Bioinformatics, 32(15), 2016, 2361–2363
doi: 10.1093/bioinformatics/btw130
Advance Access Publication Date: 7 March 2016
Applications Note

OXFORD

Genome analysis

Pathway analysis by randomization incorporating structure—PARIS: an update

Mariusz Butkiewicz[1,2,†], Jessica N. Cooke Bailey[1,2,†,*], Alex Frase[3], Scott Dudek[3], Brian L. Yaspan[4], Marylyn D. Ritchie[3], Sarah A. Pendergrass[3] and Jonathan L. Haines[1,2]

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA, [2]Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, USA, [3]Biomedical and Translational Informatics Program, Geisinger Health System, Danville, PA, USA and [4]Department of Human Genetics, Genentech, Inc, South San Francisco, CA, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: Alfonso Valencia

Received on December 18, 2015; revised on February 23, 2016; accepted on March 3, 2016

Abstract

Motivation: We present an update to the pathway enrichment analysis tool 'Pathway Analysis by Randomization Incorporating Structure (PARIS)' that determines aggregated association signals generated from genome-wide association study results. Pathway-based analyses highlight biological pathways associated with phenotypes. PARIS uses a unique permutation strategy to evaluate the genomic structure of interrogated pathways, through permutation testing of genomic features, thus eliminating many of the over-testing concerns arising with other pathway analysis approaches.
Results: We have updated PARIS to incorporate expanded pathway definitions through the incorporation of new expert knowledge from multiple database sources, through customized user provided pathways, and other improvements in user flexibility and functionality.
Availability and implementation: PARIS is freely available to all users at https://ritchielab.psu.edu/software/paris-download.
Contact: jnc43@case.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have been successful in identifying genetic variants that have revealed new insights into the genetic architecture of complex human diseases and traits. However, most reported GWAS variants confer only incremental risk and only explain a small proportion of disease heritability (Manolio *et al.*, 2009).

Pathway analysis algorithms were created to aggregate single-variant statistical analysis results (Yaspan and Veatch, 2011) to identify pathways with enrichment of genetic associations. These build a higher level abstraction of single-variant data and collapse it into biologically informed gene sets comprising pathways.

Version 1.0 of the Pathway Analysis by Randomization Incorporating Structure (PARIS) tool was created to evaluate aggregated association signals generated from GWAS experiments across pathways of interest including the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Yaspan *et al.*, 2011). PARIS groups the SNPs into linkage disequilibrium (LD) and linkage equilibrium (LE) features, defined based on available LD information for the population of interest. These features are further grouped into pathways (defined by online or manually curated sources) and analysis is performed using a file defining the composition of the defined pathways of interest, a file containing the association statistics, and a file defining the LD block regions for the target population. The significance of a pathway is determined by permutation testing of random pathways of similar composition

(composed of LD and LE blocks of similar size) (Yaspan *et al.*, 2011). In each permutation, the features in a pathway are replaced by a randomly selected set of features of similar size and number, thus mimicking the physical features of the pathway of interest. The total number of features with a significant *P*-value is compared between the true and random pathways. If the pathway of interest contains a higher number of significant features than the random pathways, the pathway of interest will then be considered significant. This approach was chosen to minimize execution time, to allow for a more flexible user input, and to correct the analysis for genome structure. PARIS additionally allows for assessment of the contribution of each gene to the overall pathway signal, providing information about whether or not it appears many genes are contributing to a signal, or instead one gene with many significant features contributes to the signal detected for a specific pathway. This provides a faster method compared to other methods that re-compute association statistics. The method by which PARIS corrects for gene size and LD were extensively tested and were reported in the previous article (Yaspan *et al.*, 2011).

The PARIS program only requires GWAS results that meet the input requirements (i.e. include variant/position, chromosome number and *P*-value).

Here we describe PARIS v2.4, a new Python implementation of PARIS, as part of the Biofilter 2.0 software suite (Pendergrass *et al.*, 2013).

2 Results

2.1 Novel features

Figure 1A outlines the updates to PARIS 2.4, where *orange* elements represent changes to the software tool.

The previous version 1.0 of PARIS was designed as a stand-alone program. Now, we have updated and adapted PARIS 2.4 to be part of Biofilter 2.0, which, via the Library of Knowledge Integration database, allows broader access to external databases, including KEGG, Gene Ontology (GO), Reactome, BioGRID, MINT, Pfam, PharmGKB and NetPath, in addition to manually curated pathways and provides a more flexible and user-friendly command line user interface.

As a result of integrating PARIS into Biolfilter 2.0, analysis options are extended to include all of those available in the Biofilter program (Pendergrass *et al.*, 2013).

Additionally, users now have the option to prioritize the permutation space to increase execution performance and omit unnecessary calculations for low-impact variants (with *P*-value above a user-defined threshold) that do not contribute to the final result. PARIS permutations will terminate early if the user specified *P*-value threshold is reached.

As part of the improved user interface, PARIS 2.4 now allows for better error identification compared to the previous version. Input errors will print to the screen or logfile when encountered. Additionally, PARIS 2.4 now channels error-prone results to an output file that separates unused input to '.invalid.' files. One new option is the ability to correct for errors in variant position mapping. In the case that a variant was mapped to a wrong chromosome, the program automatically re-maps the variant to the correct chromosome. This change is tracked in the log file and can be checked by the user.

A crucial new aspect in PARIS version 2.4 is the extension of feature boundaries. These can now be expanded beyond physical gene boundaries to match variants within user-specified distances outside the LD feature region of interest. PARIS is now able to utilize custom LD and LE information beyond the default HapMap Northern Europeans from Utah (CEU) population.

A guided example for using the PARIS 2.4 package is included in the **Supplementary Materials**.

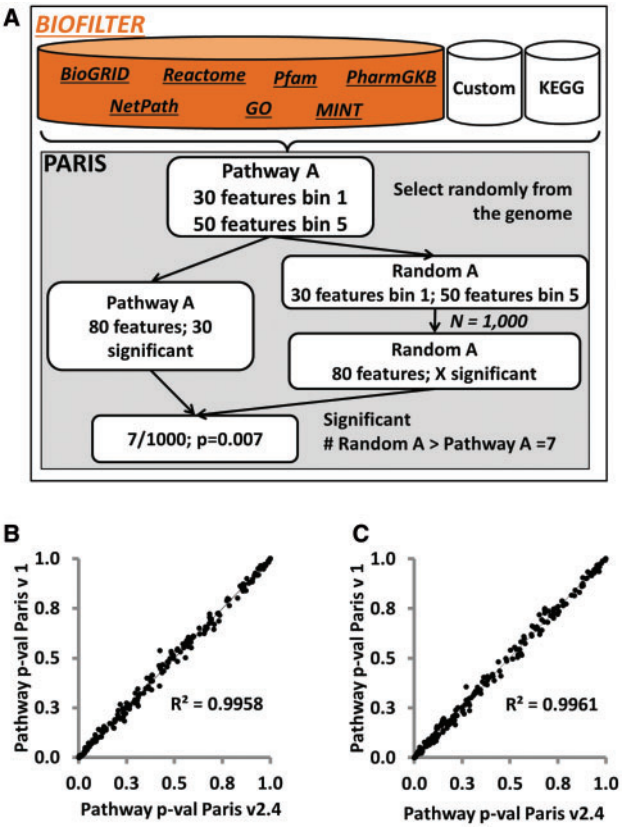

Fig. 1. (**A**) Visual representation of PARIS 2.4; *orange* elements represent updates. (**B and C**) Correlation of p-values obtained for 199 KEGG pathways investigated using gene boundary thresholds of 0 kb (**B**) and 50 kb (**C**)

Table 1. Pathway performance validation by significance thresholds in Versions 1.1.3 and 2.4.0 of PARIS

Gene boundary extension	Version	$P < 0.05$	$P \geq 0.05$
0 kb	1.1.3	26	173
	2.4.0	26	173
50 kb	1.1.3	39	160
	2.4.0	38	161

2.2 Performance validation

We executed PARIS on GWAS results from ~480 k SNPs evaluated by the NEIGHBOR consortium (Bailey *et al.*, 2014; Wiggs *et al.*, 2012). Comparing results from a prior version of PARIS (1.1.3) to this new version (2.4), results are highly correlated (Fig. 1B and C) for the 199 overlapping KEGG pathways, using gene boundary thresholds of 0 and 50 kb. To provide a comprehensive and unbiased comparison, we utilized the database and pathway definition files that were previously utilized with version 1.1.3 of PARIS (Bailey *et al.*, 2014). Pathway *P*-values were highly correlated, as indicated in Figure 1B and C. Further, we evaluated the number of pathways attaining $P < 0.05$ (Table 1) and found that with both the

0 and 50 kb gene boundary extension options, the same pathways achieved this significance threshold, except in one case, where a pathway was $P < 0.05$ in the prior version of PARIS and $P \geq 0.05$ in the updated version due to difference in random number generation between the two versions. These results validate the performance consistency between the two versions of PARIS.

3 Discussion

PARIS was originally developed to evaluate the non-random aggregation of *common* variant *single-marker* GWAS signals that do not necessarily attain GWAS-level significance but that approach significance and aggregate in biological pathways of interest, thus providing a method by which to detect the presence of concomitant biological trends across pathways and other connections between genes and biological features in terms of genetic associations.

The prior version of PARIS had limited access to outdated databases. As part of the Biofilter 2.0 package, PARIS 2.4 now accesses updated knowledge sources, thus relaying the most pertinent biological information relevant to analysis. PARIS 2.4 now incorporates a more user-friendly interface, in addition to accessing the options available in Biofilter 2.0 (Pendergrass *et al.*, 2013).

Limitations of PARIS include that results are limited to the scope of the input data, i.e. if GWAS results from an array with poor genomic coverage are evaluated by PARIS, the pathways of interest will necessarily be limited to those with variants present in the input. Additionally, PARIS 2.4 includes LD information from European populations (CEU); however, user-defined LD definitions can be used. Future versions will incorporate more diverse LD definitions.

In conclusion, PARIS version 2.4 allows for more flexible and extended utilization beyond the initial version 1.0 (Yaspan *et al.*, 2011) while also giving results that are consistent with the original implementation and thus providing an updated tool for secondary pathway analysis of genome-wide data.

Acknowledgements

The authors would like to thank the NEIGHBORHOOD consortium.

Funding

This work was supported by EY012118 and AG047133. The NEIGHBORHOOD is supported by National Institutes of Health/NEI (1R01EY022305) and by the Center for Inherited Disease Research (genotyping). J.N.C.B. is supported in part by a PhRMA Informatics Fellowship.

Conflict of Interest: none declared.

References

Bailey,J.N.C. *et al.* (2014) Hypothesis-independent pathway analysis implicates GABA and Acetyl-CoA metabolism in primary open-angle glaucoma and normal-pressure glaucoma. *Hum. Genet.*, **133**, 1319–1330.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Pendergrass,S.A. *et al.* (2013) Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min*, **6**, 25.

Wiggs,J.L. *et al.* (2012) Common variants at 9p21 and 8q22 are associated with increased susceptibility to optic nerve degeneration in glaucoma. *PLoS Genet.*, **8**, e1002654.

Yaspan, B.L. and Veatch, O.J. (2011). Strategies for pathway analysis from GWAS data. *Curr. Protoc. Hum. Genet.*, **71**, 1.20.1–1.20.15.

Yaspan,B.L. *et al.* (2011) Genetic analysis of biological pathway data through genomic randomization. *Hum. Genet.*, **129**, 563–571.