An Introduction to *QNB*

Lian Liu <liulian19860905@163.com>

Modified: 22 Jul, 2016. Compiled: October 19, 2016

1 Introduction

The *QNB* R-package has been developed for differential methylation analysis. Estimate variance and mean dependence in count data from MeRIP-seq and test for differential methylation based on a model using quadratic-negative-binomial distribution. Please don't hesitate to <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 methylation analysis based on a model using quadratic-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 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.

3 Differential Methylation Analysis

The main function of `QNB` R-package is to analyse differential 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 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. The default is per-conditon.
- 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 automaticly.

Other parameters:

- plot.dispersion - the default is TRUE. If `plot.dispersion = FALSE`, it will not save the dispersion figure.
- pvals.only - get pvalue only. If `pvals.only = TRUE`, get pvalue only in result matrix. If `pvals.only = FALSE`, get pvalue and FDR in result matrix.
- 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".

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

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

```
> head(result)
```

```
      pvalue     log2.fc          q
1 0.81951159 -0.09311204 12.2035362
2 0.39653912 -0.49033278  8.0256739
3 0.78836720 -0.44208962  3.1056776
4 0.07677246 -0.61073600 19.6557028
5 0.53619232  0.32786644  8.6458297
6 0.66628481        -Inf  0.5803058
```

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),expression,FDR(if `pvals.only=FALSE`))). 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 - log2(Fold-change). log 2 (fold enrichment) within the peak in the IP sample compared with the input sample.
- q - The expression of each methylation site.
- FDR - fdr of the methylation site, indicating the significance of the peak as an RNA differential methylation site after multiple hypothesis correction.

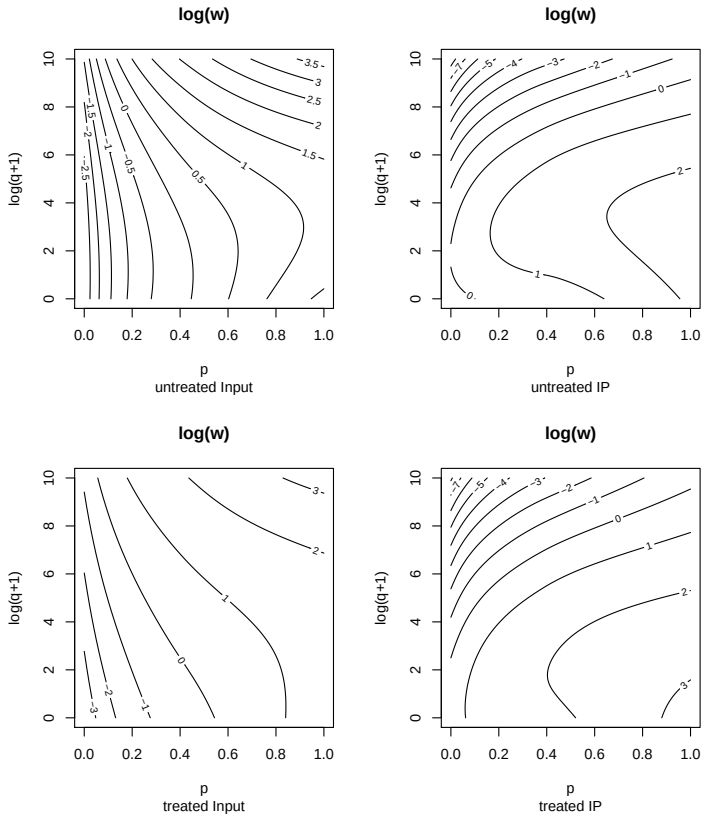

Figure 1: **The dispersion of reads count on common scale in DAA dataset.** A. The dispersion of Input samples. B. The dispersion of IP samples. In each dataset, the variance of two conditions is very similar, but there are slight difference between them. Generally, the variance increases following the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

3.2 Comparison without replicates

If you have replicates for one condition but not for the other,or there are no replicates for tow conditions, you can select "mode=blind"to estimate the dispersion.

```
> 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)
```

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

```
> head(no_rep_unmeth1)
```

```
  S
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_0.99.0     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
```