Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw392
Advance Access Publication Date: 26 June 2016
Applications Note

Gene expression

ABAEnrichment: an R package to test for gene set expression enrichment in the adult and developing human brain

Steffi Grote[1], Kay Prüfer[1], Janet Kelso[1,†] and Michael Dannemann[1,2,*,†]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany and [2]Medical Faculty, University of Leipzig, Leipzig 04103, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last 2 authors should be regarded as joint Last Authors.

Associate Editor: Oliver Stegle

Received on March 21, 2016; revised on May 27, 2016; accepted on June 16, 2016

Abstract

Summary: We present ABAEnrichment, an R package that tests for expression enrichment in specific brain regions at different developmental stages using expression information gathered from multiple regions of the adult and developing human brain, together with ontologically organized structural information about the brain, both provided by the Allen Brain Atlas. We validate ABAEnrichment by successfully recovering the origin of gene sets identified in specific brain cell-types and developmental stages.
Availability and Implementation: ABAEnrichment was implemented as an R package and is available under GPL ($\geq$ 2) from the Bioconductor website (http://bioconductor.org/packages/3.3/bioc/html/ABAEnrichment.html).
Contacts: steffi_grote@eva.mpg.de, kelso@eva.mpg.de or michael_dannemann@eva.mpg.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Functional enrichment analyses using ontologies such as the Gene Ontology have been widely used to gain insight into the functions and phenotypes influenced by sets of gene candidates that emerge from various genome-wide screens. Patterns of gene expression can be very informative about developmental processes and functions taking place in particular organs or tissues, but this requires extensive expression information, including component tissues and developmental stages, to be available for the organ/tissue of interest.

The Allen Human Brain Atlas projects have quantified gene expression in multiple regions of the human brain over the course of brain development, and provide a valuable resource that can be used to pinpoint sets of genes that are characteristic of particular brain regions and/or developmental stages (Hawrylycz *et al.*, 2012; Miller *et al.*, 2014; Allen Human Brain Atlas; human.brain-map.org and BrainSpan Atlas of the Developing Human Brain; brainspan.org). To date, the transcriptomes of multiple regions of the brains of adult humans as well as brains from individuals at various developmental stages have been measured (Supplementary Tables S1–S4). None of the tools developed to date (e.g. Brain Explorer, http://human.brain-map.org/static/brainexplorer) allow for the statistical evaluation of gene candidates in terms of expression enrichment in specific regions or stages using an ontology. Here, we demonstrate the utility of a tool to carry out enrichment analyses using the Allen Brain Atlas information.

2 Methods

In order to perform ontology gene set enrichment analyses for expression in specific brain regions it is necessary to assign to each anatomical structure the genes that are expressed in that structure.

We apply an expression threshold to identify expressed genes in this step. A workflow illustrating the entire data processing and enrichment analyses is shown in Supplementary Figure S1.

2.1 Gene expression data

We used two human brain expression datasets provided by the Allen Brain Atlas. First, gene expression in the adult human brain was measured in six adult individuals using a microarray (Microarray Survey, Oct. 2013 v.7, Supplementary Table S1), and consists of normalized expression profiles for ~16 000 genes in 414 brain regions (Microarray Data Normalization, March 2013 v.1 see Table S2). Second, the developing human brain dataset consists of normalized RNA-Seq expression measurement (RPKM, 'RNA-Seq Gencode v10 summarized to genes') of ~17 000 genes from 42 human donors classified into 31 different developmental periods ranging from 8 post-conceptual weeks to 40 years (Table S3). 16 of the 26 available brain regions were sampled in donors of at least 20 different ages, while the remaining 10 brain regions were sampled in fewer than five ages (Supplementary Tables S3 and S4, brainspan.org).

2.2 Processing of expression data

For the adult human brain datasets the normalized expression estimates for genes in brain regions were computed by averaging the expression levels across all samples and probe sets for a given gene and brain region (see Section 2.1). To increase the power to detect developmental effects we restricted to the 16 brain regions sampled in at least 20 age categories. When testing for expression enrichment across five developmental stages (Supplementary Table S5) we computed the mean RPKM expression value of a gene in all samples for a given brain region and developmental stage. The input expression datasets were provided by the Allen Brain Atlas and the processed data are available via our data package ABAData (http://bioconductor.org/packages/3.2/data/experiment/html/ABAData.html).

2.3 Ontologies

The Allen Brain Atlas provides non-overlapping ontologies that describe the developing and adult human brain and contain 3317 and 1534 brain regions, respectively. Direct expression data for the adult brain is available for 414 of the 1534 brain regions in the corresponding ontology. Genes in brain regions without direct expression measurements are defined to be expressed if at least one sub-region shows expression evidence for the given gene (Supplementary Table S6). For the developing human brain we restricted the corresponding ontology to the 16 brain regions with expression data available for at least 20 age time points. When including the superstructures inheriting expression data from these 16 brain regions, a total of 47 brain regions were used in the enrichment analysis (Supplementary Table S6).

2.4 Enrichment analyses

Genes are annotated to brain regions using default or user-defined expression cut-offs. Default cut-offs are the 10%-steps of expression quantiles across all brain regions. The enrichment analysis is then performed by the core function *aba_enrich* using either a hypergeometric test or a Wilcoxon-rank sum test as implemented in FUNC (Prüfer *et al.*, 2007, supplementary Figure 1). The full functionality of our package is described in our Bioconductor R vignette: https://bioconductor.org/packages/release/bioc/vignettes/ABAEnrichment/inst/doc/ABAEnrichment.html

3 Evaluation

To determine whether our approach accurately identifies genes that have previously been reported to show enrichment in different brain regions we tested the software on two datasets (Cahoy *et al.*, 2008; Kang *et al.*, 2011, supplementary data file). Cahoy *et al.* report a set of marker genes for oligodendrocytes. We selected the top 40 markers as candidates and tested for their enrichment in the adult brain using ABAEnrichment. Oligodendrocytes have been reported to be predominantly present in the white matter of the brain (Hagan *et al.*, 2012). Expression enrichment analysis indicated that the 40 marker genes are significantly enriched in multiple regions (min FWER < 0.05) of which the top five enriched regions represent all white matter tissues available (Supplementary Table S7). Secondly, we used genes that have been shown to be differentially expressed at different developmental stages in multiple human brain regions (Supplementary Table S8, Kang *et al.*, 2011). From the six brain regions Kang *et al.* reported, five were measured directly for the Allen Brain Atlas and one (neocortex) has substructures with direct expression measurements (Supplementary Table S9). Strikingly, when we restricted our analysis to regions where we have direct expression evidence in all cases the region Kang *et al.* reported had the lowest mean FWER across default expression cut-offs (Supplementary Table S10). For genes reported to be specifically expressed in neocortex we find that the brain region with the lowest mean FWER is in all cases a direct sub-region of neocortex.

4 Conclusions

ABAEnrichment provides a method for analyzing gene set expression enrichment in the brain regions assayed by the Allan Brain Atlas projects, and is the first software package to perform genome-wide analyses using this resource. We show that it is possible to use the fine-grained expression profiles provided by the Allen Brain Atlas projects to determine whether sets of candidate genes identified through analyses such as screens for positive selection (Scheinfeldt and Tishkoff, 2013), genome-wide association studies (Visscher *et al.*, 2012) and analyses of archaic introgression (Vattathil and Akey, 2015) are enriched in particular brain regions in the adult or developing brain. These expression patterns might provide insights into the processes in which these genes act, and this information can then be used to guide further functional experiments in cell lines, organoids or model organisms.

Acknowledgements

We would like to thank Udo Stenzel for technical assistance and the Max Planck Society for financial support.

Conflict of Interest: none declared.

References

Cahoy,J.D. *et al.* (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.*, **28**, 264–278.

Hagan,C.E. *et al.* (2012) 20 – nervous System. In: *Comparative Anatomy and Histology*. Academic Press, San Diego, pp. 339–394.

Hawrylycz,M.J. *et al.* (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, **489**, 391–399.

Kang,H.J. *et al*. (2011) Spatio-temporal transcriptome of the human brain. *Nature*, **478**, 483–489.

Miller,J.A. *et al*. (2014) Transcriptional landscape of the prenatal human brain. *Nature*, **508**, 199–206.

Prüfer,K. *et al*. (2007) FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics*, **8**, 41.

Scheinfeldt,L.B. and Tishkoff,S.A. (2013) Recent human adaptation: genomic approaches, interpretation and insights. *Nat. Rev. Genet*., **14**, 692–702.

Vattathil,S. and Akey,J.M. (2015) Small Amounts of Archaic Admixture Provide Big Insights into Human History. *Cell*, **163**, 281–284.

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