rnaseqcomp: A Benchmark for RNA-seq Quantification Pipelines

Mingxiang Teng mxteng@jimmy.harvard.edu
Rafael A. Irizarry rafa@jimmy.harvard.edu
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
Department of Biostatistics, Harvard T.H. Chan School Public Health, Boston, MA, USA
2015-05-25

Contents

1 Introduction

RNA sequencing (RNA-seq) has been utilized as the standard technology for measuring the expression abundance of genes, transcripts, exons or splicing junctions. Numerous quantification methods were proposed to quantify such abundances with combination of some RNA-seq read aligner. Unfortunately, it is currently difficult to evaluate the performance of the best method, due in part to the high costs of running assessment experiments as well as the computational requirements of running these algorithms. We have developed a series of statistical summaries and data visualization techniques to evaluate the performance of transcript quantification, particularly specificity and sensitivity.

The `rnaseqcomp` R-package performs comparisons and provides direct plots on these statistical summaries. It requires the inputs as an quantification table (or two, depending on which statistical comparisons is performed) by comapred pipelines on a paire of RNA-seq samples . With nessesary meta information on these pipelines (*e.g.* names), a two step analysis will generate the desired evaluations.

1. Data filtering and data preparation. In this step, options are provided for any filtering and calibration operations on the raw data. A S4 class `rnaseqcomp` object will be generated form next step.
2. Statistical summary evaluation and visualization. Functions are provided for specificity and sensitivity evaluations.

2 Getting Started

Load the package in R

```
library(rnaseqcomp)
```

3 Preparing Data

As the benchmark evaluation is performed on a pair of RNA-seq replicates, a quantification table should contain $2n$ columns (n corresponding to the numebr of pipeline compared), with each column representing a sample and each row representing a feature (*i.e.* genes, transcripts, exons, splicing junctions, etc.). The function matrixFilter takes this table as one of the inputs, with extra options such as meta information of pipelines, features for evaluation and features for calibration, and returns a S4 rnaseqcomp object that contains everything for downstream evaluation.

There are several reasons why we need extra options in this step:

1. Meta information of pipelines basically is a factor to check the sanity of table columns, and to provide unique names of pipelines for downstream analysis.
2. Since there might be dramatic quantification difference between different features, *e.g.* between protein coding genes and lincRNA genes, evaluations based on a subset of features can provide stronger robustness than using all involved features. Thus, an option is offered for selecting subset of features.
3. Due to different pipelines reports different units of quantification, such as FPKM (fragments per kilobases per million), RPKM (reads per kilobases per million), TPM (transcripts per million) etc. Calibrations across different units are necessary. Options are provided in the way that on which features the calibrations are based and to what pipeline the signals are mapped.

We show here an example of selecting house-keeping genes(Eisenberg and Levanon 2013) for calibration and filtering protein coding genes for evaluation. In this vignette, we will use enbedded dataset encodeCells as examples to illustrate this package. This dataset contains two cell-line quantifications, GM12878 and K562, each with two technical replicates by ENCODE project(ENCODE). In total, quantifications from 9 pipelines are included. Here, 9 pipelines are made up with 6 quantification methods (RESM(Li and Dewey 2011), Cufflinks(Trapnell et al. 2010), FluxCapacitor(Montgomery et al. 2010), Sailfish(Patro, Mount, and Kingsford 2014), eXpress(Roberts and Pachter 2013) and Naive) in conjunction to 2 mapping algorithms (STAR and TopHat2) and different tuning parameters.

```
# load the dataset in this package
data(encodeCells)
ls()
## [1] "arrayFC"  "genemeta" "gm12878"  "k562"     "repInfo"
```

Here, gm12878 and k562 are both quantification tables; repInfo is the meta information of pipelines; genemeta is the meta information for features: gene type and if house-keeping gene; arrayFC is fold change information between GM12878 and K562 cell lines from microarray platform(Ernst et al. 2011).

In order to fit into funtion 'matrixFilter', necessary transformation to logical vectors are needed for extra options.

```
txFIdx <- genemeta$type == "protein_coding"
hkIdx <- genemeta$housekeeping
unitFIdx <- grepl("Cufflinks",repInfo)
```

Generic function show is provided for bird-eye view of S4 rnaseqcomp object.

```
dat1 <- matrixFilter(gm12878,repInfo,txFIdx,hkIdx,unitFIdx)
class(dat1)
## [1] "rnaseqcomp"
## attr(,"package")
## [1] "rnaseqcomp"
show(dat1)
## rnaseqcomp: Benchmark for RNA-seq quantification pipelines
##
## Reps:
##  RSEM_Bowtie_TPM RSEM_Bowtie_TPM RSEM_Bowtie_pmeTPM RSEM_Bowtie_pmeTPM RSEM_STAR_TPM RSEM_STAR_TPM Cuff
##
## Calibration subset log2Median:
```

```
##  2.950468 3.0268 2.971773 3.048759 2.69488 2.790772 3.840821 3.815581 3.951611 3.933167 4.636201 4.5983
##
## Detrened signal scaler:
##  3.886994
##
## Quantification data has  20387  rows and  18  columns:
##                 RSEM_Bowtie_TPM RSEM_Bowtie_TPM RSEM_Bowtie_pmeTPM
## ENSG00000237613               0               0                  0
## ENSG00000268020               0               0                  0
## ENSG00000186092               0               0                  0
## ENSG00000237683               0               0                  0
## .                           ...             ...                ...
## ENSG00000198886          193.42          204.08             192.95
## ENSG00000198786          111.38          140.69             111.11
## ENSG00000198695           78.13          137.17              77.95
## ENSG00000198727          188.82          212.64             188.36
##                 RSEM_Bowtie_pmeTPM   . eXpress_Bowtie_RPKM
## ENSG00000237613                  0 ...                   0
## ENSG00000268020                  0 ...                   0
## ENSG00000186092                  0 ...                   0
## ENSG00000237683                  0 ...                   0
## .                              ... ...                 ...
## ENSG00000198886             203.46 ...               15844
## ENSG00000198786             140.26 ...               22341
## ENSG00000198695             136.76 ...                5133
## ENSG00000198727                212 ...               31182
##                 eXpress_Bowtie_RPKM Naive_TopHat_RPKM Naive_TopHat_RPKM
## ENSG00000237613                   0                 0                 0
## ENSG00000268020                   0                 0                 0
## ENSG00000186092                   0                 0                 0
## ENSG00000237683                   0                 0                 0
## .                               ...               ...               ...
## ENSG00000198886               13667  209.800876254042   224.87603164446
## ENSG00000198786               23797  431.365376013138  566.405423605872
## ENSG00000198695                7443  74.4834403686513  116.134432983972
## ENSG00000198727               29223  477.737072649147  557.854403640662
```

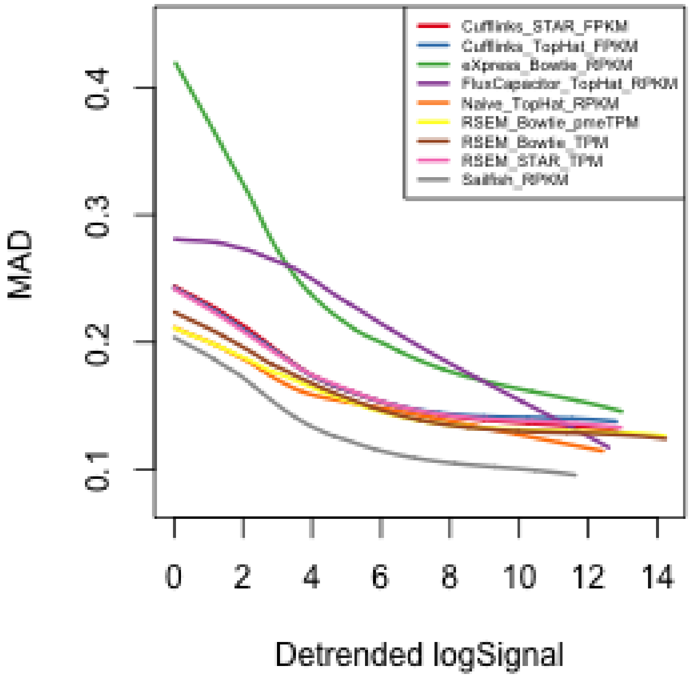
4 Visualizing Benchmarks

Three type of QC metrics can be evaluated by this package. More details please refer to our paper(Teng).

4.1 Specificity on expressed features.

This metric is evaluated by the quantification deviations between RNA-seq technical replicates. Basically lower deviations indicate higher specificity. Both one number statistics and deviation stratified by express signals are provided.

```
plotMAD(dat1)
## One number statistics: MAD
##       Cufflinks_STAR_FPKM     Cufflinks_TopHat_FPKM
##                     0.174                     0.174
##       eXpress_Bowtie_RPKM FluxCapacitor_TopHat_RPKM
##                    0.235                     0.236
##        Naive_TopHat_RPKM        RSEM_Bowtie_pmeTPM
##                    0.160                     0.164
##          RSEM_Bowtie_TPM             RSEM_STAR_TPM
##                    0.169                     0.174
##            Sailfish_RPKM
##                    0.137
```



Detrended signals shown in the plot are actually the signals with the same scales as Cufflinks pipelines, as we selected `unitFIdx` as signals from Cufflinks. In this case, FPKM by Cufflinks.

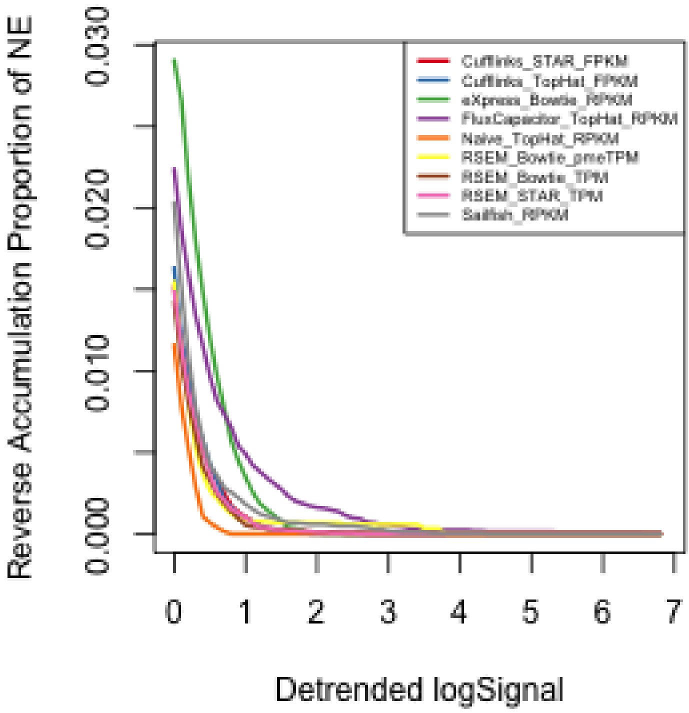
4.2 Specificity on non-expressed features

The proportions of non-expressed features is another important statistics. However, two types of non-expressed features should be analyzed seperately:

4.2.1 Features expressed in one technical replciate but not the other.

The reverse accumulated propotions of such either-or expressed features are plotted stratefied by the detrended signals as described previously. Basically, a lower curve indicates higher specificity on these features.

```
nonexpress <- plotNE(dat1)
```



4.2.2 Features expressed in neither replciates, and others.

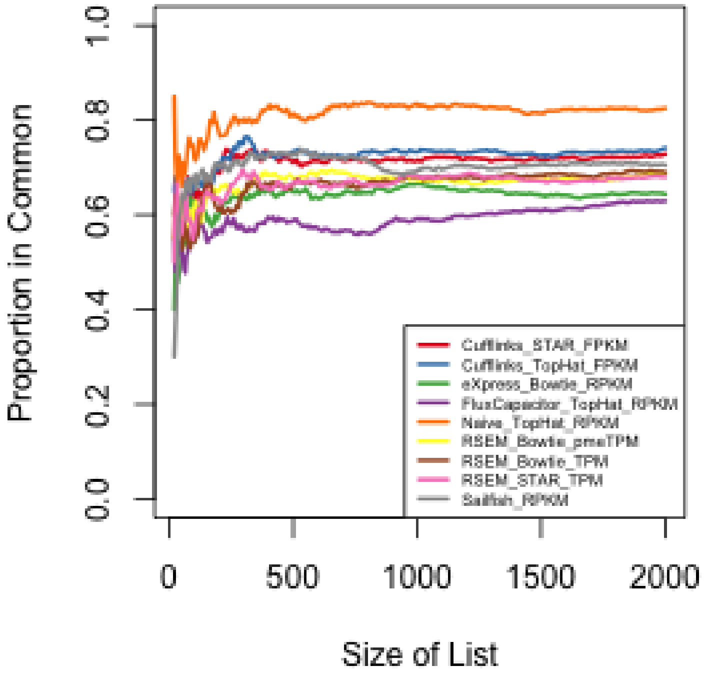
Here, proportions of both expressed, both non-expressed and either-or expressed features are list as a table.

```
nonexpress
##                             pEE   pNE   pNN
## Cufflinks_STAR_FPKM       0.502 0.015 0.483
## Cufflinks_TopHat_FPKM     0.500 0.016 0.484
## eXpress_Bowtie_RPKM       0.496 0.029 0.475
## FluxCapacitor_TopHat_RPKM 0.492 0.022 0.486
## Naive_TopHat_RPKM         0.530 0.011 0.459
## RSEM_Bowtie_pmeTPM        0.520 0.016 0.464
## RSEM_Bowtie_TPM           0.510 0.015 0.475
## RSEM_STAR_TPM             0.501 0.015 0.484
## Sailfish_RPKM             0.547 0.020 0.433
```

4.3 Specificity in differential analysis

We calculate the fold change of features between two different cell-lines and compare the fold change concordance between two technical replicates. A stratefy that summarizes the overlapped proportions among top differential expressed features is used, as we described before(Irizarry et al. 2005).

```
dat2 <- matrixFilter(k562,repInfo,txFIdx,hkIdx,unitFIdx)
plotCAT(dat1,dat2)
##       Cufflinks_STAR_FPKM     Cufflinks_TopHat_FPKM
##                 0.7181102                 0.7316770
##       eXpress_Bowtie_RPKM FluxCapacitor_TopHat_RPKM
##                 0.6449438                 0.5903614
##         Naive_TopHat_RPKM        RSEM_Bowtie_pmeTPM
##                 0.8231738                 0.6752351
##           RSEM_Bowtie_TPM             RSEM_STAR_TPM
##                 0.6771930                 0.6722973
##             Sailfish_RPKM
##                 0.7058824
```



Basically higher curve indicates better specificity. `plotCAT` also provides a one number summary of such specificity, which is the median of all overlap proportions plotted. In addition, constant is allowed for a more robust estimation of fold change.

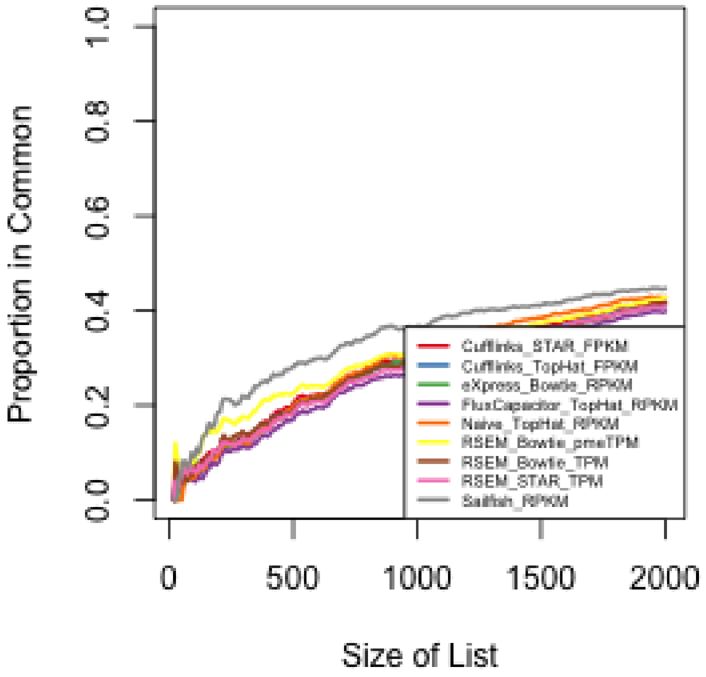
```
plotCAT(dat1,dat2,constant=1)
```

4.4 Sensitivity in differential analysis

There are other platforms provide the same quantifications such as microarray. We thus compare differential analysis of RNA-seq and other technology to evaluate sensitivity of pipelines. We have documented an object `arrayFC` which has been estimated from microarray technology(Ernst et al. 2011). We don't document the steps how we calculated microarray fold change here, since it is beyond the scope of this vignette.

```
genes <- genemeta[genemeta$type == "protein_coding", 1]
microarray <- arrayFC[match(genes,names(arrayFC))]
plotCAT(dat1,dat2,microarray=microarray)
```

```
##       Cufflinks_STAR_FPKM     Cufflinks_TopHat_FPKM
##                 0.3064039                 0.2910891
##       eXpress_Bowtie_RPKM FluxCapacitor_TopHat_RPKM
##                 0.2980392                 0.2811881
##         Naive_TopHat_RPKM        RSEM_Bowtie_pmeTPM
##                 0.3223881                 0.3180000
##           RSEM_Bowtie_TPM              RSEM_STAR_TPM
##                 0.2945274                 0.2941176
##             Sailfish_RPKM
##                 0.3661836
```



By comparing with microarray differential analysis, CAT plots will be plotted as higher curve indicates better sensitivity.

References

Eisenberg, E., and E. Y. Levanon. 2013. "Human Housekeeping Genes, Revisited." *Trends Genet* 29 (10): 569–74.

ENCODE. "Https://www.encodeproject.org/."

Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, et al. 2011. "Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types." *Nature* 473 (7345): 43–49.

Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, et al. 2005. "Multiple-Laboratory Comparison of Microarray Platforms." *Nat Methods* 2 (5): 345–50.

Li, B., and C. N. Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or Without a Reference Genome." *BMC Bioinformatics* 12: 323.

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis. 2010. "Transcriptome Genetics Using Second Generation Sequencing in a Caucasian Population." *Nature* 464 (7289): 773–7.

Patro, R., S. M. Mount, and C. Kingsford. 2014. "Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms." *Nat Biotechnol* 32 (5): 462–4.

Roberts, A., and L. Pachter. 2013. "Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments." *Nat Methods* 10 (1): 71–73.

Teng, Rafael A., Mingxiang; Irizarry. "Rnaseqcomp: A Benchmark for RNA-Seq Quantification Pipelines Based on a Minimal Dataset." In preparation.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. 2010. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation." *Nat Biotechnol* 28 (5): 511–5.