Bioconductor's PADOG package

Adi L. Tarca[1,2,3] Zonghui Xu[4] Gaurav Bhatti[1]

October 19, 2026

[1]Center for Molecular Medicine and Genetics, Wayne State University
[2]Department of Obstetrics and Gynecology, Wayne State University
[3]Department of Computer Science, Wayne State University
[4]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School

1 Overview

This package implements the ***P**athway **A**nalysis with **D**own-weighting of **O**verlapping **G**enes* (**PADOG**) algorithm described in Tarca et al. (2012) and a pathway / gene set analysis benchmark Tarca et al. (2013). The PADOG method can be applied to analyze any type of gene sets yet in here it is illustrated using KEGG pathways. The method computes a gene set score as the mean of absolute values of weighted moderated gene t-scores. The gene weights are chosen to favor genes appearing in few pathways versus genes that appear in many pathways. The significance of pathway scores is evaluated using sample/array labels permutation that preserve the gene-gene correlation structure. The package also contains a benchmark for gene set analysis in general and allows any new gene set analysis method to be benchmarked against PADOG or other existing methods (e.g. GSA). The original benchmark used 42 different data sets, each involving a disease (e.g. Colorectal Cancer) for which there is a KEGG or Metacore pathway with the same name. The only assumption we make (proven to hold in Tarca et al. (2012) and Tarca et al. (2013)) is that the KEGG's or Metacore's pathway with the same name as the disease under the study should be found significant and/or ranked near the top by gene set analysis methods when analyzing a dataset that studies the disease phenotype in question (cases vs controls).

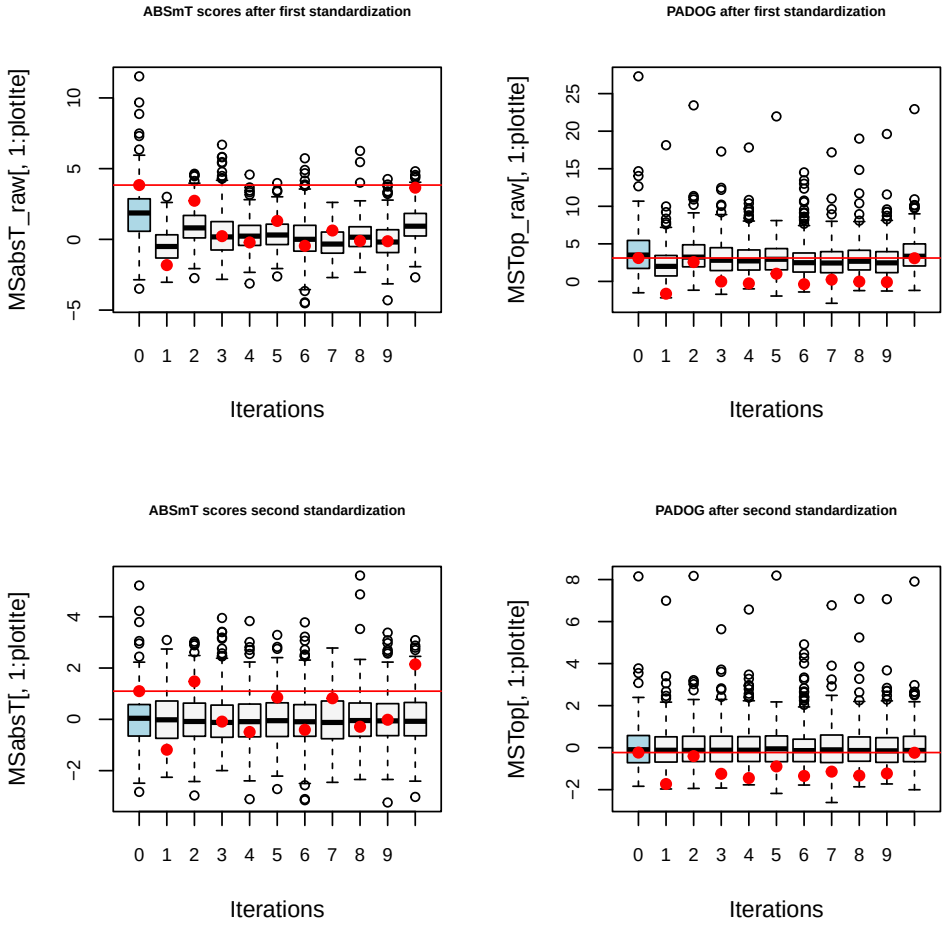
2 Pathway / gene set analysis with PADOG package

This document provides basic introduction on how to use the `PADOG` package. For extended description of the methods used by this package please consult Tarca et al. (2012) and Tarca et al. (2013). We demonstrate the functionality of this package using a colorectal cancer dataset obtained using Affymetrix GeneChip technology and available through GEO (GSE9348) and incorporated in the `KEGGdzPathwaysGEO` package. This experiment contains 12 normal samples and 70 colorectal cancer samples and is described in Hong et al. (2010). The RMA preprocessed data using the `affy` package is the entry point for the `padog` function:

```
> library(PADOG)
> set = "GSE9348" # this set is from "KEGGdzPathwaysGEO" collection
> #write a function to extract required info into a list
> getdataaslist = function(x) {
+     loadCompPADOGBenchmark(x)
+     x=get(x)
+     exp = experimentData(x)
+     dataset = exp@name
+     disease = notes(exp)$disease
+     dat.m = exprs(x)
+     ano = pData(x)
+     design = notes(exp)$design
+     annotation = paste(x@annotation, ".db", sep = "")
+     targetGeneSets = notes(exp)$targetGeneSets
+     list = list(dataset, disease, dat.m, ano, design, annotation, targetGeneSets)
+     names(list) = c("dataset", "disease", "dat.m", "ano", "design", "annotation",
+         "targetGeneSets")
+     return(list)
+ }
> dlist = getdataaslist(set)
> #run padog function on KEGG pathways
>
> #prepare human KEGG pathways as gene list for enrichment analysis with PADOG
> library(KEGGREST)
> res <- keggLink("pathway", "hsa")
> a=data.frame(path=gsub(paste("path:","hsa",sep=""),"",res),
+              gns=gsub(paste("hsa",":",sep=""),"",names(res)))
> KEGGpathways=tapply(a$gns,a$path,function(x){as.character(x)})
> KEGGpathwaysNames=keggList("pathway", "hsa")[paste("hsa",names(KEGGpathways),sep="")]
> names(KEGGpathwaysNames)<-names(KEGGpathways)
> #run padog on this dataset
> myr = padog(esetm = dlist$dat.m,
+     group = dlist$ano$Group,
+     paired = dlist$design == "Paired",
+     block = dlist$ano$Block,
+     targetgs = dlist$targetGeneSets,
+     annotation = dlist$annotation,
+     gslist = KEGGpathways,gs.names=KEGGpathwaysNames,
+     verbose = FALSE,
+     Nmin = 3,
+     NI = 100,
+     plots = TRUE,
+     dseed = 1
+ )
> myr$Name = substr(gsub("- Homo sapiens (human)", "", myr$Name, fixed = TRUE),1,36)
```

```
> myr[1:20,-c(4,5)]
```

	Name	ID	Size	PmeanAbsT	Ppadog
04976	Bile secretion	04976	77	0.0001	1e-04
01523	Antifolate resistance	01523	30	0.0500	1e-04
04981	Folate transport and metabolism	04981	28	0.1200	1e-04
04923	Regulation of lipolysis in adipocyte	04923	58	0.0300	1e-04
01521	EGFR tyrosine kinase inhibitor resis	01521	78	0.0900	1e-04
03008	Ribosome biogenesis in eukaryotes	03008	74	0.0200	1e-02
03013	Nucleocytoplasmic transport	03013	101	0.0200	1e-02
04156	Integrated stress response (ISR) sig	04156	51	0.0400	1e-02
00670	One carbon pool by folate	00670	34	0.0300	1e-02
00601	Glycosphingolipid biosynthesis - lac	00601	28	0.0500	1e-02
05222	Small cell lung cancer	05222	92	0.0600	1e-02
05210	Colorectal cancer	05210	86	0.0300	1e-02
04912	GnRH signaling pathway	04912	89	0.0800	1e-02
04110	Cell cycle	04110	157	0.0200	2e-02
03083	Polycomb repressive complex	03083	78	0.0300	2e-02
01212	Fatty acid metabolism	01212	55	0.0800	2e-02
00920	Sulfur metabolism	00920	10	0.0600	2e-02
00450	Selenocompound metabolism	00450	17	0.1600	2e-02
00071	Fatty acid degradation	00071	42	0.0400	3e-02
03020	RNA polymerase	03020	33	0.0800	3e-02



Note that for this colorectal cancer dataset it is reasonable to expect that the KEGG's Colorectal cancer pathway will be found significant and/or ranked close to the top. PmeanAbsT corresponds to the p-value obtained without using gene weights and hence the result is worse (higher p-value) compared to pPadog obtained by using the gene weights that are inversely related to how often the genes appear accross all gene sets to be analyzed. The plot created when `plots=TRUE` in the call to `padog` shows how gene weighting improves the gene set analysis for the target pathway set via the `targetgs` argument. Figure above shows the distribution of pathway/gene set scores (y axis) for PADOG and ABSmT (which is PADOG without gene weights) after the first standardization (row randomization) and after second standardization (between gene sets standardization). The x axis represents the number of iterations. Iteration 0 uses true class labels, all others used randomly permuted labels. The target pathway (set via the `targetgs` argument) in this dataset is the *Colorectal Cancer pathway* (KEGG ID 05210). Its score is shown with a red bullet throughout all 4 panels, and a red horizontal line marks its level when obtained with the true class labels ($ite = 0$, x-axis). The box plots of scores obtained with the true class labels are also highlighted in blue. With PADOG, after the second standardization, the target pathway scores obtained from permutations are less frequently above the red line (more extreme) than for teh raw scores ABSmT.

To run PADOG in parallel, you need to have package `doParallel` installed, and set `parallel = TRUE` in the call to `padog`:

3 Benchmark of gene set analysis methods

The entire collection of 42 datasets is available from data packages `KEGGdzPathwaysGEO` and `KEGGandMetacoreDzPatl` These datasets can be used to benchmark PADOG against existing approaches. The subset of 24 datasets part of the KEGGdzPathwaysGEO collection is given in Table 1:

Table 1: **The 24 datasets part of the KEGGdzPathwaysGEO collection**

GEOID	Pubmed	Ref.	Disease/Target pathway	KEGGID	Tissue
GSE1297	14769913	Blalock et al. (2004)	Alzheimer's Disease	hsa05010	Hippocampal CA1
GSE5281	17077275	Liang et al. (2007)	Alzheimer's Disease	hsa05010	Brain, Entorhinal Cortex
GSE5281	17077275	Liang et al. (2007)	Alzheimer's Disease	hsa05010	Brain, hippocampus
GSE5281	17077275	Liang et al. (2007)	Alzheimer's Disease	hsa05010	Brain, Primary visual cortex
GSE20153	20926834	Zheng et al. (2010)	Parkinson's disease	hsa05012	Lymphoblasts
GSE20291	15965975	Zhang et al. (2005)	Parkinson's disease	hsa05012	Postmortem brain putamen
GSE8762	17724341	Runne et al. (2007)	Huntington's disease	hsa05016	Lymphocytes (blood)
GSE4107	17317818	Hong et al. (2007)	Colorectal Cancer	hsa05210	Mucosa
GSE8671	18171984	Sabates-Bellver et al. (2007)	Colorectal Cancer	hsa05210	Colon
GSE9348	20143136	Hong et al. (2010)	Colorectal Cancer	hsa05210	Colon
GSE14762	19252501	Wang et al. (2009)	Renal Cancer	hsa05211	Kidney
GSE781	14641932	Lenburg et al. (2003)	Renal Cancer	hsa05211	Kidney
GSE15471	19260470	Badea et al. (2008)	Pancreatic Cancer	hsa05212	Pancreas
GSE16515	19732725	Pei et al. (2009)	Pancreatic Cancer	hsa05212	Pancreas
GSE19728		-	Glioma	hsa05214	Brain
GSE21354		-	Glioma	hsa05214	Brain, Spine
GSE6956	18245496	Wallace et al. (2008)	Prostate Cancer	hsa05215	Prostate
GSE6956	18245496	Wallace et al. (2008)	Prostate Cancer	hsa05215	Prostate
GSE3467	16365291	He et al. (2005)	Thyroid Cancer	hsa05216	Thyroid
GSE3678		-	Thyroid Cancer	hsa05216	Thyroid
GSE9476	17910043	Stirewalt et al. (2008)	Acute myeloid leukemia	hsa05221	Blood, Bone marrow
GSE18842	20878980	Sanchez-Palencia et al. (2010)	Non-Small Cell Lung Cancer	hsa05223	Lung
GSE19188	20421987	Hou et al. (2010)	Non-Small Cell Lung Cancer	hsa05223	Lung
GSE3585	17045896	Barth et al. (2006)	Dilated cardiomyopathy	hsa05414	Heart

All datasets used to compare the pathway analysis methods were obtained from GEO.

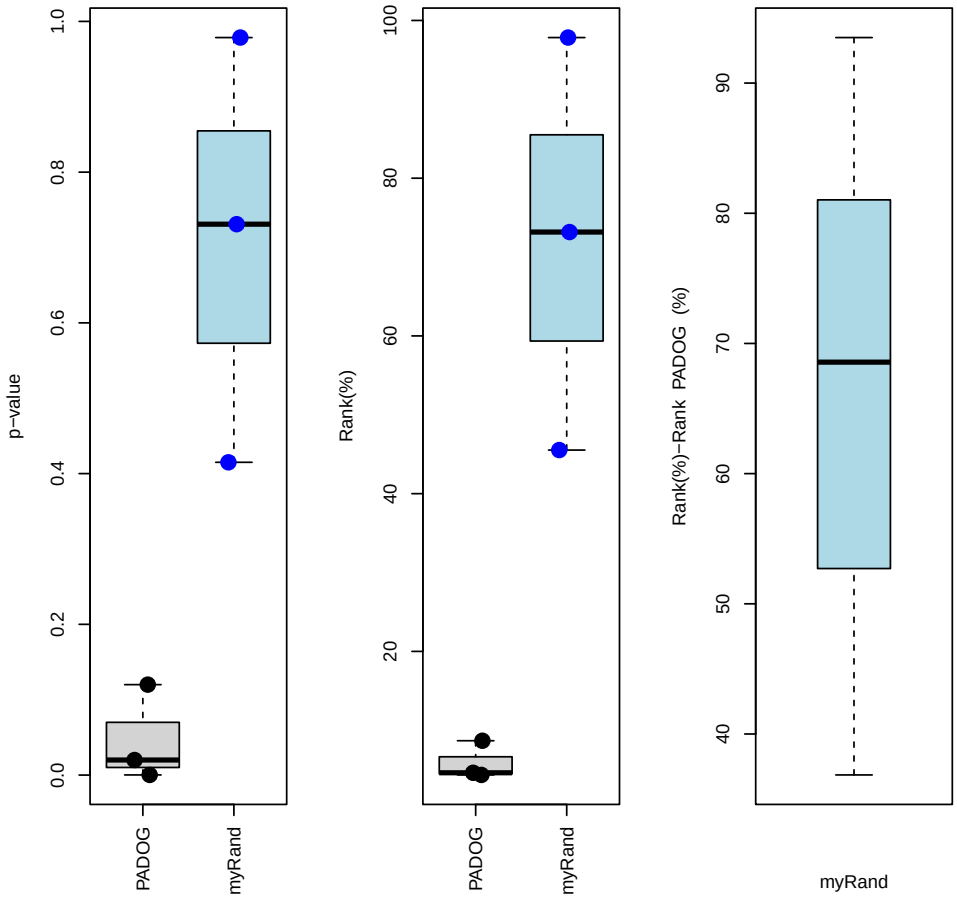
To illustrate how to compare PADOG against a user defined gene set analysis method we create a function called `randomF` that assigns random uniform P-values to gene sets. The user defined function has to take in 3 arguments:

1. `set`: the name of a dataset available from `KEGGdzPathwaysGEO` or `KEGGandMetacoreDzPathwaysGEO`;
2. `mygslist`: a list with elements being vectors of gene ids for a given geneset
3. `minsize`: minimum number of genes in a geneset to be considered for analysis

The output should be a dataframe with columns: ID, P, Rank, Dataset, Method for the geneset(s) considered to be relevant in that dataset (targetGeneSets).

```
> randFun = function(mname = "myRand") {
+     #a helper function to pass additional variables to your method
+     getdataaslist = getdataaslist
+     return(function(set, mygslist, minsize) {#your method function
+         #load the dataset and extract the required info using the function getdataaslist
+         dlist = getdataaslist(set)
+         #get rid of duplicates probesets per ENTREZ ID by keeping the probeset
+         #with smallest p-value (computed using limma)
+         aT1 = filteranot(esetm = dlist$dat.m, group = dlist$ano$Group,
+             paired = dlist$design == "Paired", block = dlist$ano$Block,
+             annotation = dlist$annotation)
+         #create an output dataframe for this toy method with random gene set p-values
+         mygslistSize = unlist(lapply(mygslist, function(x) {
+             length(intersect(aT1$ENTREZID, x))}))
+         res = data.frame(ID = names(mygslist), P = runif(length(mygslist)),
+             Size = mygslistSize, stringsAsFactors = FALSE)
+         #drop genesets with less than minsize genes in the current dataset
+         res = res[res$Size >= minsize,]
+         res$FDR = p.adjust(res$P,"fdr")
+         #compute ranks
+         res$Rank = rank(res$P) / dim(res)[1]*100
+         #needed to compare ranks between methods; must be the same as given
+         #in mymethods argument "list(myRand="
+         res$Method = mname
+         #needed because comparisons of ranks between methods is paired at dataset level
+         res$Dataset = dlist$dataset
+         #output only result for targetGeneSets (expected to be relevant)
+         return(res[res$ID %in% dlist$targetGeneSets,])
+         }
+     )
+ }
> randomF = randFun()
> #compare myRand method against PADOG on 3 datasets only
> mysets = c("GSE9348","GSE8671","GSE1297")
> out = compPADOG(datasets = mysets, existingMethods = c("PADOG"),
+     mymethods = list(myRand = randomF),
+     gslist = KEGGpathways, Nmin = 3, NI = 50, plots = TRUE,
+     verbose=FALSE, parallel = FALSE, dseed = 1, pkgs = NULL)
> print(out)

$summary
       Method    p geomean        p med % p.value<0.05 % q.value<0.05
PADOG   PADOG 0.007829....         0.02          66.67          33.33
myRand myRand 0.667067.... 0.730932....             0              0
         rank mean     rank med p Wilcox.     p LME coef. LME
PADOG  5.871725.... 4.607046....         1 1.0000000   0.00000
```

```
myRand 72.17705.... 73.17073....         1 0.9900699  66.30533

$ranks
$ranks$PADOG
[1] 8.672087 4.336043 4.607046

$ranks$myRand
[1] 45.52846 97.83198 73.17073


$pvalues
$pvalues$PADOG
[1] 0.1200 0.0002 0.0200

$pvalues$myRand
[1] 0.4149973 0.9785582 0.7309329


$qvalues
$qvalues$PADOG
[1] 0.963560831 0.004341176 0.410000000

$qvalues$myRand
[1] 0.9115120 0.9974737 0.9890592
```



Details about the meaning of the columns in the out table are given in Tarca et al. (2012). The better the method, the smaller the p-values and ranks for the target pathways, since these are supposed to be significant to their respective datasets.
For a more meaningful benchmark one can compare the same three methods on all 24 datasets of the KEGGdzPathwaysGEO collection + the 10 of the 18 datasets from the KEGGandMetacoreDzPathwaysGEO collection that had a KEGG target pathway (34 datasets total).

```
>  allsets=loadCompPADOGBenchmark(dataset = NULL, collection = "all",assign=FALSE)
>  metacoresets=c("GSE11906","GSE19420","GSE20164","GSE22780","GSE30153","GSE38666_epithelia","G
>  keggsets=setdiff(allsets,metacoresets)
>  out = compPADOG(datasets = keggsets, collection="all",
+                 existingMethods = c("GSA","PADOG"),
+     mymethods = list(myRand = randomF), gslist = KEGGpathways,
+     Nmin = 3, NI = 1000, plots = TRUE, verbose=TRUE,
+     parallel = TRUE, dseed = 1, pkgs = NULL)
```

Note, in the original benchmark in Tarca et al. (2013) 8 of the 18 datasets from the KEGGandMetacoreDzPathwaysGEO collection had a target pathway defined in the Metacore database instead of

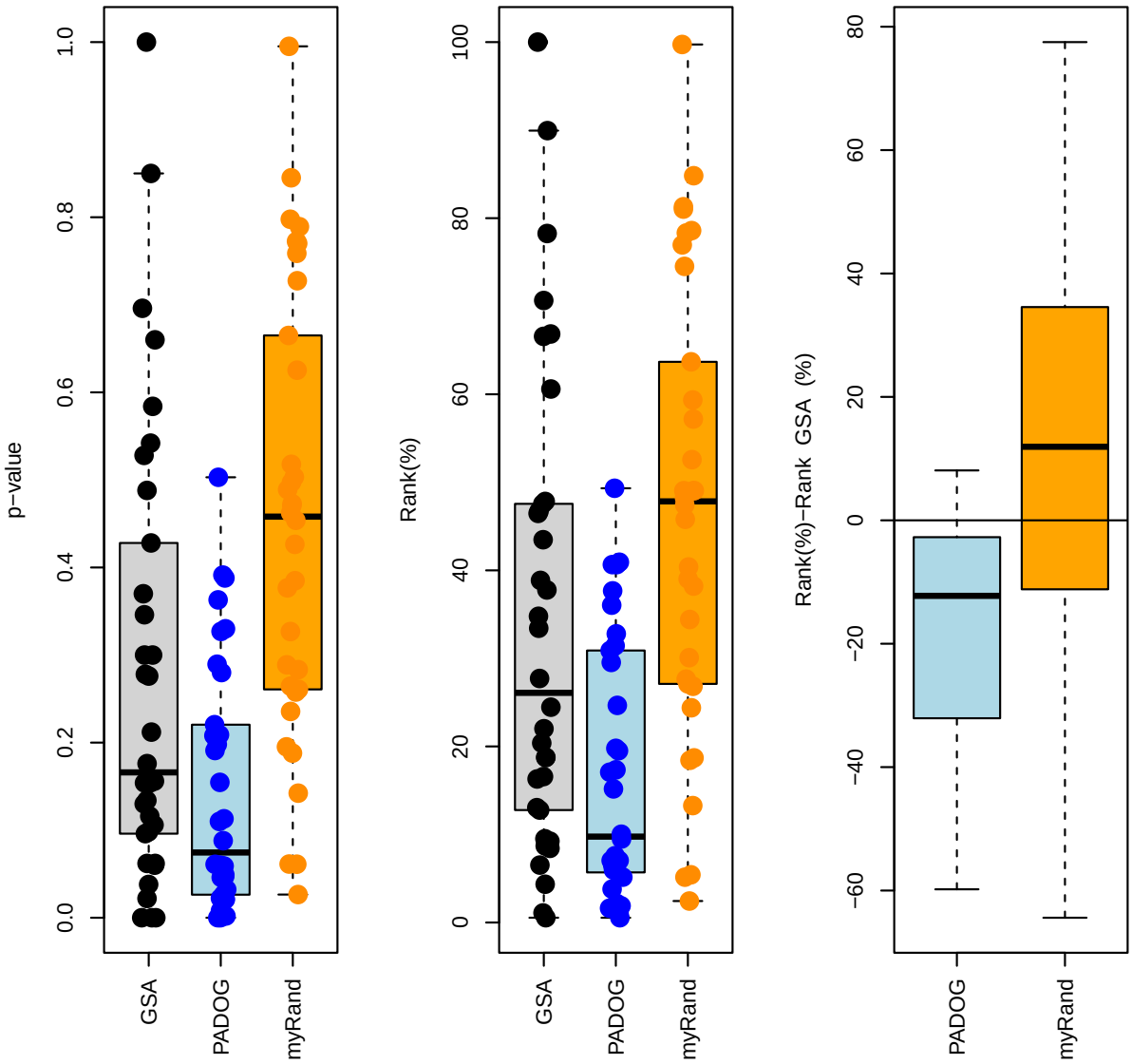


Figure 1: Comparison of PADOG, GSA and myRand enrichment methods on 34 datasets using 1,000 permutations per method.

KEGG, and hence can not be considered herein since the Metacore pathway definitions are not publicly available. In the benchmark above, the emerging trend is that PADOG returns smaller p-values and better rankings based on those p-values for pathways expected to be relevant in the specific datasets involved, relative to GSA and the random enrichment function (myRand).

4 Changes in version 1.6

The main change in this version is that loading data from the original KEGGandMetacoreDzPathwaysGEO and KEGGdzPathwaysGEO data packages through the function `loadCompPADOGBenchmark()`.

References

L. Badea, V. Herlea, S. O. Dima, T. Dumitrascu, and I. Popescu. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology*, 55:2016–2027, 2008.

A. S. Barth, R. Kuner, A. Buness, M. Ruschhaupt, S. Merk, L. Zwermann, S. Kaab, E. Kreuzer, G. Steinbeck, U. Mansmann, A. Poustka, M. Nabauer, and H. Sultmann. Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *J. Am. Coll. Cardiol.*, 48:1610–1617, Oct 2006.

E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. U.S.A.*, 101:2173–2178, Feb 2004.

H. He, K. Jazdzewski, W. Li, S. Liyanarachchi, R. Nagy, S. Volinia, G. A. Calin, C. G. Liu, K. Franssila, S. Suster, R. T. Kloos, C. M. Croce, and A. de la Chapelle. The role of microRNA genes in papillary thyroid carcinoma. *Proc. Natl. Acad. Sci. U.S.A.*, 102:19075–19080, Dec 2005.

Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin. Cancer Res.*, 13:1107–1114, Feb 2007.

Y. Hong, T. Downey, K. W. Eu, P. K. Koh, and P. Y. Cheah. A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clin. Exp. Metastasis*, 27:83–90, Feb 2010.

J. Hou, J. Aerts, B. den Hamer, W. van Ijcken, M. den Bakker, P. Riegman, C. van der Leest, P. van der Spek, J. A. Foekens, H. C. Hoogsteden, F. Grosveld, and S. Philipsen. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*, 5: e10312, 2010.

M. E. Lenburg, L. S. Liou, N. P. Gerry, G. M. Frampton, H. T. Cohen, and M. F. Christman. Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*, 3:31, Nov 2003.

W. S. Liang, T. Dunckley, T. G. Beach, A. Grover, D. Mastroeni, D. G. Walker, R. J. Caselli, W. A. Kukull, D. McKeel, J. C. Morris, C. Hulette, D. Schmechel, G. E. Alexander, E. M.

Reiman, J. Rogers, and D. A. Stephan. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics*, 28:311–322, Feb 2007.

H. Pei, L. Li, B. L. Fridley, G. D. Jenkins, K. R. Kalari, W. Lingle, G. Petersen, Z. Lou, and L. Wang. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer Cell*, 16:259–266, Sep 2009.

H. Runne, A. Kuhn, E. J. Wild, W. Pratyaksha, M. Kristiansen, J. D. Isaacs, E. Regulier, M. Delorenzi, S. J. Tabrizi, and R. Luthi-Carter. Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood. *Proc. Natl. Acad. Sci. U.S.A.*, 104:14424–14429, Sep 2007.

J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti, J. Luz, T. V. Ranalli, V. Gomes, A. Pastorelli, R. Faggiani, M. Anti, J. Jiricny, H. Clevers, and G. Marra. Transcriptome profile of human colorectal adenomas. *Mol. Cancer Res.*, 5:1263–1275, Dec 2007.

A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell, and M. E. Farez-Vidal. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer*, Sep 2010.

D. L. Stirewalt, S. Meshinchi, K. J. Kopecky, W. Fan, E. L. Pogosova-Agadjanyan, J. H. Engel, M. R. Cronk, K. S. Dorcy, A. R. McQuary, D. Hockenbery, B. Wood, S. Heimfeld, and J. P. Radich. Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer*, 47:8–20, Jan 2008.

A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bionformatics*, 13(136), 2012.

A. L. Tarca, G. Bhatti, and R. Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *Plos One*, 8(11), 2013.

T. A. Wallace, R. L. Prueitt, M. Yi, T. M. Howe, J. W. Gillespie, H. G. Yfantis, R. M. Stephens, N. E. Caporaso, C. A. Loffredo, and S. Ambs. Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res.*, 68:927–936, Feb 2008.

Y. Wang, O. Roche, M. S. Yan, G. Finak, A. J. Evans, J. L. Metcalf, B. E. Hast, S. C. Hanna, B. Wondergem, K. A. Furge, M. S. Irwin, W. Y. Kim, B. T. Teh, S. Grinstein, M. Park, P. A. Marsden, and M. Ohh. Regulation of endocytosis via the oxygen-sensing pathway. *Nat. Med.*, 15:319–324, Mar 2009.

Y. Zhang, M. James, F. A. Middleton, and R. L. Davis. Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 137B:5–16, Aug 2005.

B. Zheng, Z. Liao, J. J. Locascio, K. A. Lesniak, S. S. Roderick, M. L. Watt, A. C. Eklund, Y. Zhang-James, P. D. Kim, M. A. Hauser, E. Grunblatt, L. B. Moran, S. A. Mandel, P. Riederer, R. M.

Miller, H. J. Federoff, U. Wullner, S. Papapetropoulos, M. B. Youdim, I. Cantuti-Castelvetri, A. B. Young, J. M. Vance, R. L. Davis, J. C. Hedreen, C. H. Adler, T. G. Beach, M. B. Graeber, F. A. Middleton, J. C. Rochet, and C. R. Scherzer. PGC-1?, a potential therapeutic target for early intervention in Parkinson's disease. *Sci Transl Med*, 2:52ra73, Oct 2010.