metaSeq: Meta-analysis of RNA-seq count data

Koki Tsuyuzaki[1], and Itoshi Nikaido[2].

September 26, 2013

[1]Department of Medical and Life Science, Tokyo University of Science.
[2]Bioinformatics Research Unit, Advanced Center for Computing and Communication, RIKEN.

k.t.the-answer@hotmail.co.jp

Contents

1 Introduction

This document provides the way to perform meta-analysis of RNA-seq data using *metaSeq* package. Meta-analysis is a attempt to integrate multiple data in different studies and retrieve much reliable and reproducible result. In our package, the probability of one-sided *NOISeq* [1] is applied in each study. This is because the numbers of reads are often different depending on its study and *NOISeq* is robust method against its difference (see the next section).

2 RSE: Read-Size Effect

In many cases, the number of reads are depend on study. For example, here we prepared multiple RNA-Seq count data designed as Breast Cancer cell lines vs Normal cells measured in 4 different studies (this data is also accessible by data(BreastCancer)).

ID in this vignette	Accession (SRA / ERA Accession)	Experimental Design
StudyA	SRP008746	Breast Cancer (n=3) vs Normal (n=2)
StudyB	SRP006726	Breast Cancer (n=1) vs Normal (n=1)
StudyC	SRP005601	Breast Cancer (n=7) vs Normal (n=1)
StudyD	ERP000992	Breast Cancer (n=2) vs Normal (n=1)

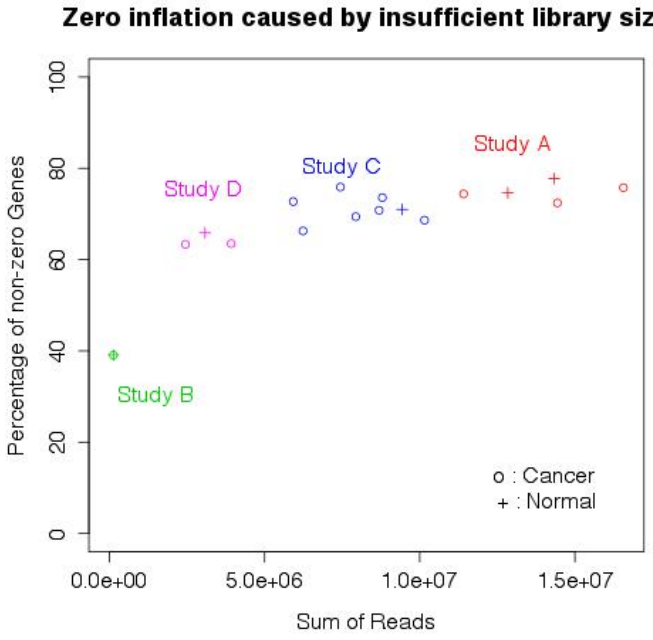

Figure 1: Difference of the number of reads

As shown in the figure 1, the number of reads in StudyA, B, C, and D are relatively different. Generally, statistical test is influenced by the number of reads; the more the number of reads is large, the more the statistical tests are tend to be significant (see the next section). Therefore, in meta-analysis of RNA-seq data, data may be suffered from this bias. Here we call this bias as RSE (Read Size Effect).

3 Robustness against RSE

In the point of view of robustness against RSE, we evaluated five widely used method in RNA-seq; *DESeq* [2], *edgeR* [3], *baySeq* [4], and *NOISeq* [1]. Here we used only StudyA data. All counts in the matrix are repeatedly down-sampled in accordance with distributions of binomial (the probability equals 0.5). 1 (original), 1/2, 1/4, 1/8, 1/16, and 1/32-fold data are prepared as low read size situation. In each read size, four methods are conducted (figure 2.A, this data is also accessible by data(StudyA) and data(pvals)), then we focussed on how top500 genes of original data in order of significance will change its members, influenced by low read size (figure 2.B).

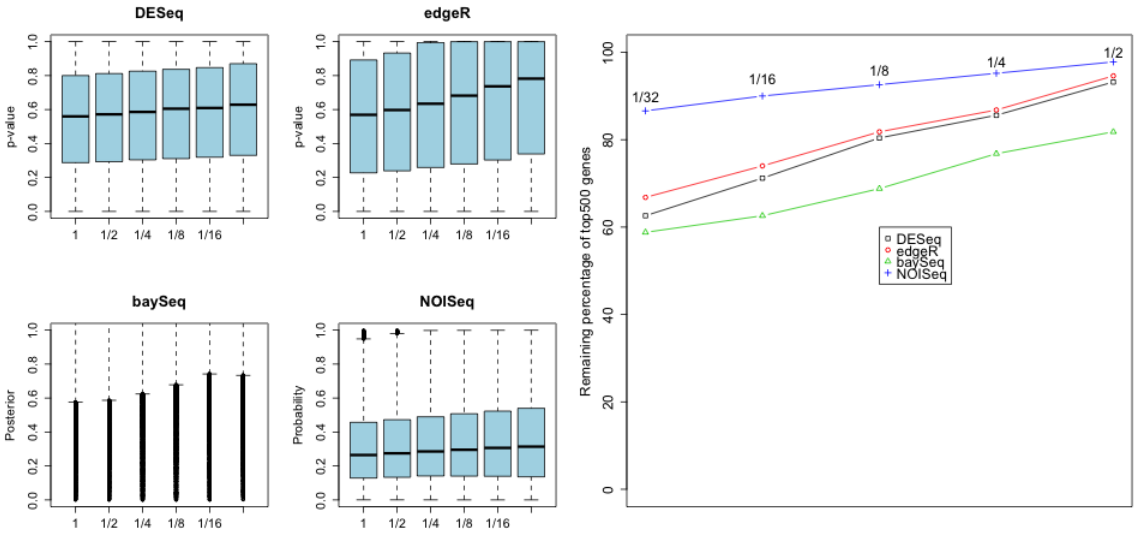

Figure 2: A(left): RSE in each RNA-Seq method, B(right): Top 500 genes in order of significance

Ideal method will returns same result regardless of read size, because same data was used. As shown in figure 2, *NOISeq* is not almost affected by the number of reads and robustlly detects same genes as DEGs. Therefore, we concluded that *NOISeq* is suitable method at least in the point of view of meta-analysis. Note that probability of *NOISeq* is not equal to p-value; it is the probability that a gene is differentially expressed [1]. Our package integrates its probability by Fisher's method [5] or Stouffer's method (inverse normal method) [6]. In regard to Stouffer's method, weighting by the number of replicates (sample size) is used.

4 Getting started

At first, install and load the *metaSeq* and *snow*.

```
> library("metaSeq")
> library("snow")
```

The RNA-seq expression data in breast cancer cell lines and normal cells is prepared. The data is measured from 4 different studies.The data is stored as a matrix (23368 rows $\times$ 18 columns).

```
> data(BreastCancer)
```

We need to prepare two vectors. First vector is for indicating the experimental condition (e.g., 1: Cancer, 2: Normal) and second one is for indicating the source of data (e.g., A: StudyA, B: StudyB, C: StudyC, D: StudyD).

```
> flag1 <- c(1,1,1,0,0, 1,0, 1,1,1,1,1,1,1,0, 1,1,0)
> flag2 <- c("A","A","A","A","A", "B","B", "C","C","C","C","C","C","C","C", "D","D","D")
```

Then, we use meta.readData to create R object for meta.oneside.noiseq.

```
> cds <- meta.readData(data = BreastCancer, factor = flag1, studies = flag2)
```

oneside.noiseq is performed in each studies and each probabilities are summalized as member of list object.

```
> ## This is very time consuming step.
> # cl <- makeCluster(4, "SOCK")
> # result <- meta.oneside.noiseq(cds, k = 0.5, norm = "tmm", replicates = "biological",
> #  factor = flag1, conditions = c(1, 0), studies = flag2, cl = cl)
> # stopCluster(cl)
>
> ## Please load pre-calculated result (Result.Meta)
> ## by data function instead of scripts above.
> data(Result.Meta)
> result <- Result.Meta
```

Fisher's method and Stouffer's method can be applied to the result of meta.oneside.noiseq.

```
> F <- Fisher.test(result)
> S <- Stouffer.test(result)
```

These outputs are summalized as list whose length is 3. First member is the probability which means a gene is upper-regulated genes, and Second member is lower-regulated genes. Weight in each study is also saved as its third member (weight is used only by Stouffer's method).

```
> head(F[[1]])
```

```
[1] 0.3799577 0.5249790 0.5308748        NA 0.1356067 0.2263183

> head(F[[2]])

[1] 0.8431580 0.6293839 0.4216739        NA 0.3793319 0.6147443

> F[[3]]

Study 1 Study 2 Study 3 Study 4
      5       2       8       3

> head(S[[1]])

[1] 0.3696589 0.2615383 0.2680294        NA 0.2933731 0.3009430

> head(S[[2]])

[1] 0.6303411 0.7384617 0.7319706        NA 0.7066269 0.6990570

> S[[3]]

Study 1 Study 2 Study 3 Study 4
      5       2       8       3
```

5 Meta-analysis by non-NOISeq method

For some reason, we may want to use non-NOISeq method like *DESeq*, *edgeR*, or even cuffdiff [7]. We prepared other.oneside.noiseq as optional function for such methods. Returned object can be directlly applied for Fisher.test and Stouffer.test.

```
> ## Assume this matrix as one-sided p-values
> ## generated by non-NOISeq method (e.g., cuffdiff)
> upper <- matrix(runif(300), ncol=3, nrow=100)
> lower <- 1 - upper
> weight <- c(3,6,8)
> ## other.oneside.pvalues function return a matrix
> ## which can input Fisher.test or Stouffer.test
> result <- other.oneside.pvalues(upper, lower, weight)
> ## Fisher's method (without weighting)
> F <- Fisher.test(result)
> str(F)

List of 3
 $ Upper : num [1:100] 0.374 0.786 0.624 0.656 0.116 ...
 $ Lower : num [1:100] 0.705 0.22 0.343 0.493 0.809 ...
 $ Weight: Named num [1:3] 3 6 8
  ..- attr(*, "names")= chr [1:3] "Exp 1" "Exp 2" "Exp 3"
```

```
> F

$Upper
  [1] 0.373921626 0.785959734 0.623507593 0.656084398 0.115580409
  [6] 0.744318017 0.236865169 0.228607367 0.840797249 0.944115324
 [11] 0.323932658 0.955202246 0.377893610 0.029408862 0.005070636
 [16] 0.277467321 0.306845245 0.134910718 0.287945491 0.428492176
 [21] 0.371410156 0.710363338 0.764393053 0.704558231 0.375650436
 [26] 0.122396784 0.161651656 0.581837723 0.465271110 0.645725905
 [31] 0.928593351 0.697507710 0.318883998 0.723862302 0.951534534
 [36] 0.586982042 0.634121592 0.049983175 0.010346514 0.094424305
 [41] 0.819881482 0.222256405 0.126463567 0.673684010 0.388407690
 [46] 0.917571030 0.338442599 0.889467072 0.648099965 0.336817892
 [51] 0.908571517 0.545775557 0.745681297 0.380088404 0.255532233
 [56] 0.814240244 0.414719071 0.602375871 0.547237577 0.642729740
 [61] 0.300595022 0.665999636 0.895110374 0.419753096 0.125839057
 [66] 0.648308774 0.747700548 0.860729715 0.217027114 0.941961184
 [71] 0.827676357 0.353651395 0.142337516 0.040982980 0.551910809
 [76] 0.344598637 0.538936376 0.341745634 0.294217293 0.426750183
 [81] 0.958850371 0.206639438 0.583483476 0.653142970 0.245281534
 [86] 0.170035054 0.902229235 0.503418512 0.900990640 0.903935531
 [91] 0.547094552 0.789488701 0.605071988 0.539465010 0.025224182
 [96] 0.993738815 0.494504319 0.607413698 0.565280520 0.307778284

$Lower
  [1] 0.70532414 0.22022428 0.34322392 0.49332519 0.80867395
  [6] 0.30301521 0.75937687 0.76684120 0.32956739 0.04089980
 [11] 0.84774873 0.11309581 0.55782439 0.69995781 0.99645946
 [16] 0.88707048 0.78026358 0.70801032 0.68188946 0.64895708
 [21] 0.80064841 0.06460170 0.29585195 0.18151305 0.81588230
 [26] 0.45729611 0.85612055 0.62825193 0.79733066 0.32838244
 [31] 0.23688576 0.36151776 0.16231282 0.53971010 0.17068871
 [36] 0.37541240 0.16545228 0.87192237 0.94262011 0.97530367
 [41] 0.15658471 0.57586728 0.91393651 0.04668811 0.70783244
 [46] 0.18062139 0.56211818 0.13959904 0.62929422 0.83966187
 [51] 0.25343714 0.40981854 0.24835624 0.79974473 0.57498069
 [56] 0.15341543 0.41446115 0.34939793 0.58850238 0.18020250
 [61] 0.10831022 0.56351604 0.02341733 0.68294893 0.65290793
 [66] 0.21108691 0.51928040 0.28460642 0.44356392 0.20135386
 [71] 0.18008467 0.79323104 0.22762222 0.80058529 0.43329411
 [76] 0.87362700 0.68383164 0.86408131 0.89412948 0.82974583
 [81] 0.04095062 0.89922765 0.69494025 0.09793815 0.90405667
 [86] 0.35345565 0.19604553 0.25495580 0.27300109 0.12544388
 [91] 0.23037629 0.24634253 0.31649713 0.31808437 0.61710078
 [96] 0.02236594 0.44140321 0.31291503 0.48282970 0.18290291
```

```
$Weight
Exp 1 Exp 2 Exp 3
    3     6     8

> ## Stouffer's method (with weighting by sample-size)
> S <- Stouffer.test(result)
> str(S)

List of 3
 $ Upper : num [1:100] 0.478 0.824 0.528 0.512 0.121 ...
 $ Lower : num [1:100] 0.522 0.176 0.472 0.488 0.879 ...
 $ Weight: Named num [1:3] 3 6 8
  ..- attr(*, "names")= chr [1:3] "Exp 1" "Exp 2" "Exp 3"

> S

$Upper
  [1] 0.478317321 0.824040645 0.528239570 0.512205146 0.120759731
  [6] 0.757444501 0.130228417 0.322220276 0.673462046 0.955627538
 [11] 0.315661092 0.900196037 0.234029678 0.172469320 0.007707666
 [16] 0.160769965 0.166133245 0.183056739 0.151353418 0.532104682
 [21] 0.245674816 0.810021885 0.650810699 0.885756716 0.233114089
 [26] 0.295051496 0.167482818 0.481677629 0.294622199 0.528287782
 [31] 0.819232676 0.782029738 0.760843648 0.537088709 0.898769813
 [36] 0.733765860 0.832469263 0.074980112 0.051201771 0.067824272
 [41] 0.678969627 0.133314402 0.115092991 0.935355956 0.340867522
 [46] 0.900951336 0.235024021 0.908024119 0.479939878 0.323799579
 [51] 0.867283469 0.376814986 0.581332083 0.385629536 0.309904080
 [56] 0.890777774 0.581238551 0.422500108 0.354776864 0.887751900
 [61] 0.282004894 0.621872433 0.906509568 0.269512877 0.332516095
 [66] 0.534667351 0.625127627 0.853352898 0.639613803 0.844227274
 [71] 0.738126702 0.393279005 0.636839309 0.206118492 0.735439860
 [76] 0.254157622 0.382833076 0.216550597 0.239575466 0.271725129
 [81] 0.909961512 0.097669186 0.422868544 0.904463543 0.149734018
 [86] 0.674437846 0.826784784 0.419095682 0.776166727 0.938493090
 [91] 0.756869104 0.856738156 0.761766794 0.457111938 0.100926993
 [96] 0.983056650 0.673380007 0.446470045 0.434966023 0.838434280

$Lower
  [1] 0.52168268 0.17595935 0.47176043 0.48779485 0.87924027
  [6] 0.24255550 0.86977158 0.67777972 0.32653795 0.04437246
 [11] 0.68433891 0.09980396 0.76597032 0.82753068 0.99229233
 [16] 0.83923004 0.83386675 0.81694326 0.84864658 0.46789532
 [21] 0.75432518 0.18997811 0.34918930 0.11424328 0.76688591
 [26] 0.70494850 0.83251718 0.51832237 0.70537780 0.47171222
```

```
 [31] 0.18076732 0.21797026 0.23915635 0.46291129 0.10123019
 [36] 0.26623414 0.16753074 0.92501989 0.94879823 0.93217573
 [41] 0.32103037 0.86668560 0.88490701 0.06464404 0.65913248
 [46] 0.09904866 0.76497598 0.09197588 0.52006012 0.67620042
 [51] 0.13271653 0.62318501 0.41866792 0.61437046 0.69009592
 [56] 0.10922223 0.41876145 0.57749989 0.64522314 0.11224810
 [61] 0.71799511 0.37812757 0.09349043 0.73048712 0.66748390
 [66] 0.46533265 0.37487237 0.14664710 0.36038620 0.15577273
 [71] 0.26187330 0.60672099 0.36316069 0.79388151 0.26456014
 [76] 0.74584238 0.61716692 0.78344940 0.76042453 0.72827487
 [81] 0.09003849 0.90233081 0.57713146 0.09553646 0.85026598
 [86] 0.32556215 0.17321522 0.58090432 0.22383327 0.06150691
 [91] 0.24313090 0.14326184 0.23823321 0.54288806 0.89907301
 [96] 0.01694335 0.32661999 0.55352996 0.56503398 0.16156572

$Weight
Exp 1 Exp 2 Exp 3
    3     6     8
```

6 Setup

This vignette was built on:

```
> sessionInfo()

R Under development (unstable) (2013-09-25 r63985)
Platform: x86_64-apple-darwin10.8.0 (64-bit)

locale:
[1] ja_JP.UTF-8/ja_JP.UTF-8/ja_JP.UTF-8/C/ja_JP.UTF-8/ja_JP.UTF-8

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

other attached packages:
[1] metaSeq_0.99.0     snow_0.3-12        NOISeq_2.0.0
[4] Biobase_2.21.7     BiocGenerics_0.7.5

loaded via a namespace (and not attached):
[1] tools_3.1.0
```

References

[1] Tarazona, S. and Garcia-Alcalde, F. and Dopazo, J. and Ferrer, A. and Conesa, A. Genome Research *Differential expression in RNA-seq: A matter of depth*, 21(12): 2213-2223, 2011.

[2] Simon Anders and Wolfgang Huber Genome Biology *Differential expression analysis for sequence count data.*, 11: R106, 2010.

[3] Robinson, M. D. and McCarthy, D. J. and Smyth, G. K. Bioinformatics *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.*, 26: 139-140, 2010

[4] Thomas J. Hardcastle R package version 1.14.1. *baySeq: Empirical Bayesian analysis of patterns of differential expression in count data.*, 2012.

[5] Fisher, R. A. Statistical Methods for Research Workers, 4th edition, Oliver and Boyd, London, 1932.

[6] Stouffer, S. A. and Suchman, E. A. and DeVinney, L. C. and Star, S. A. and Williams, R. M. Jr. The American Soldier, Vol. 1 - Adjustment during Army Life. Princeton, Princeton University Press, 1949

[7] Trapnell, C. and Williams, B. A. and Pertea, G. and Mortazavi, A. and Kwan, G. and Baren, M. J. and Salzberg, S. L. and Wold, B. J. and Pachter, L. Nature biotechnology *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiaiton*, 28: 511-515, 2010.