Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2016, 12, 477

ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization

Guangchuang Yu[ab] and Qing-Yu He*[a]

Received 5th October 2015,
Accepted 20th November 2015

DOI: 10.1039/c5mb00663e

www.rsc.org/molecularbiosystems

Reactome is a manually curated pathway annotation database for unveiling high-order biological pathways from high-throughput data. *ReactomePA* is an R/Bioconductor package providing enrichment analyses, including hypergeometric test and gene set enrichment analyses. A functional analysis can be applied to the genomic coordination obtained from a sequencing experiment to analyze the functional significance of genomic loci including *cis*-regulatory elements and non-coding regions. Comparison among different experiments is also supported. Moreover, *ReactomePA* provides several visualization functions to produce highly customizable, publication-quality figures. The source code and documents of *ReactomePA* are freely available through Bioconductor (http://www.bioconductor.org/packages/ReactomePA).

Introduction

High-throughput experiments generating large and complex datasets are routinely performed in biomedical research to identify and unravel the underlying pathways of complex diseases. However, characterizing disease signatures from omics data is challenging. Statistical analyses are employed to investigate gene-pathway associations to help explore biological questions in a pathway context.

Reactome[1] is a manually curated resource that describes chemical reactions, biological processes and pathways. It has been applied to expression[2] and proteomic data analysis[3] and helped identify altered pathways in cancer research. Although many webservers, tools and packages have been implemented for gene ontology (GO),[4] disease ontology (DO)[5] and the Kyoto Encyclopedia of Genes and Genomes (KEGG),[4] only Cytoscape plugins[6,7] have been developed for reactome pathway analysis. All of these Cytoscape plugins emphasize pathway exploration and visualization. In addition, most of the existing tools for functional analyses are gene-based, which ignore non-coding regulatory information. The Genomic Regions Enrichment of Annotations Tool (GREAT) proposed the idea of analysing the functional significance of *cis*-regulatory elements.[8] Such a sequence-based method incorporates the impact of non-coding regions, which are critical for many gene expression regulations. However, there is no existing tool that supports functional analysis of the reactome pathway at the sequence level.

To fill this gap, we present *ReactomePA*, which evaluates pathway associations with gene lists or the genomic coordination obtained from high-throughput genomic and proteomic studies.

Implementation and examples

ReactomePA is developed using the R statistical language and is released within the Bioconductor project.[9] It extends from the in-house developed package *DOSE*[5] and supports the hypergeometric testing and gene set enrichment analysis (GSEA)[10] of the reactome pathway. The *enrichPathway* function allows users to select an appropriate background of genes as the baseline. The *gsePathway* function supports GSEA to evaluate the enriched reactome pathways of high-throughput data. These approaches can be used to verify interesting altered pathways and to identify unanticipated pathway associations. Bonferroni, Benjamini, False Discovery rate and *q*-values are incorporated for multiple comparison corrections. *ReactomePA* provides several high-quality visualization functions to help interpret analysis results, including *barplot* and *dotplot* for summarizing enrichment results, *cnetplot* for visualizing the gene-pathway association network, the *enrichMap* function for viewing the enriched pathway network and *gseaplot* which visualizes the running sum of the enrichment scores and its association with the phenotype. Our in-house developed package, *ChIPseeker*,[11] was incorporated with *ReactomePA* and the pathway analyses can be applied to genomic coordinations obtained from RNA-seq or ChIP-seq experiments.

Case study 1: enrichment analysis of GTEx data

To demonstrate the utilities of *ReactomePA*, here we apply it to analyse data from the Genotype-Tissue Expression Project (GTEx) to identify pathway signatures of transcriptome variation.

[a] *Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, Institute of Life and Health Engineering, College of Life Science and Technology, Jinan University, Guangzhou, 510632, China.*
E-mail: tqyhe@jnu.edu.cn; Fax: +86-20-85227039; Tel: +86-20-85227039
[b] *State Key Laboratory of Emerging Infectious Disease, School of Public Health, The University of Hong Kong, Hong Kong SAR, China*

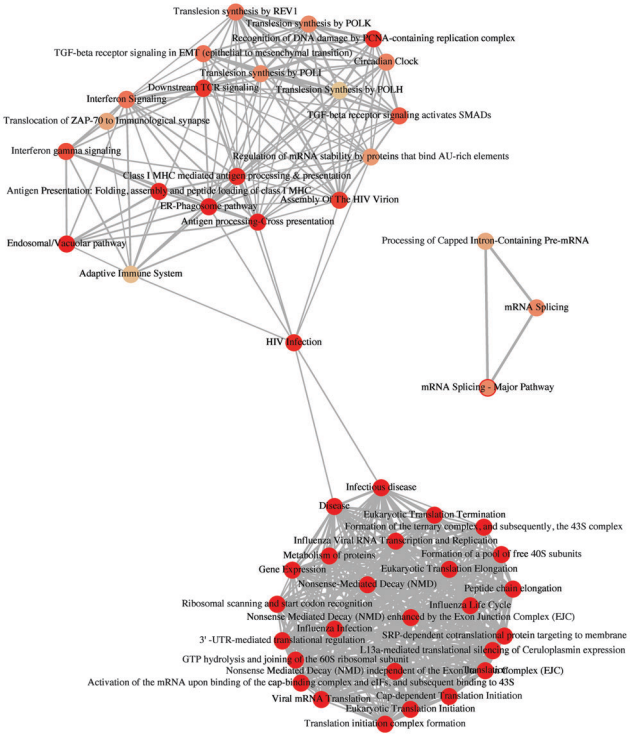

Fig. 1 Reactome enrichment analysis of GTEx data.

Variation in splicing plays critical roles in defining the individual and tissue phenotypes and associates with disease development.[12] A previous study demonstrated that genes with a high splicing variability among individuals are enriched in ribosomal proteins and related to translation by GO and KEGG analysis.[12] We used the top 2% of genes with a high variation in splicing to perform the hypergeometric test using *ReactomePA*. The final result was visualized by the *enrichMap* function, as shown in Fig. 1. The enrichment result contains three components. The highly condense component is translation and ribosomal protein related. This component is consistent with the discovery of the GO and KEGG analysis. In addition to this component, *ReactomePA* reports another two components that are not discovered by the GO and KEGG analysis. An isolated component that has mRNA splicing related pathways and, more importantly, a component containing several signalling pathways that are related to the translation component. These pathways, including *TGF-beta receptor signalling in EMT*[13] and *TGF-beta receptor signalling activated SMADs*,[14] are associated with splicing events. Immune related pathways that aren't reported in the paper[12] are also related to a high splicing variability.[15]

We analyse the whole data using GSEA *via* the *gsePathway* function. It reports a similar result with more detailed pathways. For example, *Signalling by Notch* is a conserved pathway in many cell types and at various stages during development, including in regulating adult neural stem cells.[16] The alternative splicing of Notch2 has been identified in a large fraction of AML patients.[17] The running enrichment scores and their association with a phenotype can be visualized simultaneously *via* the *gseaplot* function, as shown in Fig. 2.

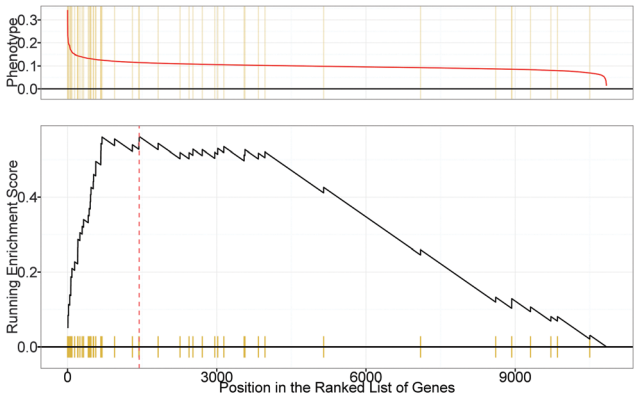

Fig. 2 Running enrichment scores and the phenotype association of the *Signalling by Notch* pathway.

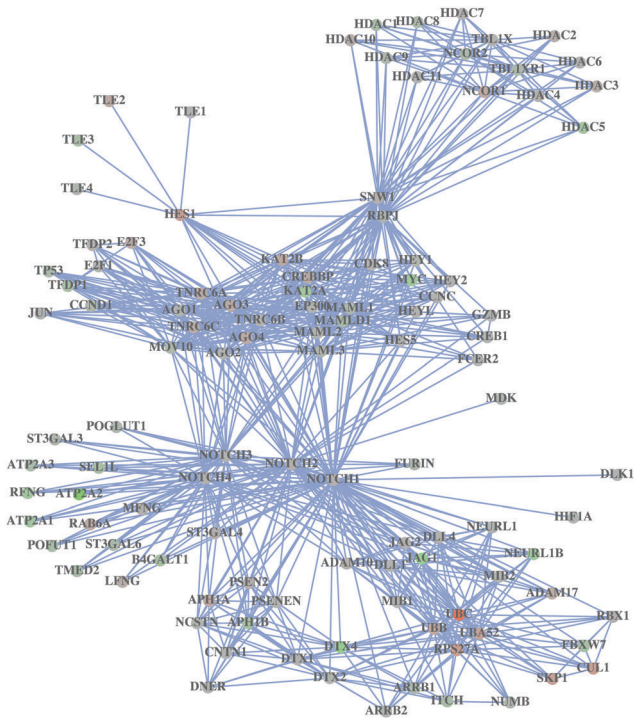

Fig. 3 *Signalling by Notch* pathway.

The corresponding pathway can be displayed *via* the *viewPathway* function, as demonstrated in Fig. 3.

Case study 2: enrichment analysis of NGS data

To demonstrate the usage of combining the reactome pathway analysis with NGS data by integrating the analysis with our in-house developed package *ChIPseeker*,[11] we collect ChIP-seq data of the CTCF binding sites in HMF (GEO accession number: GSM749665) and AoAF (GSM749666) cell lines[18] and Polycomb ortholog CBX6 (GSM1295076) and CBX7 (GSM1295077) binding sites in fibroblasts.[19] *ChIPseeker* provides a *seq2gene* function that links genomic regions to genes by many-to-many mapping. It considers the host gene (exon/intron), promoter region and flanking genes from the intergenomic region that may be under

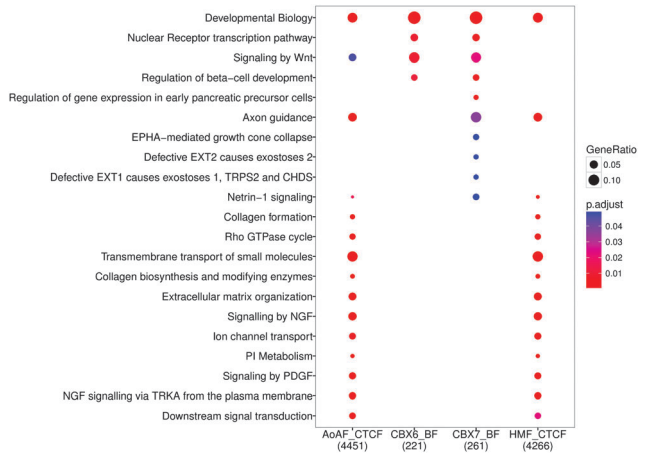

Fig. 4 Comparison of the enriched pathways from different experiments.

control *via cis*-regulation. We firstly annotate these ChIP-seq data to genes with this many-to-many scheme and then use *enrichPathway* to unveil the enriched pathways. Functional profiles of the different experiments or conditions can be compared *via* our in-house developed *clusterProfiler* package.[4] The final enrichment results of these four ChIP-seq data were visualized by the *clusterProfiler* as illustrated in Fig. 4.

Conclusions

ReactomePA is developed as an R package and released under the Artistic-2.0 License. It extends from *DOSE* to support the functional analysis of the reactome pathway and works seamlessly with *ChIPseeker* to analyse the functional pathway using variable NGS data. Comparison of different datasets is also supported *via clusterProfiler*. We provide these packages to work as an integrated pipeline for high-throughput analysis. *ReactomePA* supports several model organisms including *celegans*, *fly*, *human*, *mice*, *rat*, *yeast* and *zebrafish*. More importantly, *ReactomePA* provides users with the ability to produce highly customizable, publication quality figures. With these visualization methods, the results produced by *ReactomePA* are more interpretable.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (21271086).

Notes and references

1 D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein and P. D'Eustachio, *Nucleic Acids Res.*, 2014, **42**, D472–D477.
2 S. Jupe, A. Fabregat and H. Hermjakob, *Current Protocols in Bioinformatics*, ed. A. Baxevanis Al, 2015, vol. 49, p. 8.20.1–9.
3 R. Haw, H. Hermjakob, P. D'Eustachio and L. Stein, *Proteomics*, 2011, **11**, 3598–3613.
4 G. Yu, L.-G. Wang, Y. Han and Q.-Y. He, *OMICS: J. Integr. Biol.*, 2012, **16**, 284–287.
5 G. Yu, L.-G. Wang, G.-R. Yan and Q.-Y. He, *Bioinformatics*, 2015, **31**, 608–609.
6 W. P. A. Ligtenberg, D. Bošnački and P. A. J. Hilbers, *J. Bioinf. Comput. Biol.*, 2013, **11**, 1350004.
7 G. Wu, E. Dawson, A. Duong, R. Haw and L. Stein, *ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis*, F1000Research, 2014, vol. 3, p. 146.
8 C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger and G. Bejerano, *Nat. Biotechnol.*, 2010, **28**, 495–501.
9 W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron and M. Morgan, *Nat. Methods*, 2015, **12**, 115–121.
10 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.
11 G. Yu, L.-G. Wang and Q.-Y. He, *Bioinformatics*, 2015, **31**, 2382–2383.
12 M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, T. Gte. Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie and R. Guigó, *Science*, 2015, **348**, 660–665.
13 S. Lamouille, J. Xu and R. Derynck, *Nat. Rev. Mol. Cell Biol.*, 2014, **15**, 178–196.
14 S. Tao and K. Sampath, *Dev., Growth Differ.*, 2010, **52**, 335–342.
15 A. Belicha-Villanueva, J. Blickwedehl, S. McEvoy, M. Golding, S. O. Gollnick and N. Bangia, *Immunol. Res.*, 2010, **46**, 32–44.
16 J. L. Ables, J. J. Breunig, A. J. Eisch and P. Rakic, *Nat. Rev. Neurosci.*, 2011, **12**, 269–283.
17 C. Lobry, P. Oh, M. R. Mansour, A. T. Look and I. Aifantis, *Blood*, 2014, **123**, 2451–2459.
18 H. Wang, M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers and J. A. Stamatoyannopoulos, *Genome Res.*, 2012, **22**, 1680–1688.
19 H. Pemberton, E. Anderton, H. Patel, S. Brookes, H. Chandler, R. Palermo, J. Stock, M. Rodriguez-Niedenführ, T. Racek, L. de Breed, A. Stewart, N. Matthews and G. Peters, *Genome Biol.*, 2014, **15**, R23.