, peaks are significantly enriched with overall taSNPs. Here, we use the whole genome as background and binomial test as the testing method. More results can be found in the Supplementary Material.

Acknowledgements

We thank Drs. Yun Li, Yijuan Hu, Di Wu and Hao Wu for helpful discussion, improving the quality of the manuscript.

Funding

LC and ZS Qin are supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P01 GM085354.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Maurano,M.T. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.

Roadmap Epigenomics,C. *et al*. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Schaub,M.A. *et al*. (2012) Linking disease associations with regulatory information in the human genome. *Genome Res*., **22**, 1748–1759.

Stranger,B.E. *et al*. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Welter,D. *et al*. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., **42**, D1001–D1006.