Data Quality Control – Outlier detection

Aim of lecture: The importance of quality control to make sure your data is of high quality

Kjell Petersen
Introduction to High Throughput Genomics
December 2011
The HT genomics pipeline

1. Biological question
2. Experimental design
3. Microarray experiment
4. NextGenSeq experiment
5. Quality control
6. Preprocessing
7. Statistical analysis
8. Biological verification and interpretation

Decision point:
- Yes: Proceed to next step
- No: Revisit previous steps
Outlier detection

- As long as the majority of the data are assumed to be acceptable, outlier detection methods can be used for quality control
- Methods
  - Multivariate methods (PCA, CA)
  - Clustering (Hierarchical clustering)
Batch effects

• Technical variance added to batches of samples
• Randomisation and balancing with respect to biology of interest in the experiment is important
  – The impact of the technical variance is reduced
  – Possible to separate technical variation from biological variation
  – May be able to remove technical variation
Batch effects

• Identification of possible batch effects
  – Group samples according to known batches and give each group a unique colour

• Methods
  – Multivariate methods (PCA, CA)
  – Clustering (Hierarchical clustering)
  – Box plots/distribution plots
Why use CA-plot for outlier detection?

Samples swapped?

Outlier?
How to check for batch effects

Dataset 1: Grouped according to biology
Dataset 1: Grouped according to low RIN values and scanner errors
Dataset 2: Grouped according to biology
Dataset 2: Grouped according to labelling dates
How to spot an outlier and when to remove it

Dataset 3: Grouped according to biology
Dataset 3: Grouped according to biology
Outliers- can we remove them?

Dataset 3: Grouped according to biology
Boxplot: to look at the distribution of the data

Un-normalized

Quantile normalized

Dataset 3

Look at un-normalized data first
QC – outlier detection summarized

• Always check for batch effects
  – Keep track of all batches and problems with samples
• Plot both un-normalized and normalized data
  – CA plot
  – Hierarchical clustering
• If outliers
  – Remove if you have a good reason to remove it (technical failures etc.)
  – Keep if you can't find any particular reason for that sample to be an outlier
QC – in R

From the R-lumi package