An Introduction to *QNB*

Lian Liu <liulian19860905@163.com>

Modified: 6 Nov, 2016. Compiled: November 10, 2016

1 Introduction

With high throughput sequencing techniques, such as MeRIP-Seq, transcriptome-wide RNA methylation profile is now available as count-based measurements, with which it is often of special interests to study the dynamics in post-transcriptional regulation via differential RNA methylation analysis. However, the sample size of RNA methylation experiment is usually very small due to its cost, making the inference very difficult. Meanwhile, there exist a large number of very lowly expressed genes whose methylation level cannot be accurately estimated due to limited sequencing depth. The *QNB* R-package is a statistical approach for differential RNA methylation analysis with count-based small-sample sequencing data. The method is based on 4 independent negative binomial distributions, and with their variances and means linked by local regressions, respectively. Please don't hesitate to contact <liulian19860905@163.com> if you have any problem. The inputs of the main function `qnbtest` are four reads count matrix for IP samples of two conditions and Input samples of two conditions.The *QNB* package fullfills the following one key function:

- differential RNA methylation analysis based on a model using quad-negative-binomial distribution

We will in the next see how the the main functions can be accomplished in a single command.

2 Input data

As input, the *QNB* package expects count data from two conditions (e. g., treated and untreated) as obtained, e. g., from MeRIP-Seq, in the form of two rectangular tables of integer values for each condition, one is Input control sample and another is IP sample. The table cell in the i-th row and the j-th column of the table shows the reads count of the methylation site i in sample j.

The count values must be raw counts of sequencing reads. So, please do not supply other quantities, such as (rounded) normalized counts – this will lead to nonsensical results.

In this vignette, we will work with DAA dataset. The original DAA raw data in SRA format was obtained directly GEO (GSE48037), which consists of 3 IP and 3 Input MeRIP-Seq replicates obtained under wild type condition and after DAA treatment, respectively (a total of 12 libraries). The short sequencing reads are firstly aligned to human genome assembly hg19 with Tophat2 [1], and then get RNA N6-methyl-adenosine (m6A) sites using *exomePeak* R/Bioconductor package [2] with UCSC gene annotation database [3]. In the peak calling step, to obtain a consensus RNA methylation site set between two experimental conditions (wild type and DAA treatment), we merged 6 IP samples and 6 Input samples, respectively. Then we used Bioconductor packages on R platform to obtain the reads count matrix. In the matrix, it includes the reads counts of m6A methylation sites from IP and Input samples (each with 3 replicates) under two conditions.

3 Differential Methylation Analysis

The main function of *QNB* R-package is to analyse differential RNA methylation. Meths are the reads count matrix of IP samples from two conditions, and unmeths are Input control samples from two condition. To get the differential RNA methylation, we estimate the dispersion for each site between treated (including IP and Input control sample) and untreated (including IP and Input control sample). In addtion, IP and Input control samples must be the same replicates, but it is may be the different replicates under two conditons.

To estimate the dispersion, there are four ways how the empirical dispersion can be computed:

- pooled - Use the samples from all conditions with replicates to estimate a single pooled empirical dispersion value, called `"pooled"`, and assign it to all samples.
- per-condition - For each condition with replicates, compute an empirical dispersion value by considering the data from samples for this condition.
- blind - Ignore the sample labels and compute an empirical dispersion value as if all samples were replicates of a single condition. This can be done even if there are no biological replicates.
- auto - Select mode according to the size of samples automatically. The default is auto.

Other parameters:

- plot.dispersion - The default is TRUE. If `plot.dispersion=FALSE`, it will not save the dispersion figure.
- output.dir - The saved file path. The default is NA. If `output.dir=NA`, the path is the current path.

Let us firstly load the package and get the toy data (came with the package) ready.

```
> library(QNB)
> f1 = system.file("extdata", "meth1.txt", package="QNB")
> f2 = system.file("extdata", "meth2.txt", package="QNB")
> f3 = system.file("extdata", "unmeth1.txt", package="QNB")
> f4 = system.file("extdata", "unmeth2.txt", package="QNB")
> meth1 = read.table(f1, header=TRUE)
> meth2 = read.table(f2, header=TRUE)
> unmeth1 = read.table(f3, header=TRUE)
> unmeth2 = read.table(f4, header=TRUE)
> head(meth1)

  S1 S2 S3
1  7  9  5
2  1  6  3
3  2  0  0
4  3  6  5
5  7  1  4
6  0  0  0

> head(unmeth1)

  S1 S2 S3
1  8  2  1
2  0  5  0
3  0  0  1
4  5  2  5
5  1  2  1
6  0  1  0
```

3.1 Standard comparison between two experimental conditions

When there are replicates under two conditions, we could select `mode="per-condition"` or `mode="pooled"` to estimate the dispersion. The default is `auto`. If `mode="per-condition"`, we estimate one dispersion for each condition, respectively. If `mode="pooled"`, we combind all replicates to generate one dataset from control samples and IP samples, then estimate one dispersion for two conditions.

```
> result = qnbtest(meth1, meth2, unmeth1, unmeth2, mode="per-condition")

[1] "Estimating dispersion for each RNA methylation site, this will take a while ..."

> head(result)
```

```
       pval4         fc         q0      padj
1 0.78490363 -0.1240921 12.2742967 0.9727062
2 0.36272519 -0.5152295  8.0316299 0.9727062
3 0.79032881 -0.4734347  3.1432272 0.9727062
4 0.06613825 -0.6550325 19.9338001 0.8216957
5 0.57085286  0.3269663  8.7317667 0.9727062
6 0.40214557       -Inf  0.5874658 0.9727062
```

The results will be saved in the specified output directory, including the dispersion figure(if `plot.dispersion=TRUE`) and the result table (including 4 columns (pvalue, log2(fold-change), q, FDR)). The following figure is the dispersion figure. The first row is the wild type dispersion of Input and IP samples, and the second row is the DAA dispersion of Input and IP samples.

- pvalue - Indicate the significance of the methylation site as an RNA differential methylation site
- log2.fc - The log2 odds ratio between IP sample and Input sample.
- q - The standardized feature abundance, which is proportional to the expression level of the RNA transcript.
- FDR - FDR of the methylation site, indicating the significance of the peak as an RNA differential methylation site after multiple hypothesis correction usting BH method.

3.2 Comparison without replicates

Proper replicates are essential to interpret a biological experiment. After all, any attempt to work without replicates will lead to conclusions of very limited reliability. But the *QNB* package can deal with them.

If you have replicates for one condition but not for the other, or without any replicates for two conditions, you can select `mode="blind"` to estimate the dispersion. We combind all samples under two conditions to generate replicates for two conditions. Then we estimate one dispersion for two conditions.

```
> f1 = system.file("extdata", "no_rep_meth1.txt", package="QNB")
> f2 = system.file("extdata", "no_rep_meth2.txt", package="QNB")
> f3 = system.file("extdata", "no_rep_unmeth1.txt", package="QNB")
> f4 = system.file("extdata", "no_rep_unmeth2.txt", package="QNB")
> no_rep_meth1 = read.table(f1, header=TRUE)
> no_rep_meth2 = read.table(f2, header=TRUE)
> no_rep_unmeth1 = read.table(f3, header=TRUE)
> no_rep_unmeth2 = read.table(f4, header=TRUE)
> head(no_rep_meth1)
```

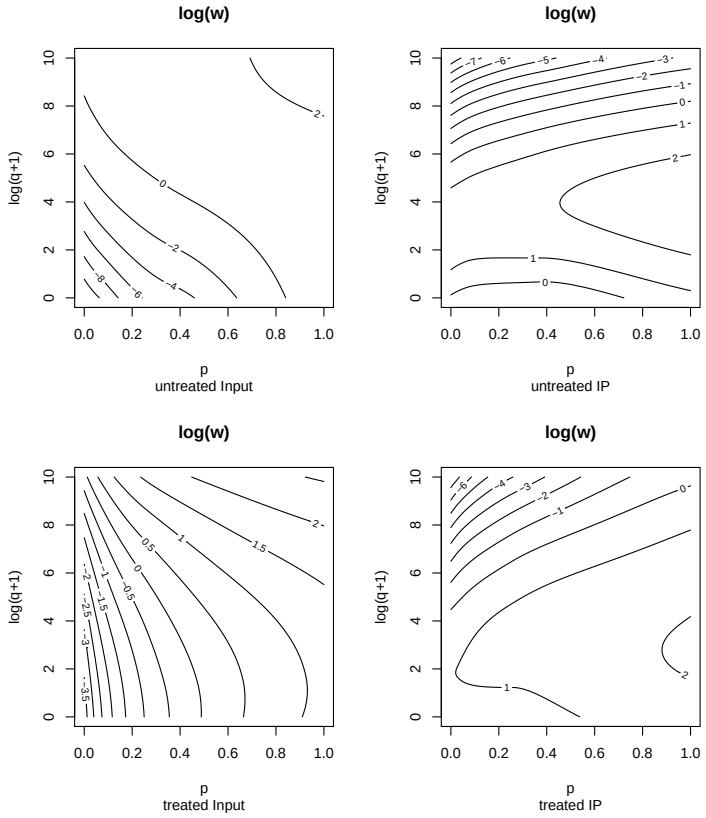

Figure 1: **The dispersion of reads count on common scale in DAA dataset.** The first row is the wild type dispersion of Input and IP samples, and the second row is the DAA dispersion of Input and IP samples. In each dataset, the variance of two conditions is very similar, but there are slight difference between them. Generally, the variance increases following the feature abundance $\log(\mathrm{q}+1)$ and absolute methylation level p.

```
  x
1 7
2 1
3 2
4 3
5 7
6 0
```

```
> head(no_rep_unmeth1)
```

```
  x
1 8
2 0
3 0
4 5
5 1
6 0
```

```
> result = qnbtest(no_rep_meth1,
+                  no_rep_meth2,
+                  no_rep_unmeth1,
+                  no_rep_unmeth2,
+                  mode="blind")
```

```
[1] "Estimating dispersion for each RNA methylation site, this will take a while ..."
```

3.3 Select mode automatically

If you could not decide which mode to estimate dispersion, `mode="auto"` will select suitable way to estimate dispersion according to the replicates.

```
> result = qnbtest(meth1, meth2, unmeth1, unmeth2)
```

```
[1] "Estimating dispersion for each RNA methylation site, this will take a while ..."
```

4 Session Information

```
> sessionInfo()
```

```
R version 3.2.2 (2015-08-14)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 8 x64 (build 9200)

locale:
[1] LC_COLLATE=Chinese (Simplified)_China.936
[2] LC_CTYPE=Chinese (Simplified)_China.936
[3] LC_MONETARY=Chinese (Simplified)_China.936
```

```
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_China.936

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

other attached packages:
[1] QNB_1.0.0      locfit_1.5-9.1

loaded via a namespace (and not attached):
[1] tools_3.2.2     grid_3.2.2      lattice_0.20-33
```

References

[1] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):295–311, 2013.

[2] J. Meng, X. Cui, M. K. Rao, Y. Chen, and Y. Huang. Exome-based analysis for rna epigenome sequencing data. *Bioinformatics*, 29(12):1565–7, 2013.

[3] D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent. The ucsc genome browser database: 2014 update. *Nucleic Acids Research*, 42(D1):D764–D770, 2014.