Cluster analysis of replicated alternative polyadenylation data using shrinkage canonical correlation analysis – the PASCCA package

2018-09-07

1 Overview

The package *PASCCA* is an easy-to-use R package for analyses of APA related gene expression, including the characterization of poly(A) sites, quantification of association between genes with/without repeated measurements, clustering of APA-related genes to infer significant APA specific gene modules, and the evaluation of clustering performance with a variety of indexes. By providing a better treatment of the noise inherent in repeated measurements and taking into account multiple layers of poly(A) site data, *PASCCA* could be a general tool for clustering and analyzing APA-specific gene expression data.

2 Installation

You can install the *PASCCA* package using the following commands on the R console:

```
>install.packages("devtools")
>library(devtools)
>install_github("BMILAB/PASCCA")
>library(PASCCA)
```

3 Preparations

PASCCA needs APA-related gene expression matrix with/without repeated measurements as input, the first column is poly(A) or exon names, the second column is gene names, and the remaining column is sample names under different biological conditions such as different tissues, cell types and developmental stages. If the set of samples have repeated measurements, the order of the samples in data must be arranged in from big to small according to the number of replicates. **Table 1** illustrates the input data.

Table 1. The example data as PASCCA input.

ChrStrandCoord	Gene	dry_seed1	dry_seed2	dry_seed3	embryo1	embryo2	shoot1
Chr1+10817	LOC_Os01g01010	29	20	22	33	40	44
Chr1+8793	LOC_Os01g01010	15	11	5	1	2	0
Chr1+19858	LOC_Os01g01040	14	5	7	12	10	4
Chr1+20319	LOC_Os01g01040	20	25	46	31	28	34
Chr1+19709	LOC_Os01g01040	0	8	4	1	2	1
Chr1+79765	LOC_Os01g01150	3	0	0	0	2	18
Chr1+77040	LOC_Os01g01150	0	1	0	7	2	0
Chr1+78646	LOC_Os01g01150	15	4	14	15	7	0
Chr1+77944	LOC_Os01g01150	3	8	7	3	1	0
Chr1+79018	LOC_Os01g01150	5	15	7	9	6	0

4 Standard analysis work-flow

PASCCA consists of three main function *PAprocess*, *PASCCA*, *PASCCluster* to data preprocessing, distance matrix computation and hierarchical clustering, respectively.

4.1 Data preprocessing

Data preprocessing is an important step in the data mining process. First, we use the function *PAprocess* to do preliminary processing of APA-related gene expression data. It provides two steps taken to pre-process data.

- Data cleaning – To filter out genes with one poly(A) site from gene expression data.
- Data transformation – To transform the count data to the log2 scale by setting the parameter *log=TRUE* (default: TRUE)

```
> ##Loading example data
> data(polyA_example_data2)
> dim(data2)
[1] 200  44
> class(data2)
[1] "data.frame"
> data2[1:3,1:4]
       chrstrandcoord          gene dry_seed1 dry_seed2
1 AAAA02035470.1-50840 BGIOSGA037882         0         0
```

```
2 AAAA02035470.1-51650 BGIOSGA037882        5        3
3 AAAA02035470.1-49375 BGIOSGA037883       42       44
>
> ##Data preprocessing
> pre_data <- PAprocess(data2,log=TRUE)
> dim(pre_data)
[1] 101  44
> pre_data[1:3,1:4]
        chrstrandcoord          gene dry_seed1 dry_seed2
1 AAAA02035470.1-50840 BGIOSGA037882  0.000000  0.000000
2 AAAA02035470.1-51650 BGIOSGA037882  2.321928  1.584963
7 AAAA02035470.1+35963 BGIOSGA037889  0.000000  0.000000
```

4.2 Distance matrix computation

Second, based on processed expression data, to calculate the distance between genes with multiple poly(A) sites by the function *PASCCA* with seven parameters (*data, alpha, repli, tissues, tiss*).

1) *data*: the APA-related gene expression, the first column is poly(A) or exon names, the second column is gene names, and the remaining column is sample names under different biological conditions such as different tissues, cell types and developmental stages. If the set of samples have repeated measurements, the order of the samples in data must be arranged in from big to small according to the number of replicates.
2) *alpha:* the cut-off value of the significance level. We accept the null hypothesis if the significance level is above the cut-off value. It means the confidence interval is 95% when the alpha is 0.05. The default value of alpha is 0.05.
3) *repli:* the numbers of replicates per biological condition such as different tissues, cell types and developmental stages. Note that it needs to be in the same order as the input.
4) *tissues:* the total number of biological conditions. If the input data consists of root with three biological replicates, seed with three biological replicates and flower with two biological replicates, the *tissues* will be three because there are three conditions (root\seed\flower).
5) *tiss:* the frequency of the first type of repetition. If the input data consists of root with three biological replicates, seed with three biological replicates and flower with two biological replicates, the *tiss* will be two since both root and seed have three biological replicates.

```
> ##Getting information of the samples
> sample_name <- colnames(pre_data)[3:ncol(pre_data)]
> sample_name <- strsplit(sample_name,"\\d$")
> sample_name <- paste("",lapply(sample_name,"[[",1),sep="");
> table(sample_name)
sample_name
anther dry_seed embryo endosperm  husk
  3      3         3        3      3
 imbibed_seed leaf_20days leaf_60days mature_pollen pistil 3
    3           3           3           3
  root_5days   root_60days   shoot   stem_60days
       3           3           3           3

> ##Getting the number of repetitions per sample
> sample_replicates <- as.numeric(table(sample_name))
> sample_replicates <- sample_replicates[order(sample_replicates,
+                           decreasing = TRUE)]
>
> ##Calculationg PASCCA distance matrix
> gene_dist <- PASCCA(pre_data, alpha = 0.05,
+                 repli=sample_replicates,
+                 tissues=length(unique(sample_name)),
+                 tiss=sum(sample_replicates==sample_replicate[1]))
> str(gene_dist)
 num [1:46, 1:46] 0 0.687 0.273 0 0 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:46] "BGIOSGA000003" "BGIOSGA000004" "BGIOSGA000006"
"BGIOSGA000007" ...
  ..$ : chr [1:46] "BGIOSGA000003" "BGIOSGA000004" "BGIOSGA000006"
"BGIOSGA000007" ...
> gene_dist[1:3,1:3]
            BGIOSGA000003 BGIOSGA000004 BGIOSGA000006
BGIOSGA000003    0.0000000    0.6872573    0.2728280
BGIOSGA000004    0.6872573    0.0000000    0.4044428
BGIOSGA000006    0.2728280    0.4044428    0.0000000
> #or
> gene_dist <- PASCCA(pre_data, alpha = 0.05,repli=c(rep(3,14)), ti
ssues=14, tiss=14)
```

4.3 Clustering analysis

Distances of all gene pairs obtained by the function *PASCCA*, then the distance matrix is further used for clustering by the function *PASCCluster* with three parameters (*dist, nc, plot*). We adopted the widely-used clustering method, hierarchical clustering, to cluster genes, which was implemented by the R function using *hclust* default parameters. *PASCCluster* returns a list, including an object of class *hclust* which describes the tree produced by the clustering process and a vector with group memberships by *cutree*. Besides, when the parameter *plot* is TRUE, it will generate the following dendrogram **Figure 1.**

1) *dist:* a dissimilarity matrix as produced by the function *PASCCA*.
2) *nc:* numeric scalar (OR a vector) with the number of clusters the tree should be cut into.
3) *plot:* plot clustering tree of a hierarchical clustering if the value is TRUE (default: FALSE)

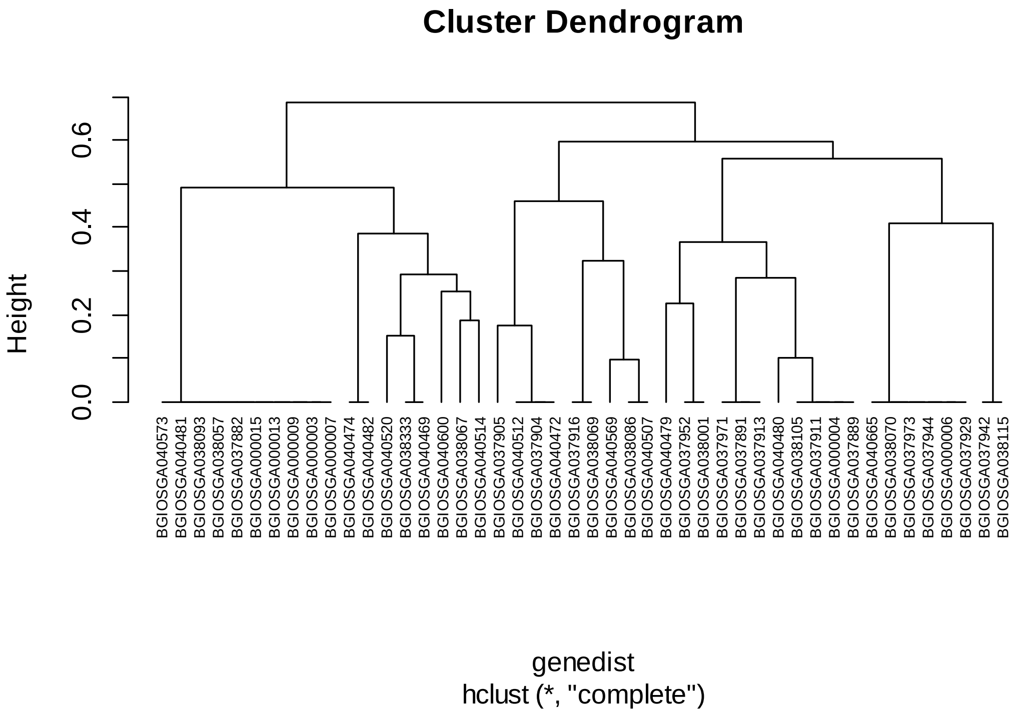

Figure 1. Visualization clusters.

```
> gene_cluster <- PASCCluster(gene_dist,nc=5,plot = TRUE)
> str(gene_cluster)
List of 2
 $ hclust:List of 7
  ..$ merge     : int [1:45, 1:2] -1 -5 -6 -7 -8 -23 -28 -37 -44 -2
 ...
  ..$ height    : num [1:45] 0 0 0 0 0 0 0 0 0 0 ...
  ..$ order     : int [1:46] 44 37 28 23 8 7 6 5 1 4 ...
  ..$ labels    : chr [1:46] "BGIOSGA000003" "BGIOSGA000004" "BGIOS
GA000006" "BGIOSGA000007" ...
  ..$ method    : chr "complete"
  ..$ call      : language hclust(d = genedist, method = "complete
")
```

```
  ..$ dist.method: NULL
  ..- attr(*, "class")= chr "hclust"
 $ cutree: Named int [1:46] 1 2 3 1 1 1 1 1 2 2 ...
  ..- attr(*, "names")= chr [1:46] "BGIOSGA000003" "BGIOSGA000004"
"BGIOSGA000006" "BGIOSGA000007" ...
> head(gene_cluster$cutree)
BGIOSGA000003 BGIOSGA000004 BGIOSGA000006 BGIOSGA000007 BGIOSGA000
009 BGIOSGA000013
  1             2             3             1             1             1
> table(gene_cluster$cutree)
 1  2  3  4  5
10 11  8  9  8
```

5 Session information

```
> sessionInfo()
R version 3.5.0 (2018-04-23)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1
Matrix products: default

locale:
[1] LC_COLLATE=Chinese (Simplified)_People's Republic of China.936
[2] LC_CTYPE=Chinese (Simplified)_People's Republic of China.936
[3] LC_MONETARY=Chinese (Simplified)_People's Republic of China.93
6
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_People's Republic of China.936

attached base packages:
[1] parallel  stats   graphics grDevices utils    datasets  methods
[8] base

other attached packages:
[1] PASCCA_0.1.0 plyr_1.8.4

loaded via a namespace (and not attached):
[1] compiler_3.5.0 tools_3.5.0     withr_2.1.2     yaml_2.1.19
[5] memoise_1.1.0   Rcpp_0.12.17    digest_0.6.15   devtools_1.13.5
```

References

[1] Long,Y. Lin,Q. Wu,X. et al.,2018 PASCCA: clustering poly(A) site data with repeated measurements based on shrinkage canonical correlation analysis.

[2] Hong, S., X. Chen, L. Jin and M. Xiong, 2013 Canonical correlation analysis for RNA-seq co-expression networks. Nucleic Acids Research 41: e95-e95.

[3] Yao, J., C. Chang, M. L. Salmi, Y. S. Hung, A. Loraine et al., 2008 Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. BMC Bioinformatics 9: 1471-2105.