PowerExplorer Manual

Xu Qiao, Laura Elo

2018-02-19

Contents

Abstract

This vignette demonstrates R package `PowerExplorer` as a power and sample size estimation tool for RNA-Seq and quantitative proteomics data.

`PowerExplorer` contains the following main features:

- Estimation of power based on the current data
- Prediction of power corresponding to the increased sample sizes
- Result visualizations

Introduction

Power and sample size estimation is one of the important principles in designing next-generation sequencing experiments to discover differential expressions. `PowerExplorer` is a power estimation and prediction tool currently applicable to RNA-Seq and quantitative proteomics experiments.

The calculation procedure starts with estimating the distribution parameters of each gene or protein. With the obtained prior distribution of each feature, a specified amount of simulations are executed to generate data (read counts for RNA-Seq and protein abundance for proteomics) repetitively for each entry based on null and alternative hypotheses. Furthermore, the corresponding statistical tests (t-test or Wald-test) are performed and the test statistics are collected. Eventually the statistics will be summarized to calculate the statistical power.

Input Data Preparation

For both RNA-Seq (gene expression levels) and quantitative proteomics (protein abundance levels) datasets, the data matrix should be arranged as genes/proteins in rows and samples in columns. Here we show a RNA dataset as an example:

```
library(PowerExplorer)
data("exampleRNASeqData")
head(exampleRNASeqData$dataMatrix[,1:6])
#>        Sample_A_1 Sample_A_2 Sample_A_3 Sample_A_4 Sample_A_5 Sample_B_1
#> Gene_1        469        324         38       1059         64        496
#> Gene_2         84        276        263        182        181        737
#> Gene_3        293        173        272        123        475        169
#> Gene_4        310        209        550        212        394       1064
#> Gene_5         82        141        216        202        494        293
#> Gene_6        583         98        137        179        214        884
```

A grouping vector indicating the sample groups to which all the samples belong should also be created, for example:

```
show(exampleProteomicsData$groupVec)
#>  [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C"
```

The sample groups corresponding to the data:

```
colnames(exampleProteomicsData$dataMatrix)
#>  [1] "Sample_A_1" "Sample_A_2" "Sample_A_3" "Sample_A_4" "Sample_A_5"
#>  [6] "Sample_B_1" "Sample_B_2" "Sample_B_3" "Sample_B_4" "Sample_B_5"
#> [11] "Sample_C_1" "Sample_C_2" "Sample_C_3" "Sample_C_4" "Sample_C_5"
```

Note that the grouping vector length should be equal to the column number of the data matrix.

Power Estimation

Here we use a randomly generated RNASeq dataset `exampleRNASeqData` as an example to estimate the current power of the dataset. The input dataset is named as `dataMatrix` and the grouping vector as `groupVec`.

To run the estimation, apart from the input, we still need to specify the following parameters:

- `isLogTransformed`: FALSE; the input data is not log-transformed.
- `dataType`: "RNA-Seq"; the datatype can be declared as "Proteomics" or "RNA-Seq".
- `minLFC`: 0.5; the threshold of Log2 Fold Change, proteins with lower LFC will be discarded.
- `alpha`: 0.05; the controlled false positive (Type I Error) rate.
- `ST`: 50; the simulation of each gene will be run 50 times (ST>50 is recommended).
- `seed`: 345; optional, a seed value for the random number generator to maintain the reproducibility.
- `showProcess`: FALSE; no detailed processes will be shown, set to TRUE if debug is needed.
- `saveSimulatedData`: FALSE; if TRUE, save the simulated data in ./savedData directory.

The results will be summaried in barplot, boxplot and summary table.

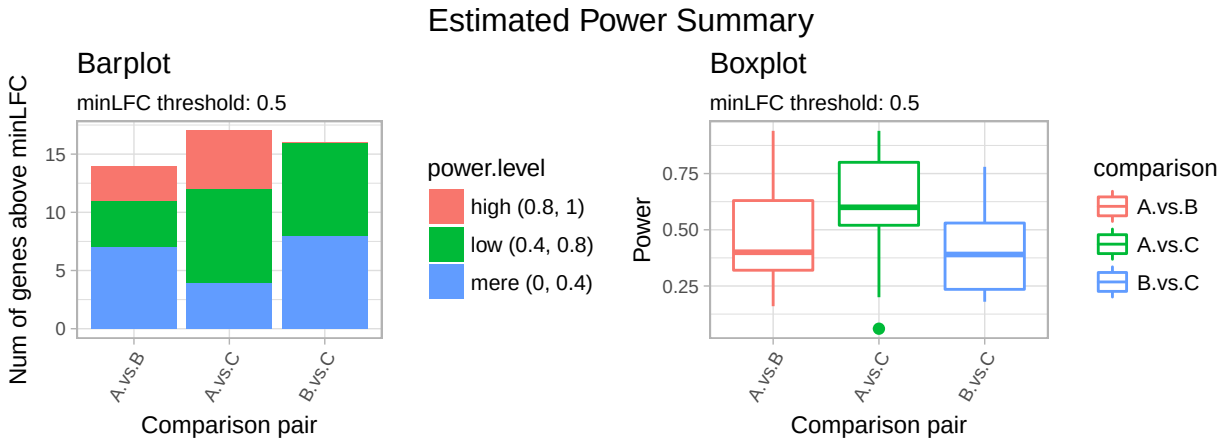
```
library(PowerExplorer)
data("exampleRNASeqData")
res <- estimatePower(inputObject = exampleRNASeqData$dataMatrix,
                     groupVec = exampleRNASeqData$groupVec,
                     isLogTransformed = FALSE,
                     dataType = "RNASeq",
                     minLFC = 0.5,
                     alpha = 0.05,
                     ST = 50,
                     seed = 345)
#> ##------ Mon Feb 19 14:43:34 2018 ------##
#> Num. of groups:   3
#> Num. of replicates:   5
#> Num. of simulations:  50
#> Min. Log Fold Change:     0.5
#> False Postive Rate:   0.05
#> Transformed:  FALSE
#>
#> ##------ Mon Feb 19 14:43:34 2018 ------##
#> 0 of 110 entries are filtered due to excessive zero counts
#> Estimating distribution parameters...
#>
#> Estimating NB parameters by DESeq2...
#>
#> [A.vs.B] 14 of 110 genes are over minLFC threshold 0.5:
#>
#> [A.vs.B] Log2 Fold Change Quantiles:
#>   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
#> 0.00 0.05 0.08 0.13 0.16 0.20 0.26 0.31 0.39 0.58 1.80
#>
#> [A.vs.C] 17 of 110 genes are over minLFC threshold 0.5:
#>
#> [A.vs.C] Log2 Fold Change Quantiles:
#>   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
#> 0.01 0.04 0.08 0.12 0.18 0.23 0.27 0.33 0.46 0.89 2.19
#>
#> [B.vs.C] 16 of 110 genes are over minLFC threshold 0.5:
```

```
#>
#> [B.vs.C] Log2 Fold Change Quantiles:
#>   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
#> 0.00 0.03 0.08 0.12 0.15 0.23 0.27 0.33 0.41 0.56 2.18
#>
#> Simulation in process, it may take a few minutes...
#>
#> Power Estimation between groups A.vs.B:
#>
#> OVERALL ESTIMATED POWER: 0.4986
#>
#>
#> Simulation in process, it may take a few minutes...
#>
#> Power Estimation between groups A.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.5894
#>
#>
#> Simulation in process, it may take a few minutes...
#>
#> Power Estimation between groups B.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.4275
```

Visualization

The estimated results can be summarized using `plotEstPwr`, the only input needed is the `estimatedPower`, which should be the estimated power object returned from `estimatePower`.

```
plotEstPwr(res)
```



Comp.	Gene Num.	Avg. Power	H (0.8, 1)	L (0.4, 0.8)	M (0, 0.4)
A.vs.B	14	0.50	3 (21%)	4 (29%)	7 (50%)
A.vs.C	17	0.59	5 (29%)	8 (47%)	4 (24%)
B.vs.C	16	0.43	0 (0%)	8 (50%)	8 (50%)

The graph contains 3 plots, the `barplot` vertically shows the number of genes/proteins above the minLFC threshold, columns indicates the comparison pairs, each column presents the proportions of three power levels in three colours as indicated in the legend `power.level`; The boxplot shows the overall power distribution of each comparsion; And the summary table summarize the power in a numerical way with the same information shown in the previous two plots.

Power Predictions

With the same dataset, to run a prediction, a different parameter is needed:

- `rangeSimNumRep`: the power of replicate number 5 to 20 will be predicted.

Similar to the estimation process, however, the simulations will be excuted with each sample size specified in `rangeSimNumRep`. (Note: the term sample size in this vignette refers to the replicate number of each group/case)

It is possible to append the prediction results within the same object by using the same result object as an input.

```
data("exampleRNASeqData")
res <- predictPower(inputObject = res,
                    groupVec = exampleRNASeqData$groupVec,
                    isLogTransformed = FALSE,
                    dataType = "RNASeq",
                    rangeSimNumRep = c(5, 10, 15, 20),
                    minLFC = 1,
                    alpha = 0.05,
                    ST = 50,
                    seed = 345)
#> ##------ Mon Feb 19 14:43:57 2018 ------##
#> Num. of groups:   3
#> Num. of replicates:   5, 10, 15, 20
#> Num. of simulations:  50
#> Min. Log Fold Change:     1
#> False Postive Rate:   0.05
#> Transformed:  FALSE
#>
#> ##------ Mon Feb 19 14:43:57 2018 ------##
#> 0 of 110 entries are filtered due to excessive zero counts
#> Estimating distribution parameters...
#>
#> Estimating NB parameters by DESeq2...
#>
#> [A.vs.B] 4 of 110 genes are over minLFC threshold 1:
#>
#> [A.vs.B] Log2 Fold Change Quantiles:
#>   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
#> 0.00 0.05 0.08 0.13 0.16 0.20 0.26 0.31 0.39 0.58 1.80
#>
#> [A.vs.C] 10 of 110 genes are over minLFC threshold 1:
#>
#> [A.vs.C] Log2 Fold Change Quantiles:
#>   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
#> 0.01 0.04 0.08 0.12 0.18 0.23 0.27 0.33 0.46 0.89 2.19
#>
#> [B.vs.C] 5 of 110 genes are over minLFC threshold 1:
#>
#> [B.vs.C] Log2 Fold Change Quantiles:
#>   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
#> 0.00 0.03 0.08 0.12 0.15 0.23 0.27 0.33 0.41 0.56 2.18
#>
```

```
#> ##--Simulation with 5 replicates per group--##
#>
#> [repNum:5] Simulation in process, it may take a few minutes...
#>
#> [repNum:5] Power Estimation between groups A.vs.B:
#>
#> OVERALL ESTIMATED POWER: 0.73
#>
#>
#> [repNum:5] Simulation in process, it may take a few minutes...
#>
#> [repNum:5] Power Estimation between groups A.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.65
#>
#>
#> [repNum:5] Simulation in process, it may take a few minutes...
#>
#> [repNum:5] Power Estimation between groups B.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.66
#>
#>
#> ##--Simulation with 10 replicates per group--##
#>
#> [repNum:10] Simulation in process, it may take a few minutes...
#>
#> [repNum:10] Power Estimation between groups A.vs.B:
#>
#> OVERALL ESTIMATED POWER: 0.96
#>
#>
#> [repNum:10] Simulation in process, it may take a few minutes...
#>
#> [repNum:10] Power Estimation between groups A.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.882
#>
#>
#> [repNum:10] Simulation in process, it may take a few minutes...
#>
#> [repNum:10] Power Estimation between groups B.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.872
#>
#>
#> ##--Simulation with 15 replicates per group--##
#>
#> [repNum:15] Simulation in process, it may take a few minutes...
#>
#> [repNum:15] Power Estimation between groups A.vs.B:
#>
#> OVERALL ESTIMATED POWER: 0.995
```

```
#>
#>
#> [repNum:15] Simulation in process, it may take a few minutes...
#>
#> [repNum:15] Power Estimation between groups A.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.966
#>
#>
#> [repNum:15] Simulation in process, it may take a few minutes...
#>
#> [repNum:15] Power Estimation between groups B.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.964
#>
#>
#> ##--Simulation with 20 replicates per group--##
#>
#> [repNum:20] Simulation in process, it may take a few minutes...
#>
#> [repNum:20] Power Estimation between groups A.vs.B:
#>
#> OVERALL ESTIMATED POWER: 1
#>
#>
#> [repNum:20] Simulation in process, it may take a few minutes...
#>
#> [repNum:20] Power Estimation between groups A.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.99
#>
#>
#> [repNum:20] Simulation in process, it may take a few minutes...
#>
#> [repNum:20] Power Estimation between groups B.vs.C:
#>
#> OVERALL ESTIMATED POWER: 0.984
```

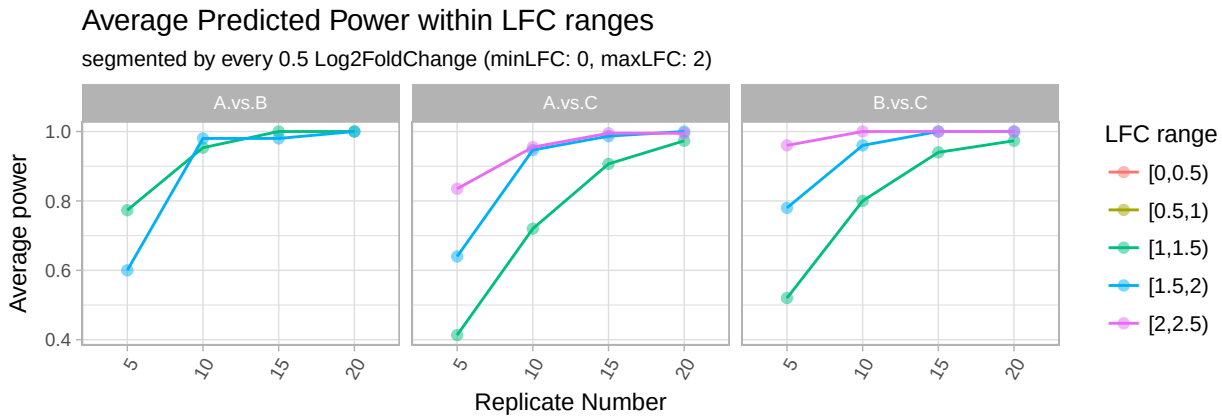
Visualization

The predicted results can be summaried using `plotPredPwr`. The input should be the predicted power object returned from `predictPower`, the summary can be optionally visualized by setting the following parameters:

- `PEObject`: A PEObject returned from PowerExplorer as input
- `minLFC` and `maxLFC`: to observe power in a specific range of LFC
- `LFCscale`: to determine the LFC scale of the observation

Lineplot (LFCscale = 0.5):

```
plotPredPwr(res, LFCscale = 0.5)
```



	repNum:5	repNum:10	repNum:15	repNum:20
A.vs.B	0.69	0.97	0.99	1
A.vs.C	0.63	0.87	0.96	0.99
B.vs.C	0.75	0.92	0.98	0.99

The output figure contains a lineplot and a summary table. For each comparison, the lineplot shows the power tendency across every Log2 Fold Change segment resulted from a complete LFC list divided by a specified `LFCscale`. Each dot on the lines represents the average power (y-axis) of the genes/proteins at a certain sample size (x-axis) within different LFC ranges. In addition, a summary table below displays the average power of each comparison across the sample sizes.

For instance, the line plot here shows the average power at four different sample sizes (5 to 30, with increment of 5) in LFCscale of 0.5. The LFC ranges from 0 to 5, and within each LFC segment, the graph shows the average power of the genes/proteins. Here, the higher LFC shows higher power, the average power of each LFC range increases with the larger sample sizes, as expected.