Visualization and assessment of RNA-Seq data normalization - the Normalization Visualization Tool

Thomas Eder[1,2], Florian Grebien[1], Thomas Rattei[2]

[1] Ludwig Boltzmann Institute for Cancer Research,
Währinger Straße 13A, 1090 Vienna, Austria.
[2] CUBE Division of Computational Systems Biology,
Department of Microbiology and Ecosystem Science,
University of Vienna, Althanstraße 14, 1090 Vienna, Austria.

May 9, 2016

Abstract

Measuring differential gene expression is a common task in the analysis of RNA-Seq data. To identify differentially expressed genes between two samples, it is crucial to normalize the data sets. While multiple normalization methods are available, all of them are based on certain assumptions that may or may not be suitable for the type of data they are applied on. Researchers therefore need to select an adequate normalization strategy for each RNA-Seq experiment. This selection includes exploration of different normalization methods as well as their comparison. Methods that agree with each other most likely represent realistic assumptions under the particular experimental conditions. We developed the *NVT* package, which provides a fast and simple way to analyze and evaluate multiple normalization methods via visualization and representation of correlation values, based on a user-defined set of uniformly expressed genes. This vignette explains the use of the package and demonstrates a typical work flow.

NVT version: 1.0

Contents

1 Installation

The latest NVT package can be downloaded from: https://github.com/Edert/NVT/releases. The installation of the package can be done directly in R.

```
install.packages(file.path("/home/user/Downloads/","NVT_1.0.tar.gz"),
repos=NULL, type="source")
```

2 Standard work flow

Initially the library has to be loaded, which provides access to the data examples 2.1.1 and the *NVT* functions.

```
library("NVT")
```

2.1 Load data

For the demonstration of the *NVT* functions we need gene expression data, so we load the example data provided with the package. We use the two human expression data sets *GSM1275862* and *GSM1275863* from the airway[1] package.

2.1.1 Load expression data

We simply load the provided data sets *myexp1* from *GSM1275862*, *myexp2* from *GSM1275863* and the length data per gene as *mylen*.

```
data(mylen)
data(myexp1)
data(myexp2)

#show just the first four elements of the loaded data
head(mylen,4)
```

```
##                  Length
## ENSG00000000003   3688
## ENSG00000000005   1881
## ENSG00000000419   4032
## ENSG00000000457  10784
```

```
head(myexp1,4)
```

```
##                  GSM1275862
## ENSG00000000003        679
## ENSG00000000005          0
## ENSG00000000419        467
## ENSG00000000457        260
```

In order to evaluate the expression between the two RNA-seq expression samples, we have to define a list of human housekeeping genes *GAPDH, ALB, ACTB, HPRT1, ADA, POLR2L, VCP, GPI, HBS1L* and *SDHA*. For the evaluation with the XY-Scatter or MA-plots a linear model is calculated in order to achieve reasonable results it is recommended to use at least 10 [2] data points.

```
mylist1<-c("ENSG00000111640","ENSG00000163631","ENSG00000075624",
           "ENSG00000165704","ENSG00000196839","ENSG00000177700",
           "ENSG00000165280","ENSG00000105220","ENSG00000112339",
           "ENSG00000073578")
```

But there is also a list of human housekeeping genes supplied by the *NVT* tool, which can be retrieved the following way:

```
data(mylist_hs)

mylist_hs
```

```
##  [1] "ENSG00000111640" "ENSG00000163631" "ENSG00000075624" "ENSG00000165704"
##  [5] "ENSG00000196839" "ENSG00000177700" "ENSG00000165280" "ENSG00000105220"
##  [9] "ENSG00000112339" "ENSG00000073578"
```

2.1.2 Load gene length data

Instead of using the length data generated directly in R or using a simple flat file, it is also possible to load the gene or exon length data directly from an annotation file in gtf or gff format. This function utilizes the GenomicRanges[3] and rtracklayer[4] packages.

- gff-version: version of the provided gff file $[gff1, gff2, gff3, gtf]$
- gff-feature: feature to use $[default : exon]$
- gff-name: name to use $[default : gene\_id]$

```
#this line gets the path to the gff file provided in the NVT package
#(annotation from Chlamydia trachomatis, ACCESSION: NC_000117 )
mygffpath <- system.file("extdata", "Ctr-D-UW3CX.gff", package = "NVT")

#this function loads the gff file from the gffpath
mylen1 <- NVTloadgff(mygffpath,"gff3","gene","locus_tag")

head(mylen1)
```

```
##       Length
## CT001    273
## CT002    303
## CT003   1476
## CT004   1467
## CT005   1092
## CT006    570
```

2.2 Generate NVTdata objects

In the first step you need to generate an *NVTdata* object with the *NVTinit* function. Here you have to provide the list of housekeeping genes, the two gene expression samples and the normalization method. Optionally you can also add the gene or exon length data.

The normalization methods are:

- N = No normalization
- TC = Total count normalization
- Med = Median normalization
- TMM = Trimmed Mean of M-values normalization
- UQ = Upper Quartile normalization
- UQ2 = Upper Quartile normalization (from NOISeq[5])
- Q = Quantile normalization implemented in limma[6]
- RPKM = Reads Per Kilobase per Million mapped reads normalization
- RPM = Reads Per Million mapped reads normalization
- TPM = Transcripts Per Million normalization
- DEQ = Relative log expression method included in the DESeq[7] package
- G = Use the provided genes to normalize

For the most methods no length information is required.

```
mynvt <- NVTinit(mylist1,myexp1,myexp2,"TMM")
```

But for RPKM and TPM normalization the length data has to be provided.

```
mynvt <- NVTinit(mylist1,myexp1,myexp2,"RPKM",mylen)
```

2.2.1 Normalize the NVTdata

The now initialized *NVTdata* object can be normalized in the next step.

```
mynorm <- NVTnormalize(mynvt)
```

```
## [1] "RPKM normalization!"
```

The text output can be suppressed via verbose=FALSE.

```
mynvt <- NVTnormalize(mynvt,verbose=F)
```

If required the now normalized data can be retrieved easily.

```
mynvalues <- show(mynorm)

head(mynvalues)
```

```
##                 GSM1275862 GSM1275863
## ENSG00000000003  8.9209656  5.8859979
## ENSG00000000005  0.0000000  0.0000000
## ENSG00000000419  5.6121512  6.1889890
## ENSG00000000457  1.1682249  0.9480595
## ENSG00000000460  0.2395964  0.2196300
```

```
## ENSG00000000938  0.0000000  0.0000000
```

2.3 Visualize expression data

One of the key features of *NVT* is the plotting of the XY-Scatter-Plots and MA-Plots. This can be done with the functions: *NVTplot*, *NVTadvancedplot*, *NVTmaplot* and *NVTadvancedmaplot*.

2.3.1 Simple plot functions

The normalized *NVTdata* object can be visualized with the plot function. With the second parameter you can modify the size-ratio of the text and the data points. All expression values are shown as grey dots and the linear model as red line, it is calculated with the data from the housekeeping genes only. The dashed grey line highlights the diagonal, so the perfect correlation between the two data sets.

```
NVTplot(mynorm,0.4)
```

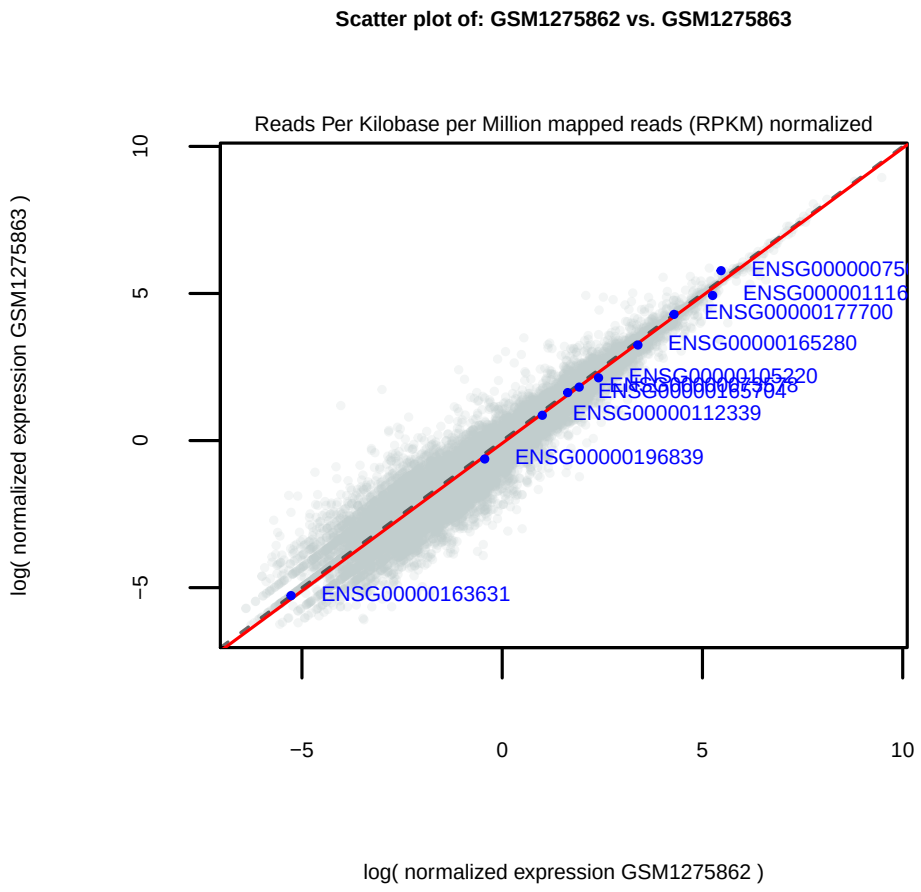

Figure 1: **XY-Scatter-Plot.** The resulting plot from the simple NVTplot function

The *NVTplotma* function works similar but produces an MA-Plot instead of a Scatter-Plot.

```
NVTmaplot(mynorm,0.4)
```

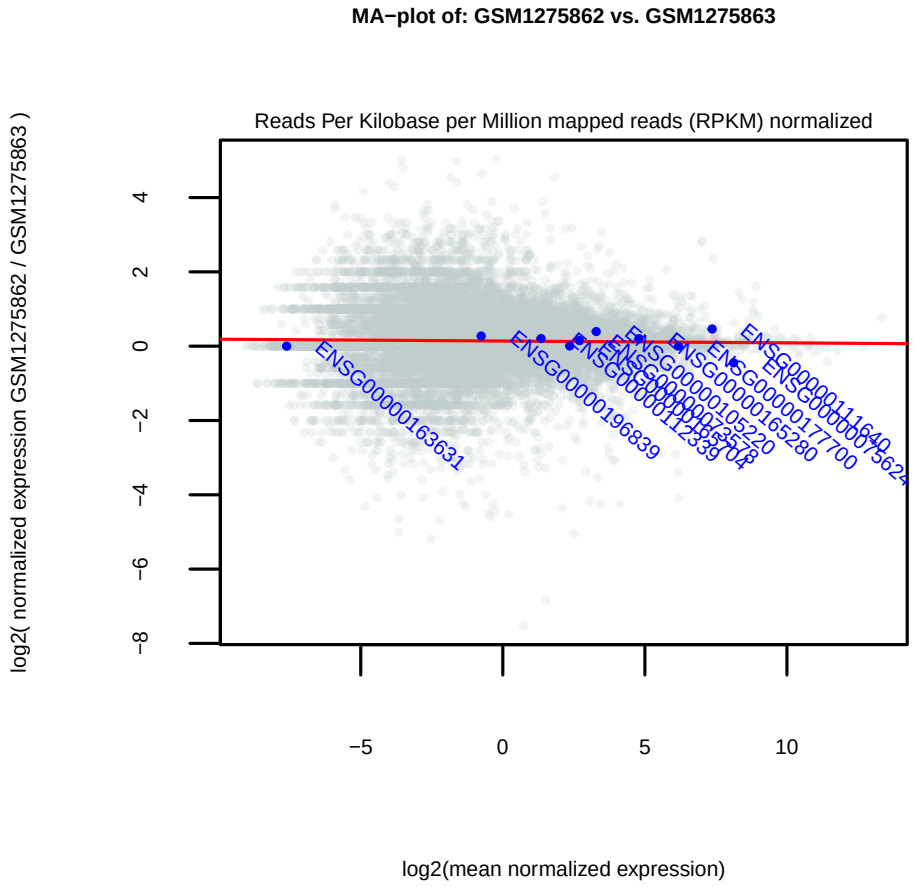

Figure 2: **MA-Plot.** The resulting plot from the simple NVTmaplot function

2.3.2 Advanced plot functions

The *NVTadvancedplot* plots via ggplot2[8] and the size parameters control data points, text and labels separately. Here we use the default values of 1 for the data points and the text and increase the labels of the housekeeping genes to 2. Again the grey dots represent the expression data and their density is visualized by the rug in dark red for x- and y-axis. The linear model of the housekeeping gene data is shown in red and the diagonal is highlighted as grey dashed line. As we want to compare different methods we now normalize with the TMM method.

```
mynvt <- NVTinit(mylist1,myexp1,myexp2,"TMM",mylen)
mynorm <- NVTnormalize(mynvt)

## [1] "Trimmed Mean of M-values normalization!"

NVTadvancedplot(mynorm,1,1,1)
```

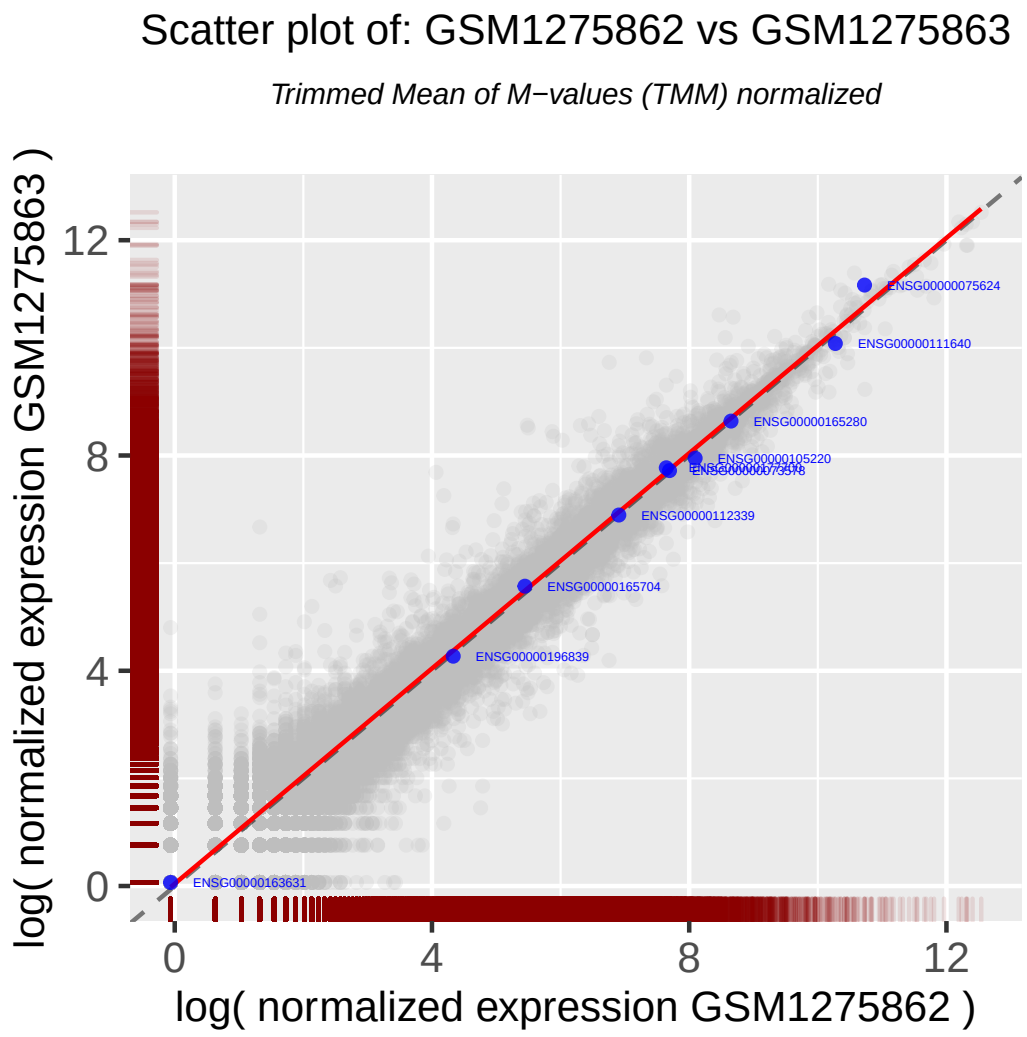

Figure 3: **Advanced XY-Scatter-Plot.** The NVTadvancedplot function results in a plot produced by ggplot2 with a rug on both axis, indicating the density of the data-points

As alternative there is also the possibility to create an MA-Plot with the *NVTadvancedmaplot* function.

```
NVTadvancedmaplot(mynorm,1,1,1)
```

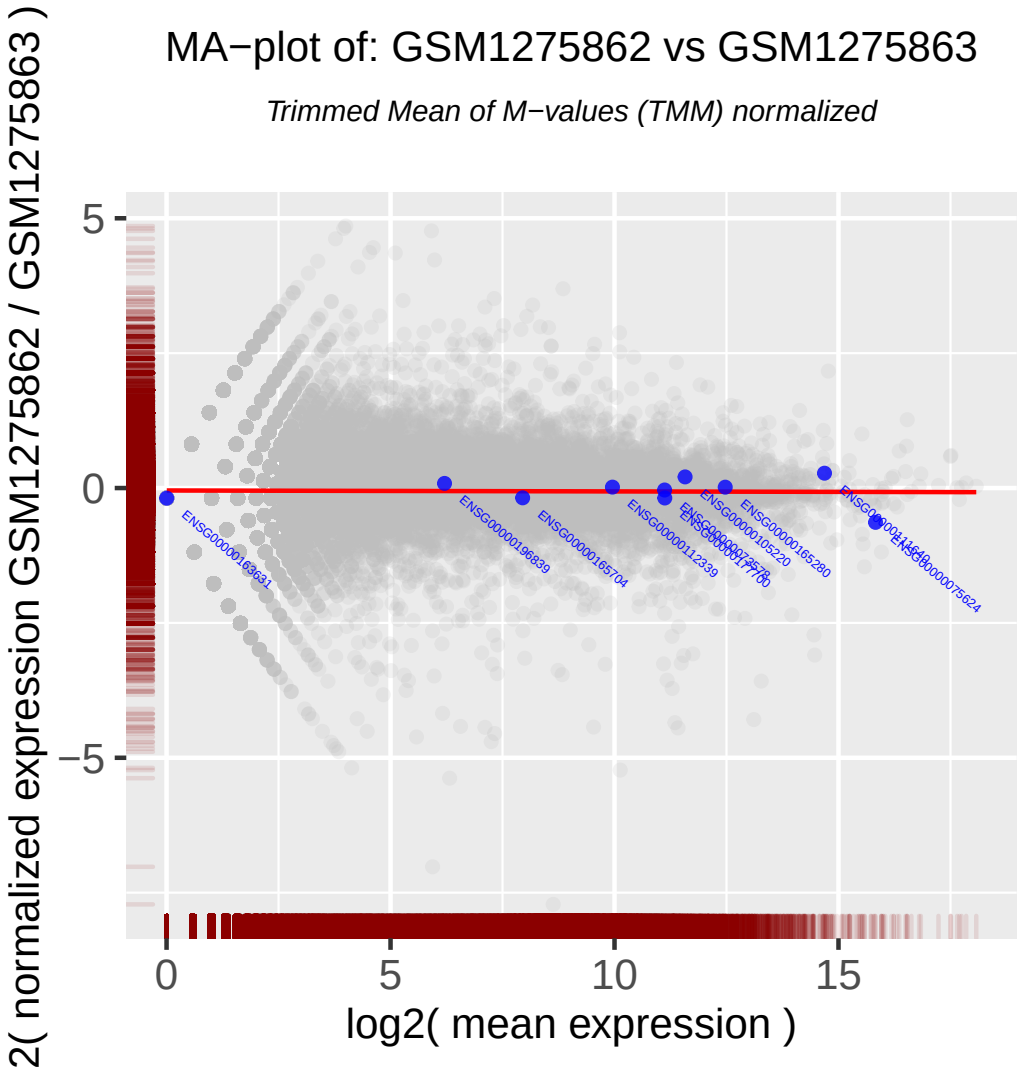

Figure 4: **Advanced MA-Plot.** The NVTadvancedmaplot function results in a plot produced by ggplot2 with a rug on both axis, indicating the density of the data-points

2.3.3 Linear model

If required, the linear model plotted as red line can be retrieved with the *NVTlm* function.

```
mylm <- NVTlm(mynorm)

summary(mylm)
```

```
##
## Call:
## lm(formula = sample2 ~ sample1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23427 -0.09105 -0.03633  0.08405  0.39746
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04327    0.15208   0.284    0.783
## sample1      1.00002    0.02006  49.846 2.91e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1891 on 8 degrees of freedom
## Multiple R-squared:  0.9968,Adjusted R-squared:  0.9964
## F-statistic:  2485 on 1 and 8 DF,  p-value: 2.905e-11
```

2.4 Correlation values

In addition to the graphical representation of the gene expression data, the correlation coefficients of the housekeeping genes of the two samples can be calculated with the functions *NVTpearson*, *NVTrmsd* and *NVTmae*.

2.4.1 Pearson correlation

The Pearson correlation coefficient of the normalized expression of the housekeeping genes is calculated with the following command, using an already normalized *NVTdata* object.

```
NVTpearson(mynorm)
```

```
##      pearson      p-value
## 9.645647e-01 6.608992e-06
```

2.4.2 RMSD and MEA correlation

The root mean square error (RMSE) also called the root mean square deviation (RMSD) is calculated with the *NVTrmsd* function.

```
NVTrmsd(mynorm)
```

```
##    RMSD
## 8137.41
```

And the mean absolute error (MAE) can be calculated and retrieved with the *NVTmae* function.

```
NVTmae(mynorm)
```

```
##      MAE
## 3114.019
```

2.5 Test all methods

To test more normalization methods on the provided data sets in one single step, the correlation coefficients of all implemented normalization methods can be calculated with the *NVTtestall* function. It requires a normalized *NVTdata* object and the correlation coefficient you are interested in. The method can be defined with:

- p = Pearson correlation
- rmsd = root mean square deviation
- mae = mean absolute error

```
NVTtestall(mynorm,"p")
```

```
## [1] "No normalization!"
## [1] "Total count normalization!"
## [1] "Median normalization!"
## [1] "Trimmed Mean of M-values normalization!"
## [1] "Upper Quartile normalization!"
## [1] "Upper Quartile normalization (from NOISeq)!"
## [1] "Quantile normalization!"
## [1] "RPKM normalization!"
## [1] "RPM normalization!"
## [1] "DESeq normalization!"
## [1] "Using DESeq"
## [1] "Input counts are not integer, converting them!"
## [1] "TPM normalization!"
## [1] "Normalization by given gene-set!"
##  [1] "ENSG00000111640" "ENSG00000163631" "ENSG00000075624" "ENSG00000165704"
##  [5] "ENSG00000196839" "ENSG00000177700" "ENSG00000165280" "ENSG00000105220"
##  [9] "ENSG00000112339" "ENSG00000073578"
##        pearson      p-value
## Q    0.9724102 2.452009e-06
## TC   0.9645647 6.608992e-06
## N    0.9645647 6.608992e-06
## G    0.9645647 6.608992e-06
## Med  0.9645647 6.608992e-06
## TMM  0.9645647 6.608992e-06
## UQ   0.9645647 6.608992e-06
```

```
## UQ2  0.9645647 6.608992e-06
## RPM  0.9645647 6.608992e-06
## DEQ  0.9645647 6.608992e-06
## TPM  0.9515575 2.272016e-05
## RPKM 0.9515575 2.272016e-05
```

It is also possible to oppress text output via verbose=FALSE (default: verbose=TRUE)

```
NVTtestall(mynorm,"p",verbose=F)
```

```
##        pearson      p-value
## Q    0.9724102 2.452009e-06
## TC   0.9645647 6.608992e-06
## N    0.9645647 6.608992e-06
## G    0.9645647 6.608992e-06
## Med  0.9645647 6.608992e-06
## TMM  0.9645647 6.608992e-06
## UQ   0.9645647 6.608992e-06
## UQ2  0.9645647 6.608992e-06
## RPM  0.9645647 6.608992e-06
## DEQ  0.9645647 6.608992e-06
## TPM  0.9515575 2.272016e-05
## RPKM 0.9515575 2.272016e-05
```

2.6 Example use case

To show a typical use case, the RNA-Seq data of GEO accession: GSE60052 was used. The following data sets of healthy control lung cells and small cell lung carcinoma cells were used: healthy *GSM1464282*, *GSM1464283 GSM1464284* and small cell lung cancer *GSM1464289*, *GSM1464290*, *GSM1464291*. We expect that the overall expression patterns will be quite diverse if we compare healthy with cancer cell expression and it is strongly recommended to use genes which are stably expressed.

First we load the *NVT*library, the gene length data, the housekeeping gene list and the expression data.

```
library("NVT")
data(myusecaseexp)
data(myusecaselen)
data(mylist_hs)

head(myusecaseexp)
```

```
##                 GSM1464282 GSM1464283 GSM1464284 GSM1464289 GSM1464290 GSM1464291
## ENSG00000000003        186         54         28          8         22        534
## ENSG00000000005        100          4         52          0          0          0
## ENSG00000000419         66         96         22          2         22        143
## ENSG00000000457         93        234         50        112          0        234
## ENSG00000000460         68        130         38         98          0        426
## ENSG00000000938        203        794        347         24          0         28
```

Now we create an NVTobject with the expression of the first healthy and the first small cell lung cancer

samples, add the gene length data and set the first normalization method to test to RPKM. We also apply the normalization and have a look at the resulting expression values.

```
mynvt <- NVTinit(mylist_hs,myusecaseexp[,1,drop=FALSE],myusecaseexp[,4,drop=FALSE],
"RPKM",myusecaselen)
mynorm <- NVTnormalize(mynvt)
```

```
## [1] "RPKM normalization!"
```

```
mynvalues <- show(mynorm)
head(mynvalues)
```

```
##                 GSM1464282 GSM1464289
## ENSG00000000003  3.4284335  0.1474595
## ENSG00000000005  5.1919942  0.0000000
## ENSG00000000419  4.5708476  0.1385105
## ENSG00000000457  1.1294454  1.3601924
## ENSG00000000460  0.9526052  1.3728722
## ENSG00000000938  4.8845696  0.5774861
```

Then we create the first plot.

```
NVTplot(mynorm,0.4)
```

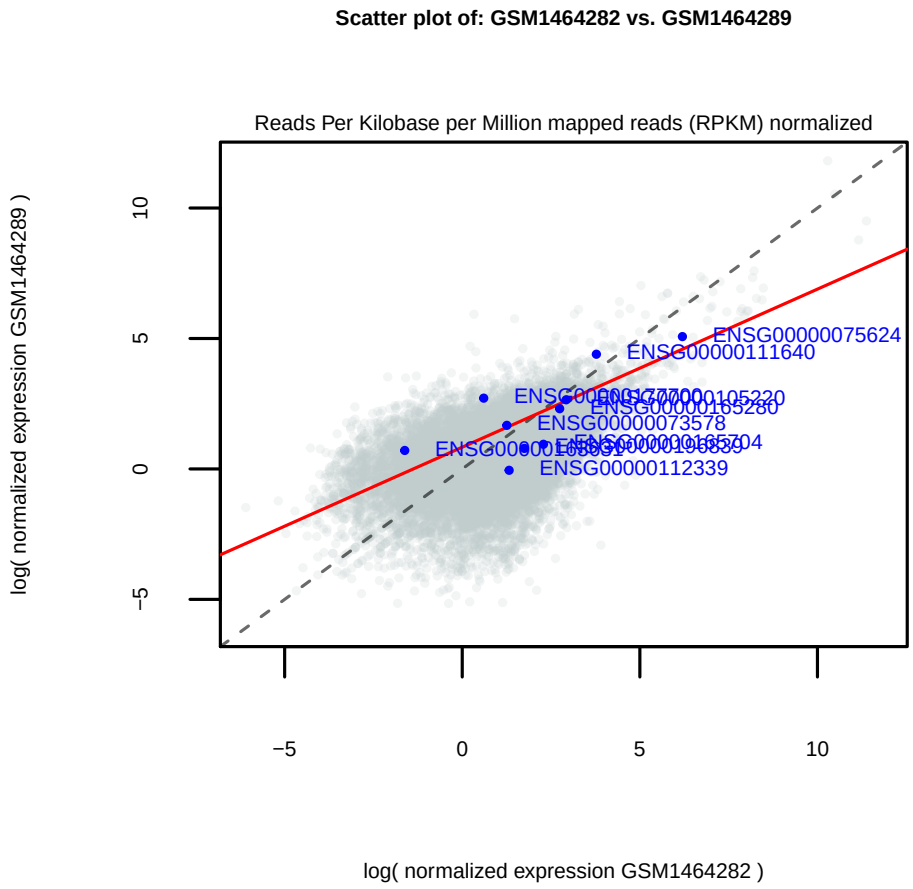

Figure 5: **XY-Scatter-Plot.** The resulting plot from the simple NVTplot function shows a broad cloud of expression values and the linear model is not very close to the diagonal

In addition we also create the MA-plot.

```
NVTadvancedmaplot(mynorm,1,1,1)
```

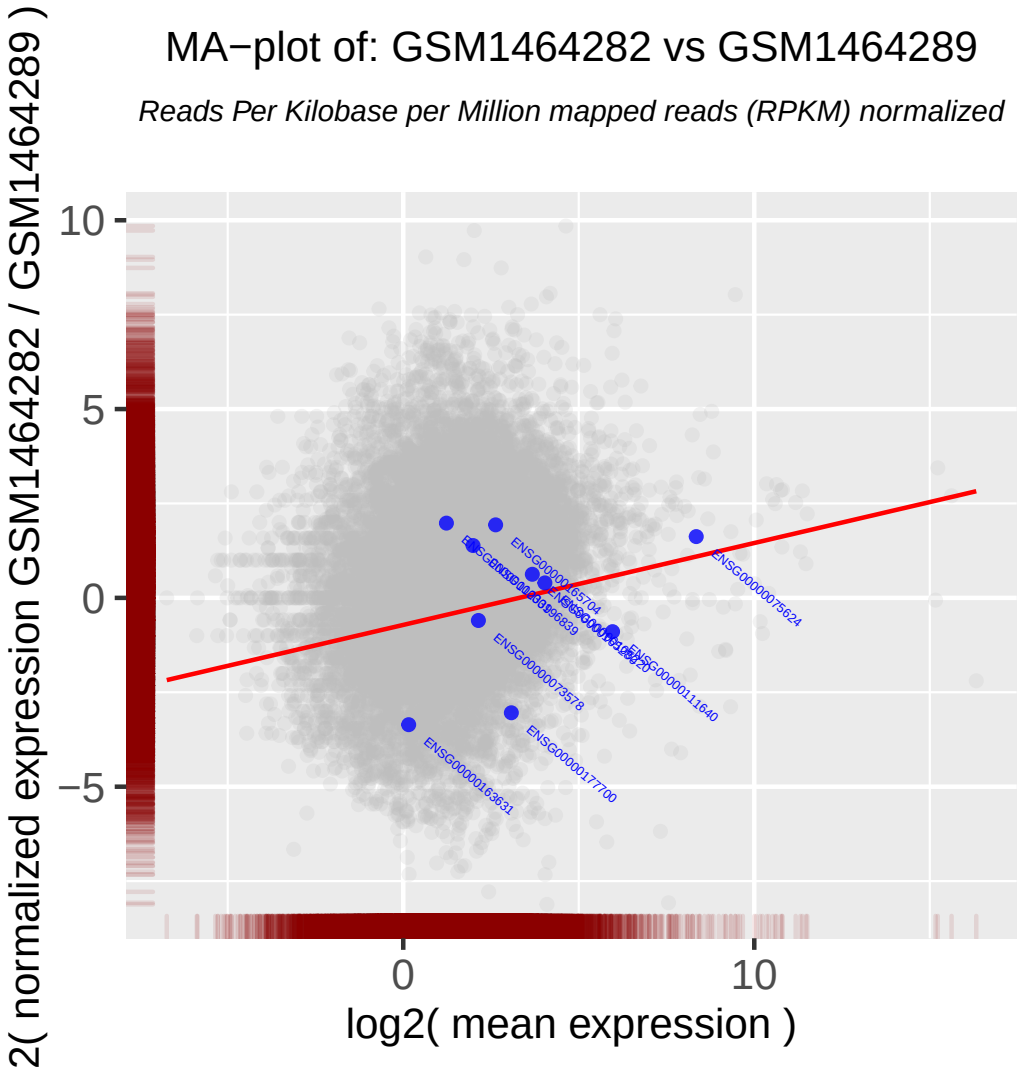

Figure 6: **Advanced MA-Plot.** The NVTadvancedmaplot function shows also that the expression of the genes is quite diverse resulting in this broad cloud and also the chosen genes are widely distributed and some are very distant to the horizontal 0-line indicating that their expression is strongly varying.

Despite we have seen that the chosen gene set is not optimal, we test all implemented normalization methods. We apply the NVTtestall function, which indicates that Q is the best normalization method based on the selected genes.

```
NVTtestall(mynorm,"p")
```

```
## [1] "No normalization!"
## [1] "Total count normalization!"
## [1] "Median normalization!"
## [1] "Trimmed Mean of M-values normalization!"
## [1] "Upper Quartile normalization!"
```

```
## [1] "Upper Quartile normalization (from NOISeq)!"
## [1] "Quantile normalization!"
## [1] "RPKM normalization!"
## [1] "RPM normalization!"
## [1] "DESeq normalization!"
## [1] "Using DESeq"
## [1] "Input counts are not integer, converting them!"
## [1] "TPM normalization!"
## [1] "Normalization by given gene-set!"
##  [1] "ENSG00000111640" "ENSG00000163631" "ENSG00000075624" "ENSG00000165704"
##  [5] "ENSG00000196839" "ENSG00000177700" "ENSG00000165280" "ENSG00000105220"
##  [9] "ENSG00000112339" "ENSG00000073578"
##         pearson      p-value
## Q    0.9670602 4.949840e-06
## DEQ  0.9265992 1.161464e-04
## G    0.9265992 1.161464e-04
## TC   0.9265992 1.161464e-04
## TMM  0.9265992 1.161464e-04
## UQ   0.9265992 1.161464e-04
## RPM  0.9265992 1.161464e-04
## N    0.9265992 1.161464e-04
## Med  0.9265992 1.161464e-04
## UQ2  0.9265992 1.161464e-04
## RPKM 0.9183081 1.763889e-04
## TPM  0.9183081 1.763889e-04
```

In order to apply this test to all data sets we can use a simple loop to go through all healthy versus cancer comparisons.

```
for (dataset1 in 1:3){
  dataset2 <- dataset1+3
  mynvt <- NVTinit(mylist_hs,myusecaseexp[,dataset1,drop=FALSE],
  myusecaseexp[,dataset2,drop=FALSE],"N",myusecaselen)
  mynorm <- NVTnormalize(mynvt)

  print(NVTtestall(mynorm,"p",verbose=F))
}
```

```
## [1] "No normalization!"
##         pearson      p-value
## Q    0.9670602 4.949840e-06
## DEQ  0.9265992 1.161464e-04
## G    0.9265992 1.161464e-04
## TC   0.9265992 1.161464e-04
## TMM  0.9265992 1.161464e-04
## UQ   0.9265992 1.161464e-04
## RPM  0.9265992 1.161464e-04
## N    0.9265992 1.161464e-04
## Med  0.9265992 1.161464e-04
```

```
## UQ2  0.9265992 1.161464e-04
## RPKM 0.9183081 1.763889e-04
## TPM  0.9183081 1.763889e-04
## [1] "No normalization!"
##        pearson      p-value
## Q    0.9903064 3.818231e-08
## TPM  0.9869293 1.257041e-07
## RPKM 0.9869293 1.257041e-07
## Med  0.9772349 1.143249e-06
## RPM  0.9772349 1.143249e-06
## N    0.9772349 1.143249e-06
## TC   0.9772349 1.143249e-06
## UQ   0.9772349 1.143249e-06
## UQ2  0.9772349 1.143249e-06
## G    0.9772349 1.143249e-06
## TMM  0.9772349 1.143249e-06
## DEQ  0.9772349 1.143249e-06
## [1] "No normalization!"
##        pearson      p-value
## RPKM 0.9995837 1.313739e-13
## TPM  0.9995837 1.313739e-13
## RPM  0.9984395 2.589449e-11
## DEQ  0.9984395 2.589449e-11
## G    0.9984395 2.589449e-11
## TC   0.9984395 2.589449e-11
## TMM  0.9984395 2.589449e-11
## UQ   0.9984395 2.589449e-11
## UQ2  0.9984395 2.589449e-11
## N    0.9984395 2.589449e-11
## Med  0.9984395 2.589449e-11
## Q    0.9967627 4.786437e-10
```

3 Session Info

- R version 3.2.5 (2016-04-14), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=de_AT.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=de_AT.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=de_AT.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=de_AT.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: knitr 1.12.3, NVT 1.0
- Loaded via a namespace (and not attached): annotate 1.48.0, AnnotationDbi 1.32.3, Biobase 2.30.0, BiocGenerics 0.16.1, BiocParallel 1.4.3, BiocStyle 1.8.0, Biostrings 2.38.4, bitops 1.0-6, colorspace 1.2-6, DBI 0.3.1, DESeq 1.22.1, evaluate 0.8.3, formatR 1.3, futile.logger 1.4.1, futile.options 1.0.0, genefilter 1.52.1, geneplotter 1.48.0, GenomeInfoDb 1.6.3, GenomicAlignments 1.6.3, GenomicRanges 1.22.4, ggplot2 2.1.0, grid 3.2.5, gtable 0.2.0, highr 0.5.1, IRanges 2.4.8, labeling 0.3, lambda.r 1.1.7, lattice 0.20-33, limma 3.26.9, magrittr 1.5, Matrix 1.2-5, munsell 0.4.3, NOISeq 2.14.1, parallel 3.2.5, plyr 1.8.3, RColorBrewer 1.1-2, Rcpp 0.12.4, RCurl 1.95-4.8, Rsamtools 1.22.0, RSQLite 1.0.0, rtracklayer 1.30.4, S4Vectors 0.8.11, scales 0.4.0, splines 3.2.5, stats4 3.2.5, stringi 1.0-1, stringr 1.0.0, SummarizedExperiment 1.0.2, survival 2.39-2, tools 3.2.5, XML 3.98-1.4, xtable 1.8-2, XVector 0.10.0, zlibbioc 1.16.0

References

[1] Himes, B. E., Jiang, X., Wagner, P., Hu, R., Wang, Q., Klanderman, B., Whitaker, R. M., Duan, Q., Lasky-Su, J., Nikolos, C., Jester, W., Johnson, M., Panettieri, R. A., Tantisira, K. G., Weiss, S. T., Lu, and Q. RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells. *PLoS ONE*, 9(6):e99625, 2014. doi:10.1371/journal.pone.0099625.

[2] Frank E. Harrell. *Regression Modeling Strategies*. Springer Series in Statistics. Springer International Publishing, Cham, 2015. URL: http://link.springer.com/10.1007/978-3-319-19425-7.

[3] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013. doi:10.1371/journal.pcbi.1003118.

[4] Michael Lawrence, Robert Gentleman, and Vincent Carey. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842, 2009. doi:10.1093/bioinformatics/btp328.

[5] Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, and Conesa A. Differential expression in rna-seq: a matter of depth. *Genome research*, 21:4436, 2011. doi:10.1101/gr.124321.111.

[6] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi:10.1093/nar/gkv007.

[7] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. doi:10.1186/gb-2010-11-10-r106.

[8] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. URL: http://had.co.nz/ggplot2/book.