# Documentation for running Normfinder in R

August 23, 2014

## The *Normfinder* algorithm

*NormFinder* is an algorithm for identifying the optimal normalization gene among a set of candidates. It ranks the set of candidate normalization genes according to their expression stability in a given sample set and given experimental design.

The algorithm is rooted in a mathematical model of gene expression and uses a solid statistical framework to estimate not only the overall expression variation of the candidate normalization genes, but also the variation between sample subgroups of the sample set e.g. normal and cancer samples. Notably, *NormFinder* provides a stability value for each gene, which is a direct measure for the estimated expression variation enabling the user to evaluate the systematic error introduced when using the gene for normalization.

The model and statistical framework underlying *NormFinder* are described in Andersen C.L. et al., *Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets*, Cancer Res. 2004;64 5245-5250.The full text article can be downloaded from
`http://cancerres.aacrjournals.org/cgi/content/abstract/64/15/5245?etoc`

*NormFinder* can analyze expression data obtained through any quantitative method e.g. real time RT- PCR and microarray based expression analysis. As default the input data is supposed to be Ct values from a real time RT-PCR run.

# Licence

The *Normfinder* software is free to use for both academical and commercial use. Modification or redistribution of the sofware is not allowed. When publishing scientific results, where NormFinder software has been used, please cite the original article: Andersen C.L. et al.,"Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets", Cancer Res. 2004;64 5245-5250.

# Disclaimer

The R-code is provided by MDL, Molecular Diagnostic Laboratory, Dept. of Molecular Medicine, Aarhus University Hospital Skejby, Denmark. *NormFinder* is provided as is, and MDL does not make any warranty, express or implied, with respect to the use of *NormFinder*. By using *Normfinder* you accept that in no event will MDL be liable for any indirect, punitive, special, incidental or consequential damages however they may arise and even if MDL has been previously advised of the possibility of such damages. 

# Download

Create on your computer a directory named `Normfinder`, say. Download the file `r.NormOldStab4.txt` and place it in the `Normfinder` directory.

# Data file

The data is stored in a file named `Datafile.txt`, say, and placed in the `Normfinder` directory.

The file `Datafile.txt` is a text file, but it is useful to think of it as organized in a table. Each row corresponds to a gene and each column to a sample. The first row contains the sample names. A sample name must be a string with no spaces. Thus a sample name as "sample 1" is not allowed. Use instead "sample_1". Similarly, the first column contains the gene names, also in the form of a string with no spaces.

As default the last row contains an identifier for the different groups. Once more these are strings with no spaces. The first entry of this last row, corresponding to the first column of the table, must also contain a name. See details below for the case where there is only one group and the last row is not included.

As default the values in the table are ct-values from a qPCR analysis. Data on a linear scale may also be handled, see details below.

# Running the programme

Start R on your computer and make `Normfinder` your working directory. Write the order

```
source("r.NormOldStab4.txt")
```

The file `r.NormOldStab4.txt` contains an R function `Normfinder` that is now available. To apply this function to your data write

```
Result=Normfinder("Datafile.txt")
```

The analysis of the data has now been performed and the results of the analysis resides in `Result`. Different outputs can be obtained on writing `Result$name`, where `name` is one of the possible outputs: *Ordered*, *Unordered* and *PairOfGenes*.

As default it is assumed that there is more than one group in the data. When there is one group only, and the file with the data has no last row with a group identifier, the above run of the programme should be replaced by

```
Result=Normfinder("Datafile.txt",Groups=FALSE)
```

As default data is assumed to be ct-values. When instead data is on a linear scale the above run of the programme should be replaced by

```
Result=Normfinder("Datafile.txt",ctVal=FALSE)
```

# Description of output

## More than one group

When writing

```
Result$Ordered
```

a table is returned where the rows are ordered according to increasing stability value. The table has four columns. First column is the gene name, second column (`GroupDif`) is a measure of the difference between the groups (two times the maximum of the $|d_{ig}|$ terms entering equation C), third column (`GroupSD`) is the common standard deviation within a group (a weighted average of the estimated intragroup variances $\hat{\sigma}_{ig}^2$ from equation B), and the fourth column (`Stability`) contains the stability measure as given by the average of the terms in equation C. An example can be seen in Table 1.

| | GroupDif | GroupSD | Stability |
|---|---|---|---|
| HSPCB | 0.00 | 0.26 | 0.12 |
| RPS13 | 0.02 | 0.44 | 0.19 |
| FLJ20030 | 0.23 | 0.30 | 0.23 |
| ATP5B | 0.25 | 0.31 | 0.24 |
| TEGT | 0.08 | 0.54 | 0.26 |
| UBC | 0.26 | 0.36 | 0.26 |
| RPS23 | 0.27 | 0.39 | 0.27 |
| UBB | 0.20 | 0.48 | 0.27 |
| CFL1 | 0.49 | 0.31 | 0.35 |
| TPT1 | 0.43 | 0.56 | 0.39 |
| FLOT2 | 0.62 | 0.66 | 0.46 |
| S100A6 | 0.66 | 0.65 | 0.47 |
| GAPD | 0.73 | 0.43 | 0.48 |
| ACTB | 0.76 | 0.56 | 0.50 |

Table 1: Output from `Result$Ordered`.

More information can be obtained on writing

```
Result$UnOrdered
```

In this case the rows have the same order as the file containing the data. This table has the same columns as before and includes the individual standard deviations for each group (`IGroupSD`) and the individual group differences (the terms $\hat{\sigma}_{ig}$ from equation B and the terms $d_{ig}$ entering equation C). An example can be seen in Table 2.

| | GroupDif | GroupSD | Stability | IGroupSD.V1 | IGroupSD.V2 | IGroupDif.V1 | IGroupDif.V2 |
|---|---|---|---|---|---|---|---|
| ATP5B | 0.25 | 0.31 | 0.24 | 0.38 | 0.20 | 0.13 | -0.13 |
| HSPCB | 0.00 | 0.26 | 0.12 | 0.30 | 0.22 | 0.00 | 0.00 |
| S100A6 | 0.66 | 0.65 | 0.47 | 0.84 | 0.30 | -0.33 | 0.33 |
| FLOT2 | 0.62 | 0.66 | 0.46 | 0.86 | 0.30 | -0.31 | 0.31 |
| TEGT | 0.08 | 0.54 | 0.26 | 0.50 | 0.58 | -0.04 | 0.04 |
| UBB | 0.20 | 0.48 | 0.27 | 0.58 | 0.34 | -0.10 | 0.10 |
| TPT1 | 0.43 | 0.56 | 0.39 | 0.57 | 0.55 | -0.21 | 0.21 |
| CFL1 | 0.49 | 0.31 | 0.35 | 0.37 | 0.23 | 0.25 | -0.25 |
| ACTB | 0.76 | 0.56 | 0.50 | 0.73 | 0.25 | 0.38 | -0.38 |
| RPS13 | 0.02 | 0.44 | 0.19 | 0.51 | 0.35 | -0.01 | 0.01 |
| RPS23 | 0.27 | 0.39 | 0.27 | 0.47 | 0.27 | -0.13 | 0.13 |
| GAPD | 0.73 | 0.43 | 0.48 | 0.41 | 0.45 | 0.36 | -0.36 |
| UBC | 0.26 | 0.36 | 0.26 | 0.43 | 0.24 | 0.13 | -0.13 |
| FLJ20030 | 0.23 | 0.30 | 0.23 | 0.34 | 0.25 | -0.11 | 0.11 |

Table 2: Output from `Result$UnOrdered`.

To obtain information on the use of an average of two genes for normalization type

`Result$PairOfGenes`

For each combination of two genes the combined stability value from formula (1.10) in the Supplementary Material is calculated. The table produced gives the names of the two genes and the value of `Stability`. The table only contains combination of genes for which the stability value, from the first run with no genes combined, is below 0.25 (the latter value can be set by `pStabLim` in the call of Normfinder). An example can be seen in Table 3.

| | Gene1 | Gene2 | Stability |
|---|---|---|---|
| 1 | ATP5B | HSPCB | 0.15 |
| 2 | ATP5B | RPS13 | 0.17 |
| 3 | ATP5B | FLJ20030 | 0.10 |
| 4 | HSPCB | RPS13 | 0.12 |
| 5 | HSPCB | FLJ20030 | 0.14 |
| 6 | RPS13 | FLJ20030 | 0.17 |

Table 3: Output from `Result$PairOfGenes`.

## 0.1 One group only

There are two possible outputs: `Result$Ordered` and `Result$PairOfGenes`.

The output `Result$Ordered` contains two columns. The first column contains the gene name and the second column (`GroupSD`) gives the standard deviation for the gene. Rows are ordered according to increasing standard deviation.

The output `Result$PairOfGenes` contains three columns. For each combination of two genes the analysis is repeated with the average of these two genes and all the remaining genes. The table produced gives the names of the two genes and the standard deviation (`GroupSD`) for the combination. The table only contains combinations of genes for which the standard deviation, from the first run with no genes combined, is below a data driven limit.

# Example

The different outputs shown above are obtained on running *Normfinder* with the input shown below.

| Sample | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|---|
| ATP5B | 14.67 | 14.89 | 14.92 | 13.62 | 13.85 | 15.09 | 14.76 | 15.32 | 15.24 |
| HSPCB | 13.57 | 13.93 | 14.02 | 11.87 | 13.24 | 13.92 | 13.72 | 14.36 | 14.27 |
| S100A6 | 14.55 | 14.38 | 14.55 | 12.78 | 14.10 | 14.80 | 14.55 | 15.39 | 15.26 |
| FLOT2 | 15.98 | 16.00 | 15.40 | 14.21 | 15.34 | 16.42 | 15.19 | 16.65 | 16.12 |
| TEGT | 15.25 | 15.62 | 15.38 | 12.30 | 14.87 | 15.56 | 14.67 | 16.02 | 15.07 |
| UBB | 13.26 | 13.46 | 13.62 | 12.39 | 12.26 | 13.57 | 13.42 | 13.45 | 13.45 |
| TPT1 | 13.53 | 13.62 | 13.62 | 13.00 | 12.57 | 13.61 | 12.86 | 13.23 | 12.99 |
| CFL1 | 12.32 | 12.27 | 12.54 | 11.09 | 12.12 | 12.62 | 12.18 | 13.08 | 13.10 |
| ACTB | 11.73 | 11.74 | 12.24 | 10.67 | 10.95 | 11.95 | 11.69 | 12.50 | 12.45 |
| RPS13 | 13.48 | 13.80 | 13.13 | 12.42 | 12.47 | 13.43 | 12.62 | 13.92 | 13.16 |
| RPS23 | 13.00 | 13.71 | 13.63 | 11.75 | 12.03 | 13.33 | 12.66 | 13.75 | 13.45 |
| GAPD | 12.53 | 13.15 | 12.74 | 10.61 | 12.67 | 13.39 | 12.85 | 13.85 | 12.66 |
| UBC | 12.75 | 12.95 | 13.17 | 11.40 | 12.72 | 12.66 | 12.55 | 13.48 | 12.94 |
| FLJ20030 | 13.40 | 13.73 | 13.65 | 12.24 | 12.91 | 13.73 | 12.67 | 13.81 | 13.21 |
| Group | Ta | Ta | Ta | Ta | Ta | Ta | Ta | Ta | Ta |

| X10 | X11 | X12 | X13 | X14 | X15 | X16 | X17 | X18 | X19 |
|---|---|---|---|---|---|---|---|---|---|
| 14.75 | 12.12 | 15.19 | 15.44 | 14.78 | 15.15 | 15.13 | 15.28 | 14.72 | 14.68 |
| 13.80 | 11.29 | 14.35 | 14.25 | 13.90 | 13.96 | 13.67 | 13.53 | 12.10 | 13.32 |
| 14.05 | 11.54 | 14.33 | 12.21 | 14.65 | 14.05 | 14.58 | 14.81 | 12.33 | 13.32 |
| 15.98 | 12.58 | 14.92 | 14.51 | 16.11 | 16.14 | 13.85 | 14.88 | 13.69 | 15.82 |
| 15.07 | 11.97 | 15.01 | 15.31 | 15.34 | 14.83 | 14.32 | 15.78 | 14.05 | 14.88 |
| 13.35 | 11.17 | 13.45 | 12.84 | 12.04 | 13.71 | 13.25 | 12.21 | 12.50 | 13.28 |
| 13.61 | 10.72 | 13.30 | 12.41 | 12.58 | 13.13 | 13.38 | 13.23 | 11.84 | 11.45 |
| 12.12 | 11.27 | 13.48 | 13.00 | 12.60 | 13.37 | 12.61 | 12.75 | 12.30 | 12.77 |
| 11.55 | 11.57 | 13.38 | 13.31 | 11.73 | 13.31 | 11.96 | 12.26 | 11.54 | 12.36 |
| 13.57 | 10.70 | 13.95 | 13.72 | 12.98 | 14.05 | 13.44 | 12.24 | 12.01 | 12.39 |
| 13.37 | 10.45 | 13.08 | 13.08 | 12.89 | 12.73 | 13.51 | 12.80 | 11.20 | 12.23 |
| 12.86 | 11.10 | 14.51 | 13.55 | 13.47 | 13.79 | 12.90 | 13.96 | 13.13 | 12.83 |
| 12.53 | 11.36 | 13.65 | 13.60 | 12.97 | 13.02 | 12.27 | 12.97 | 12.67 | 12.60 |
| 12.62 | 10.38 | 13.69 | 13.31 | 13.49 | 13.13 | 12.59 | 13.45 | 12.58 | 12.77 |
| T2-4 | T2-4 | T2-4 | T2-4 | T2-4 | T2-4 | T2-4 | T2-4 | T2-4 | T2-4 |