HTqPCR - high-throughput qPCR analysis in R and Bioconductor

Heidi Dvinge

October 18, 2010

1 Introduction

The package *HTqPCR* is designed for the analysis of cycle threshold (Ct) values from quantitative real-time PCR data. The main areas of functionality comprise data import, quality assessment, normalisation, data visualisation, and testing for statistical significance in Ct values between different features (genes, miRNAs).
Example data is included from TaqMan Low Density Arrays (TLDA), a proprietary format of Applied Biosystems, Inc. However, most functions can be applied to any kind of qPCR data, regardless of the nature of the experimental samples and whether genes or miRNAs were measured.

```
> library("HTqPCR")
```

The package employs functions from other packages of the Bioconductor project (Gentleman et al., 2004). Dependencies include *Biobase*, *RColorBrewer*, *limma*, *statmod*, *affy* and *gplots*.

Examples from the vignette

This vignette was developed in Sweave, so the embedded R code was compiled when the PDF was generated, and its output produced the results and plots that appear throughout the document. The following commands will extract all of the code from this file:

```
> all.R.commands <- system.file("doc", "HTqPCR.Rnw",
+     package = "HTqPCR")
> Stangle(all.R.commands)
```

This will create a file called HTqPCR.R in your current working directory, and this file can then either be sourced directly, or the commands run individually.

General workflow

The main functions and their use are outlined in Figure 1. Note that the QC plotting functions can be used both before and after normalisation, in order to examine the quality of the data or look for particular trends.
For the full set of available functions type:

```
> ls("package:HTqPCR")
```

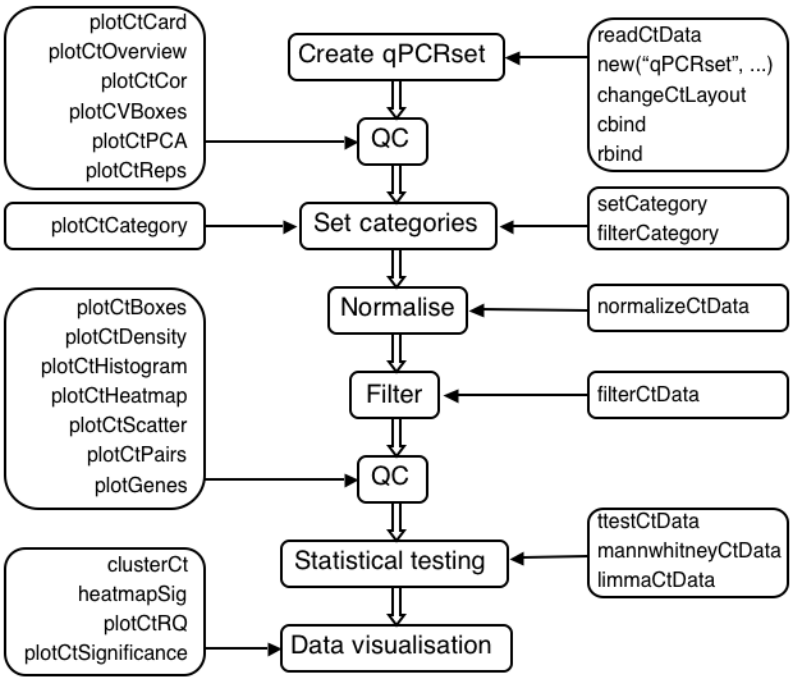


Figure 1: Workflow in *HTqPCR* analysis of qPCR data. Centre column: The main procedural steps in a typical qPCR analysis; Left: examples of visualisation functions; Right: data analysis functions.

```
 [1] "cbind.qPCRset"      "changeCtLayout"
 [3] "clusterCt"          "featureCategory"
 [5] "featureCategory<-"  "featureClass"
 [7] "featureClass<-"     "featurePos"
 [9] "featurePos<-"       "featureType"
[11] "featureType<-"      "filterCategory"
[13] "filterCtData"       "flag"
[15] "flag<-"             "getCt"
[17] "getCtHistory"       "heatmapSig"
[19] "limmaCtData"        "mannwhitneyCtData"
[21] "n.samples"          "n.wells"
[23] "normalizeCtData"    "plotCVBoxes"
[25] "plotCtBoxes"        "plotCtCard"
[27] "plotCtCategory"     "plotCtCor"
[29] "plotCtDensity"      "plotCtHeatmap"
[31] "plotCtHistogram"    "plotCtLines"
[33] "plotCtOverview"     "plotCtPCA"
[35] "plotCtPairs"        "plotCtRQ"
[37] "plotCtReps"         "plotCtScatter"
[39] "plotCtSignificance" "rbind.qPCRset"
[41] "readCtData"         "setCategory"
[43] "setCt<-"            "summary"
[45] "ttestCtData"
```

Getting help

Please send questions about *HTqPCR* to the Bioconductor mailing list.
See http://www.bioconductor.org/docs/mailList.html

2 *qPCRset* objects

The data is stored in an object of class *qPCRset*, which is similar to the *ExpressionSet* from package *Biobase* used for handling microarray data. The format is the same for raw and normalized data, and depending on how much information is available about the input data, the object can contain the following slots:

`featureNames` Object of class *character* giving the names of the features (genes, miRNAs) used in the experiment.

`sampleNames` Object of class *character* containing the sample names of the individual experiments.

`exprs` Object of class *matrix* containing the Ct values.

`flag` Object of class *data.frame* containing the flag for each Ct value, as supplied by the input files. These are typically set during the calculation of Ct values, and indicate whether the results are flagged as e.g. "Passed" or "Flagged".

`featureType` Object of class *factor* representing the different types of features on the card, such as endogenous controls and target genes.

`featurePos` Object of class *character* representing the location "well" of a gene in case TLDA cards are used, or some other method containing a defined spatial layout of features.

`featureClass` Object of class *factor* with some meta-data about the genes, for example if it is a transcription factor, kinase, marker for different types of cancers or similar. This is typically set by the user.

`featureCategory` Object of class *data.frame* representing the quality of the measurement for each Ct value, such as "OK", "Undetermined" or "Unreliable". These can be set using the function `setCategories` depending on a number of parameters, such as how the Ct values are flagged, upper and lower limits of Ct values and variations between technical and biological replicates of the same feature.

`history` Object of class *data.frame* indicating what operations have been performed on the *qPCRset* object, and what the parameters were. Automatically set when any of the functions on the upper right hand side of figure 1 are called (`readCtData, setCategory, filterCategory, normalizeCtData, filterCtData, changeCtLayout, rbind, cbind`).

Generally, information can be handled in the *qPCRset* using the same kind of functions as for *ExpressionSet*, such as `exprs`, `featureNames` and `featureCategory` for extracting the data, and `exprs<-`, `featureNames<-` and `featureCategory<-` for replacing or modifying values.
The use of `exprs` might not be intuitive to users who are not used to dealing with microarray data, and hence *ExpressionSet*. The functions `getCt` and `setCt<-` that perform the same

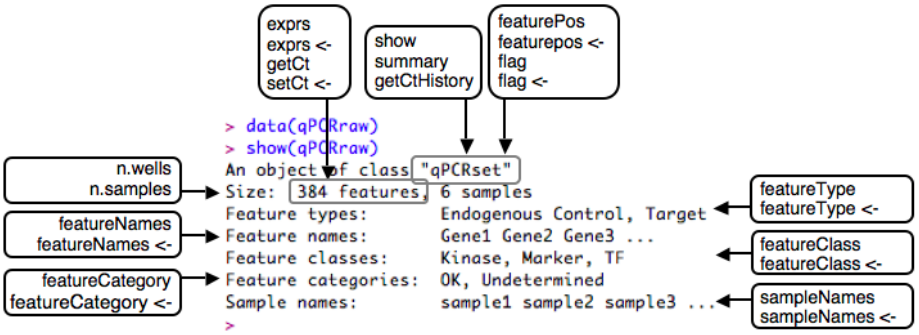


Figure 2: An example of a *qPCRset* object, and some of the functions that can be used to display and/or alter different aspects of the object, i.e. the accessor and replacement functions.

operations as `exprs` and `exprs<-` are therefore also included. For the sake of consistency, `exprs` will be used throughout this vignette for accessing the Ct values, but it can be replaced by `getCt` in all examples.
`qPCRset` objects can also be combined or reformatted in various ways (see section 13).
Two `qPCRset` test objects are included in the package: one containing raw data, and the other containing processed values. An example is shown in figure 2, along with some of the functions that can typically be used for manipulating *qPCRset* objetcs.

```
> data(qPCRraw)
> data(qPCRpros)
> class(qPCRraw)
[1] "qPCRset"
attr(,"package")
[1] ".GlobalEnv"
```

3 Reading in the raw data

The standard input consists of tab-delimited text files containing the Ct values for a range of genes. Additional information, such as type of gene (e.g. target, endogenous control) or groupings of genes into separate classes (e.g. markers, kinases) can also be read in, or supplied later.
The package comes with example input files (from Applied Biosystem's TLDA cards), along with a text file listing sample file names and biological conditions.

```
> path <- system.file("exData", package = "HTqPCR")
> head(read.delim(file.path(path, "files.txt")))
         File  Treatment
1 sample1.txt    Control
2 sample2.txt LongStarve
3 sample3.txt LongStarve
4 sample4.txt    Control
5 sample5.txt     Starve
6 sample6.txt     Starve
```

The data consist of 192 features represented twice on the array and labelled "Gene1", "Gene2", etc. There are three different conditions, "Control", "Starve" and "LongStarve", each having 2 replicates.
The input data consists of tab-delimited text files (one per sample); however, the format is likely to vary depending on the specific platform on which the data were obtained (e.g., TLDA cards, 96-well plates, or some other format). The only requirement is that columns containing the Ct values and feature names are present.

```
> files <- read.delim(file.path(path, "files.txt"))
> raw <- readCtData(files = files$File, path = path)
```

The *qPCRset* object looks like:

```
> show(raw)
An object of class "qPCRset"
Size:  384 features, 6 samples
Feature types:                  Endogenous Control, Target
Feature names:                  Gene1 Gene2 Gene3 ...
Feature classes:
Feature categories:         OK, Undetermined
Sample names:                 sample1 sample2 sample3 ...
```

NB: This section only deals with data presented in general data format. For notes regarding other types of input data, see section 12. This section also briefly deals with other types of qPCR results besides Ct data, notably the Cp values reported by the LightCycler System from Roche.

4 Data visualisation

4.1 Overview of Ct values across all groups

To get a general overview of the data the (average) Ct values for a set of features across all samples or different condition groups can be displayed. In principle, all features in a sample might be chosen, but to make it less cluttered Figure 3 displays only the first 10 features. The top plot was made using just the Ct values, and shows the 95% confidence interval across replicates within and between samples. The bottom plot represents the same values but relative to a chosen calibrator sample, here the "Control". Confidence intervals can also be added to the relative plot, in which case these will be calculated for all values compared to the average of the calibrator sample per gene.

```
> g <- featureNames(raw)[1:10]
> plotCtOverview(raw, genes = g, xlim = c(0, 50), groups = files$Treatment,
+     conf.int = TRUE, ylim = c(0, 55))
```

```
> plotCtOverview(raw, genes = g, xlim = c(0, 50), groups = files$Treatment,
+     calibrator = "Control")
```

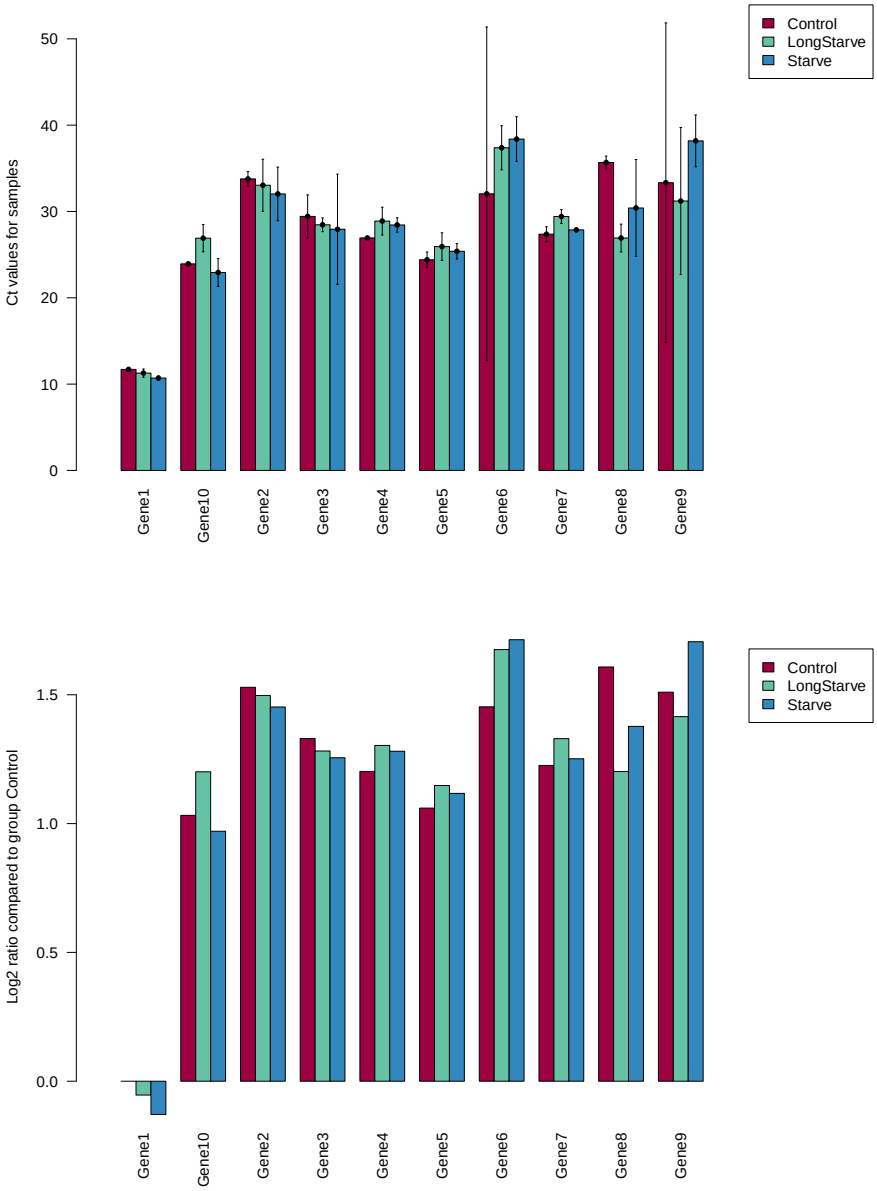


Figure 3: Overview of Ct values for the raw data.

4.2 Spatial layout

When the features are organised in a particular spatial pattern, such as the 96- or 384-well plates, it is possible to plot the Ct values or other characteristics of the features using this layout. Figure 4 shows an example of the Ct values, as well as the location of different classes of features (using random examples here), across all the wells of a TLDA microfluidic card.

```
> plotCtCard(raw, col.range = c(10, 35), well.size = 2.6)
```

```
> featureClass(raw) <- factor(c("Marker", "TF", "Kinase")[sample(c(1,
+     1, 2, 2, 1, 3), 384, replace = TRUE)])
> plotCtCard(raw, plot = "class", well.size = 2.6)
```

5 Feature categories and filtering

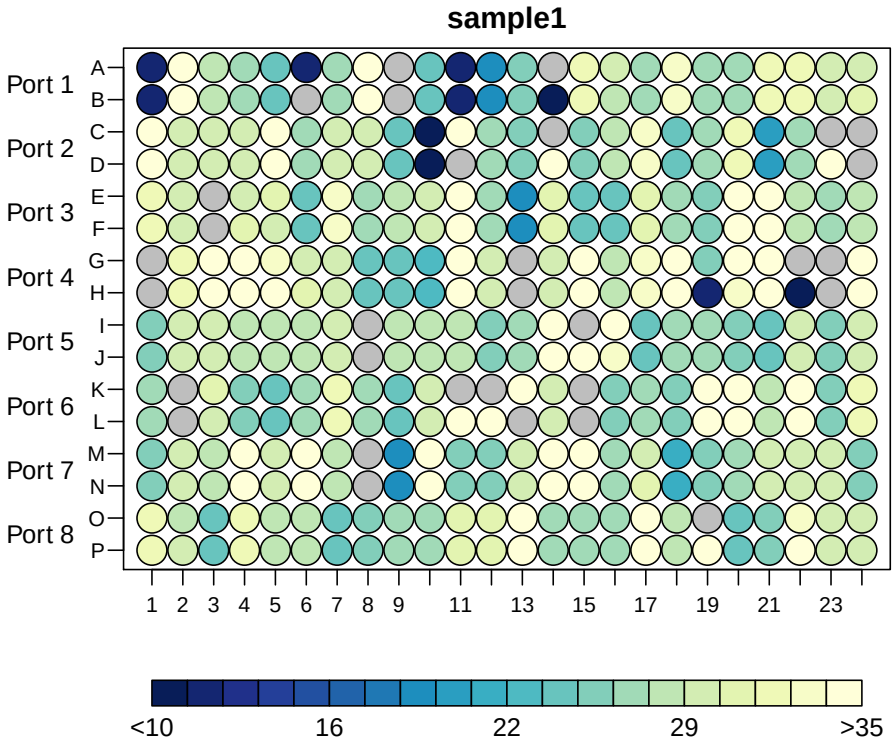
Each Ct values in HTqPCR has an associated feature category. This is an important component to indicate the reliability of the qPCR data. Aside from the "OK" indicator, there are two other categories: "Undetermined" is used to flag Ct values above a user-selected threshold, and "Unreliable" indicates Ct values that are either so low as to be estimated by the user to be problematic, or that arise from deviation between individual Ct values across replicates. By default, only Ct values labelled as "undetermined" in the input data files are placed into the "Undetermined" category, and the rest are classified as "OK". However, either before or after normalisation these categories can be altered depending on various criteria.

Range of Ct values Some Ct values might be too high or low to be considered a reliable measure of gene expression in the sample, and should therefore not be marked as "OK".

Flags Depending on the qPCR input the values might have associated flags, such as "Passed" or "Failed", which are used for assigning categories.

Biological and technical replicates If features are present multiple times within each sample, or if samples are repeated in the form of technical or biological replicates, then these values can be compared. Ct values lying outside a user-selected confidence interval (90% by default) will be marked as "Unreliable".

```
> raw.cat <- setCategory(raw, groups = files$Treatment,
+     quantile = 0.8)
Categories after Ct.max and Ct.min filtering:
             sample1 sample2 sample3 sample4 sample5 sample6
OK               313     264     327     295     296     286
Undetermined      68     119      56      86      86      96
Unreliable         3       1       1       3       2       2
Categories after standard deviation filtering:
             sample1 sample2 sample3 sample4 sample5 sample6
OK               301     254     319     274     277     275
Undetermined      68     119      56      86      86      96
Unreliable        15      11       9      24      21      13
```



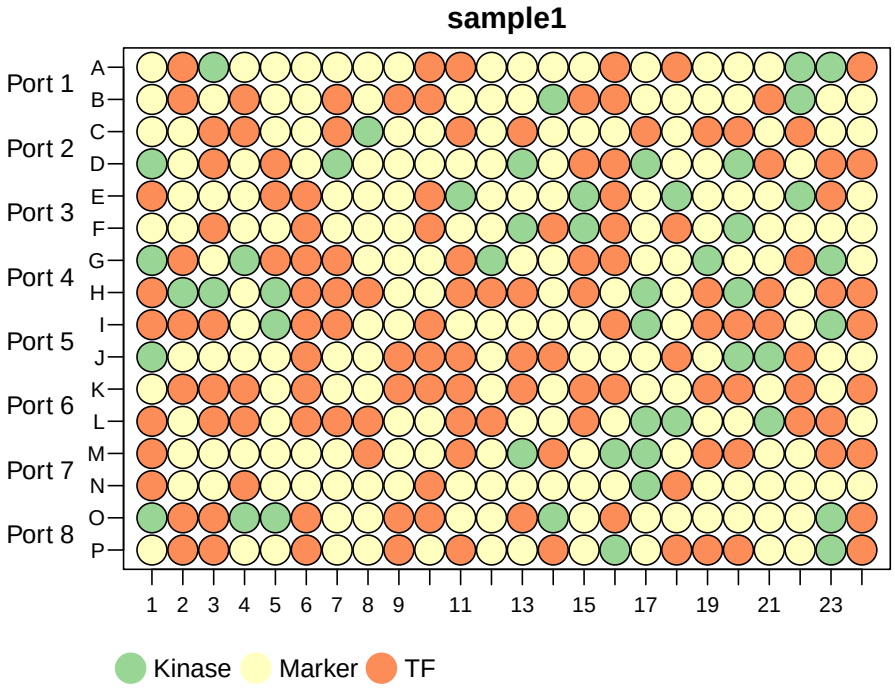


Figure 4: Ct values for the first sample (top), and the location of different feature classes (bottom). Ct values are visualised using colour intensity, and grey circles are features that were marked "undetermined" in the input file.

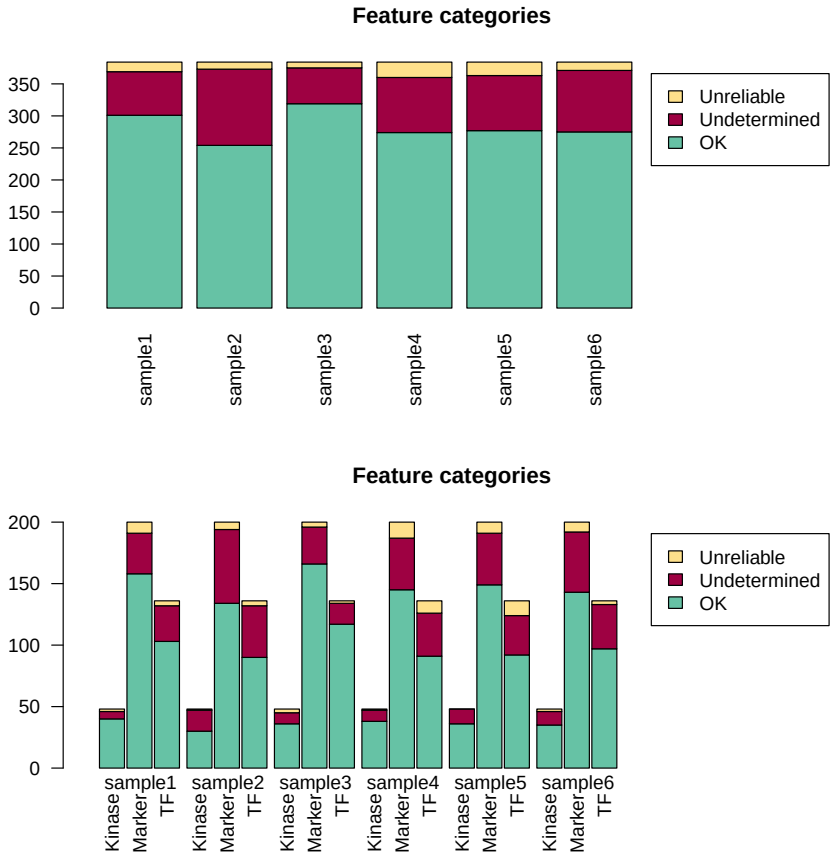


Figure 5: Summary of the categories, either for each sample individually or stratified by feature class.

A summary plot for the sample categories is depicted in Figure 5. The result can be stratified by `featureType` or `featureClass`, for example to determine whether one class of features performed better or worse than others.

```
> plotCtCategory(raw.cat)
```

```
> plotCtCategory(raw.cat, stratify = "class")
```

The results can also be shown per feature rather than averaged across samples (Figure 6).

```
> plotCtCategory(raw.cat, by.feature = TRUE, cexRow = 0.1)
```

If one doesn't want to include unreliable or undetermined data in part of the analysis, these Ct values can be set to NA using `filterCategory`. However, the presence of NAs could make the tests for differential expression less robust. When testing for differential expression the result will come with an associated category ("OK" or "Unreliable") that can instead be used to assess the quality of the results. For the final results both "Undetermined" and "Unreliable" are pooled together as being "Unreliable". However, the label for each feature can either be set according to whether half or more of the samples are unreliable, or whether only a single non-"OK" category is present, depending on the level of stringency the user wishes to enforce.

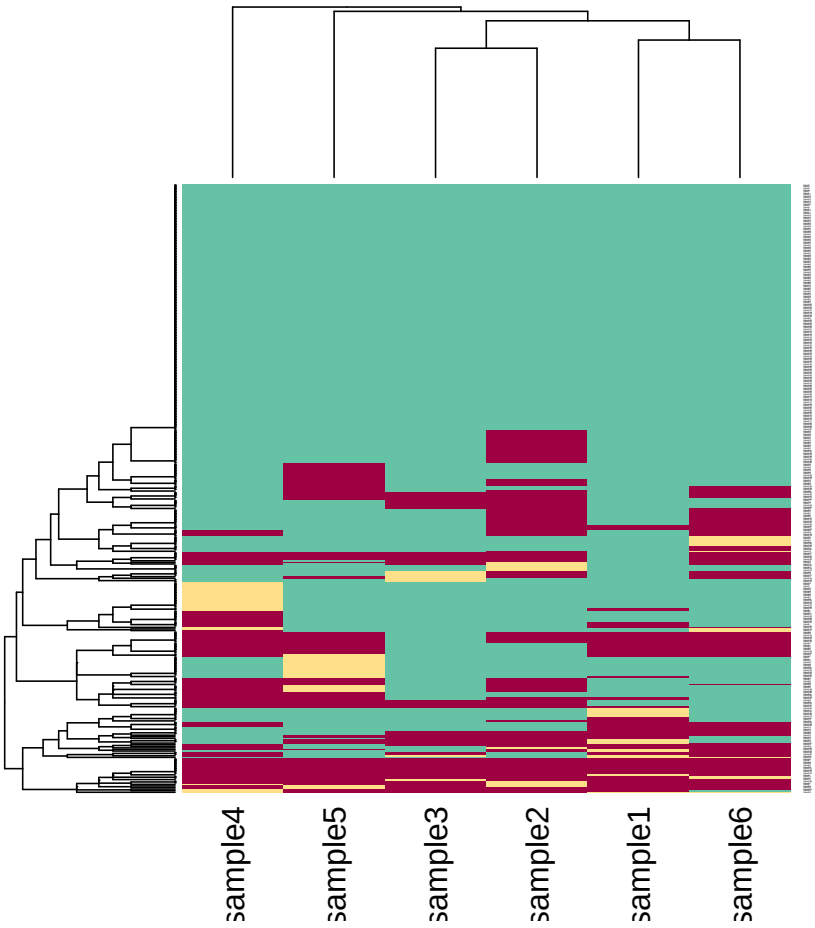


Figure 6: Summary of the categories, clustered across features.

6 Normalisation

Four different normalisation methods are currently implemented in *HTqPCR*. Two of these (`scale.rankinvariant` and `deltaCt`) will scale each individual sample by a given value, whereas the remaining two will change the distribution of Ct values.

quantile Will make the distribution of Ct values more or less identical across samples.

norm.rankinvariant Computes all rank-invariant sets of features between pairwise comparisons of each sample against a reference, such as a pseudo-mean. The rank-invariant features are used as a reference for generating a smoothing curve, which is then applied to the entire sample.

scale.rankinvariant Also computes the pairwise rank-invariant features, but then takes only the features found in a certain number of samples, and used the average Ct value of those as a scaling factor for correcting all Ct values.

deltaCt Calculates the standard deltaCt values, i.e. subtracts the mean of the chosen controls from all other values in the feature set.

For the rank-invariant normalisation methods, Ct values above a given threshold can be excluded from the calculation of a scaling factor or normalisation curve. This is useful so that a high proportion of "Undetermined" Ct values (assigned a value of 40 by default) in a given sample doesn't bias the normalisation of the remaining features.
In the example dataset, Gene1 and Gene60 correspond to 18S RNA and GADPH, and are used as endogenous controls. Normalisation methods can be run as follows:

```
> q.norm <- normalizeCtData(raw.cat, norm = "quantile")
> sr.norm <- normalizeCtData(raw.cat, norm = "scale.rank")
Scaling Ct values
        Using rank invariant genes: Gene1 Gene29
        Scaling factors: 1.00 1.06 1.00 1.03 1.00 1.00
> nr.norm <- normalizeCtData(raw.cat, norm = "norm.rank")
Normalizing Ct values
        Using rank invariant genes:
        sample1: 75 rank invariant genes
        sample2: 33 rank invariant genes
        sample3: 48 rank invariant genes
        sample4: 69 rank invariant genes
        sample5: 18 rank invariant genes
        sample6: 67 rank invariant genes
> d.norm <- normalizeCtData(raw.cat, norm = "deltaCt",
+     deltaCt.genes = c("Gene1", "Gene60"))
Calculating deltaCt values
        Using control gene(s): Gene1 Gene60
        Card 1:         Mean=14.45        Stdev=4.25
        Card 2:         Mean=15.19        Stdev=5.27
        Card 3:         Mean=14.50        Stdev=5.8
        Card 4:         Mean=14.79        Stdev=4.79
        Card 5:         Mean=14.07        Stdev=5.32
        Card 6:         Mean=13.82        Stdev=4.75
```

Comparing the raw and normalised values gives an idea of how much correction has been performed (Figure 7), as shown below for the `q.norm` object. Note that the scale on the y-axis varies.

```
> plot(exprs(raw), exprs(q.norm), pch = 20, main = "Quantile normalisation",
+     col = rep(brewer.pal(6, "Spectral"), each = 384))
```

7 Filtering and subsetting the data

At any point during the analysis it's possible to filter out both individual features or groups of features that are either deemed to be of low quality, or not of interest for a particular aspect of the analysis. This can be done using any of the feature characteristics that are included in the `featureNames`, `featureType`, `featureClass` and/or `featureCategory` slots of the data object. Likewise, the `qPCRset` object can be turned into smaller subsets, for example if only a particular class of features are to be used, or some samples should be excluded.
Simple subsetting can be done using the standard `[,]` notation of R, for example:

```
> nr.norm[1:10, ]
An object of class "qPCRset"
Size:  10 features, 6 samples
Feature types:                 Endogenous Control, Target
Feature names:                 Gene1 Gene2 Gene3 ...
Feature classes:         Kinase, Marker, TF
Feature categories:         OK, Unreliable, Undetermined
Sample names:                sample1 sample2 sample3 ...
> nr.norm[, c(1, 3, 5)]
An object of class "qPCRset"
Size:  384 features, 3 samples
Feature types:                 Endogenous Control, Target
Feature names:                 Gene1 Gene2 Gene3 ...
Feature classes:         Kinase, Marker, TF
Feature categories:         OK, Unreliable, Undetermined
Sample names:                sample1 sample3 sample5 ...
```

Filtering is done by specifying the components to remove, either by just using a single criteria, or by combining multiple filters:

```
> qFilt <- filterCtData(nr.norm, remove.type = "Endogenous Control")
Removed 4 'Endogenous Control' features based on featureType(q).
> qFilt <- filterCtData(nr.norm, remove.name = c("Gene1",
+     "Gene20", "Gene30"))
Removed 8 'Gene1/Gene20/Gene30' features based on featureNames(q).
> qFilt <- filterCtData(nr.norm, remove.class = "Kinase")
Removed 48 'Kinase' features based on featureClass(q).
> qFilt <- filterCtData(nr.norm, remove.type = c("Endogenous Control"),
+     remove.name = c("Gene1", "Gene20", "Gene30"))
Removed 4 'Endogenous Control' features based on featureType(q).
Removed 4 'Gene1/Gene20/Gene30' features based on featureNames(q).
```

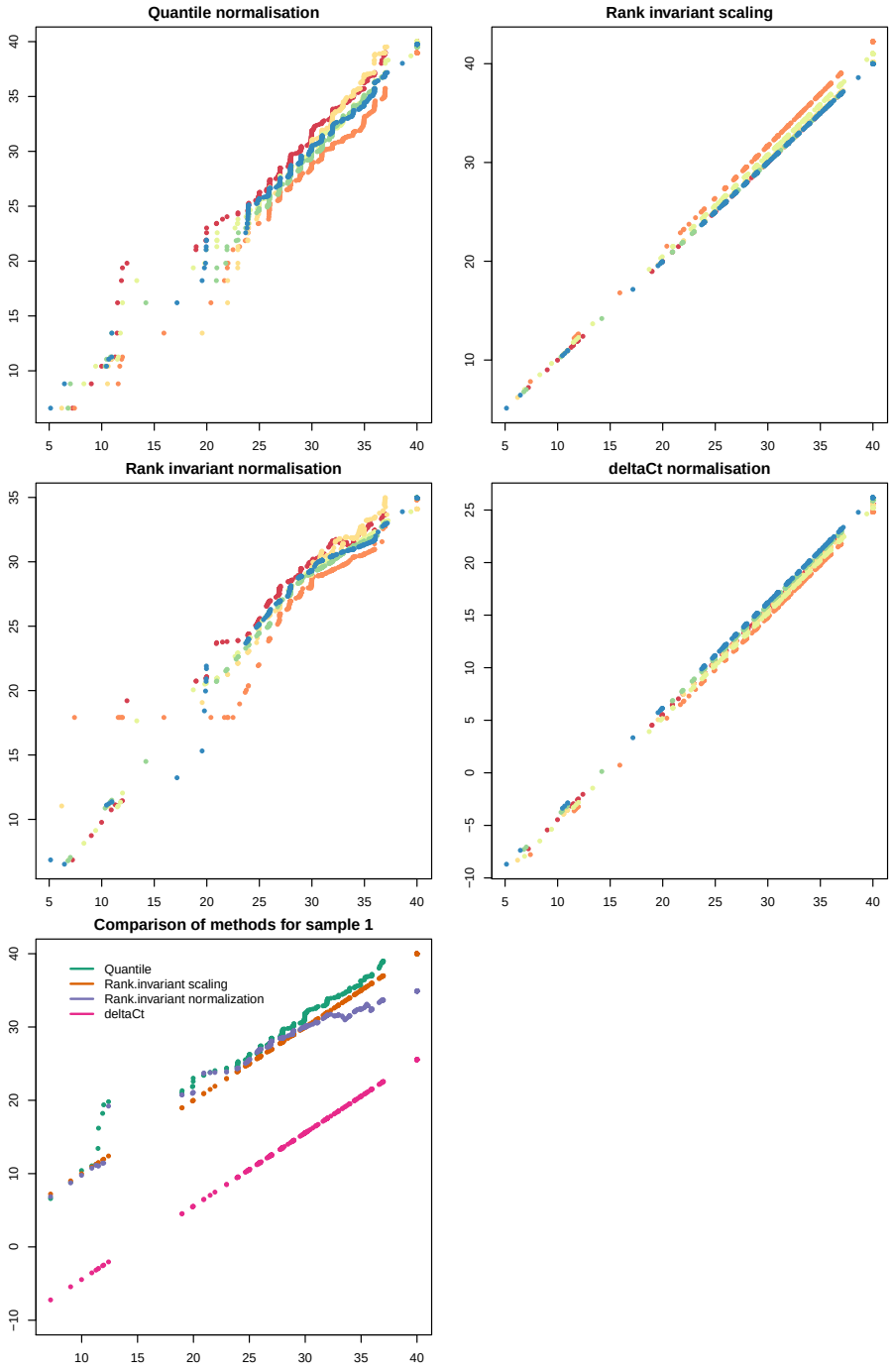


Figure 7: Normalized versus raw data, using a separate colour for each sample. The raw data is plotted along the x-axis and the normalised along y. The last plot is a comparison between normalization methods for the fist sample, still with the raw Ct values along the x-axis.

The data can also be adjusted according to feature categories. With `filterCategory` mentioned previously it's possible to replace certain Ct values with NA, but one might want to completely exclude features where a certain number of the Ct values are for example unreliable.

```
> qFilt <- filterCtData(nr.norm, remove.category = "Undetermined")
Removed 64 features with >3 'Undetermined' based on featureCategory(q).
> qFilt <- filterCtData(nr.norm, remove.category = "Undetermined",
+     n.category = 5)
Removed 10 features with >5 'Undetermined' based on featureCategory(q).
```

Another typical filtering step would be to remove features showing little or no variation across samples, prior to testing for statistical significance of genes between samples. Features with relatively constant Ct levels are less likely to be differentially expressed, so including them in the downstream analysis would cause some loss of power when adjusting the p-values for multiple testing (the feature-by-feature hypothesis testing). Variation across samples can be assessed using for example the interquartile range (IQR) values for each feature (Figure 8).

```
> iqr.values <- apply(exprs(nr.norm), 1, IQR)
> hist(iqr.values, n = 20, main = "", xlab = "IQR across samples")
> abline(v = 1.5, col = 2)
```

All features with IQR below a certain threshold can then be filtered out.

```
> qFilt <- filterCtData(nr.norm, remove.IQR = 1.5)
Removed 207 features with IQR <1.5 based on exprs(q).
```

Note that filtering prior to normalisation can affect the outcome of the normalisation procedure. In some cases this might be desirable, for example if a particular feature class are heavily biasing the results so it's preferable to split the `qPCRset` object into smaller data sets. However, in other cases it might for example make it difficult to identify a sufficient number of rank invariant features for the`norm.rankinvariant` and `scale.rankinvariant` methods. Whether to perform filtering, and if so then during what step of the analysis, depends on the genes and biological samples being analysed, as well as the quality of the data. It's therefore advisable to perform a detailed quality assessment and data comparison, as mentioned in the next section.

8 Quality assessment

8.1 Correlation between samples

The overall correlation between different samples can be displayed visually, such as shown for the raw data in Figure 9.

```
> plotCtCor(raw, main = "Ct correlation")
```

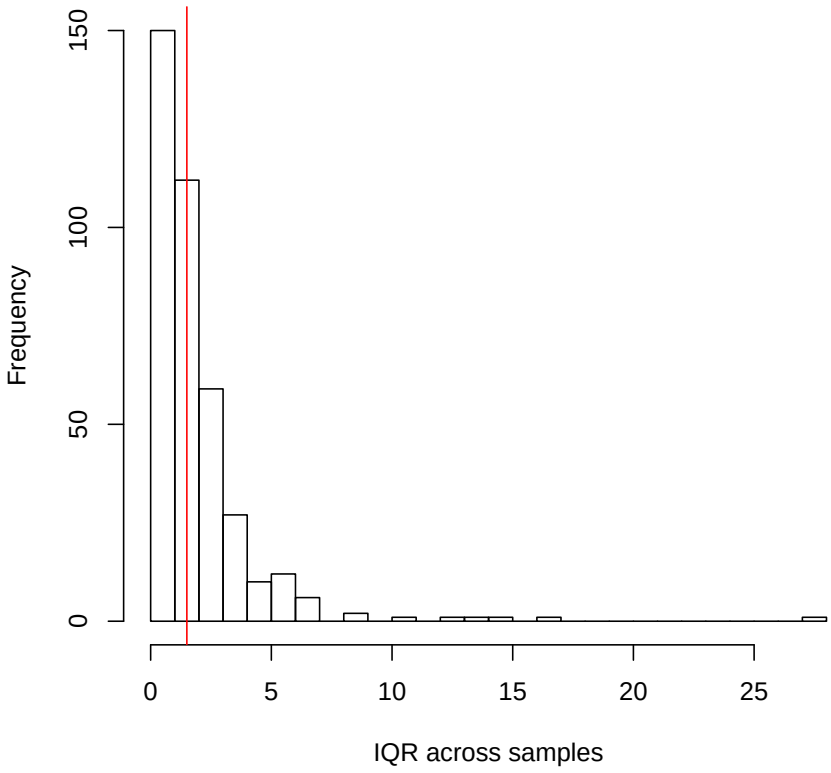


Figure 8: Histogram of the IQR values for all features across the samples, including the cut-off used in the filtering example.

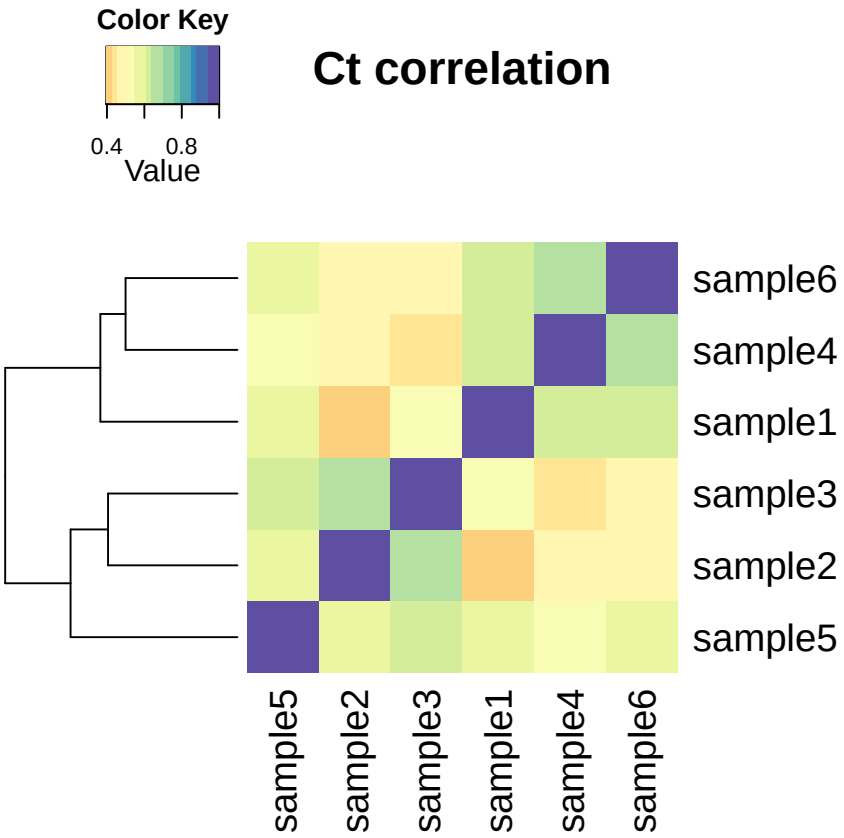


Figure 9: Correlation between raw Ct values.

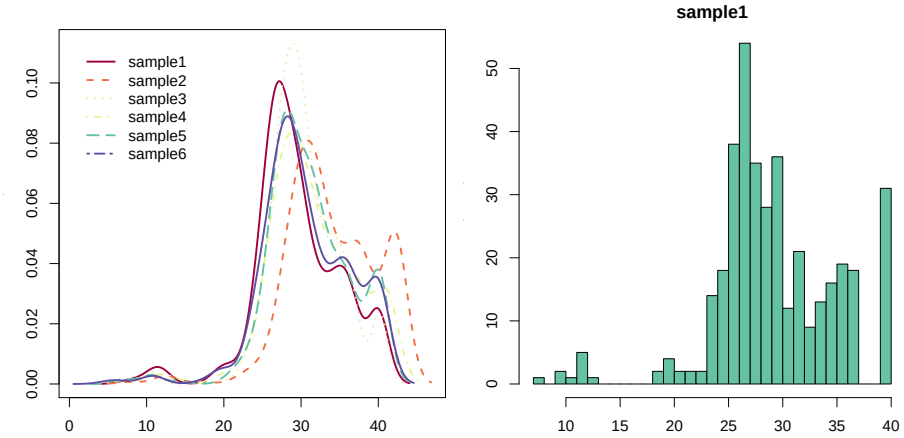


Figure 10: Distribution of Ct values for the individual samples, either using the density of all arrays (left) or a histogram of a single sample (right), after scale rank-invariant normalisation.

8.2 Distribution of Ct values

It may be of interest to examine the general distribution of data both before and after normalisation. A simple summary of the data can be obtained using `summary` as shown below.

```
> summary(raw)
        sample1  sample2  sample3 sample4  sample5  sample6
Min.    " 7.218" " 7.408" " 6.19" " 6.853" " 6.787" " 5.133"
1st Qu. "26.738" "28.855" "27.90" "26.964" "27.913" "27.514"
Median  "28.937" "30.994" "29.92" "29.943" "30.778" "29.931"
Mean    "29.542" "32.190" "30.35" "30.590" "30.995" "30.663"
3rd Qu. "33.323" "35.985" "32.98" "34.694" "34.702" "35.046"
Max.    "40.000" "40.000" "40.00" "40.000" "40.000" "40.000"
```

However, figures are often more informative. To that end, the range of Ct values can be illustrated using histograms or with the density distribution, as shown in Figure 10.

```
> plotCtDensity(sr.norm)
```

```
> plotCtHistogram(sr.norm)
```

Plotting the densities of the different normalisation methods lends insight into how they differ (Figure 11).
Ct values can also be displayed in boxplots, either with one box per sample or stratified by different attributes of the features, such as `featureClass` or `featureType` (Fig. 12).

```
> plotCtBoxes(sr.norm, stratify = "class")
```

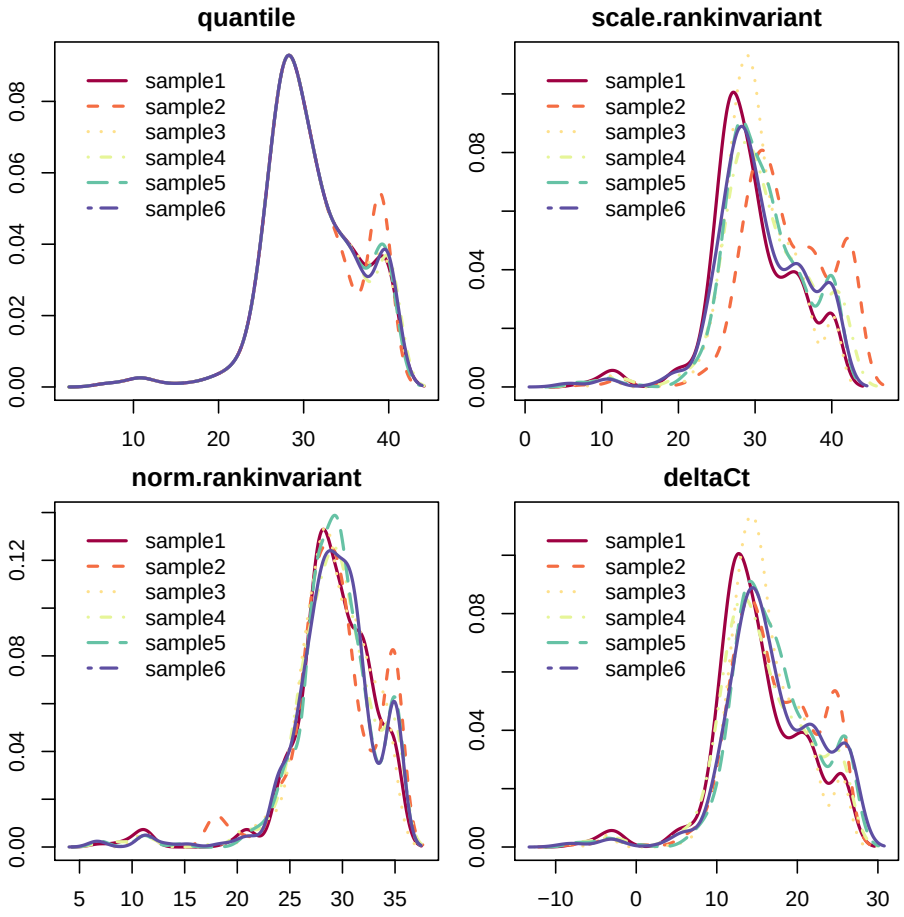


Figure 11: Densities of Ct values for all samples after each of the four normalisation methods. The peak at the high end originates from features with "Undetermined" Ct values, which are assigned the Ct value 40 by default.

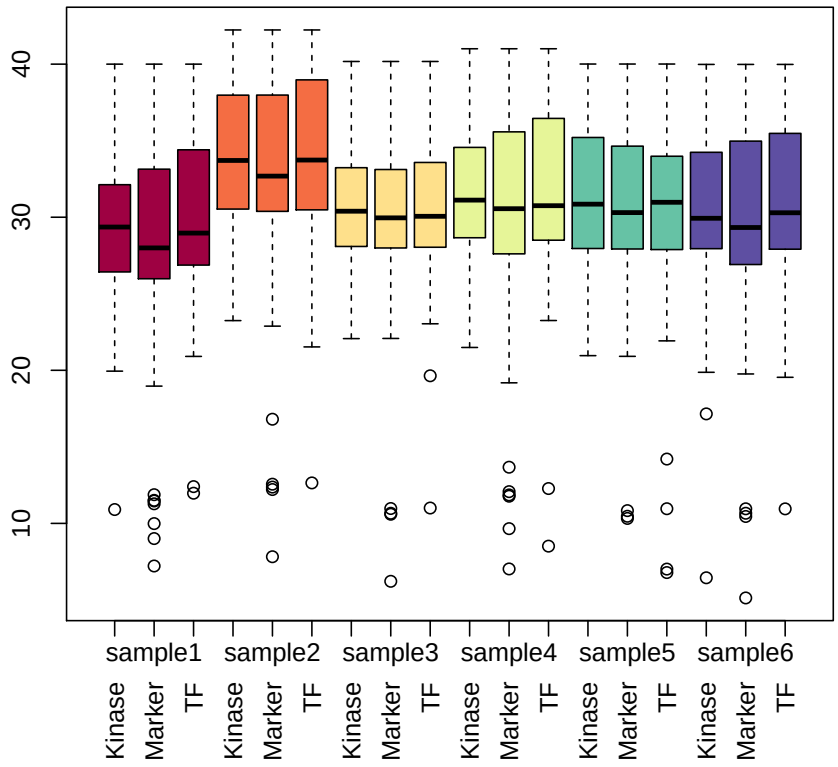


Figure 12: Boxplot of Ct values across all samples, stratified by feature classes.

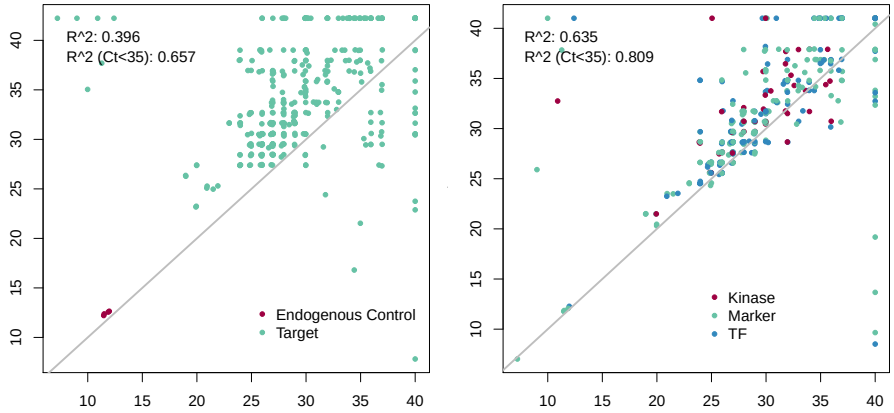


Figure 13: Scatter plot of Ct values in different samples, with points marked either by featureType (left) or featureClass (right) and the diagonal through $x = y$ marked with a grey line.

8.3 Comparison of Ct values for two samples

It is often of interest to directly compare Ct values between two samples. In Figure 13, two examples are shown for the rank-invariant normalised data: one for different biological samples, and one for replicates.

```
> plotCtScatter(sr.norm, cards = c(1, 2), col = "type",
+     diag = TRUE)
```

```
> plotCtScatter(sr.norm, cards = c(1, 4), col = "class",
+     diag = TRUE)
```

8.4 Scatter across all samples

It is also possible to generate a scatterplot of Ct values between more than the two samples shown above. In Figure 14 all pairwise comparisons are shown, along with their correlation when all Ct values <35 are removed.

```
> plotCtPairs(sr.norm, col = "type", diag = TRUE)
```

8.5 Ct heatmaps

Heatmaps provide a convenient way to visualise clustering of features and samples at the same time, and show the levels of Ct values (Figure 15). The heatmaps can be based on either Pearson correlation coefficients or Euclidean distance clustering. Euclidean-based heatmaps will focus on the magnitude of Ct values, whereas Pearson clusters the samples based on similarities between the Ct profiles.

```
> plotCtHeatmap(raw, gene.names = "", dist = "euclidean")
```

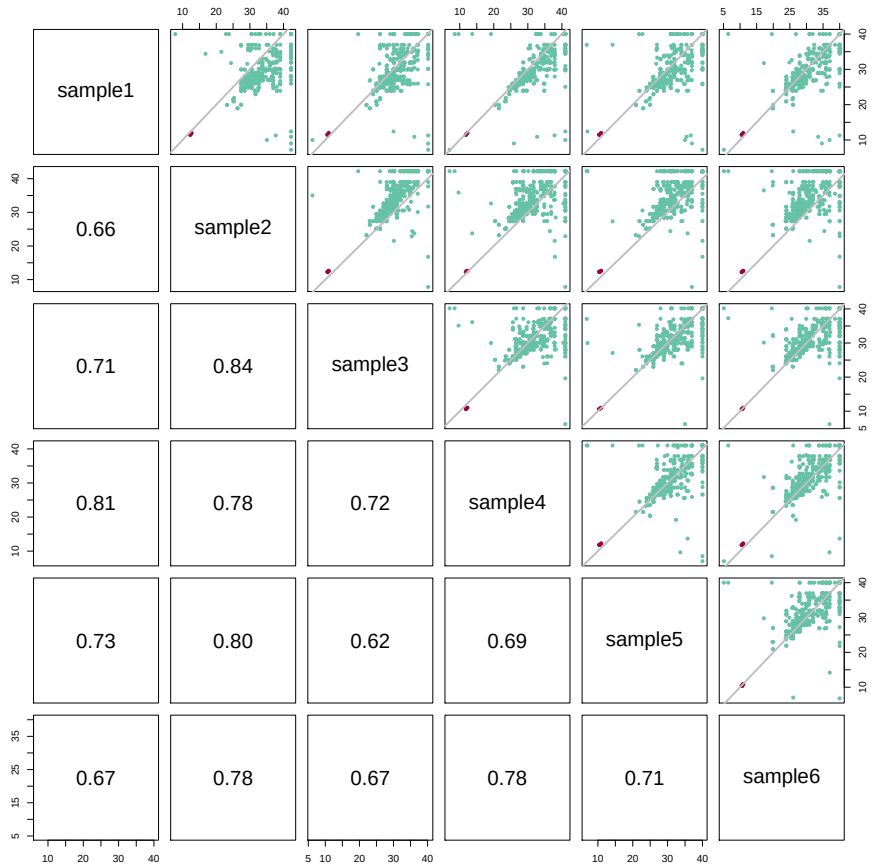


Figure 14: Scatterplot for all pairwise comparisons between samples, with spots marked depending on `featureType`, i.e. whether they represent endogenous controls or targets.

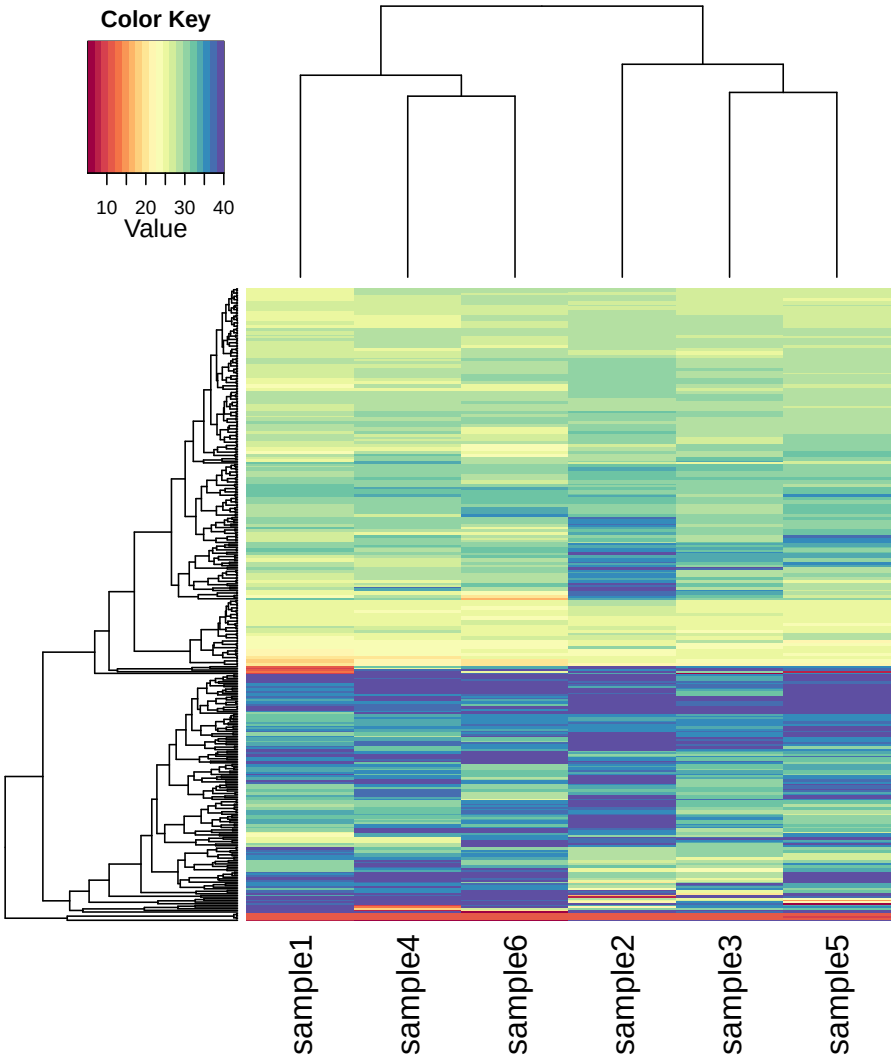


Figure 15: Heatmap for all samples and genes, based on the Euclidean distance between Ct values.

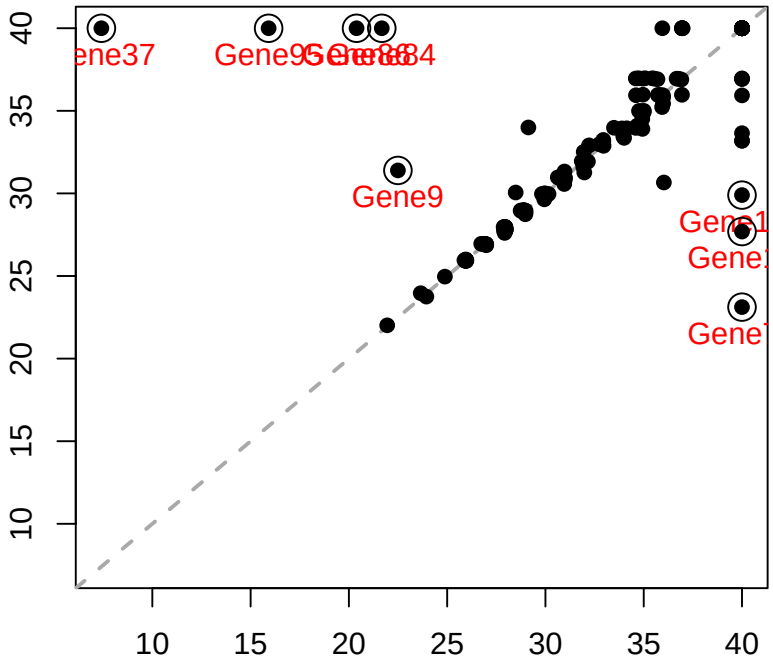


Figure 16: Concordance between duplicated Ct values in sample 2, marking features differing >20% from their mean.

8.6 Comparison of replicated features within samples

When a sample contains duplicate measurements for some or all features, the Ct values of these duplicates can be plotted against each other to measure accordance between duplicates. In Figure 16 the duplicates in sample2 are plotted against each other, and those where the Ct values differ more than 20% from the average of a given feature are marked.

```
> plotCtReps(qPCRraw, card = 2, percent = 20)
Replicates differing > 20% on card 2:
             rep1     rep2
Gene135 40.000000 29.90044
Gene14  40.000000 27.69185
Gene37   7.408248 40.00000
Gene73  40.000000 23.11949
Gene84  21.673946 40.00000
Gene86  20.389658 40.00000
Gene9   22.494440 31.39404
Gene95  15.916160 40.00000
```

Differences will often arise due to one of the duplicates marked as “Undetermined”, thus contributing to an artificially high Ct value, but other known cases exist as well.

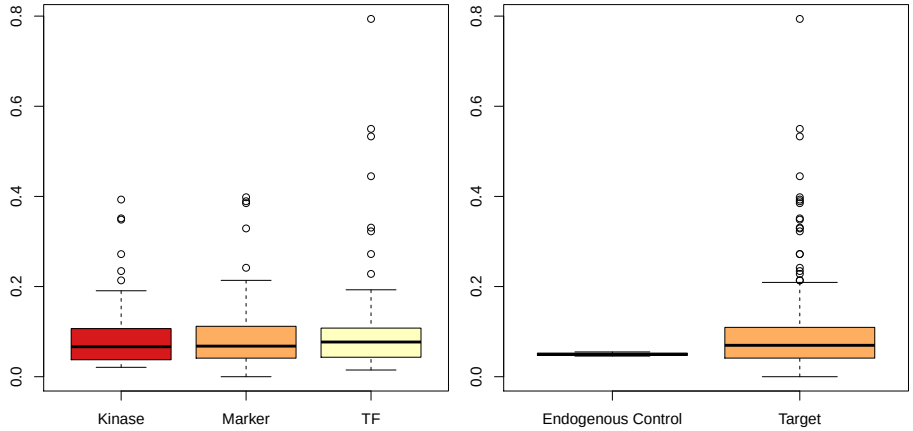


Figure 17: Coefficients of variation for each feature across all samples.

8.7 Coefficients of variation

The coefficients of variation (CV) can be calculated for each feature across all samples. Stratifying the CV values by `featureType` or `featureClass` can help to determine whether one class of features is more variable than another (Figure 17). For the example data feature classes have been assigned randomly, and the CVs are therefore similar, whereas for the feature types there's a clear difference between controls and targets.

```
> plotCVBoxes(qPCRraw, stratify = "class")
> plotCVBoxes(qPCRraw, stratify = "type")
```

9 Clustering

At the moment there are two default methods present in *HTqPCR* for clustering; hierarchical clustering and principal components analysis (PCA).

9.1 Hierarchical clustering

Both features and samples can be subjected to hierarchical clustering using either Euclidean or Pearson correlation distances, to display similarities and differences within groups of data. Individual subclusters can be selected, either using pre-defined criteria such as number of clusters, or interactively by the user. The content of each cluster is then saved to a list, to allow these features to be extracted from the full data set if desired.
An example of a clustering of samples is shown in Figure 18. In Figure 19 these data are clustered by features, and the main subclusters are marked.

```
> clusterCt(sr.norm, type = "samples")
```

```
> cluster.list <- clusterCt(sr.norm, type = "genes",
+     n.cluster = 6, cex = 0.5)
```

```
> c6 <- cluster.list[[6]]
> print(c6)
```

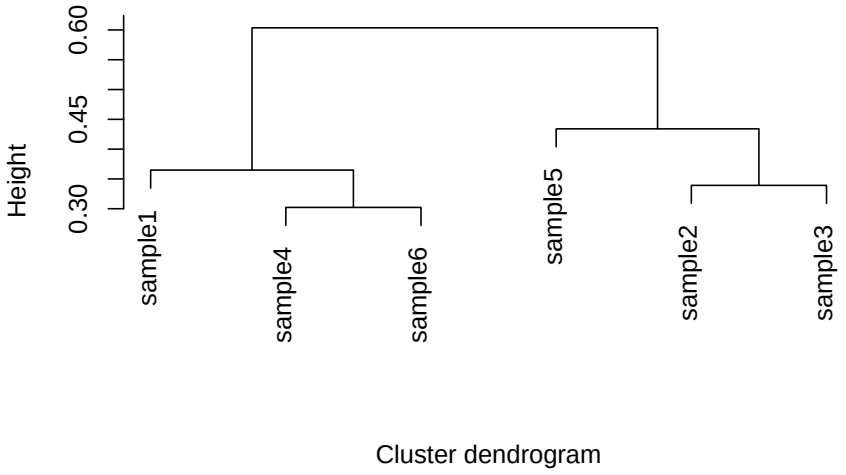


Figure 18: Hierarchical clustering of samples.

```
  Gene9  Gene13  Gene46  Gene50  Gene75  Gene93  Gene75 Gene103
      9      14      71      99     148     166     172     200
Gene132 Gene151 Gene151
    277     296     320
> show(sr.norm[c6, ])
An object of class "qPCRset"
Size:  11 features, 6 samples
Feature types:                 Endogenous Control, Target
Feature names:                 Gene9 Gene13 Gene46 ...
Feature classes:         Kinase, Marker, TF
Feature categories:         Undetermined, Unreliable, OK
Sample names:                sample1 sample2 sample3 ...
```

9.2 Principal components analysis

PCA is performed across the selected features and samples (observations and variables), and can be visualized either in a biplot, or showing just the clustering of the samples (Figure 20).

```
> plotCtPCA(qPCRraw)
> plotCtPCA(qPCRraw, features = FALSE)
```

10 Differential expression

At this stage multiple filterings might have been performed on the data set(s). To remind yourself about those, you can use the `getCtHistory` function on the *qPCRset* object.

```
> getCtHistory(sr.norm)
```

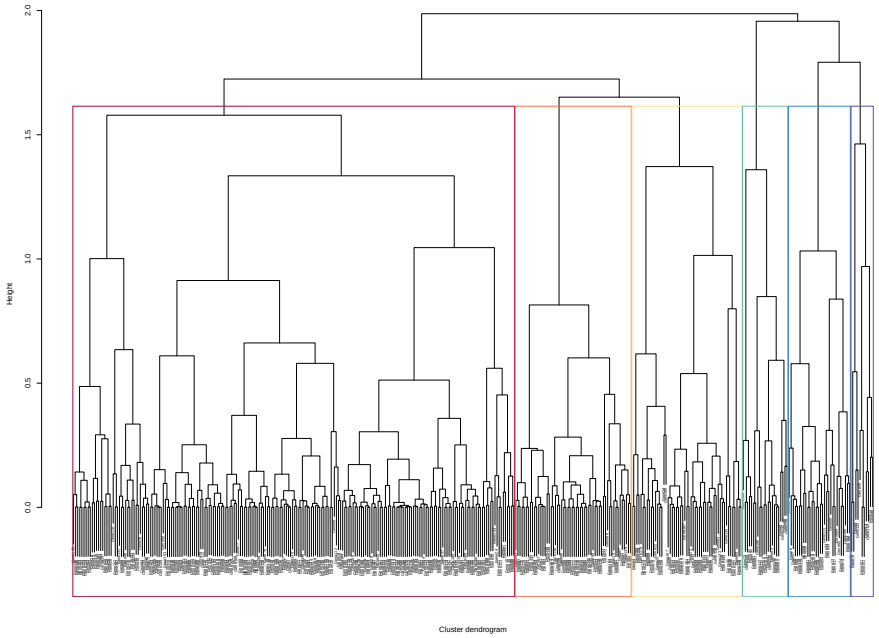


Figure 19: Hierarchical clustering of features, with subclusters marked.

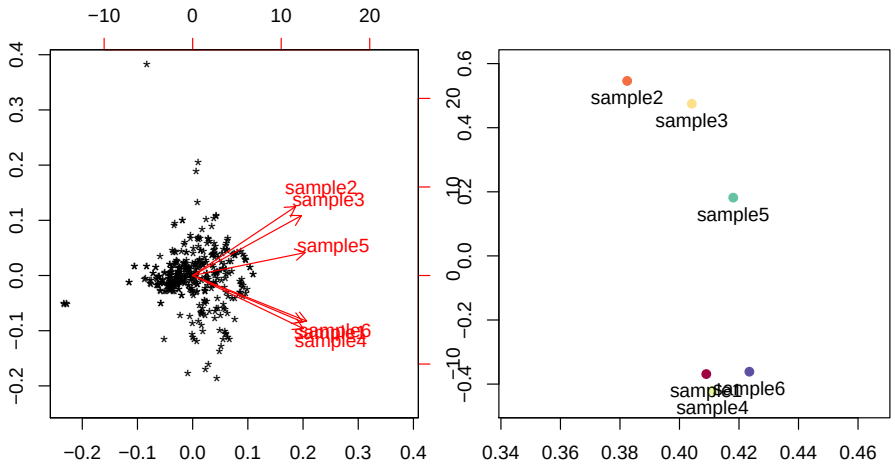


Figure 20: Left: a biplot including all features, with samples represented by vectors. Right: the same plot, including only the samples.

```
                                                                history
1                    readCtData(files = files$File, path = path)
2 setCategory(q = raw, quantile = 0.8, groups = files$Treatment)
3              normalizeCtData(q = raw.cat, norm = "scale.rank")
> getCtHistory(qFilt)
                                                                history
1                    readCtData(files = files$File, path = path)
2 setCategory(q = raw, quantile = 0.8, groups = files$Treatment)
3               normalizeCtData(q = raw.cat, norm = "norm.rank")
4                    filterCtData(q = nr.norm, remove.IQR = 1.5)
```

In general there are three approaches in *HTqPCR* for testing the significance of differences in Ct values between samples.

t-test Performing a standard t-test between two sample groups. This function will incorporate information about replicates to calculate t and p-values. This is a fairly simple approach, typically used when comparing a single treatment and control sample, and multiple pair-wise tests can be carried out one-by-one by the user.

Mann-Whitney This is a non-parametric test, also known as a two sample Wilcoxon test. Similar to the t-test, multiple pair-wise tests will have to be carried out one by one if more than two types of samples are present. This is a rank-based test that doesn't make any assumptions about the population distribution of Ct values.

limma Using a wrapper for functions from the package *limma* (Smyth, 2005) to calculate more sophisticated t and p-values for any number of groups present across the experiment. This is more flexible in terms of what types of comparisons can be made, but the users need to familiarise themselves with the *limma* conventions for specifying what contrasts are of interest.

Examples of how to use each of these are given in the next sections. In all cases the output is similar; a data frame containing the test statistics for each feature, along with fold change and information about whether the Ct values are "OK" or "Unreliable". This result can be written to a file using standard functions such as `write.table`.

10.1 Two sample types - t-test

This section shows how to compare two samples, e.g. the control and long starvation samples from the example data. A subset of the `qPCRset` data is created to encompass only these samples, and a t-test is then performed.

```
> qDE.ttest <- ttestCtData(sr.norm[, 1:4], groups = files$Treatment[1:4],
+     calibrator = "Control")
> head(qDE.ttest)
      genes feature.pos     t.test      p.value adj.p.value
2    Gene10     A10;B10 -13.705640 9.812948e-06 0.001874273
22  Gene118     I23;J23 -10.407281 8.661980e-05 0.008272191
36  Gene130     K11;L11   7.644065 3.156334e-04 0.019863762
164  Gene74       G3;H3   7.201124 4.768225e-04 0.019863762
30  Gene125       K6;L6  -7.203614 5.199938e-04 0.019863762
33  Gene128       K9;L9 -10.357650 6.826374e-04 0.020708859
         ddCt           FC meanCalibrator meanTarget
```

```
2    3.460492 9.084231e-02       24.23106   27.69156
22   8.982157 1.977430e-03       26.70531   35.68746
36  -9.433460 6.914400e+02       39.74414   30.31068
164 -5.178293 3.620942e+01       35.40699   30.22869
30   7.052337 7.534164e-03       28.28445   35.33679
33   5.210558 2.700635e-02       25.61542   30.82598
    categoryCalibrator categoryTarget
2         Undetermined            OK
22                  OK  Undetermined
36        Undetermined            OK
164       Undetermined            OK
30                  OK            OK
33                  OK            OK
```

10.2 Two sample types - Mann-Whitney

When only two samples are of interest, these can also be compared using a Mann-Whitney test. As is the section above, the function is demonstrated using the control and long starvation samples from the example data. A subset of the `qPCRset` data is created to encompass only these samples, and a Mann-Whitney test is performed.

```
> qDE.mwtest <- mannwhitneyCtData(sr.norm[, 1:4], groups = files$Treatment[1:4],
+     calibrator = "Control")
> head(qDE.mwtest)
      genes feature.pos MW.value    p.value adj.p.value
26  Gene121       K2;L2    W = 0 0.02746864  0.07842053
134  Gene47     C24;D24   W = 16 0.02842954  0.07842053
157  Gene68     E21;F21    W = 0 0.02842954  0.07842053
185  Gene93     G22;H22    W = 0 0.02842954  0.07842053
36  Gene130     K11;L11   W = 16 0.02940105  0.07842053
54  Gene147       M4;N4    W = 0 0.02940105  0.07842053
          ddCt           FC meanCalibrator meanTarget
26    2.316925 2.006947e-01       38.88749   41.20441
134 -12.205480 4.722973e+03       40.50267   28.29719
157   6.547033 1.069416e-02       34.65738   41.20441
185  25.515391 2.084980e-08       15.68902   41.20441
36   -9.433460 6.914400e+02       39.74414   30.31068
54    3.765620 7.352505e-02       35.92123   39.68685
    categoryCalibrator categoryTarget
26        Undetermined   Undetermined
134       Undetermined            OK
157       Undetermined   Undetermined
185       Undetermined   Undetermined
36        Undetermined            OK
54        Undetermined   Undetermined
```

10.3 Multiple sample types - limma

In this example all three types of treatment are compared, as well as the control against both starvation samples combined. The data is sorted by feature names, to make easier use of replicated features.

```
> design <- model.matrix(~0 + files$Treatment)
> colnames(design) <- c("Control", "LongStarve", "Starve")
> print(design)
  Control LongStarve Starve
1       1          0      0
2       0          1      0
3       0          1      0
4       1          0      0
5       0          0      1
6       0          0      1
attr(,"assign")
[1] 1 1 1
attr(,"contrasts")
attr(,"contrasts")$`files$Treatment`
[1] "contr.treatment"
> contrasts <- makeContrasts(LongStarve - Control,
+     LongStarve - Starve, Starve - Control, (Starve +
+         LongStarve)/2 - Control, levels = design)
> colnames(contrasts) <- c("LS-C", "LS-S", "S-C", "bothS-C")
> print(contrasts)
            Contrasts
Levels       LS-C LS-S S-C bothS-C
  Control      -1    0  -1    -1.0
  LongStarve    1    1   0     0.5
  Starve        0   -1   1     0.5
> sr.norm2 <- sr.norm[order(featureNames(sr.norm)),
+     ]
> qDE.limma <- limmaCtData(sr.norm2, design = design,
+     contrasts = contrasts, ndups = 2, spacing = 1)
```

The result is a list with one component per comparison. Each component is similar to the result from using `ttestCtData`.

```
> class(qDE.limma)
[1] "list"
> names(qDE.limma)
[1] "LS-C"    "LS-S"    "S-C"     "bothS-C" "Summary"
> head(qDE.limma[["LS-C"]])
      genes feature.pos   t.test      p.value adj.p.value
14   Gene11     A12;B12 7.603024 1.999665e-05 0.003839357
50  Gene142     K23;L23 6.271023 9.883760e-05 0.009488409
3    Gene10     A10;B10 5.076932 5.031756e-04 0.024152428
119  Gene32       C9;D9 5.217172 4.113509e-04 0.024152428
23  Gene118     I23;J23 4.812232 7.417071e-04 0.028481551
133  Gene45     C22;D22 4.195002 1.905657e-03 0.060981013
         ddCt           FC meanTarget meanCalibrator
14   6.051791 0.0150740287   26.23181       20.18002
50  10.752992 0.0005794645   36.48566       25.73267
3    3.460492 0.0908423086   27.69156       24.23106
```

```
119  5.645620 0.0199755612   31.37832       25.73270
23   8.982157 0.0019774303   35.68746       26.70531
133  6.108469 0.0144933090   34.42031       28.31184
    categoryTarget categoryCalibrator
14              OK       Undetermined
50    Undetermined                 OK
3               OK       Undetermined
119             OK                 OK
23    Undetermined                 OK
133             OK                 OK
```

Furthermore, there is a "Summary" component at the end where each feature is denoted with -1, 0 or 1 to respectively indicate down-regulation, no change, or up-regulation in each of the comparisons.

```
> qDE.limma[["Summary"]][21:30, ]
         Contrasts
          LS-C LS-S S-C bothS-C
  Gene116    0    0   0       0
  Gene117    0    0   0       0
  Gene118    1    0   0       0
  Gene119    0    0   0       0
  Gene12     0    0   0       0
  Gene120    0    0   0       0
  Gene121    0    0   0       0
  Gene122    0    0   0       0
  Gene123    0    0   0       0
  Gene124    0    0   0       0
```

11 Displaying the results

The results can be visualised using the generic `plotCtOverview` shown in Figure 3. However, *HTqPCR* also contains more specialised functions, for example to include information about whether differences are significant or not.

11.1 Fold changes: Relative quantification

The relative Ct levels between two groups can be plotted with the function `plotCtRQ`. Below are two examples: one the result of `ttestCtData` where the top 15 genes are selected (figure 21), and another from the first comparison in `limmaCtData` where all genes below a certain p-value are depicted (Figure 22).

```
> plotCtRQ(qDE.ttest, genes = 1:15)
```

```
> plotCtRQ(qDE.limma, p.val = 0.085, transform = "log10",
+     col = "#9E0142")
```

The hatching on the bars indicates whether the target and calibrator Ct samples were unreliable, but this also depend on whether the parameter `stringent=TRUE` or `stringent=FALSE` when testing for differential expression. See the help functions for details (`?ttestCtData` and `?limmaCtData`).

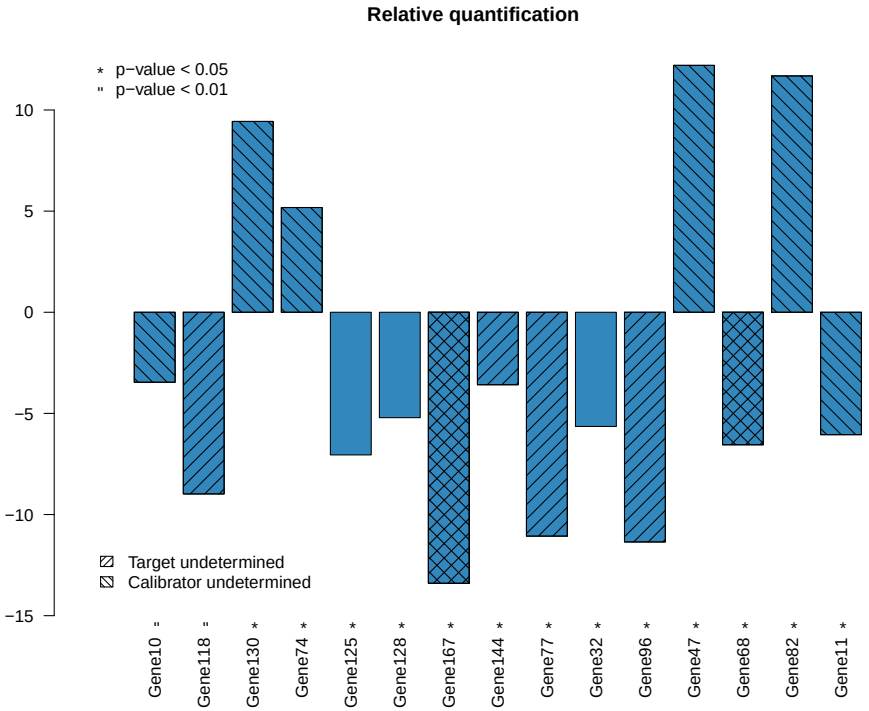


Figure 21: Relative quantification, using the top 15 features from ttestCtData.

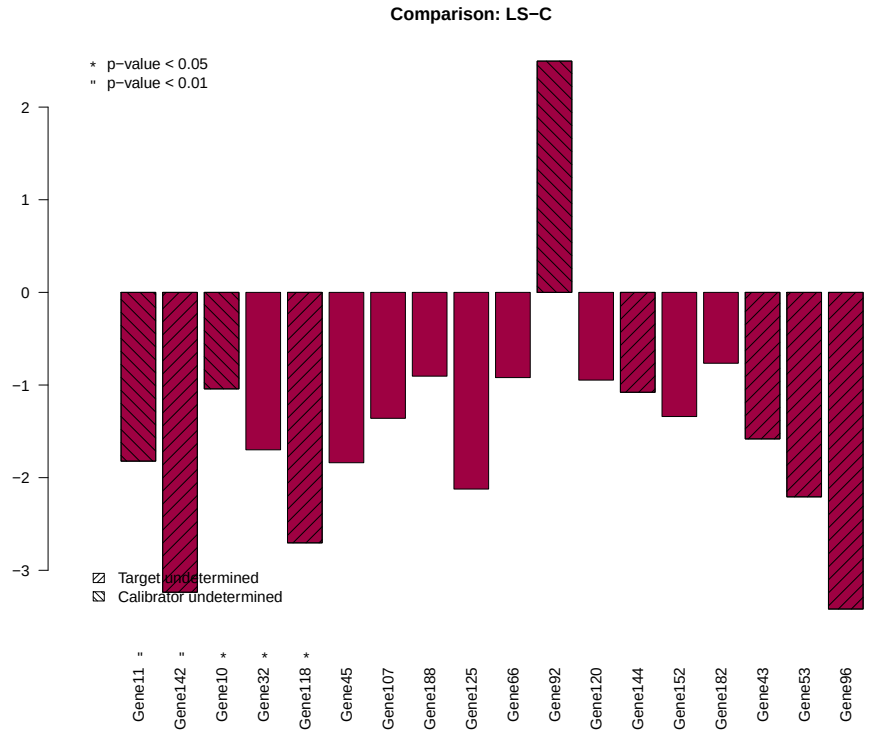


Figure 22: Relative quantification, using all features with p-value<0.085 from limmaCtData.

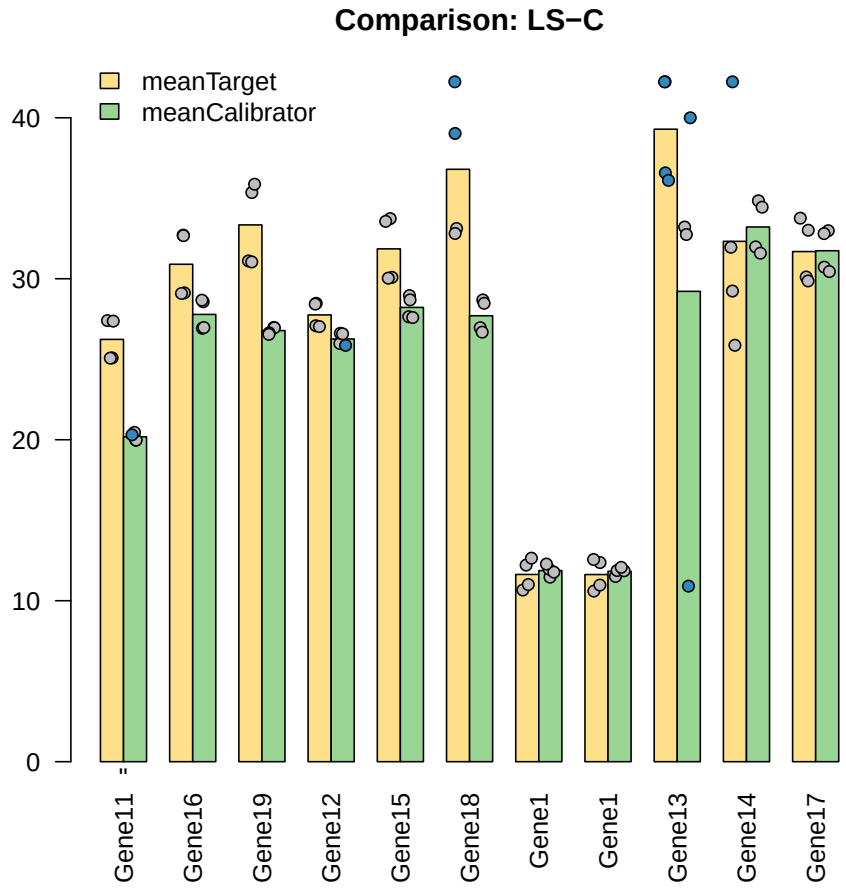


Figure 23: Ten genes from the data set, with the average Ct in two groups plotted along with the individual values. Points marked in blue are "Unreliable" or "Undetermined" whereas grey spots are "OK".

11.2 Fold changes: Detailed visualisation

In some cases it will be beneficial to more closely examine individual Ct data points from the fold changes, partly to look at the data dispersion, and partly to determine which of these values are in the "OK" versus "Unreliable"/"Undetermined" category. The function `plotCtSignificance` will take the result of `ttestCtData` or `limmaCtData`, along with the input data to these functions, and display a combined barplot showing the individual data points and marking those comparisons with a significant p-value.

```
> plotCtSignificance(qDE.limma, q = sr.norm, groups = files$Treatment,
+     target = "LongStarve", calibrator = "Control",
+     genes = featureNames(sr.norm)[11:20], un.col = "#3288BD",
+     jitter = 0.2)
```

11.3 Heatmap across comparisons

When multiple conditions are compared with `limmaCtData`, the fold changes from all comparisons can be compared to see if features cluster together in groups (Figure 24).

```
> heatmapSig(qDE.limma, dist = "euclidean")
```

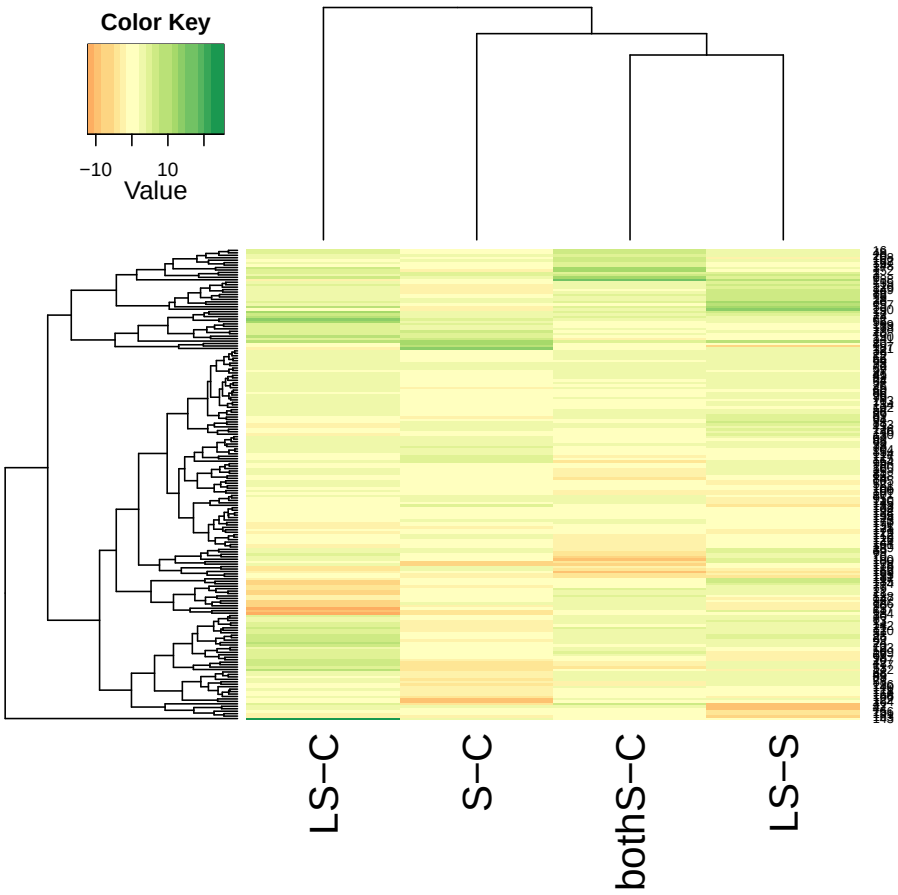


Figure 24: Fold changes across all comparisons, clustered based on Euclidean distance.

12 How to handle different input data

General information about how to read qPCR data into a *qPCRset* object is presented in section 3. Below, some more specific cases are illustrated.

12.1 Sequence Detection Systems format

The qPCR data might come from Sequence Detection Systems (SDS) software, in which case each file has a header containing some generic information about the initial Ct detection. This header varies in length depending on how many files were analysed simultaneously, and an example is shown below.

```
> path <- system.file("exData", package = "HTqPCR")
> cat(paste(readLines(file.path(path, "SDS_sample.txt"),
+     n = 19), "\n"))
SDS 2.3 RQ Results        1.2
 Filename        Testscreen analys all.sdm
 Assay Type        RQ Study
 EmbeddedFile        FileA
 Run DateTime        Fri May 15 17:10:28 BST 2009
 Operator
 ThermalCycleParams
 EmbeddedFile        FileB
 Run DateTime        Sat May 16 10:36:09 BST 2009
 Operator
 ThermalCycleParams
 EmbeddedFile        FileC
 Run DateTime        Sun May 17 13:21:05 BST 2009
 Operator
 ThermalCycleParams

 #          Plate        Pos        Flag        Sample        Detector        Task
 1        Control        A1        Passed        Sample01        Gene1        Endogenous
 2        Control        A2        Passed        Sample01        Gene2        Target
```

Only the first 7 columns are shown, since the file shown here contains >30 columns (of which many are empty). All columns for the first 20 lines can be seen in an R terminal with the command:

```
> readLines(file.path(path, "SDS_sample.txt"), n = 20)
```

For these files the parameter `SDS=TRUE` can be set in `readCtData`. The first 100 lines of each file will be scanned, and all lines preceding the actual data will be skipped (in this case 17), even when the length of the header varies between files.

12.2 Non-Ct data

Some qPCR systems, such as the LightCycler from Roche, don't provide the results as Ct values, but instead as crossing points (Cp). Ct values are measured at the exponential phase of amplification by drawing a line parallel to the x-axis of the real-time fluorescence intensity curve (fit point method), whereas Cp (second derivative method) calculates the fractional cycle where the second derivative of the real-time fluorescence intensity curve reaches the maximum value (Luu-The et al., 2005).

As long as all samples and features are quantified using the same method, it shouldn't matter whether Ct or Cp values are being used to test for significant differences between samples. The analysis of e.g. LightCycler Cp data can therefore proceed as outlined in this vignette. One thing to bear in mind though, is that for Ct values a value of 40 generally means NA, and above 35 is considered unreliable, however these numbers will be different for Cp. When using the filtering methods to set results as being "OK", "Unreliable" and "Undetermined", the parameters in e.g. `setCategory(..., Ct.max = 35, Ct.min = 10, ...)` and `readCtData(..., na.value, ...)` will need to be adjusted accordingly.

13 Manipulating qPCRset objects

Depending on the design of the qPCR card and how the data is to be analysed, it will sometimes be necessary to manipulate the `qPCRset` objects in different ways, such as by combining several objects or altering the layout of features x samples.

13.1 Multiple samples present on each plate

The result from each qPCR run of a given card typically gets presented together, such as in a file with 384 lines, one per feature, for 384 well plates. However, some cards may contain multiple samples, such as commercial cards that are designed to be loaded with two separate samples and then include 192 individual features.
Per default, `readCtData` reads each card into a *qPCRset* object as consisting of a single sample, and hence one column in the Ct data matrix. When this is not the case, the data can subsequently be split into the correct features x samples (rows x columns) dimensions using the function `changeCtLayout`. The parameter `sample.order` is a vector, that for each feature in the *qPCRset* indicates what sample it actually belongs to, and reformats all the information in the *qPCRset* (`exprs`, `featureCategory`, `flag` etc.) accordingly. The actual biological samples are likely to differ on each card, so `sample.order` merely indicates the *location* of different samples among the features present in the input data.

```
> sample2.order <- rep(c("subSampleA", "subSampleB"),
+     each = 192)
> sample4.order <- rep(c("subA", "subB", "subC", "subD"),
+     each = 96)
> qPCRnew2 <- changeCtLayout(sr.norm, sample.order = sample2.order)
> show(qPCRnew2)
An object of class "qPCRset"
Size:  192 features, 12 samples
Feature types:                 Endogenous Control, Target
Feature names:                 Gene1 Gene2 Gene3 ...
Feature classes:         Kinase, Marker, TF
Feature categories:         OK, Unreliable, Undetermined
Sample names:                 subSampleA:sample1 subSampleA:sample2 subSampleA:sample3 .
> qPCRnew4 <- changeCtLayout(sr.norm, sample.order = sample4.order)
> show(qPCRnew4)
An object of class "qPCRset"
Size:  96 features, 24 samples
Feature types:                 Endogenous Control, Target
Feature names:                 Gene1 Gene2 Gene3 ...
Feature classes:         Kinase, Marker, TF
```

```
Feature categories:         OK, Unreliable, Undetermined
Sample names:                 subA:sample1 subA:sample2 subA:sample3 ...
```

As with the other functions that manipulate the Ct data, the operation is stored in the `history` slot of the *qPCRset* for future reference.

```
> getCtHistory(qPCRnew4)
                                                              history
1                     readCtData(files = files$File, path = path)
2 setCategory(q = raw, quantile = 0.8, groups = files$Treatment)
3              normalizeCtData(q = raw.cat, norm = "scale.rank")
4      changeCtLayout(q = sr.norm, sample.order = sample4.order)
```

13.2 Combining multiple `qPCRset` objects

In some cases it might be desirable to merge multiple qPCRset objects, that have been read into R or processed individually. The *HTqPCR* package contains two functions for combining multiple `qPCRset` objects into one, by either adding columns (samples) or rows (features). This can be done for either identical samples across multiple different cards (such as a 384 well plate), or if more samples have been run on cards with the same layout.

`cbind` combines data assuming that all experiments have been carried out on identical cards, i.e. that `featureNames, featureType, featurePos` and `featureClass` is identical across all the qPCRset objects. The number of features in each object much be identical, but number of samples can vary.

`rbind` combines data assuming that the same samples have been analysed using different qPCR cards. The number of samples in each object must be identical, but the number of features can vary.

Both these functions should be used with some care; consider e.g. whether to normalize before or after joining the samples, and what method to use.
In the examples here objects with different normalisation are combined, although in a real study the `qPCRset` objects would typically contain different data.

```
> q.comb <- cbind(q.norm[, 1:3], sr.norm[, 4], nr.norm[,
+     c(1, 5, 6)])
> q.comb
An object of class "qPCRset"
Size:  384 features, 7 samples
Feature types:                 Endogenous Control, Target
Feature names:                 Gene1 Gene2 Gene3 ...
Feature classes:         Kinase, Marker, TF
Feature categories:         OK, Unreliable, Undetermined
Sample names:                sample1 sample2 sample3 ...
> q.comb2 <- rbind(q.norm, sr.norm[1:4, ], nr.norm)
> q.comb2
An object of class "qPCRset"
Size:  772 features, 6 samples
Feature types:                 Endogenous Control, Target
Feature names:                 Gene1 Gene2 Gene3 ...
```

```
Feature classes:          Kinase, Marker, TF
Feature categories:          OK, Unreliable, Undetermined
Sample names:                  sample1 sample2 sample3 ...
```

As with other functions where the `qPCRset` object is being manipulated, the information is stored in the `history` slot.

```
> getCtHistory(q.comb)
                                                                        history
1                       qPCRset1: readCtData(files = files$File, path = path)
2  qPCRset1: setCategory(q = raw, quantile = 0.8, groups = files$Treatment)
3                  qPCRset1: normalizeCtData(q = raw.cat, norm = "quantile")
4                       qPCRset2: readCtData(files = files$File, path = path)
5  qPCRset2: setCategory(q = raw, quantile = 0.8, groups = files$Treatment)
6                qPCRset2: normalizeCtData(q = raw.cat, norm = "scale.rank")
7                       qPCRset3: readCtData(files = files$File, path = path)
8  qPCRset3: setCategory(q = raw, quantile = 0.8, groups = files$Treatment)
9                 qPCRset3: normalizeCtData(q = raw.cat, norm = "norm.rank")
10                                      cbind(deparse.level, ..1, ..2, ..3)
```

14 Concluding remarks

Use the function `news` to see what bug fixes and updates have been incorporated into *HTqPCR*. For example:

```
> news(Version > 1.2, package = "HTqPCR")
Changes in version 1.3:

SIGNIFICANT USER-VISIBLE CHANGES

    o

NEW FEATURES

    o

BUG FIXES

    o   Fixed bug in plotCtOverview affecting sample names starting with "M".

    o   Corrected rbind so all data frames contain characters, not factors.

    o   Changed a "stop" to a "warning" in readCtData.

    o   Corrected qPCRset subsetting when columns are selected multiple times.

    o   Added dupcor parameter to limmaCtData.
```

This vignette was generated using:

- R version 2.12.0 RC (2010-10-11 r53293), `i386-pc-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Biobase 2.10.0, HTqPCR 1.4.0, RColorBrewer 1.0-2, limma 3.6.0, statmod 1.4.6
- Loaded via a namespace (and not attached): affy 1.28.0, affyio 1.18.0, gdata 2.8.0, gplots 2.8.0, gtools 2.6.2, preprocessCore 1.12.0, tools 2.12.0

References

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. 1

V. Luu-The, N. Paquet, E. Calvo, and J. Cumps. Improved real-time rt-pcr method for high-throughput measurements using second derivative calculation and double correction. *Biotechniques*, 38(2):287–93, 2005. 34

G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, and W. H. R. Irizarry, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005. 27