SmartSeq2 quality control report

The data in this document is generated for plate all in experiment R0759-S0001_A64199.

Additional information regarding quality control can be found in the same folder as this report:

- QC metrics and outlier information per sample
- ERCC counts for each of the ERCCs per sample
- alignment percentages per sample
- plots used in this document

The kallisto quantifications for each of the samples and merged expression matrices (estimated counts by default) are in the neighbouring directory.

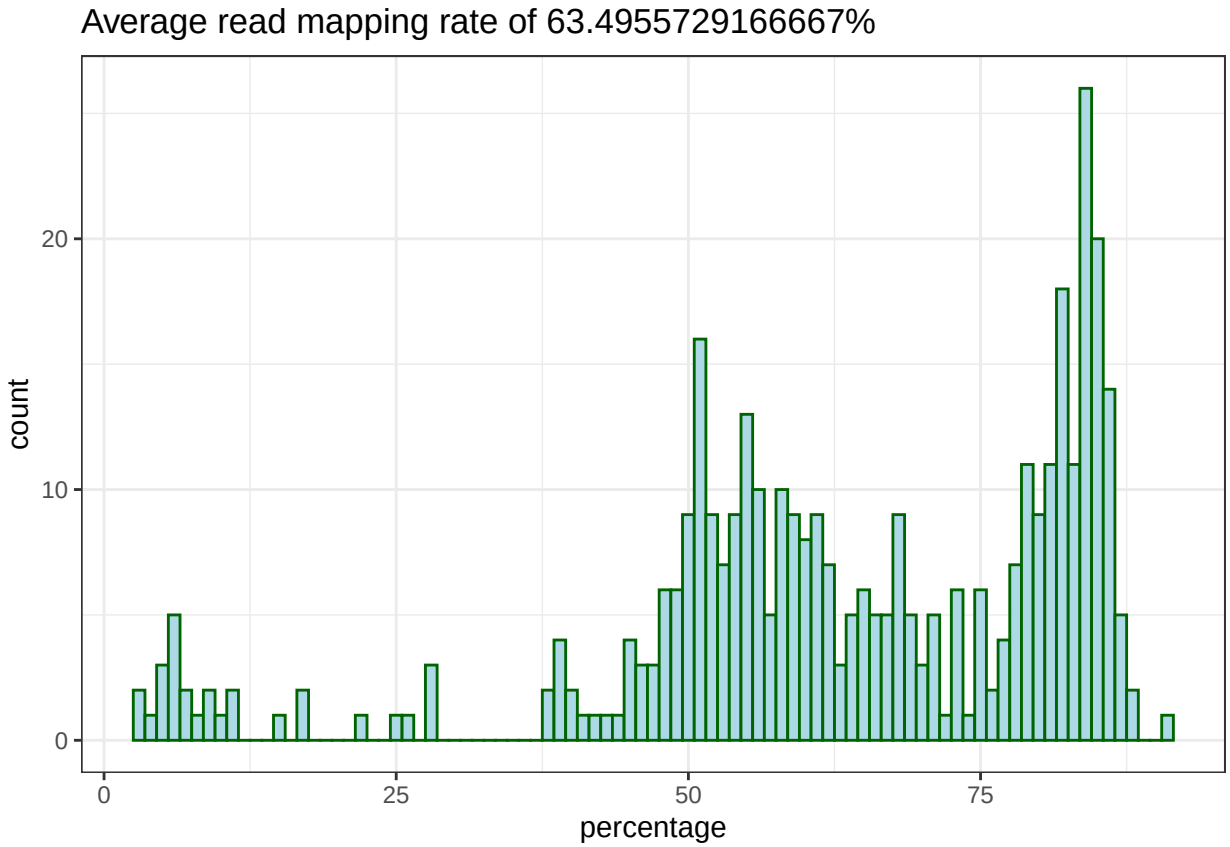
Tools used in the quality control pipeline are:

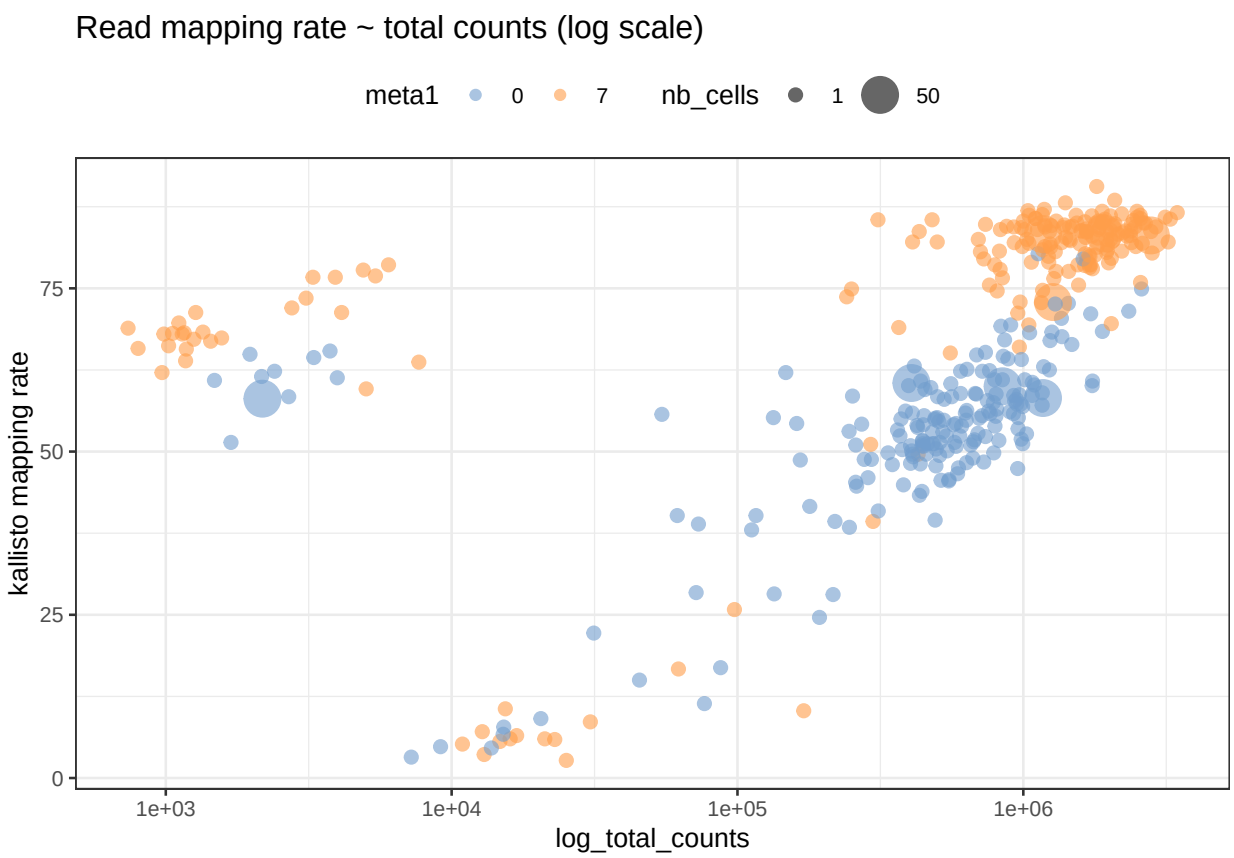
- kallisto 0.45.1
- R-4.1.2
- R markdown
- R packages: scater, ggplot2, dplyr, knitr, rjson
- nextflow 19.04.1
- singularity 2.4.2

Read mapping

Reads were mapped with the kallisto pseudo-aligner (version 0.45.1) to the reference genome with added ERCC RNA spike-in mix 1, if applicable.

Percentage of reads that map is shown for all the samples. Samples labelled as control are separated for comparison.

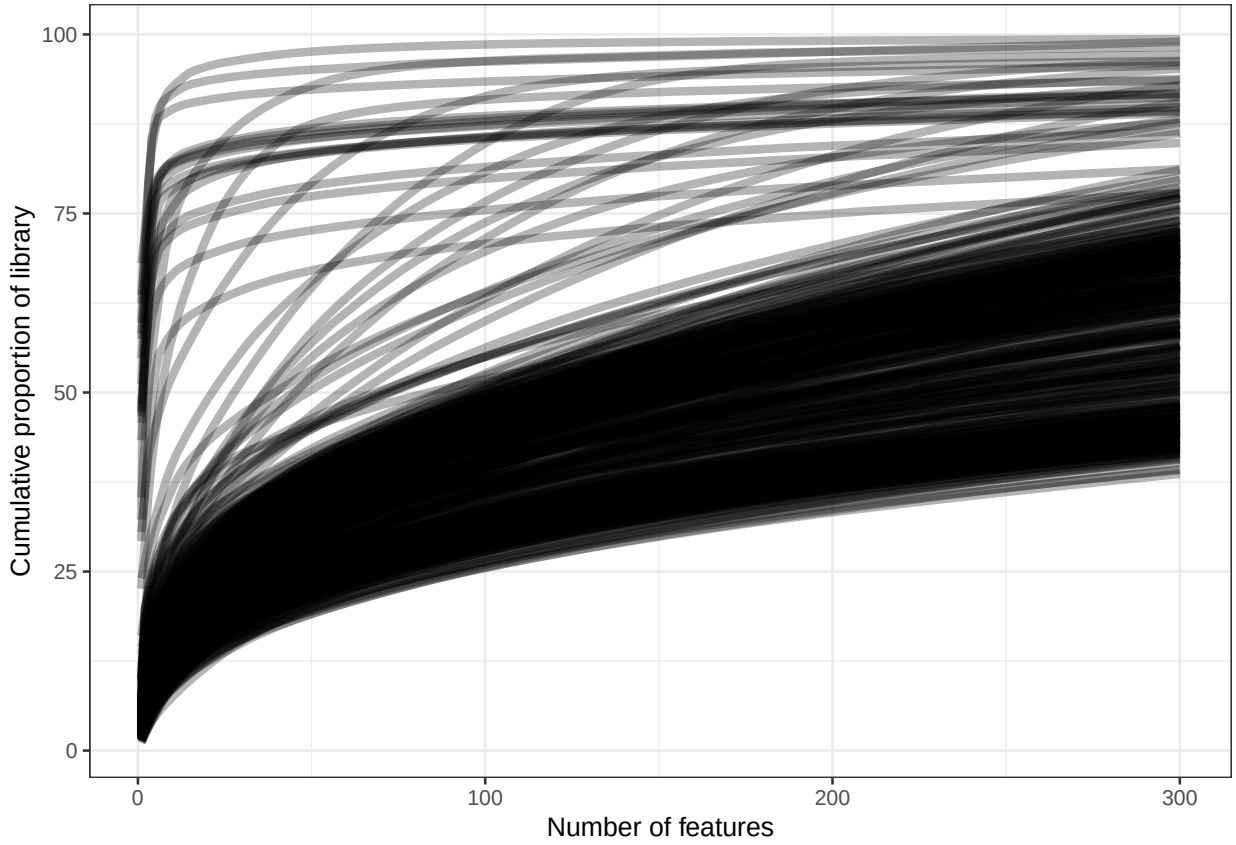

Average read mapping rate of 63.4955729166667%
count
20
10
0
0
25
50
75
percentage


Read mapping rate ~ total counts (log scale)
meta1
0
7
nb_cells
1
50
kallisto mapping rate
75
50
25
0
1e+03
1e+04
1e+05
1e+06
log_total_counts

Cumulative distribution of expression

The following plot shows, for the transcripts with the highest expression levels (top 1-300) for the sample, proportion of the sample's library (y-axis) covered by which number of those transcripts (x-axis).

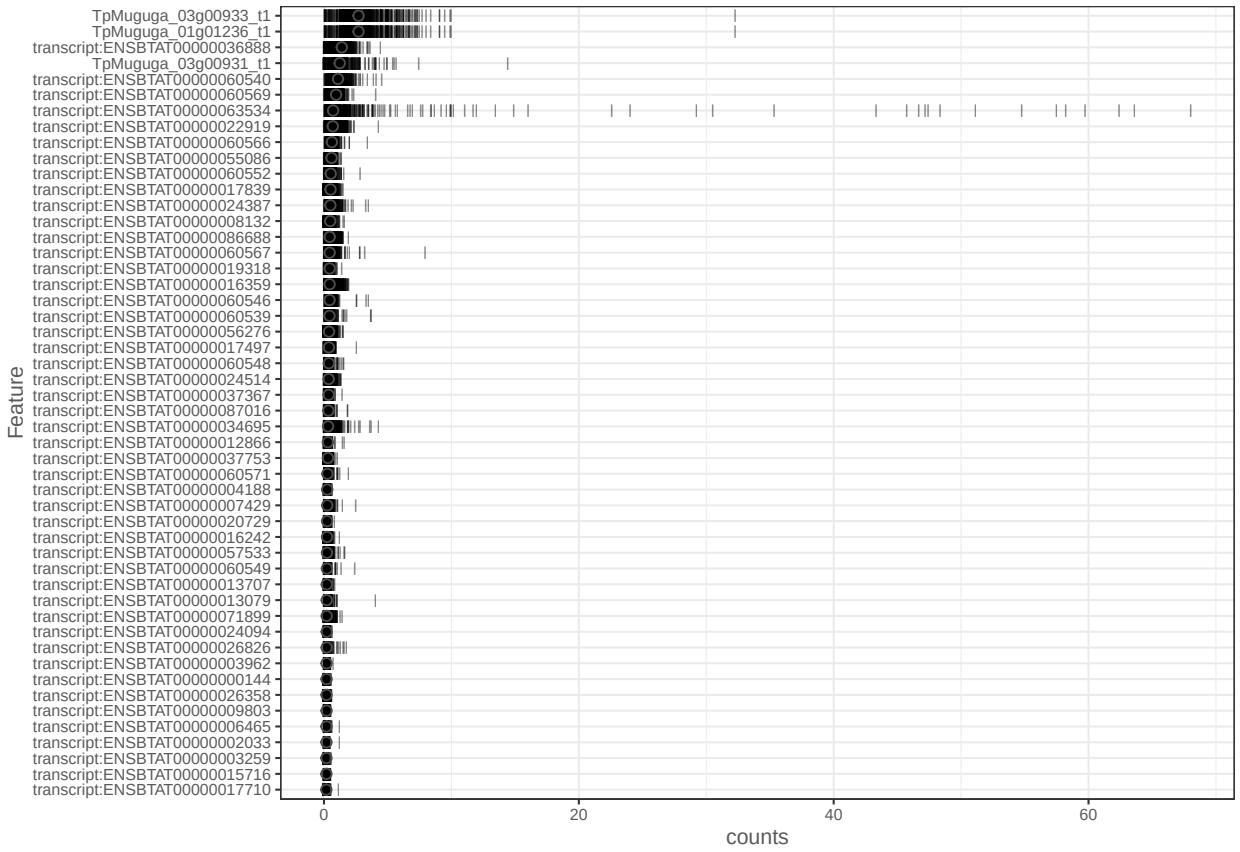
Distributions which rise high quickly are samples which are dominated by low number of transcripts while the ones having a less steep rise have counts more evenly distributed. Wells with less diverse count distribution are more similar to empty wells and might indicate lower quality or damaged cells.



Most highly expressed transcripts

Shown are top 50 features by the proportion of counts they take in all samples. Feature names (as Ensembl transcripts or annotated control features) denote samples on the y-axis with the percentage of counts they capture on the x-axis. Circle shows the proportion across all samples with proportion for each of samples shown on the same line.

If applicable, spike-ins (ERCC) and mitochondrial transcripts are labelled. Their high percentages could be indicative of poor samples although some samples have a naturally higher percentage.

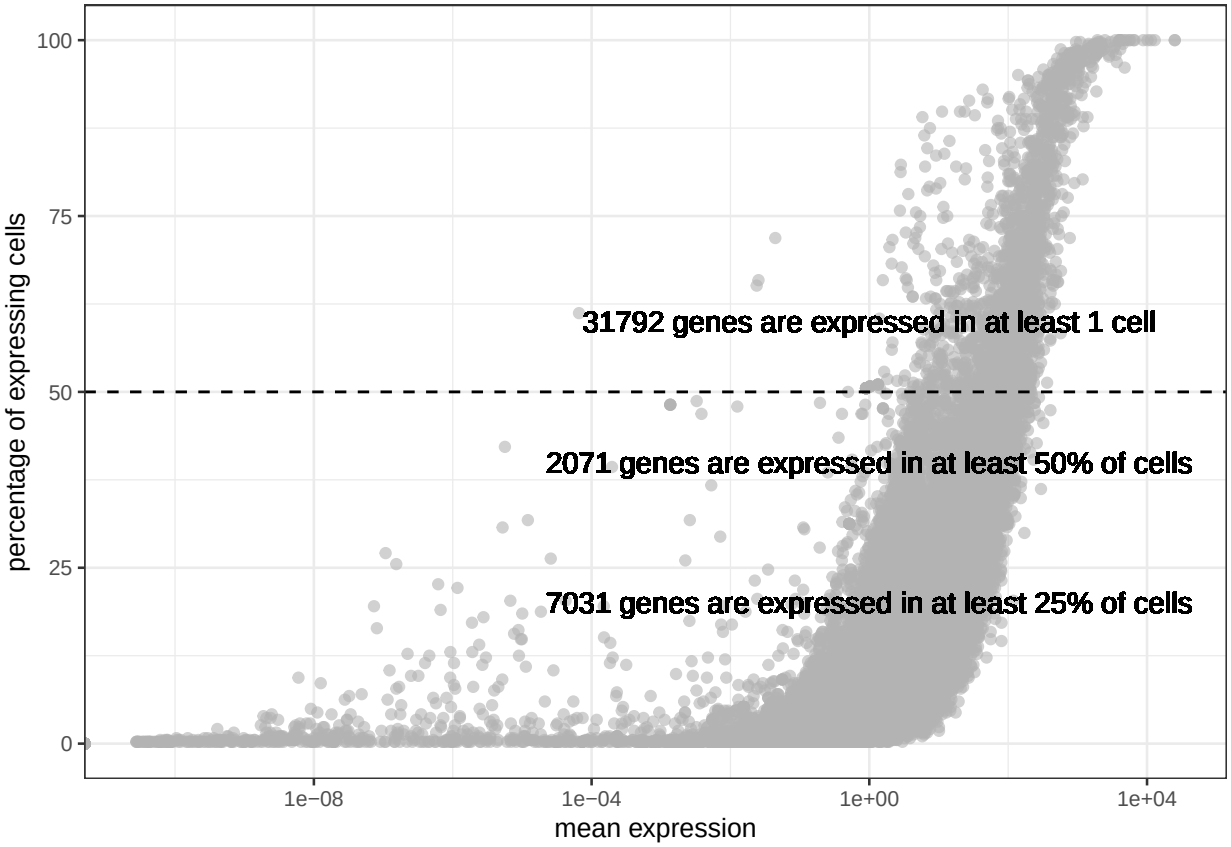


Expression frequency - mean distribution of transcripts

For all of the transcripts (features), their frequency of expression (percentage of samples expressing the specific transcript) is shown against their mean value of expression

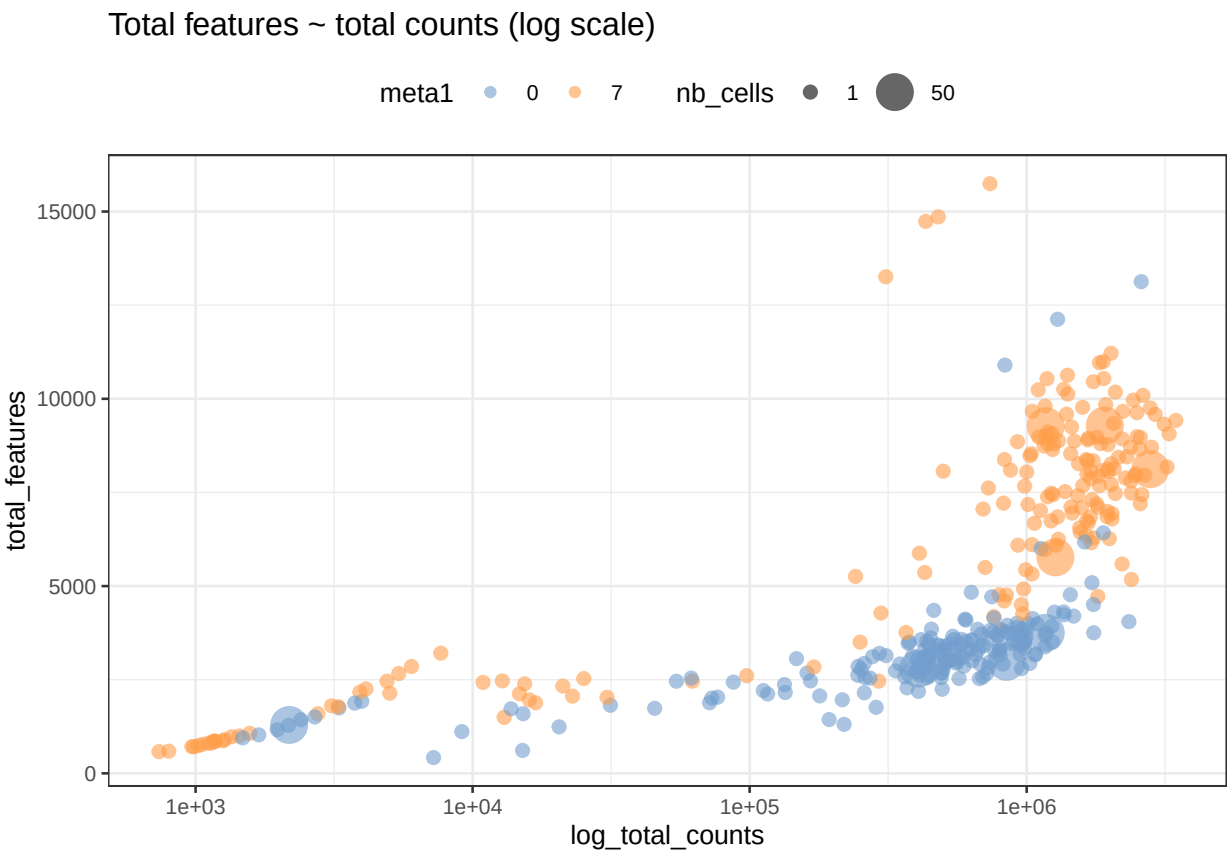
The relationship between the two variables is typically sigmoidal looking.

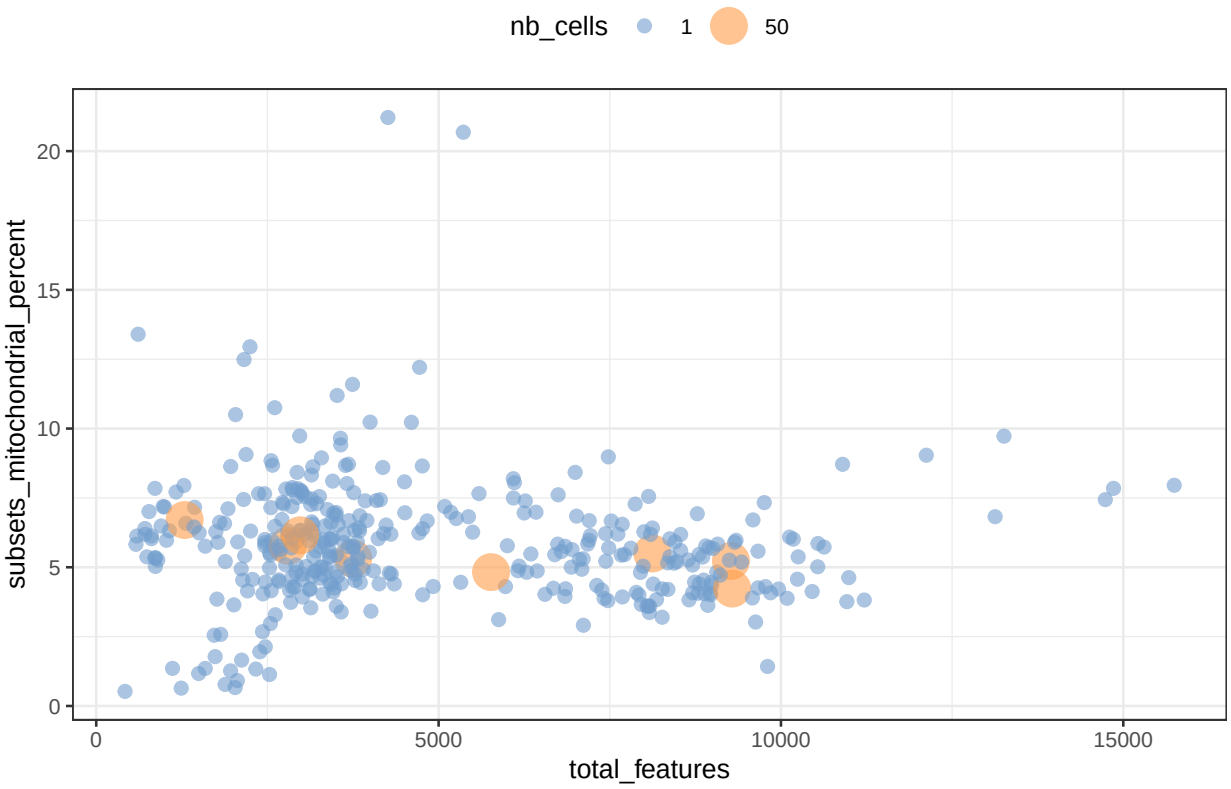
The vertical dashed line is the median of expression levels across the samples. The horizontal dashed line is at 50% of expression presence - dots above are transcripts which are expressed in more than 50% of the samples.

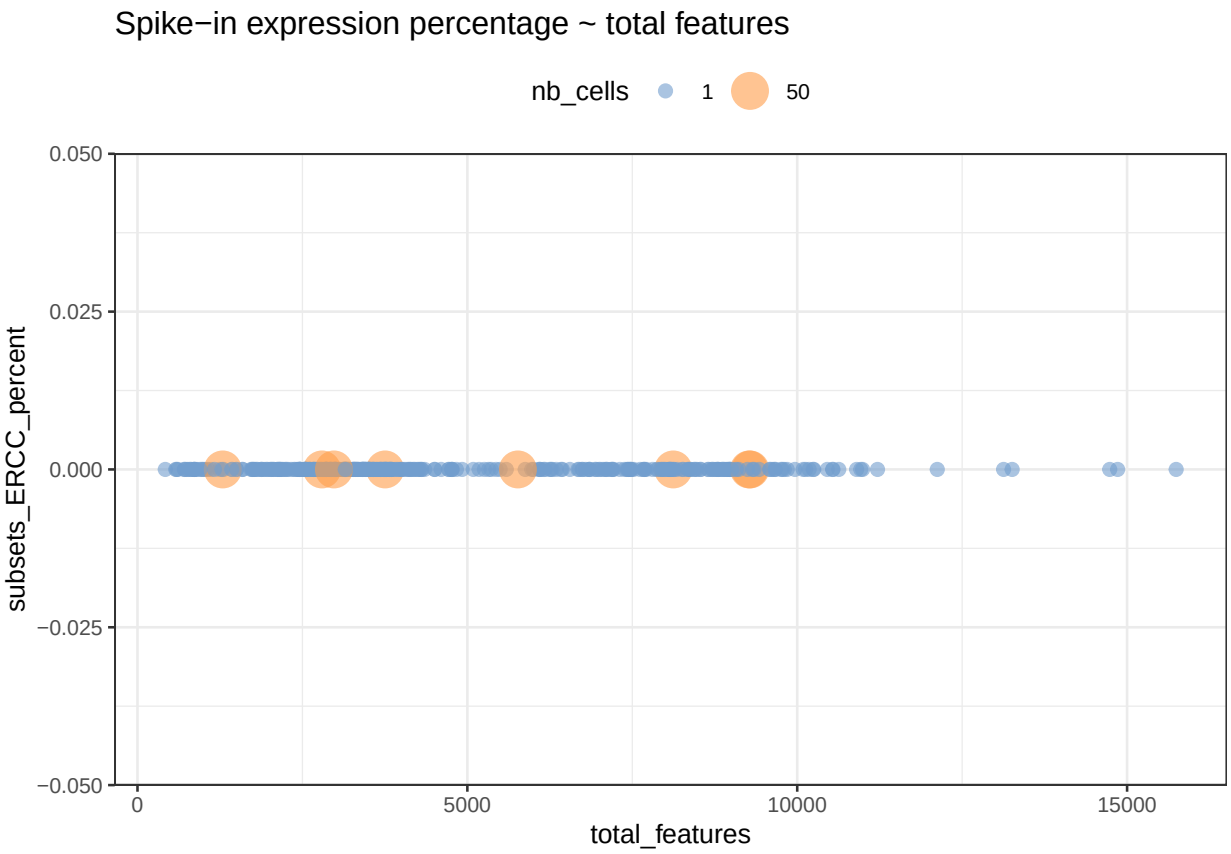


Scatterplots

Following scatterplots show, for all of the samples and their total counts, their transcript number, proportion of mitochondrial transcript and proportion of ERCC transcirpt.




Out of 48113 transcripts, Mitochondrial expression percentage ~ total features
nb_cells
1
50
subsets_mitochondrial_percent
0
5
10
15
20
0
5000
10000
15000
total_features


Spike−in expression percentage ~ total features
nb_cells
1
50
0.050
0.025
0.000
−0.025
−0.050
subsets_ERCC_percent
0
5000
10000
15000
total_features

Dimensionality reductions

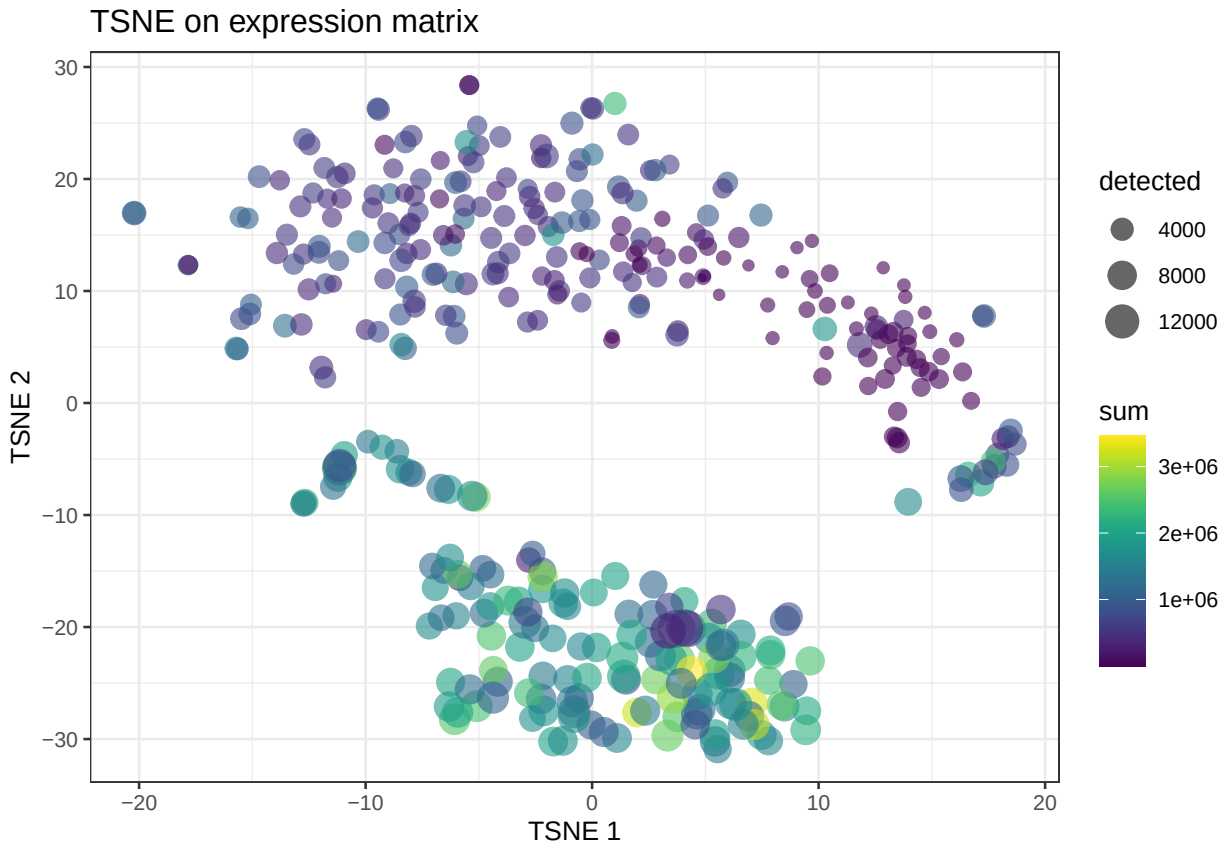
Dimensionality reductions summarise the large feature space with a smaller number of dimensions aiming to capture most of the variance in the data.

Principal component analysis (PCA) and t-SNE are one of the most common methods. PCA makes new dimensions by trying to capture as much as variance as possible in the top ones, while t-SNE tries to place the samples in 2 dimensions in a way which best captures the similarities and differences of the samples.

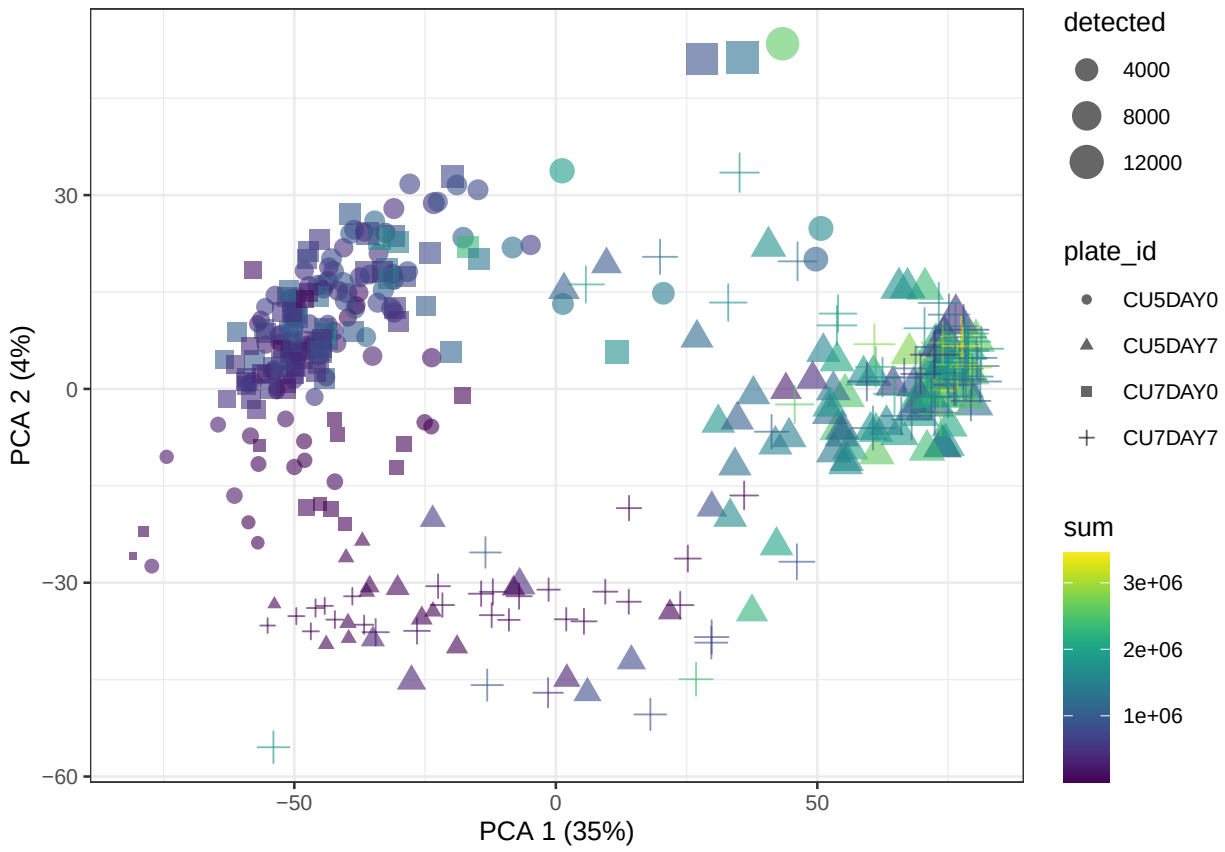
PCA/t-SNE dimensions are derived from the expression profiles of all the transcripts and visualised in the following scatterplots.

Samples generally similar to each other will be closer.

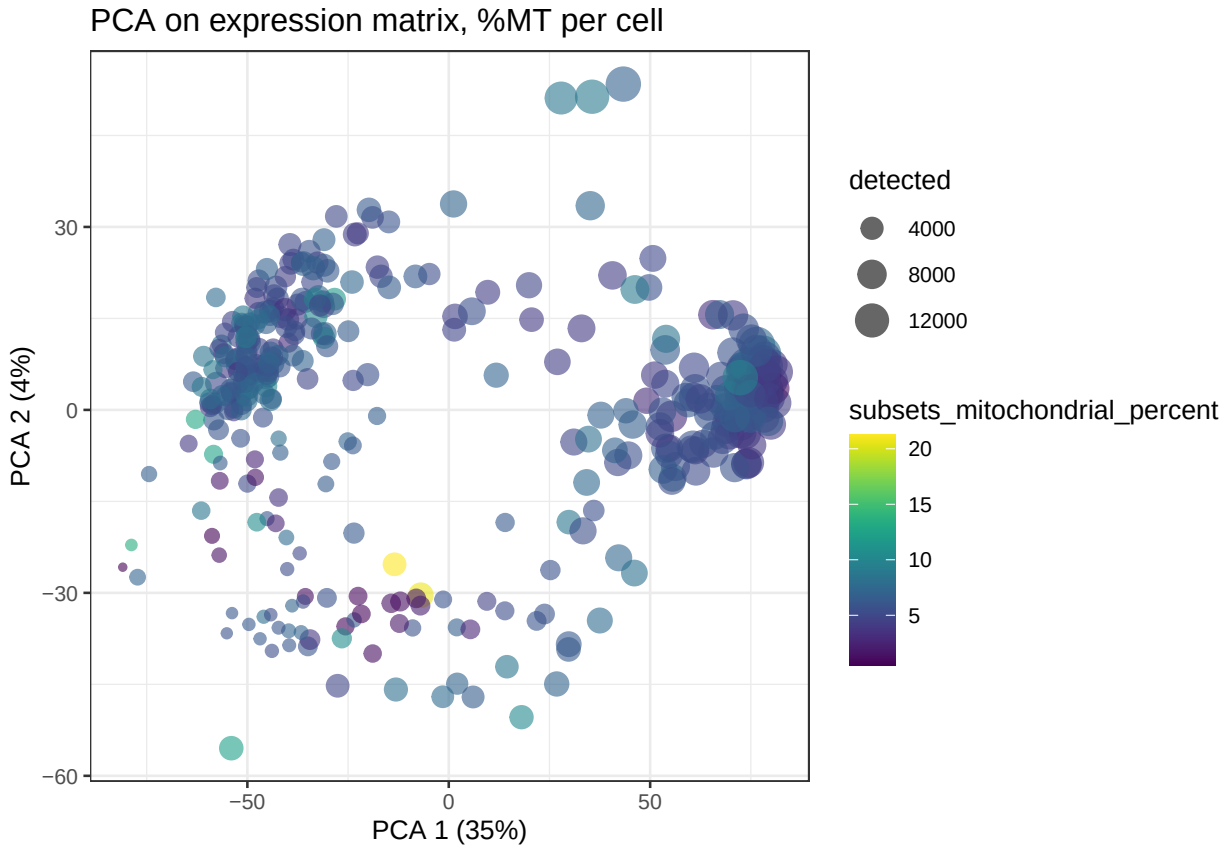
t-SNE

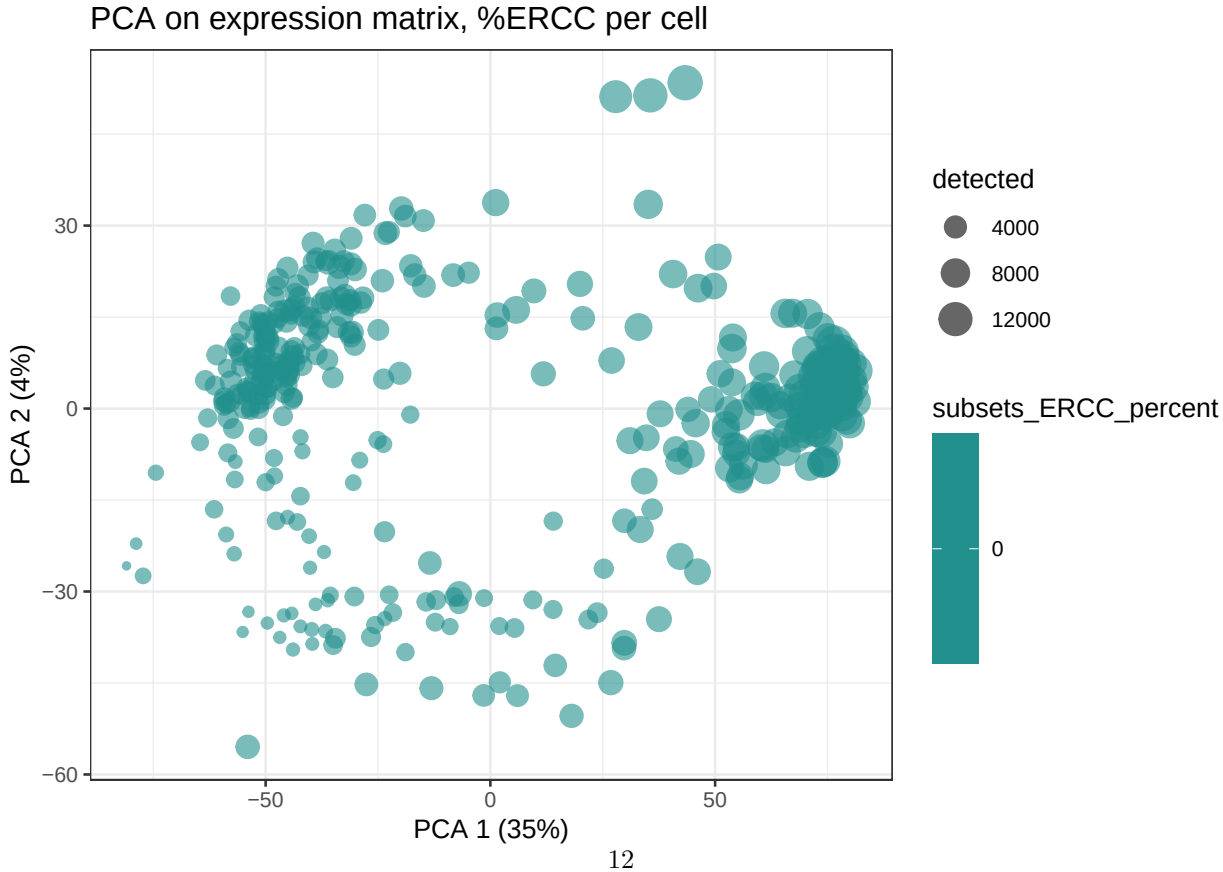


PCA


detected
4000
8000
12000
plate_id
CU5DAY0
CU5DAY7
CU7DAY0
CU7DAY7
sum
3e+06
2e+06
1e+06
PCA 2 (4%)
30
0
−30
−60
−50
0
50
PCA 1 (35%)

PCA on MT/ERCC





Outliers and key metrics per sample

Accompanying spreadsheet (`Per_sample_key_metrics.tsv`) contains the value for some of the most important metrics, position on the plate, control information and whether or not the sample is considered an outlier in the population when it comes to mapping rate, and percentage of counts which are mitochondrial.

If a sample's average read mapping rate is below 50%, or the mitochondrial percentage is over 10%, the sample is consider as an outlier.

Being an outlier on one metric is unlikely to be of much concern, but if the same sample is an outlier across all 2 it could be a sign of an unviable sample.

Following samples have been marked as **outliers**.

For 2 outlier metrics:

CU5DAY7E4, CU5DAY0F7, CU7DAY7E3, CU7DAY0H12, CU7DAY0H3

Exact information for the metrics used in the generation of this document can be found in the accompanying table - `Per_sample_key_metrics.tsv`

By default, it contains the following information:

- sample_ID
- plate position
- outlier hits - the number of times that sample has been labeled as an outlier
- outlier_pct_MT - TRUE if a sample meets the outlier metric on mitochondrial percentage
- outlier_mapping reate - TRUE if a sample meets the outlier metric on average read mapping rate
- Outlier status for the number of counts, features and proportion of mitochondrial/spike-in counts
- sum - total counts
- detected - total expressed features (transcripts)
- subsets_mitochondrial_percent - Percentage of counts belonging to mitochondrial transcripts
- subsets_ERCC_percent - Percentage of counts belonging to spike-ins
- map_rate - average read mapping rate
- is_cell_control