ScreenSeq Pipeline

Version:

- feature/simpleScripts (2020-12-23)
- Code: (https://github.com/soccin/ScreenSEQ/tree/feature/simpleScripts)

Description of results:

Counts and QC-Stats

- `<ProjectNo>____STATS.xlsx` — Overall QC stats for run; table columns are:

Column	Description
Sample	SampleId
Total	Total number of reads
Num.Processed	Number of reads that had a valid sgRNA sequence
Num.Library	Number of reads found in sgRNA library
PCT.Useable	Num.Library / Total

`PCT.Useable` gives a measure of the quality of the library. If it is low or if there is one or more samples whose values are much lower than the others this may indicate some QC issues.

- `<ProjectNo>____COUNTS.xlsx` — Raw count file.

Raw (unnormalized) counts for each sample

Column	Description
sgRNA	Sequence of sgRNA
Gene	Gene targeted
ProbeID	Unique Probe Id
LibName	Library Name
Samp1	Sample 1 counts
…	…
SampN	Sample N counts

Differential Analysis

For each statistical comparison there are two output files

- `<ProjectNo>_DiffAnalysis_<SetName>_.pdf`
- `<ProjectNo>_DiffAnalysis_<SetName>_.xlsx`

Where `<ProjectNo>` is the project number of the dataset and `<SetName>` is the name of the specific dataset for the comparison. The PDF file has 4 plots QC plots:

- Boxplot of the normalized log2 transformed data. This plots shows the distribution of datapoints for each sample using the standard `boxplot` function from `R`. The purpose of this plot is to look for potential outlier samples within a group of replicates and to look for bias/batch effects.
- Multidimensional scaling (MDS) plot, `plotMDS` from Biocoductors **edgeR** package. This plot projects the data down to 2 dimensions and is a form of clustering. Again the idea is to check of outlier samples and to see that the different samples groups are well separted.

Problems in either of these two plots could indicate potenital problems in any significance testing.

The differential analysis is done with **edgeR** in two ways; probe level significan analysis and if there are multiple probes for genes then also a gene level signifance analysis. The third plot show the:

- Scatter plot of log average intensity (normalized counts) vs the log fold change. Each dot is a probe. Significant probes are in red.

The table of significant probes is in the first sheet of the excel file named *ProbeLevel*. In that file are the following columns:

Column	Description
ProbeID	ID tag for each prob
SEQ	Sequence of Probe
LIB	(i)RNA library
FC	Fold Change in natural units (FC = Grp1/Grp2)
logFC	log (base 2) fold change
PValue	raw p-value
FDR	multiple test corrected p-value (FDR)
avgAll	average of counts of all samples
avg.`<Grp1>`	average of counts in group 1
avg.`<Grp2>`	averate of counts in group 2

The final plot shows a volcano plot of the gene level signifigance analysis. The x-axis is the logFC and the y-axis is the p-value (on log scale). Each point is a gene and the red points are significant. The significance test used for the gene level analysis is from the **edgeR** package and is the `camera` function, which combines the multiple probe data for a given gene usign a gene set analysis type method. The second sheet of the excel file has the significance table for the gene level analysis. It has the following columns.

Column	Description
NGenes	Number of probes for this gene
Direction	Aggregate direction of change
PValue	Raw p-value
FDR	correctedf p-value
logFC	log (base 2) folde change