UNIVERSITY OF CAMBRIDGE

MPhil in Computational Biology 2016/2017
Course: **Functional Genomics**

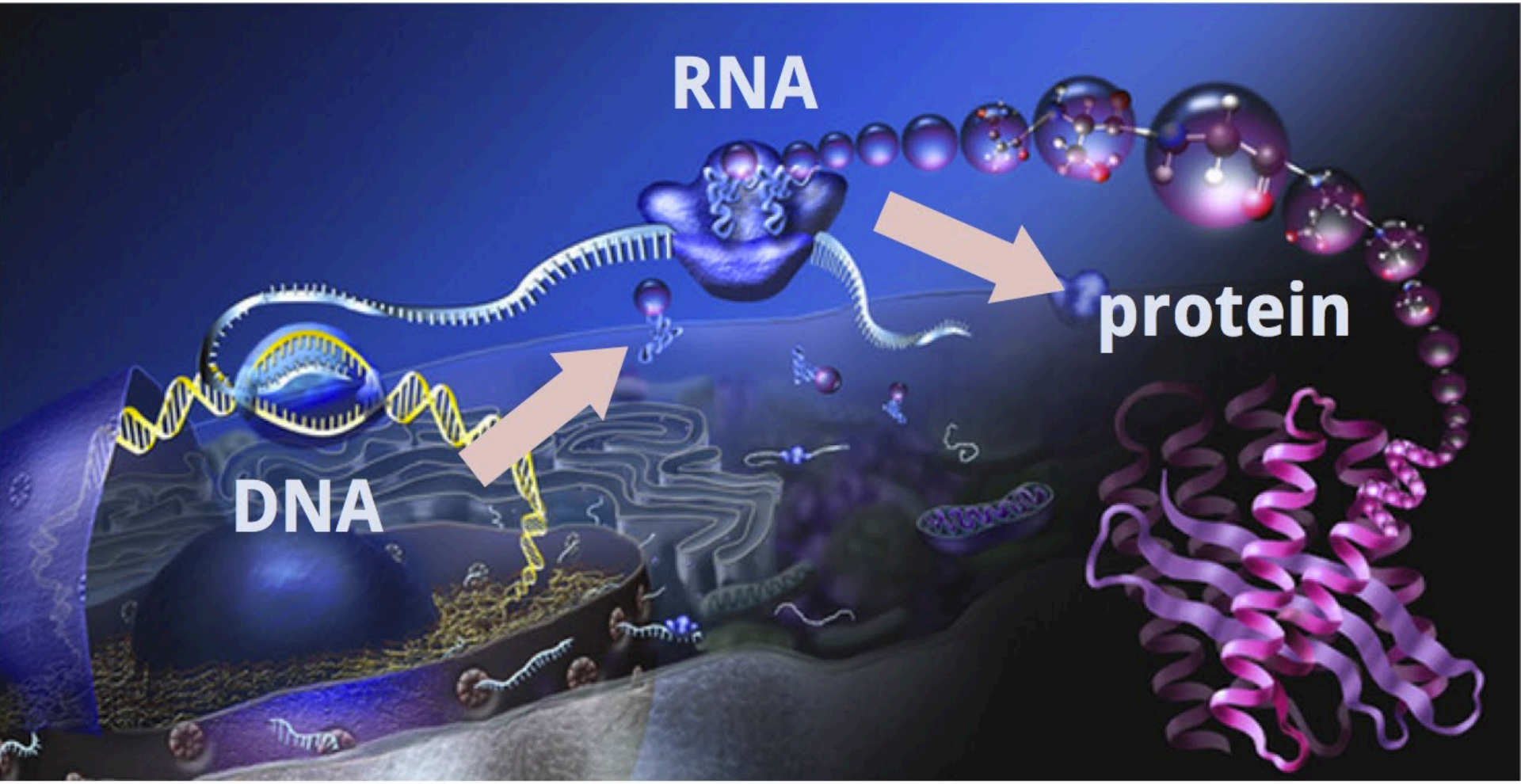
# Analysis of microarray data

*Maurizio Callari*

maurizio.callari@cruk.cam.ac.uk

Contributions by ***Mark Dunning, Shamith Samarajiwa, Benilton Carvalho, Oscar Rueda, Roslin Russell.***

# The central dogma of biology



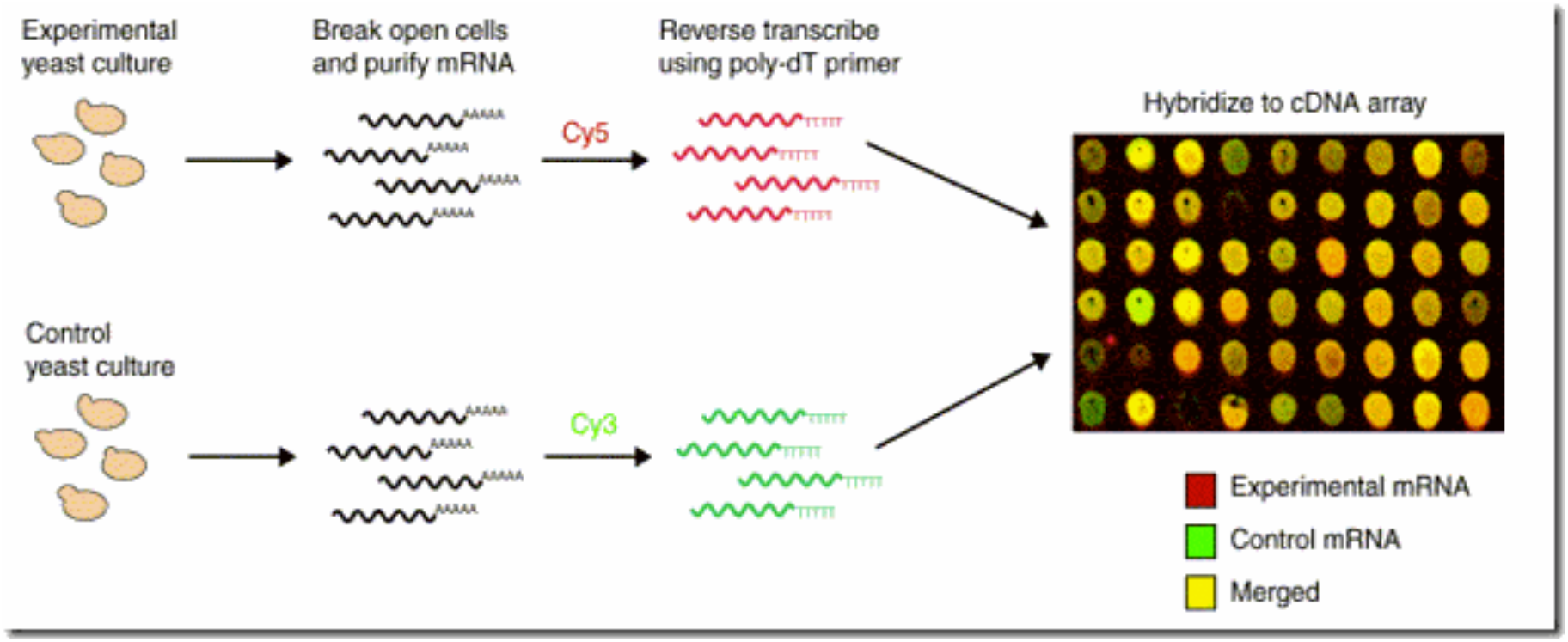

http://www.labgrab.com/

# Studying physiological and pathological processes

- DNA level
  - Mutations (SNP, CNV, aneuploidy)
  - Methylation
- RNA level
  - Gene expression
  - miRNA expression
- Protein level
  - Quantification
  - Localization
  - Post-translational modifications (e.g. phosphorylation)

# From single gene to a genome scale: microarrays

~1990: 2-colour microarrays



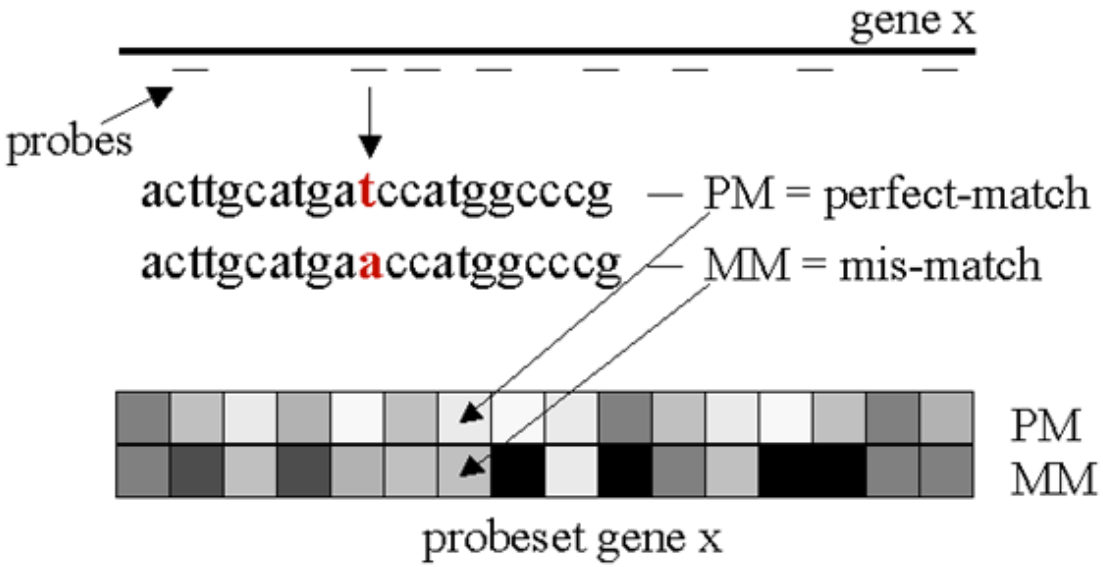

http://www.biologyexams4u.com/

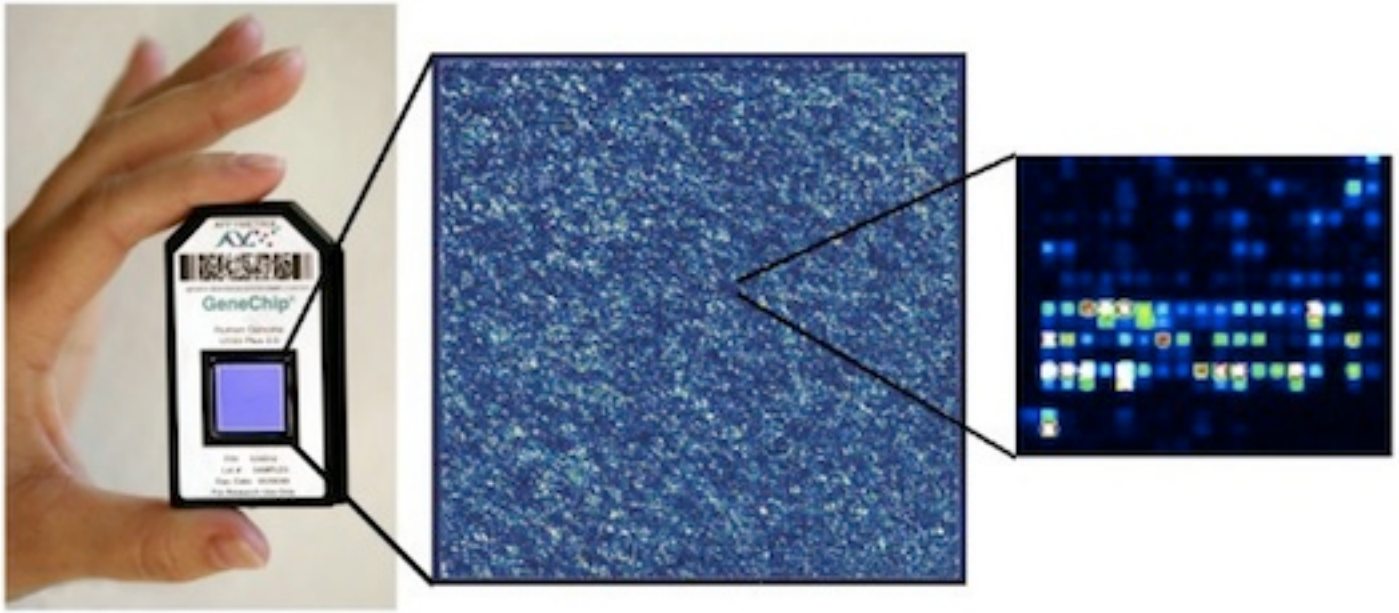
# One colour microarrays: Affymetrix

GeneChip® Human Genome U133 Plus 2.0

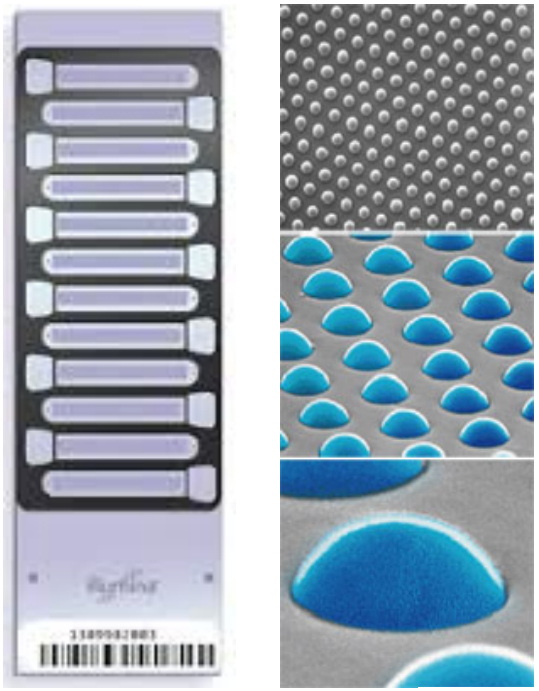
~54000 probesets (11 probes each)

Probes synthesised in situ using photolithography



# One colour microarrays: Illumina

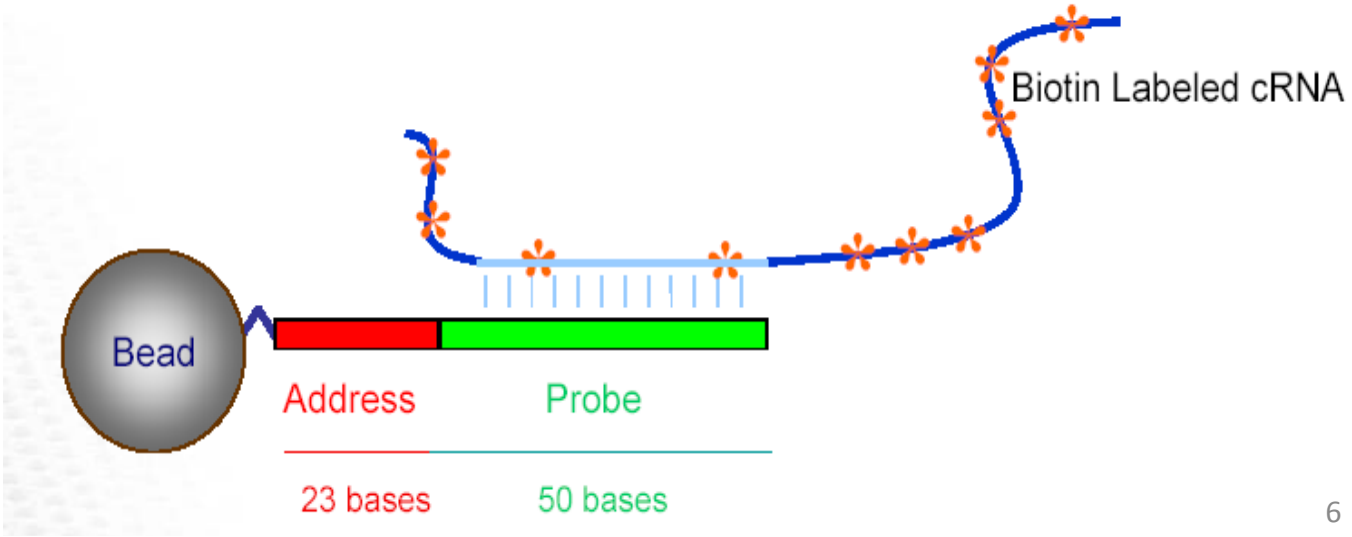


## Beadarray Technology

- each bead is coated with hundred-thousand copies of the same 73mer probe:

- 23 base address linked to 50-base gene specific probe

- 30 copies of each bead type per array: 47k genes = >1.4M beads!



# One colour microarrays: Illumina

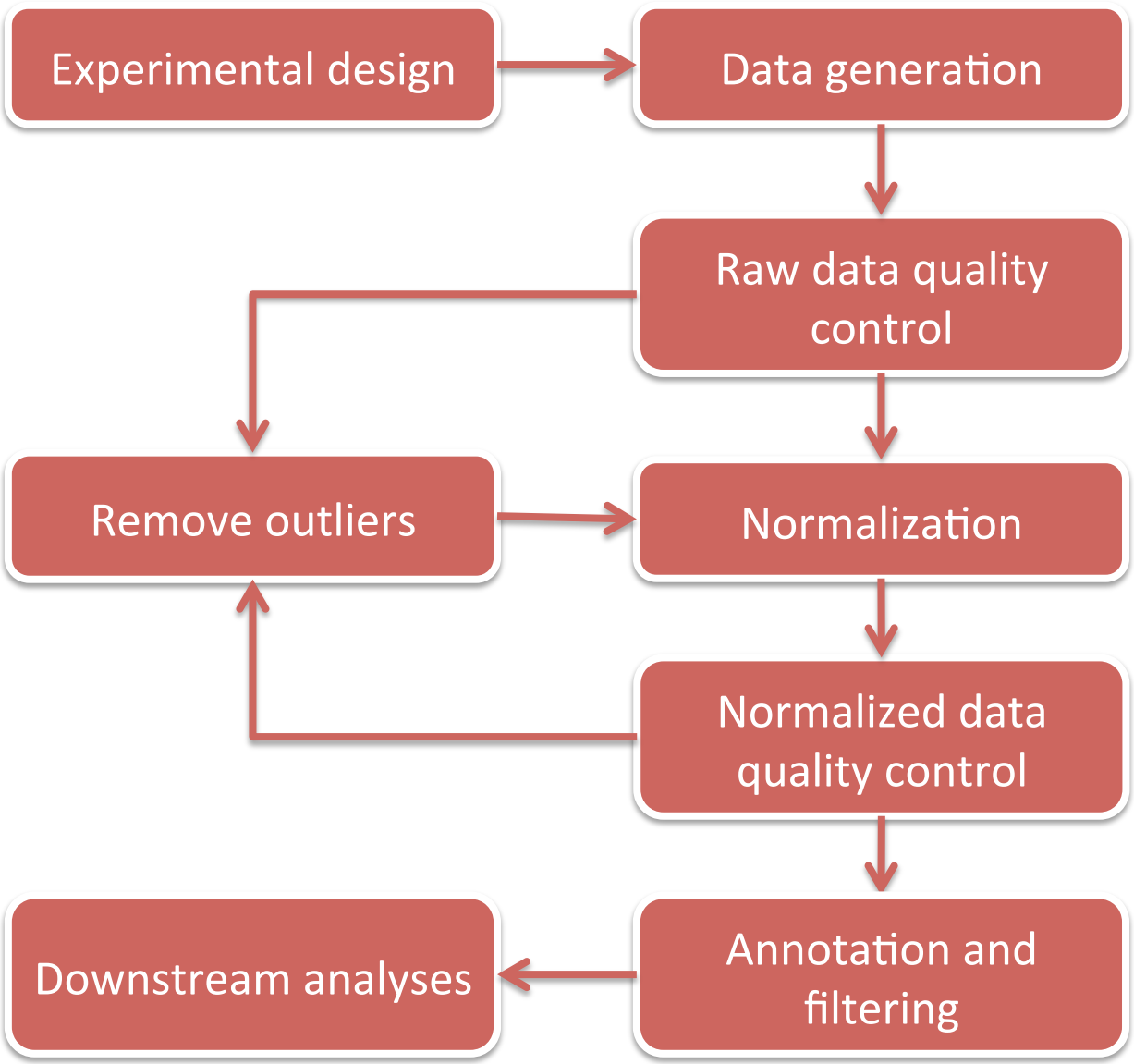
# Microarrays vs next-generation sequencing

| | RNAseq | Miroarray |
|---|---|---|
| Result Type | Rich, not limited to expression | Limited to expression only |
| Expression | Can quantify expression on exon and gene level | Can quantify expression on exon or gene level |
| Novel Discovery | Can be used for novel discovery | Can only detect what is on the chip |
| Analysis | Difficult | Easy |
| Interpretation | Difficult | Easy |
| Price for assay | Price has become comparable to microarray, however the analysis hardware and analysis time may increase the final cost | Price is stable |

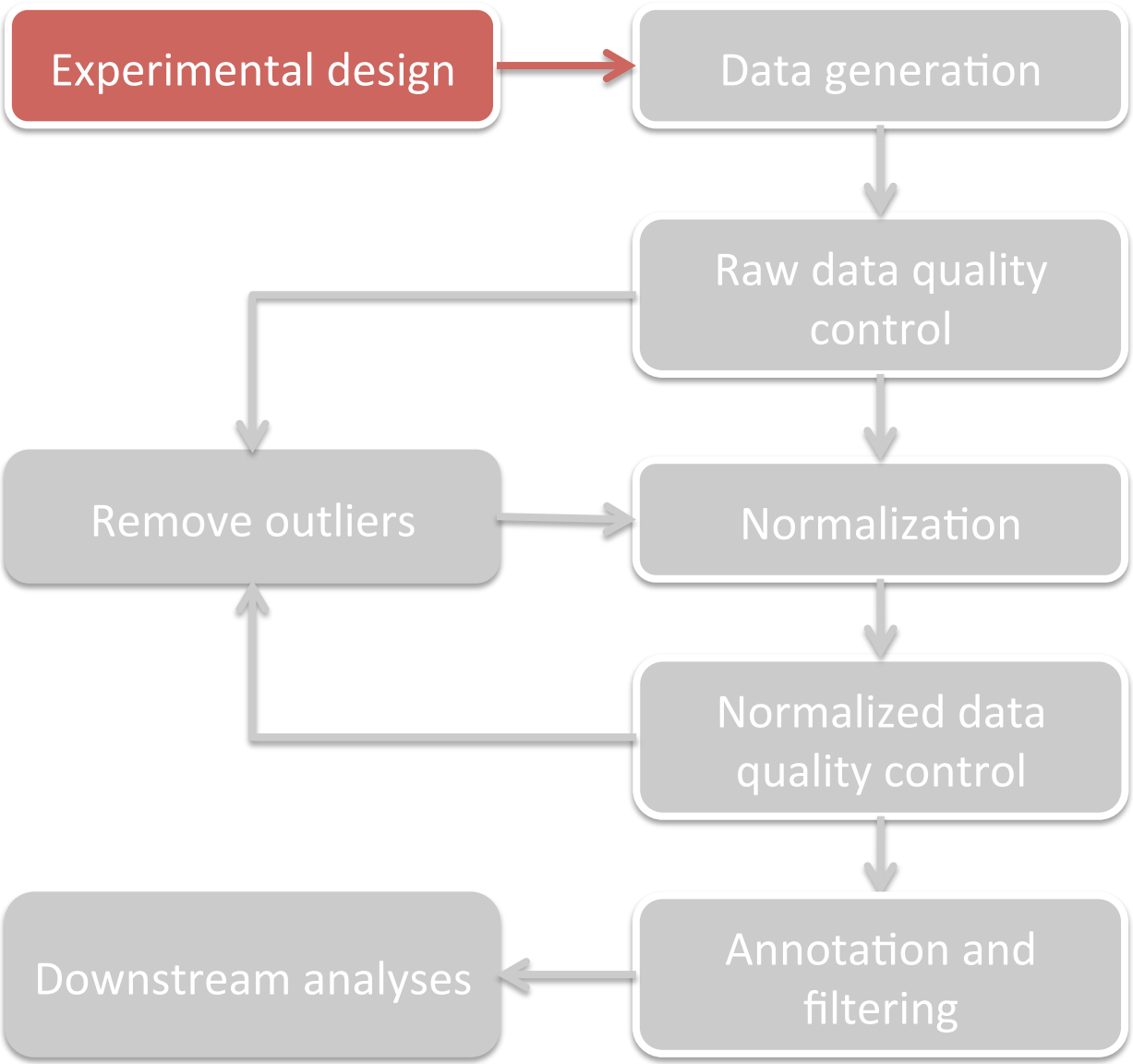
Wealth of data available online e.g. GEO and ArrayExpress

http://slideplayer.com/ (Yan Guo)

# The workflow

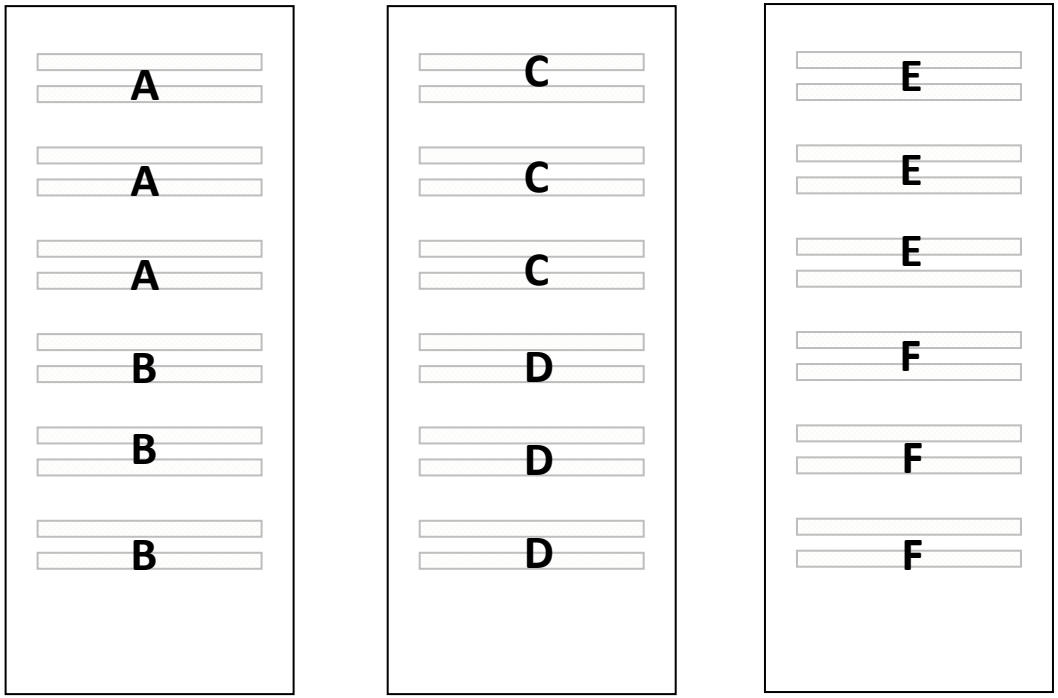


# The workflow



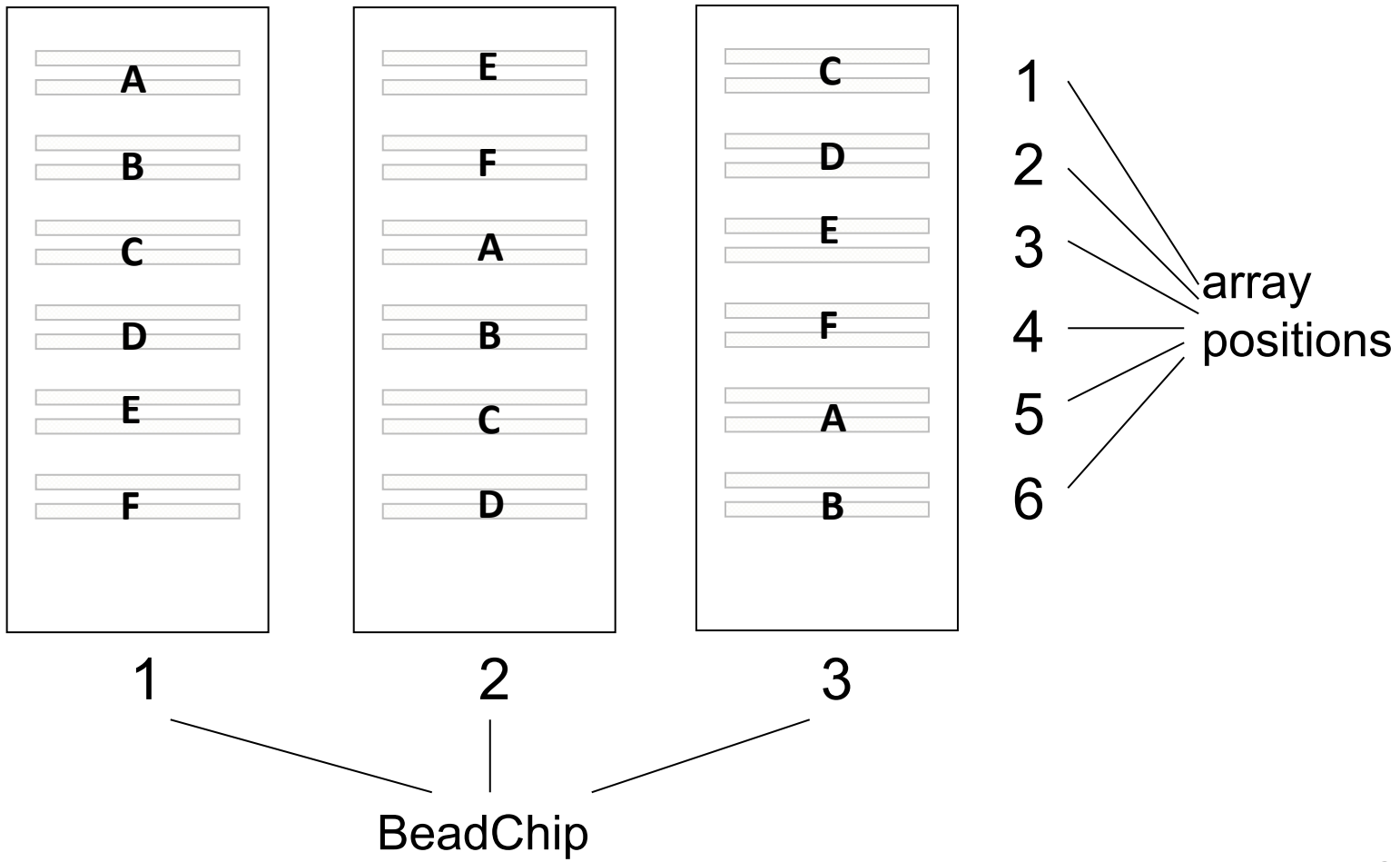
# Experimental design

6 sample types (A-F) x 3 replicates

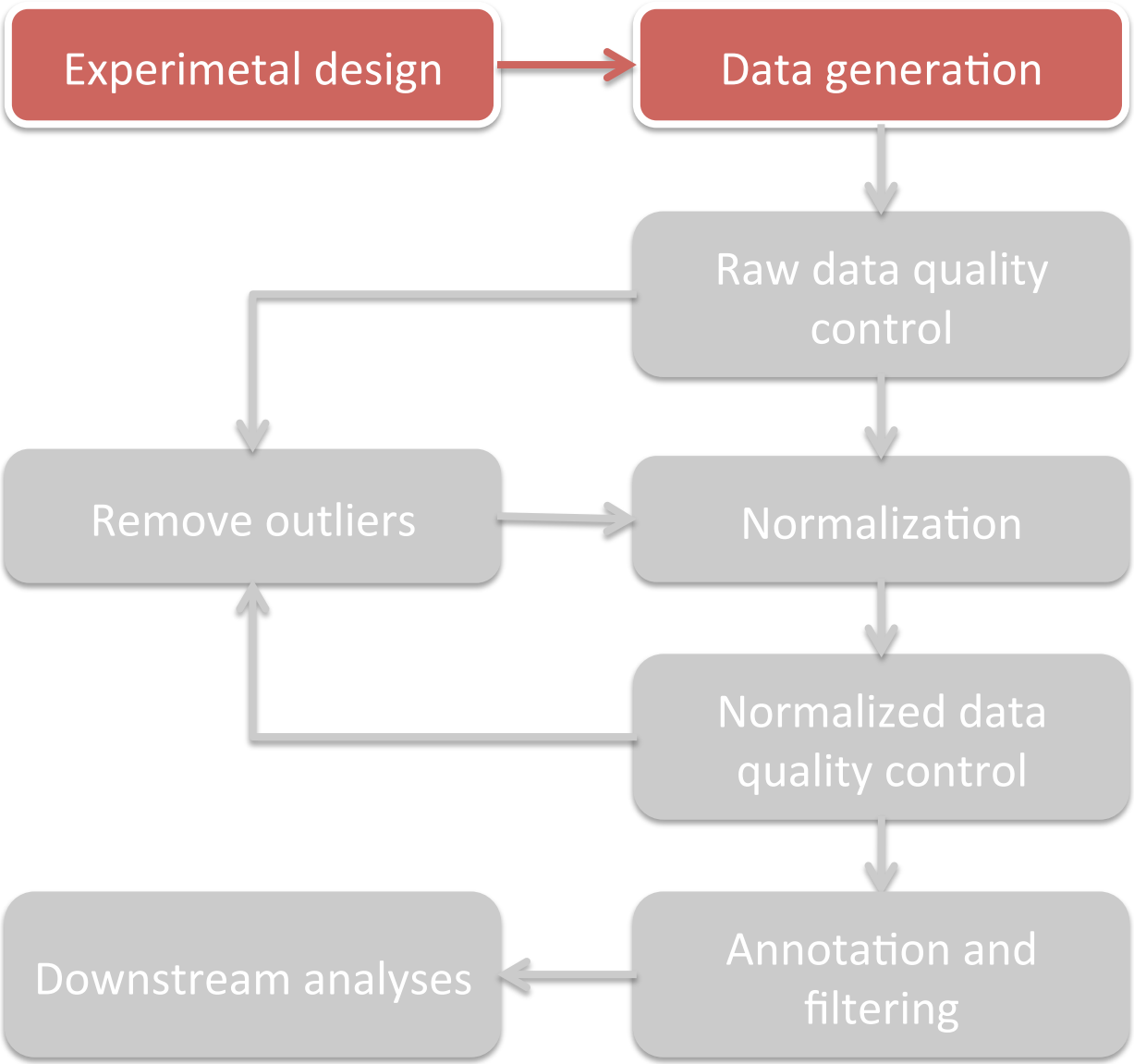


# Experimental design

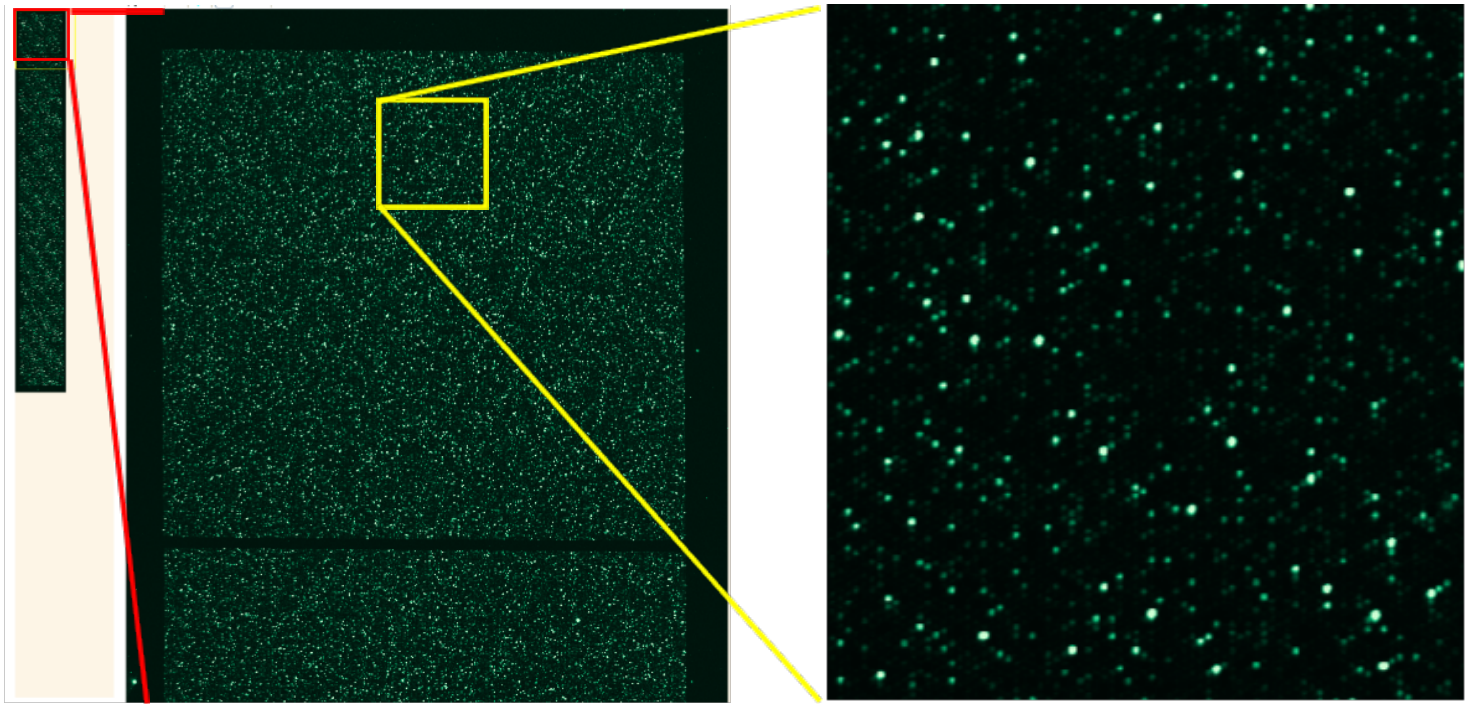
6 sample types (A-F) x 3 replicates



# The workflow



# Raw data



*R (beadarray)*

**Bead level data**

User friendly (GUI)

*GenomeStudio* → *R (beadarray)*

**Probe level data**

More flexibility

*GenomeStudio* → *R (beadarray, lumi, limma)*

**Gene level data**

# Bead level data

Example:

ProbeID

Background corrected intensity

Bead Centre

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Code | Grn | GrnX | GrnY |
| 2 | 2 | 1686 | 405.9445 | 994.7201 |
| 3 | 2 | 2148 | 1485.263 | 465.5954 |
| 4 | 2 | 2391 | 981.7433 | 710.9218 |
| 5 | 2 | 1961 | 414.4303 | 895.2175 |
| 6 | 2 | 2477 | 1026.212 | 942.4114 |
| 7 | 2 | 2659 | 720.4089 | 1370.215 |
| 8 | 2 | 1772 | 1139.226 | 816.4459 |
| 9 | 2 | 2737 | 1143.429 | 213.7267 |
| 10 | 2 | 2369 | 1110.516 | 203.423 |
| 11 | 2 | 2283 | 1483.378 | 548.7356 |
| 12 | 2 | 2371 | 895.504 | 976.541 |
| 13 | 2 | 2532 | 1667.515 | 864.9724 |
| 14 | 2 | 2558 | 1133.62 | 960.1776 |
| 15 | 2 | 1931 | 1127.286 | 1469.364 |
| 16 | 2 | 1760 | 279.3574 | 946.3187 |
| 17 | 2 | 2690 | 812.6176 | 803.8156 |
| 18 | 2 | 2583 | 1048.631 | 889.1783 |
| 19 | 2 | 2432 | 509.0219 | 1079.245 |
| 20 | 2 | 2538 | 929.3365 | 1226.301 |
| 21 | 2 | 2280 | 553.4136 | 885.7501 |
| 22 | 2 | 2077 | 714.496 | 250.4801 |
| 23 | 2 | 2551 | 536.4883 | 206.4698 |
| 24 | 2 | 1593 | 936.7546 | 543.4179 |

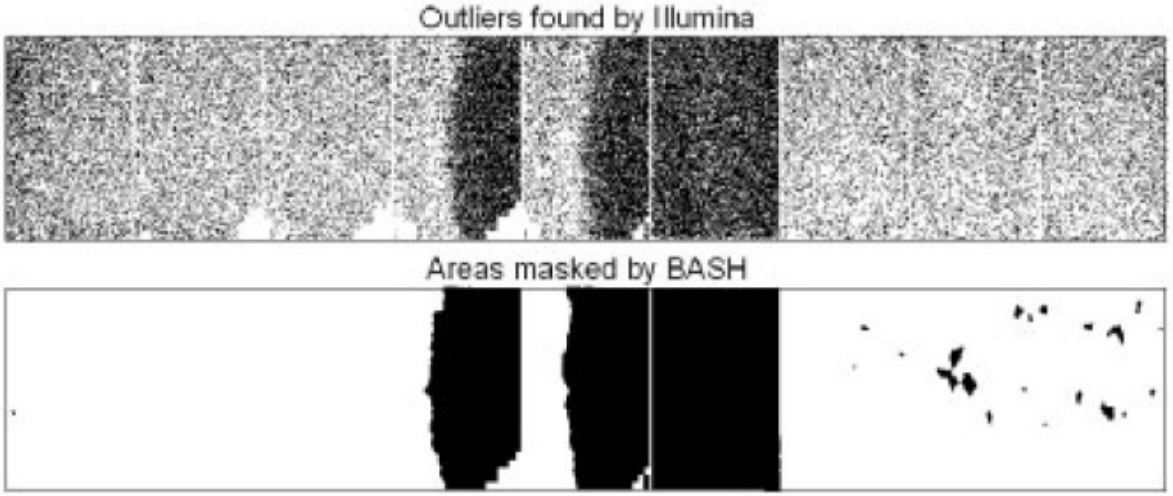
Information for **all** beads on an array. **Same format for all Illumina arrays (expression, SNP, methylation..)**

# BASH (*beadarray*)

- BASH is a method for detecting and removing artefacts from Illumina arrays
- BASH looks for clusters of outliers close together and major intensity variation across the array.
- BASH outputs weights for each bead:

  0=exclude from further analysis.

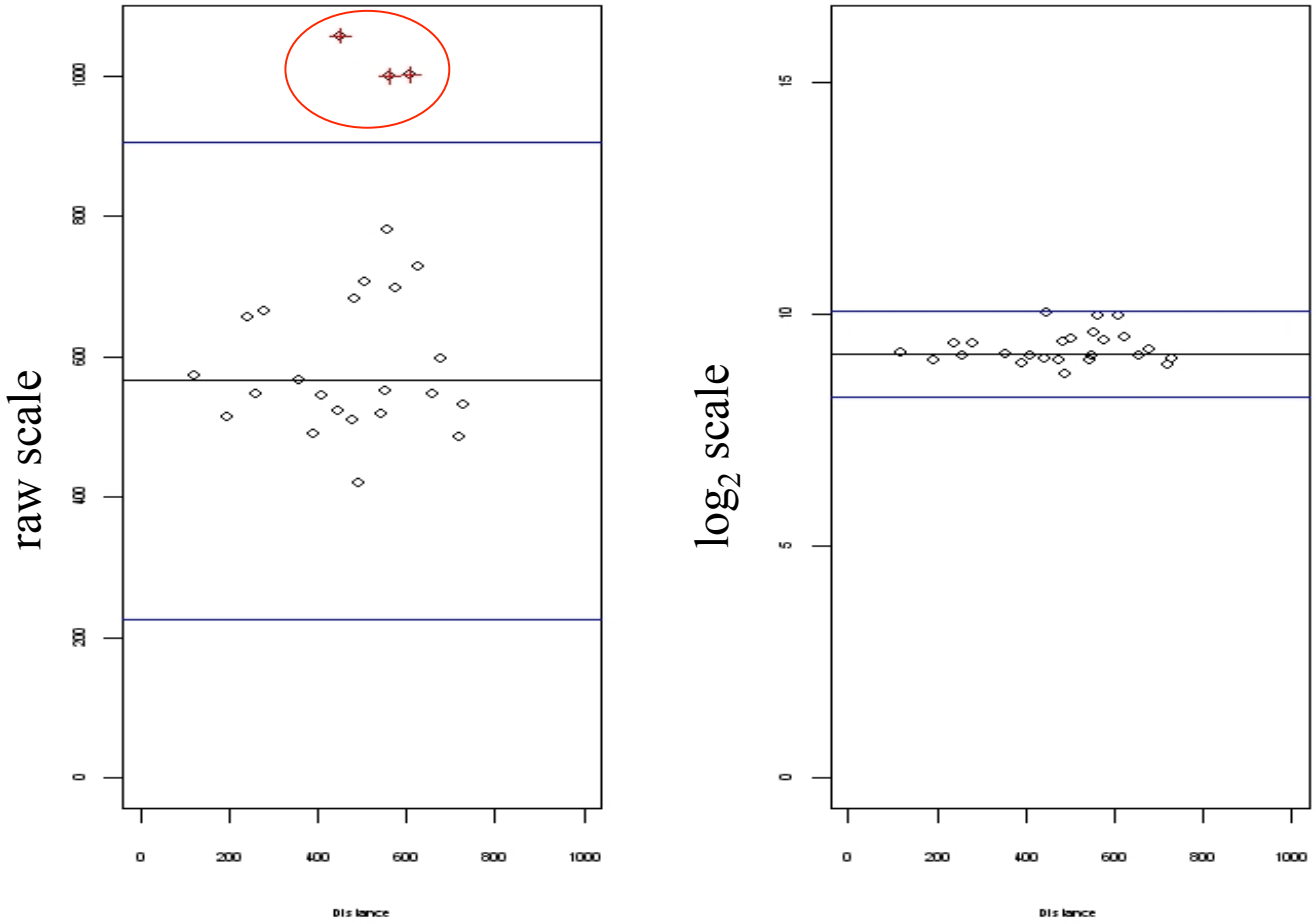
  1=include




Cairns JM et al. Bioinformatics 2008

# Summarization

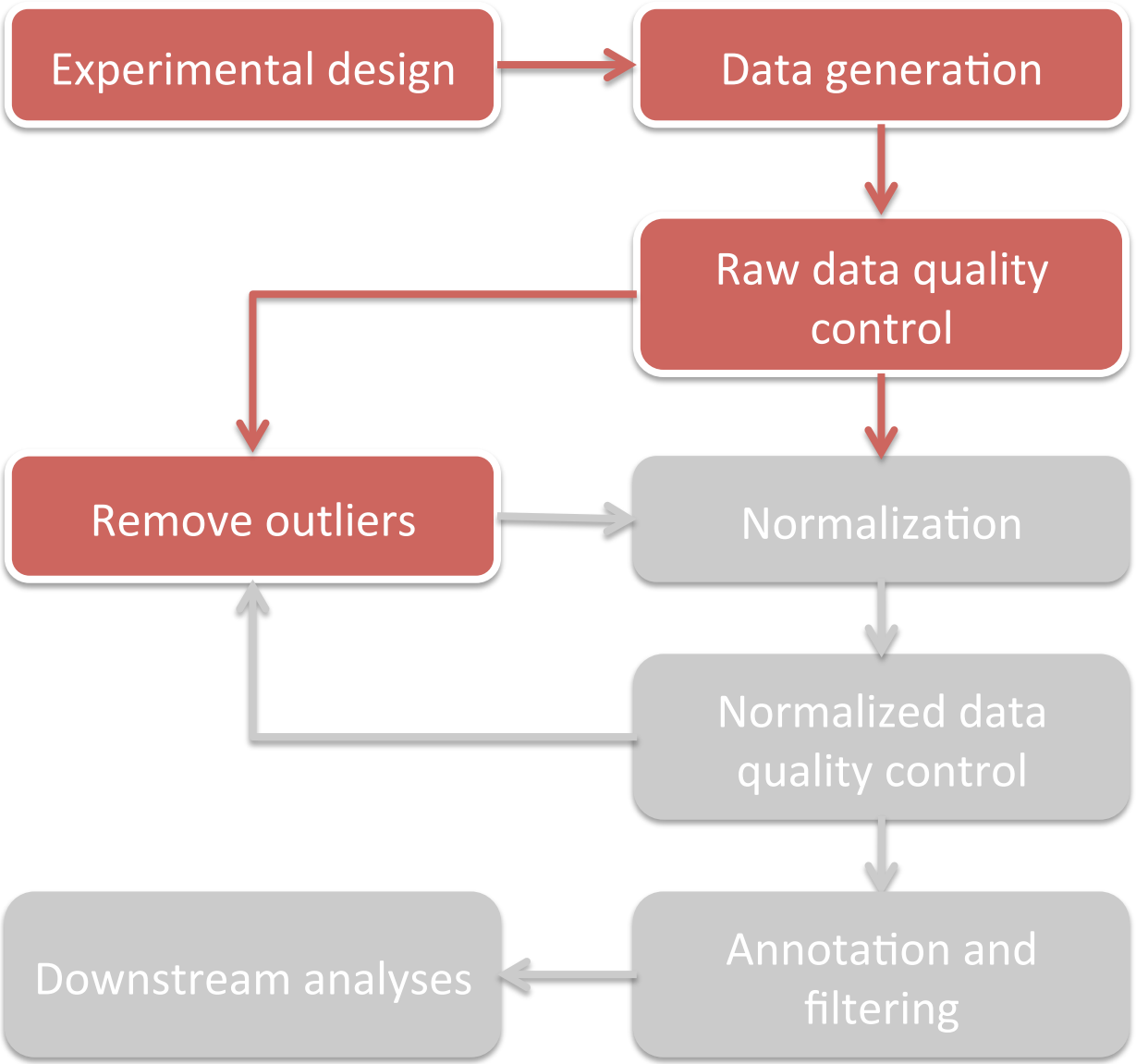
Outlier detection is sensitive to $\log_2$ transformation




> 3 median absolute deviation (MAD)

Using data on the original scale (un-logged) for outlier detection gives more outliers *above* the median

# The workflow

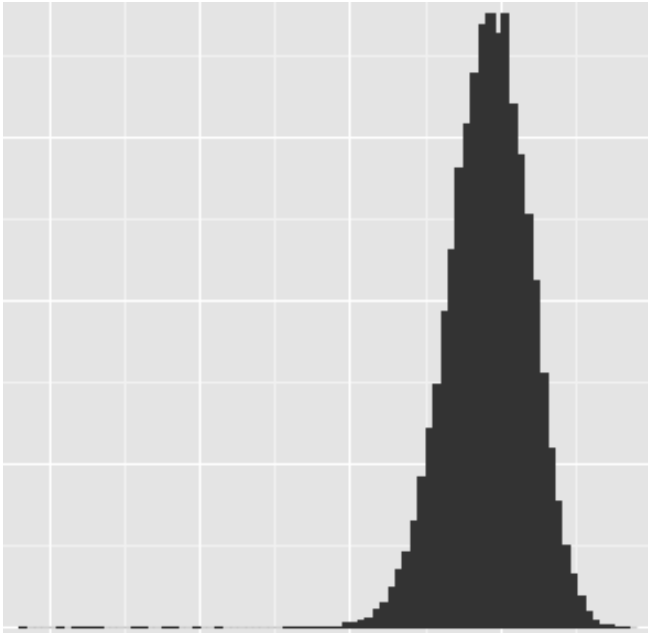


# Outlier detection

Inclusion of inaccurate data will hamper our analysis

Useful metrics:

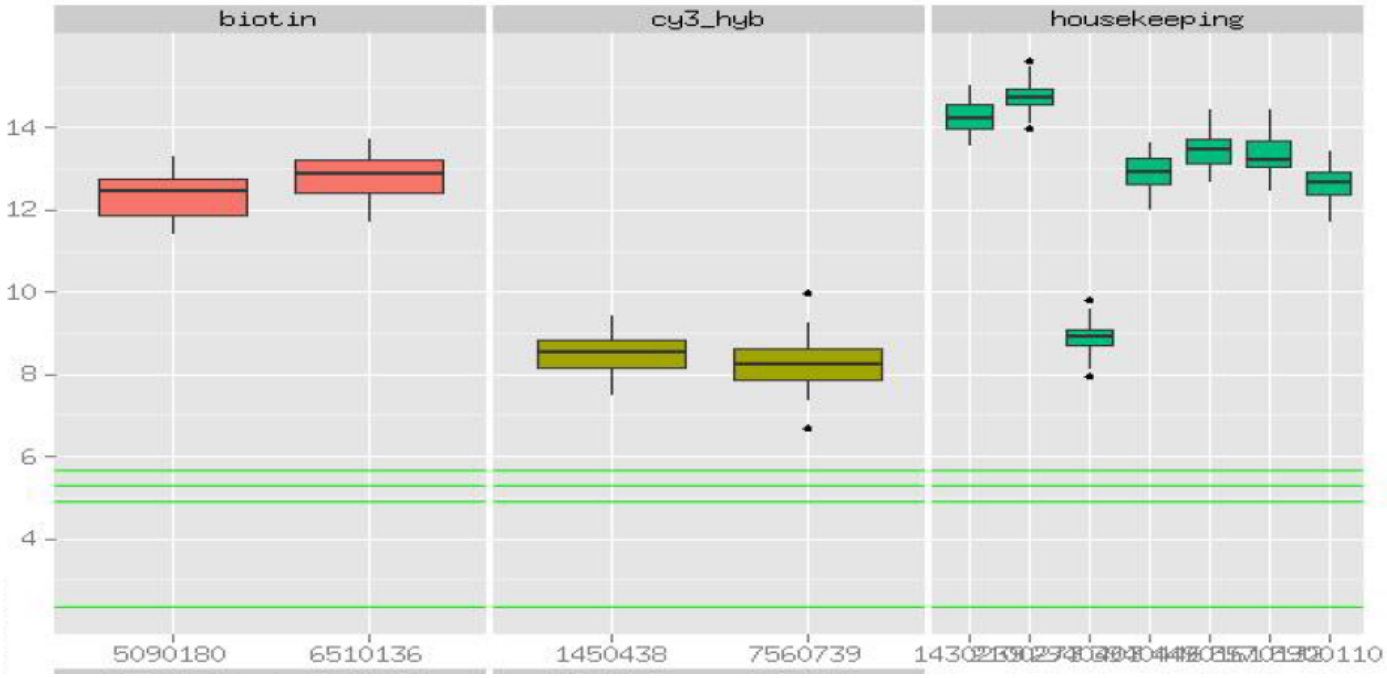
- Positive/negative controls
- Number of detected probes
- Signal distribution
- Clustering/PCA



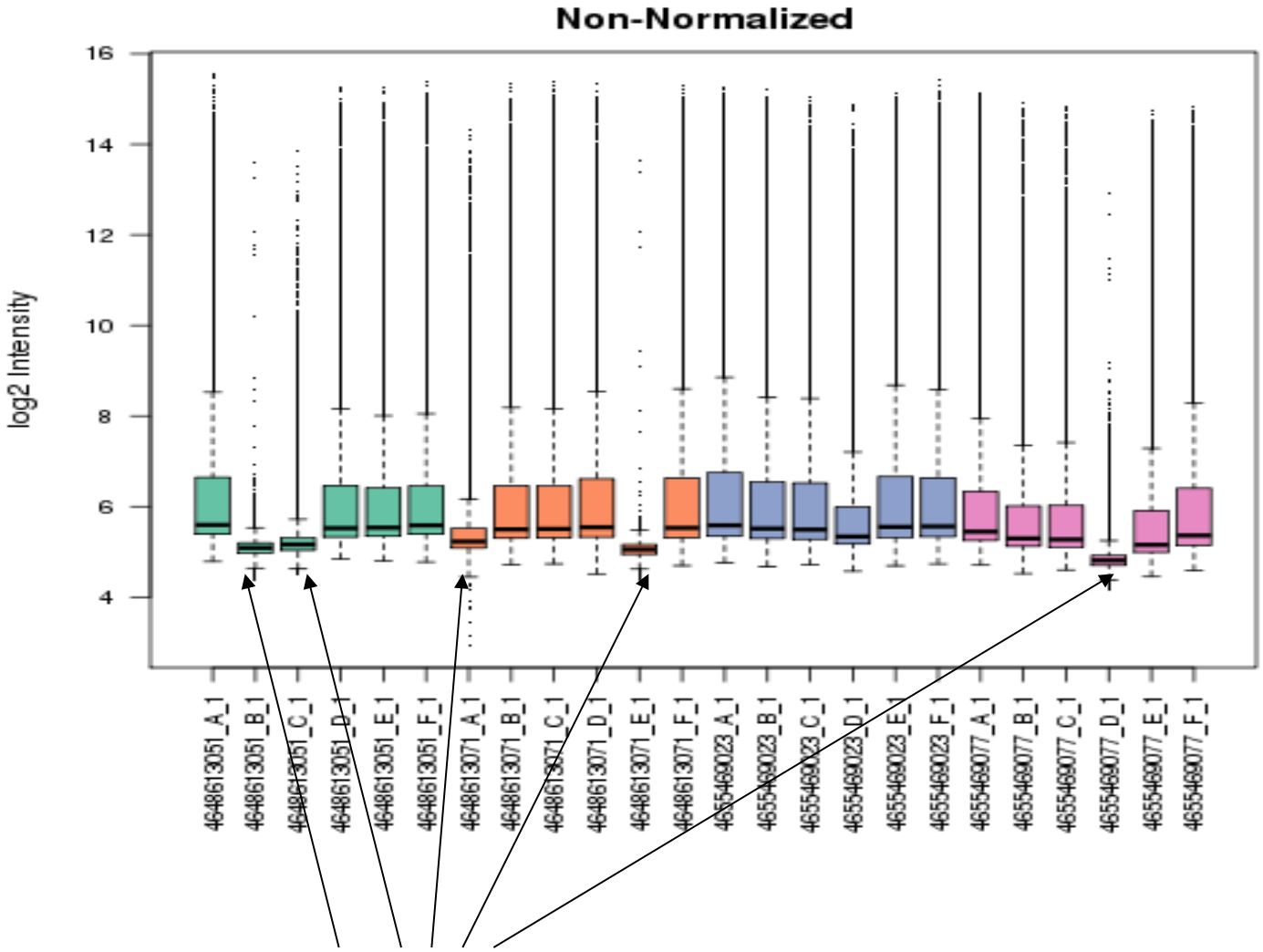
- Hard to define absolute thresholds
- Can normalization "correct" the bias?

# Control probes

- Illumina design probes for housekeeping genes (**Positive controls**)
- They also have probes that should not hybridise to the target genome (**Negative controls**)
- The negative controls are also used to assess whether a given probe is expressed / detected.



# Outlier detection

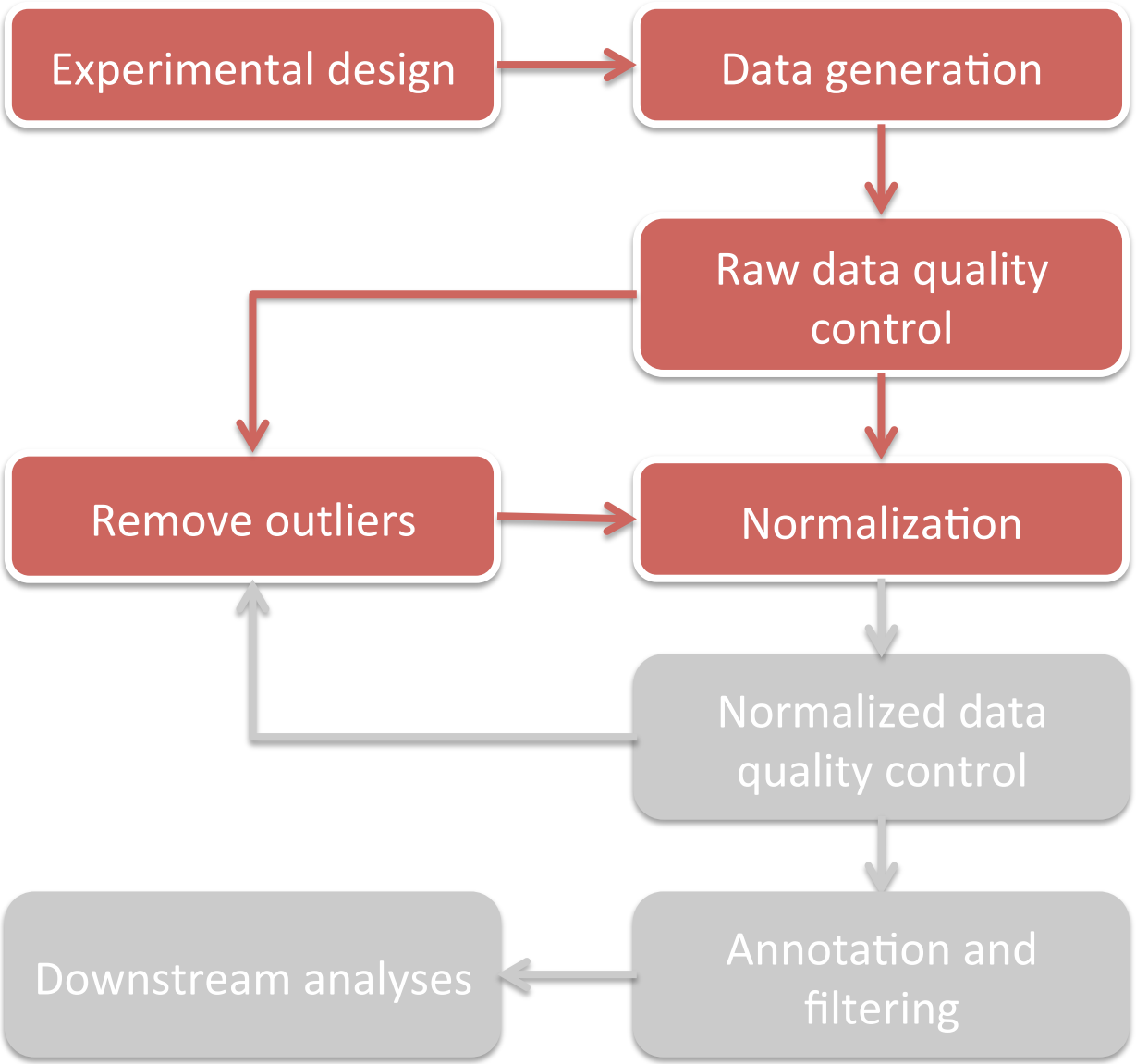



Colour indicates different sample groups

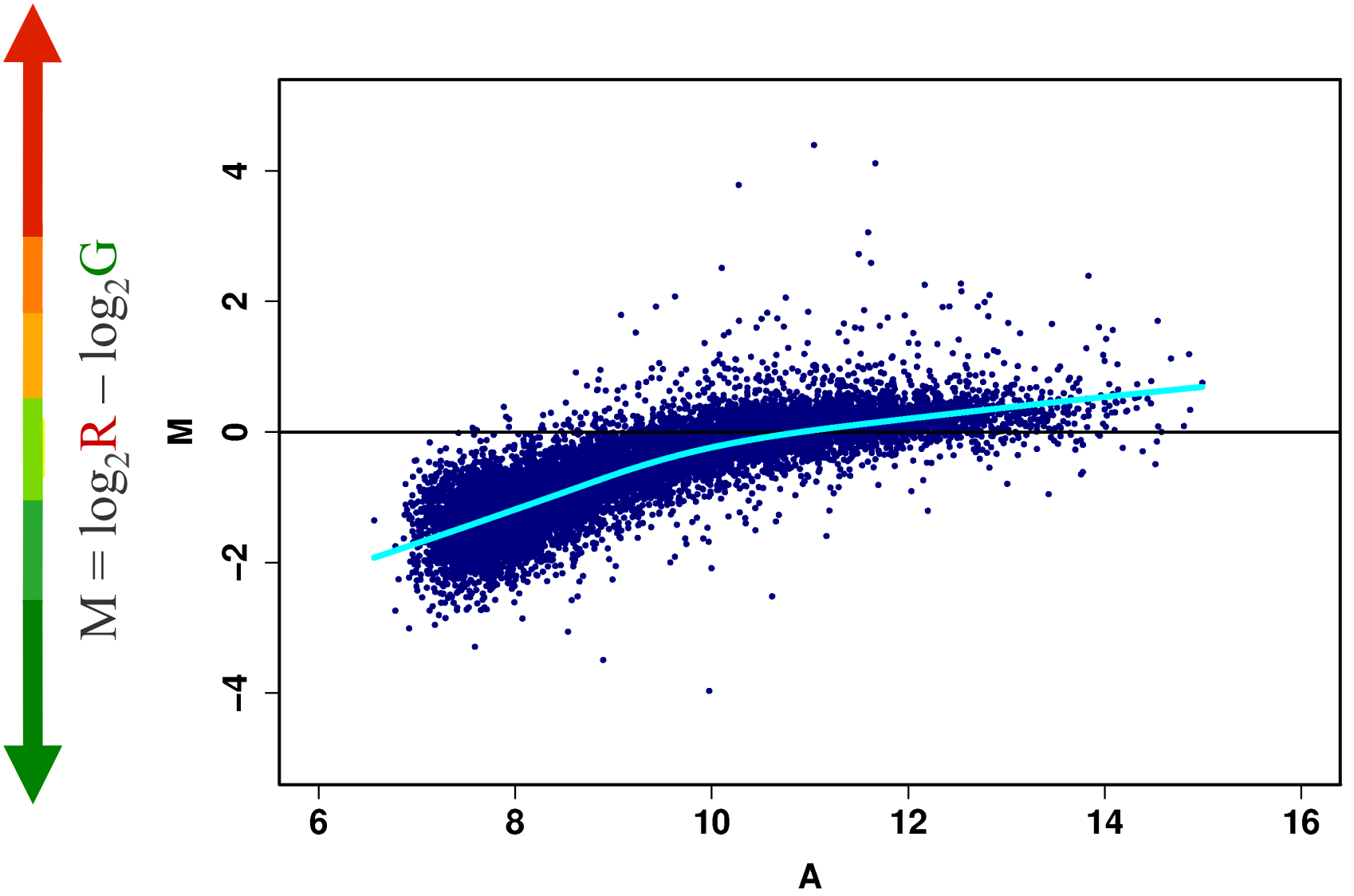
The analyst chose to remove these arrays

Example from Stew MacArthur

# The workflow



# MA-plot (two colour microarrays)

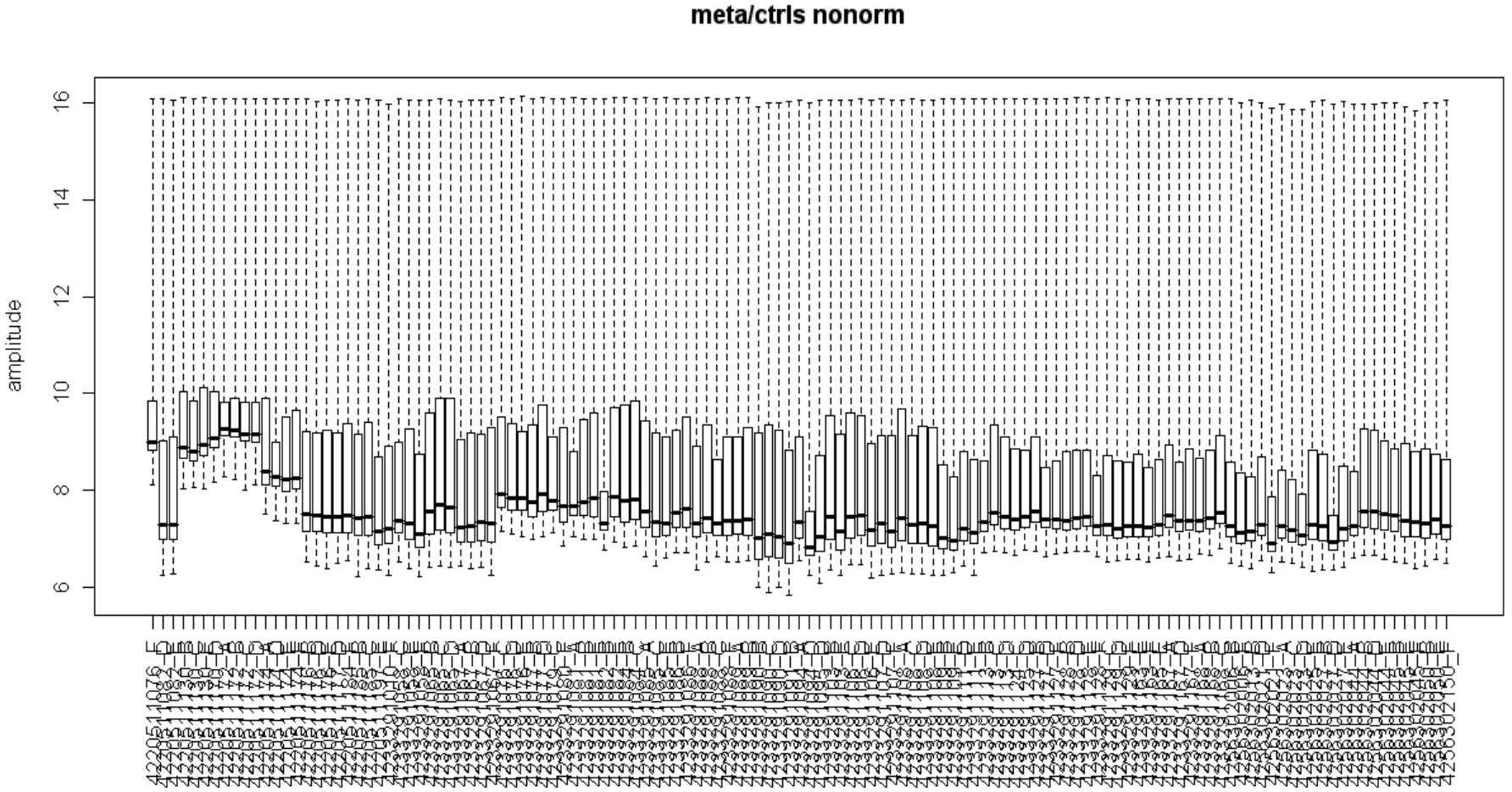



$$A = {}^{1}/_{2}\,(\log_2 R + \log_2 G)$$

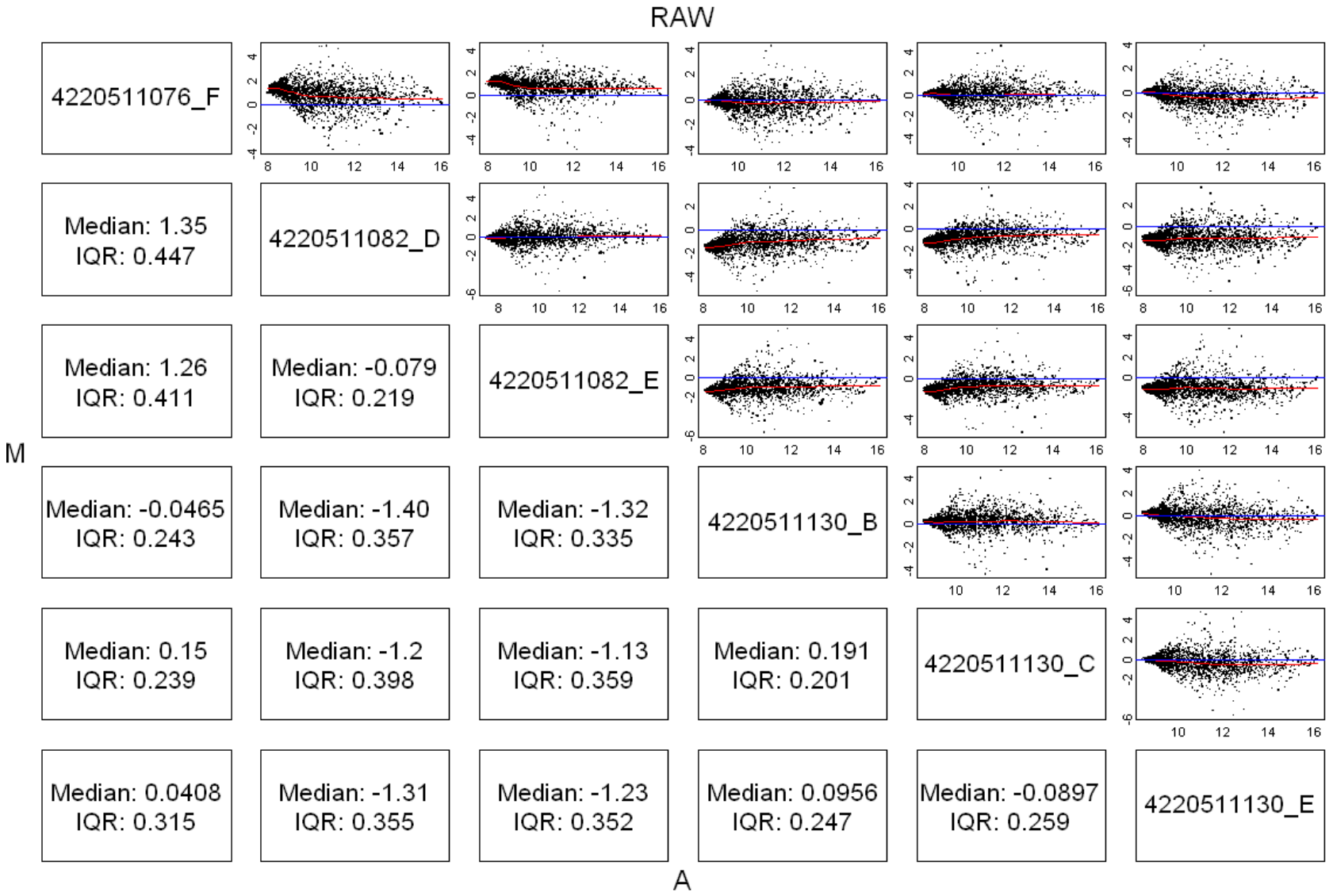
# (inter-array) normalization

**Aim**: to reduce technical noise preserving biological differences

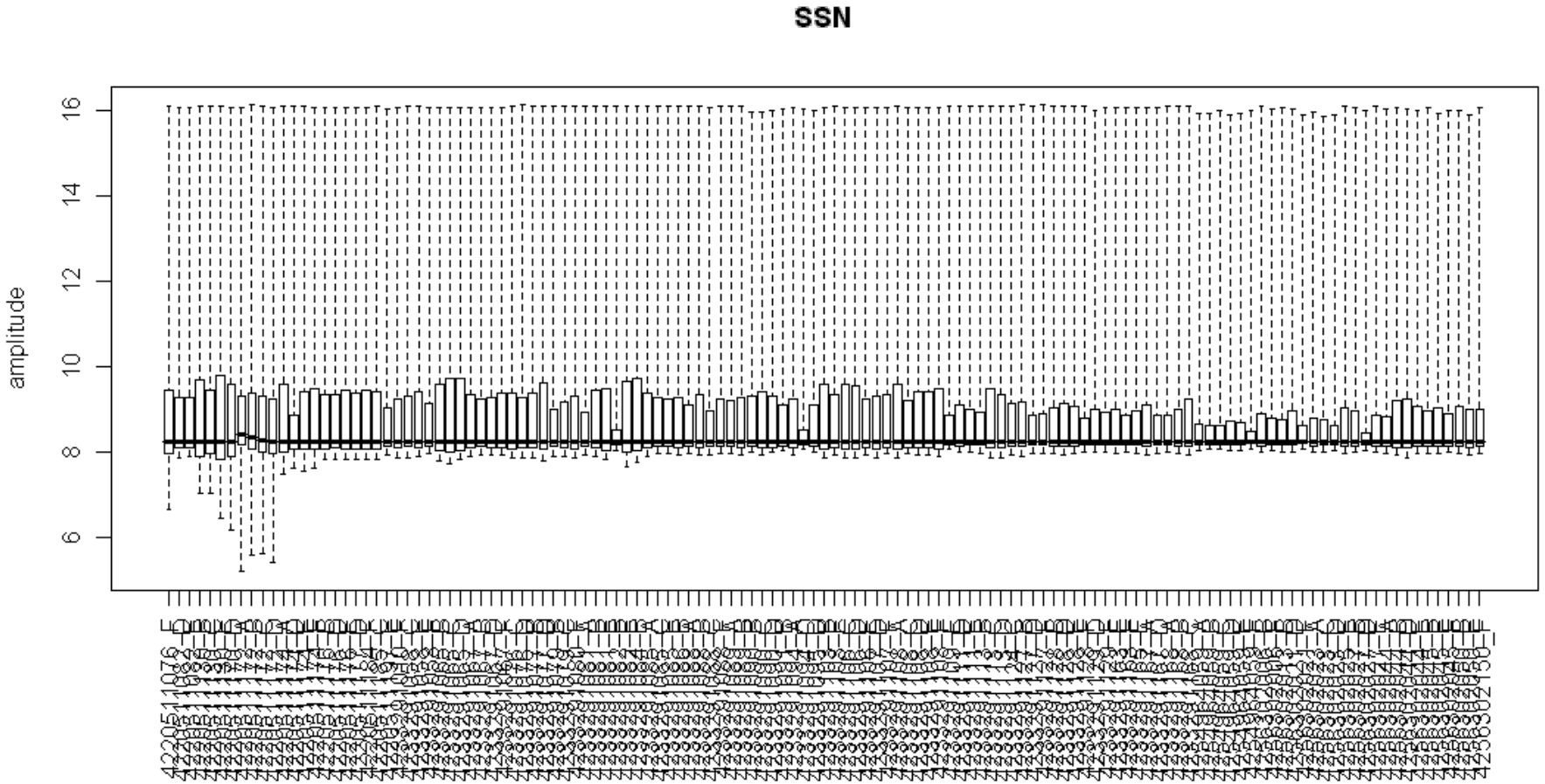
**Assumptions**: Most genes are not differentially expressed and a similar number of genes are expected to be up- and down-regulated (i.e. similar distributions expected for all samples)



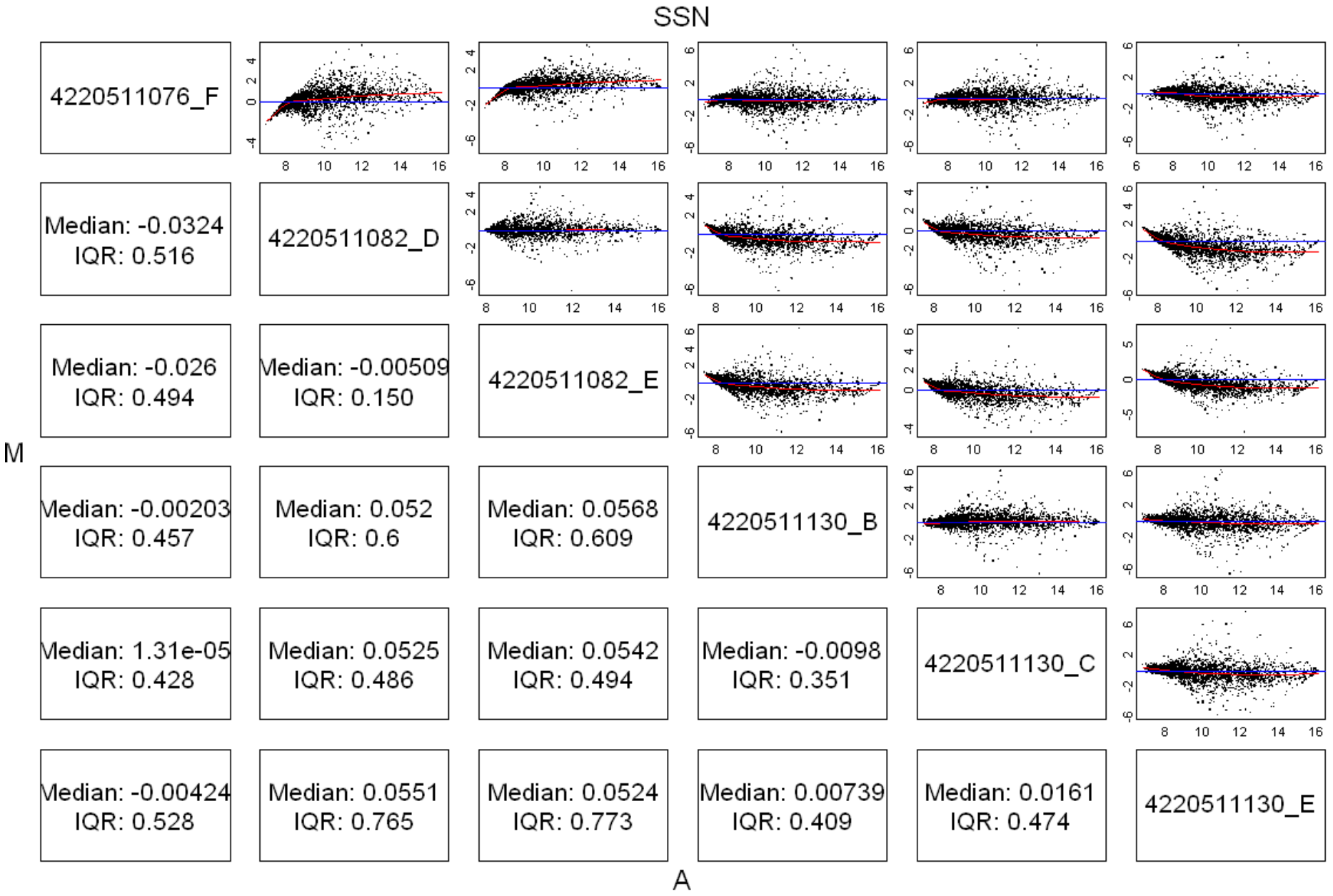
# MA-plot (one colour microarrays)



# Simple Scaling Normalization

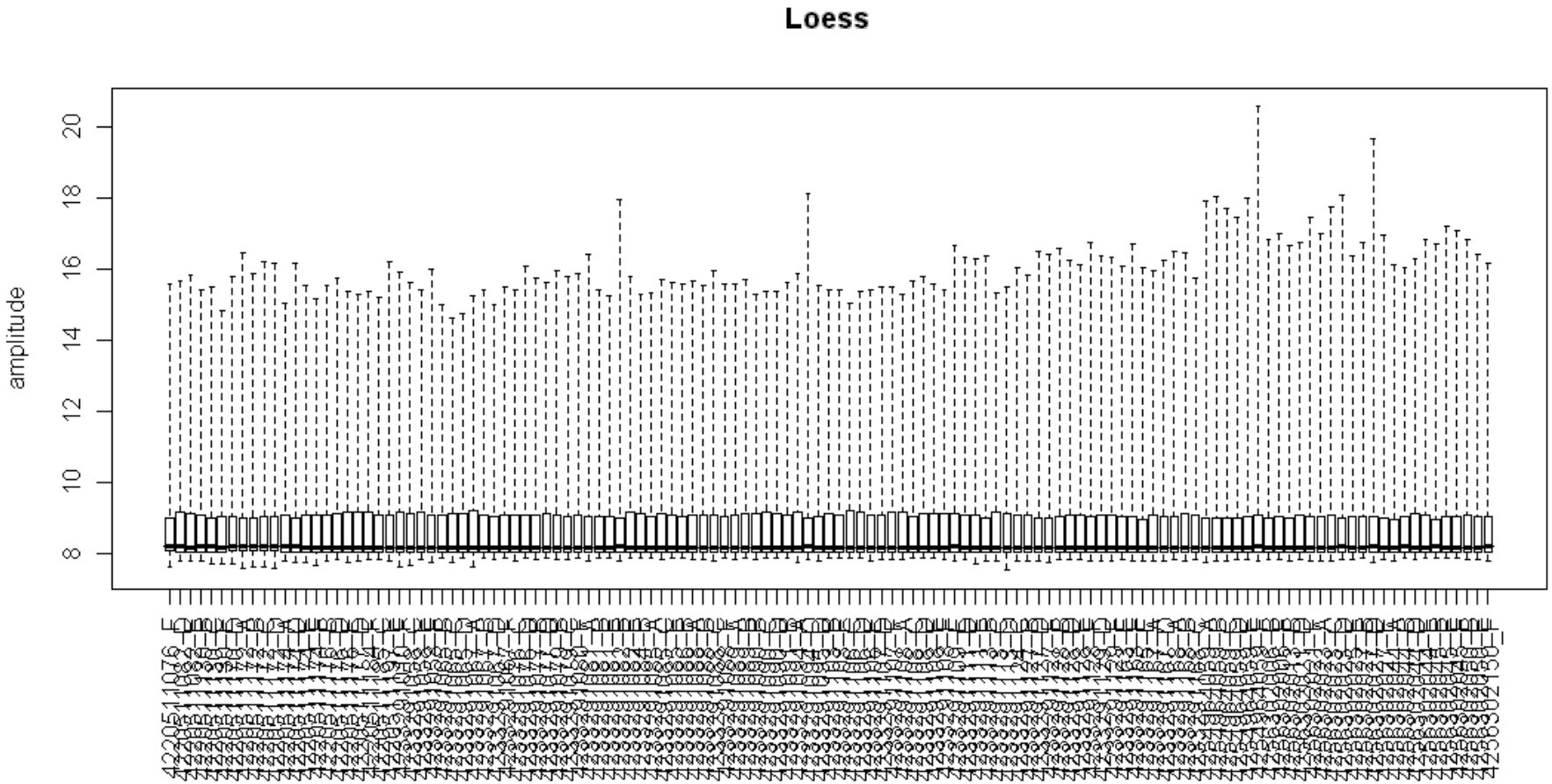


# Simple Scaling Normalization

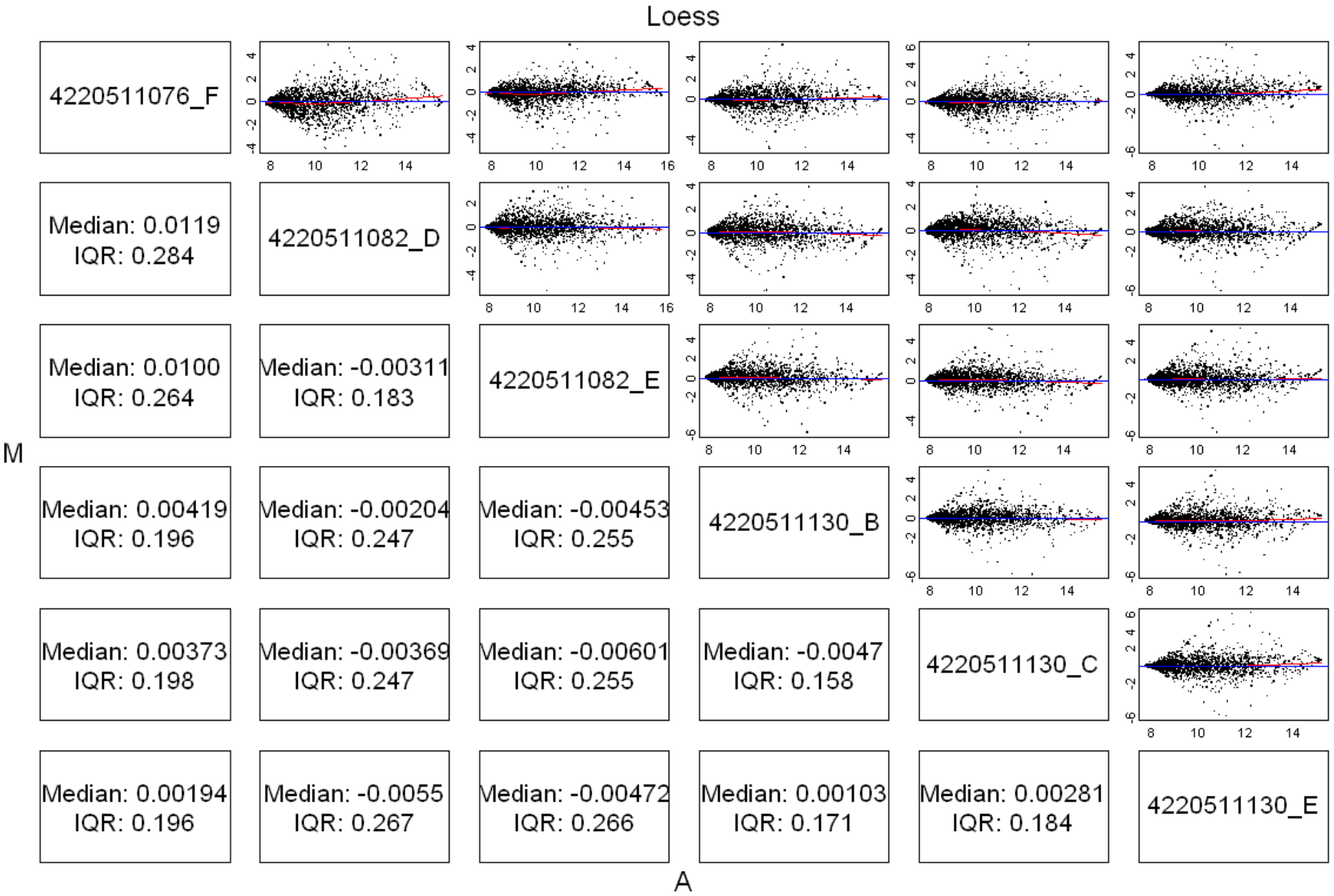


# Loess Normalization

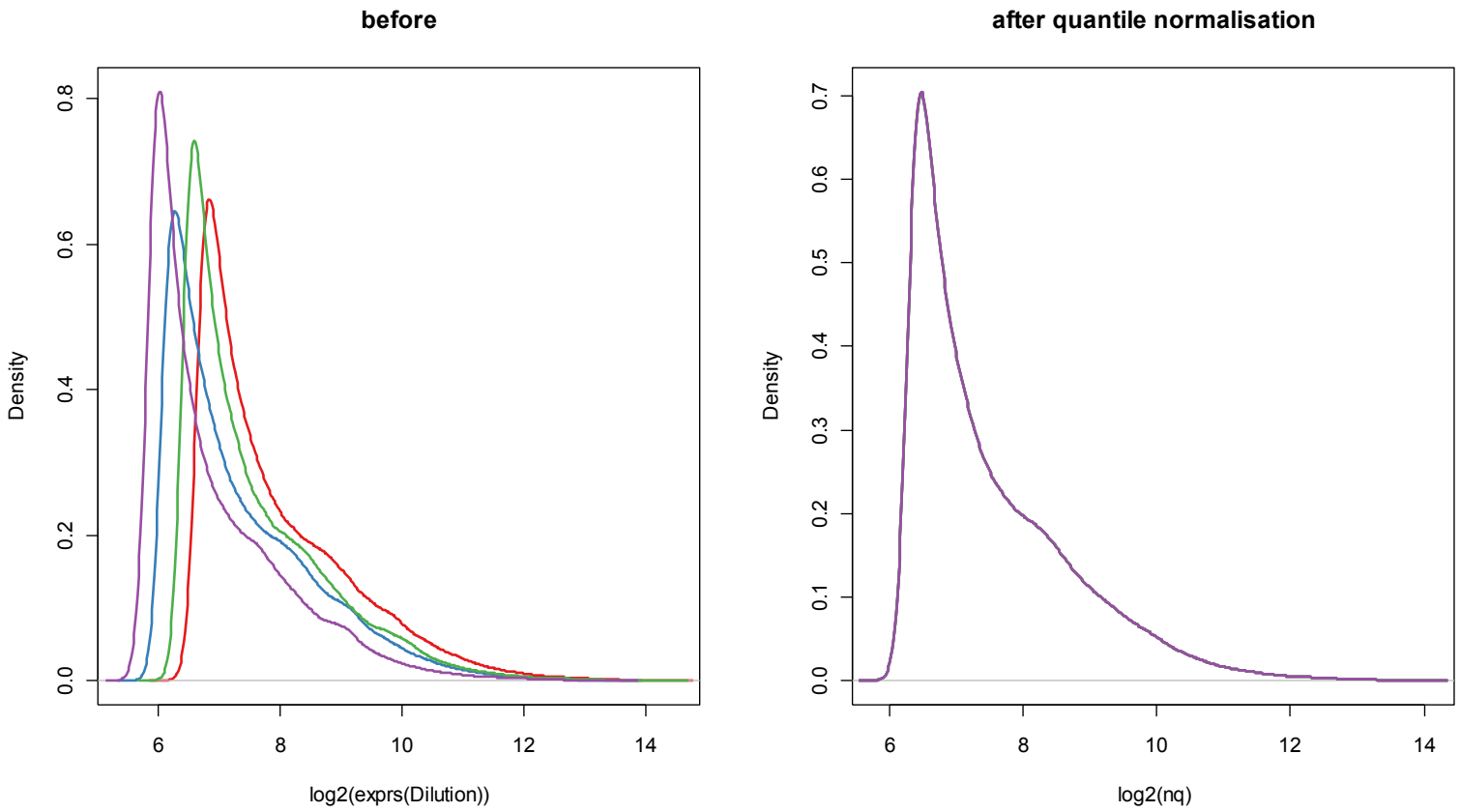
LOcally wEighted Scatterplot Smoothing



# Loess Normalization



# Quantile Normalization




Quantile normalisation is: per array rank-transformation followed by replacing ranks with values from a common reference distribution

# Quantile Normalization

```
##   Array1 Array2 Array3
## A      1      3      9
## B      3      4      1
## C      9      2      5
## D      2      1      7
## E      4      9      6
```

```
##      Array1  Array2  Array3
## [1,] "Rank1" "Rank3" "Rank5"
## [2,] "Rank3" "Rank4" "Rank1"
## [3,] "Rank5" "Rank2" "Rank2"
## [4,] "Rank2" "Rank1" "Rank4"
## [5,] "Rank4" "Rank5" "Rank3"
```

# Quantile Normalization

```
##       Array1 Array2 Array3
## [1,]      1      1      1
## [2,]      2      2      5
## [3,]      3      3      6
## [4,]      4      4      7
## [5,]      9      9      9
```

```
## Rank1 Rank2 Rank3 Rank4 Rank5
##     1     3     4     5     9
```

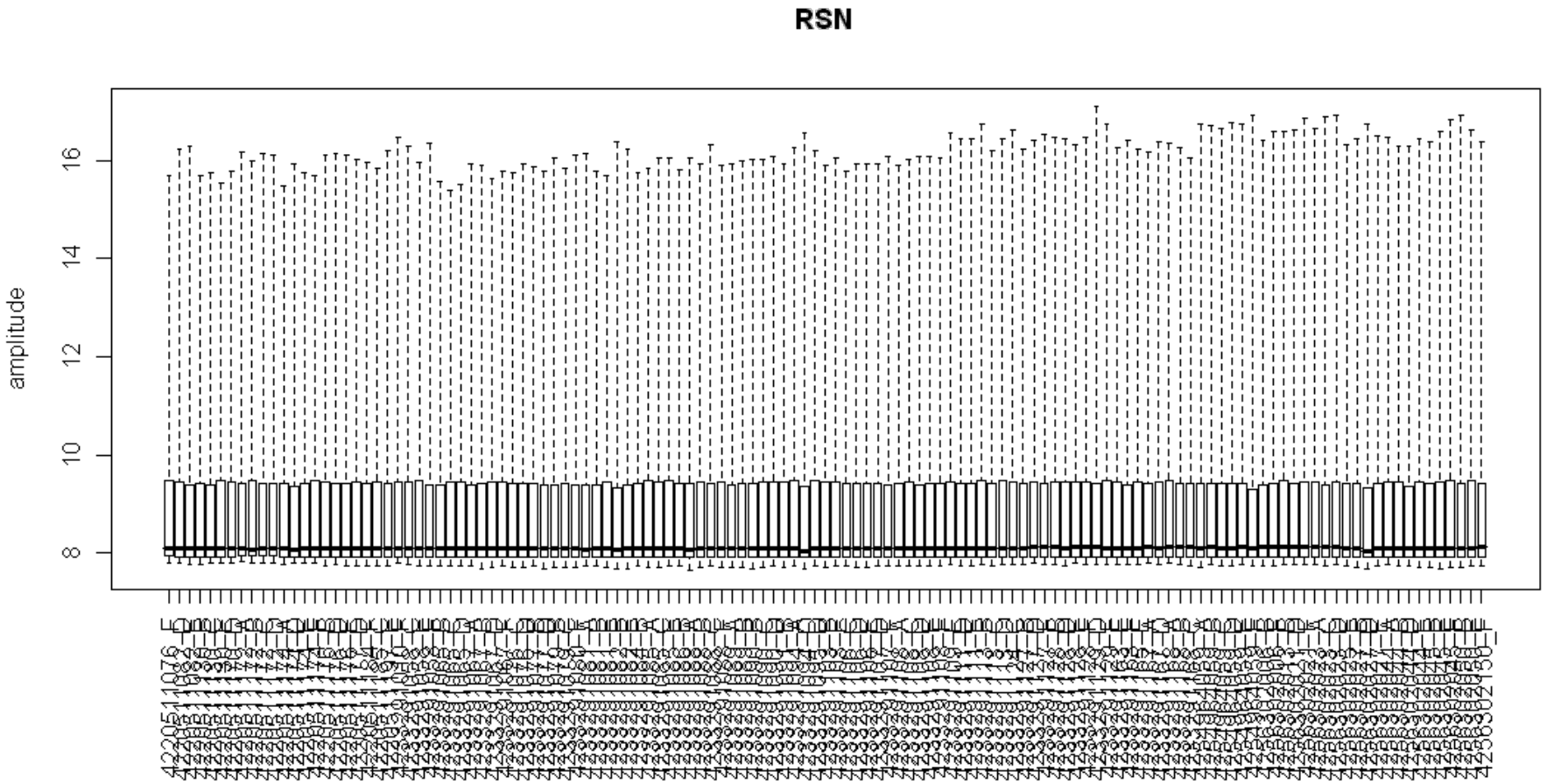
# Quantile Normalization

```
## Rank1 Rank2 Rank3 Rank4 Rank5
##     1     3     4     5     9
```

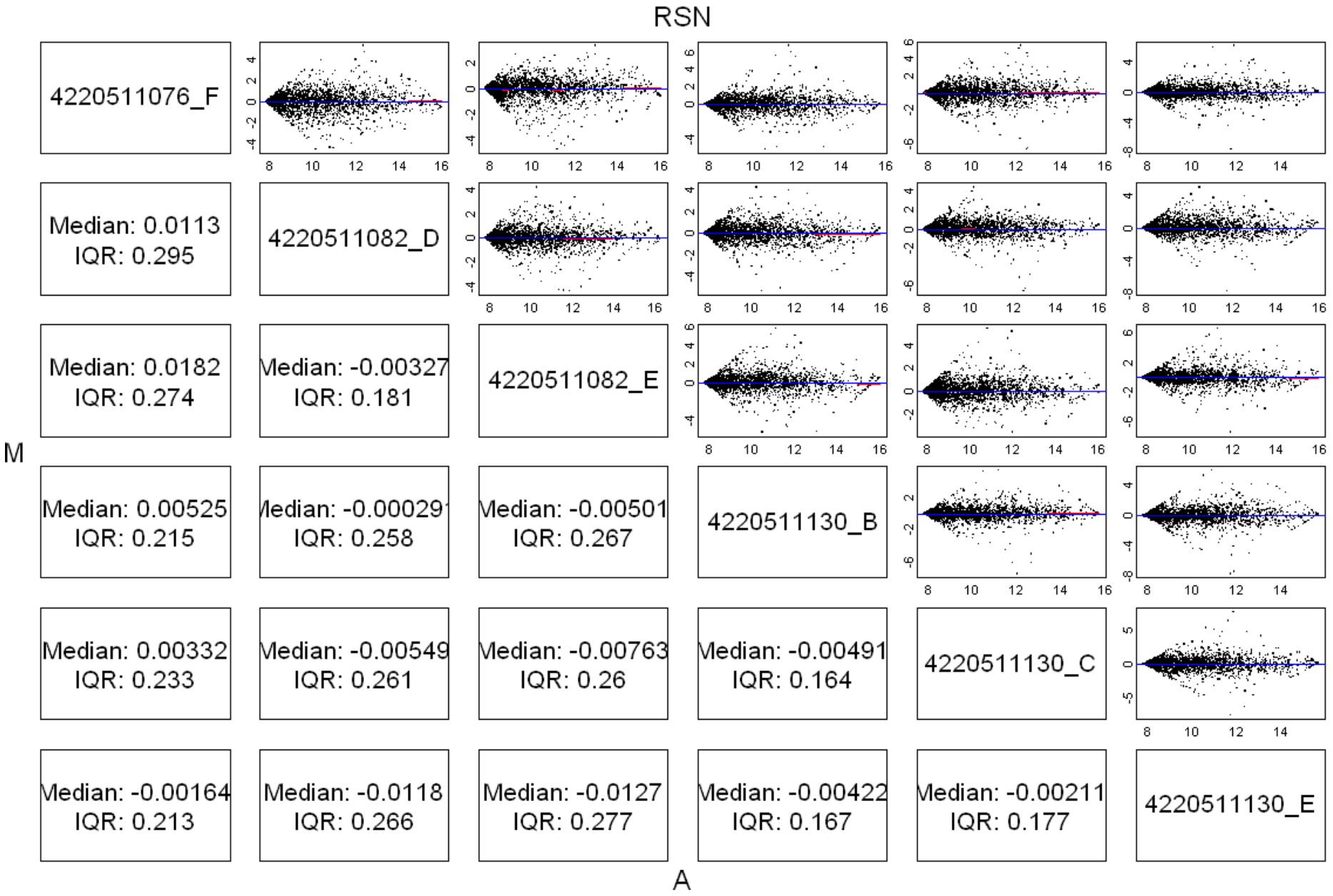
```
##       Array1  Array2  Array3
## [1,] "Rank1" "Rank3" "Rank5"
## [2,] "Rank3" "Rank4" "Rank1"
## [3,] "Rank5" "Rank2" "Rank2"
## [4,] "Rank2" "Rank1" "Rank4"
## [5,] "Rank4" "Rank5" "Rank3"
```

```
##   Array1 Array2 Array3
## A      1      4      9
## B      4      5      1
## C      9      3      3
## D      3      1      5
## E      5      9      4
```

# Robust Spline Normalization

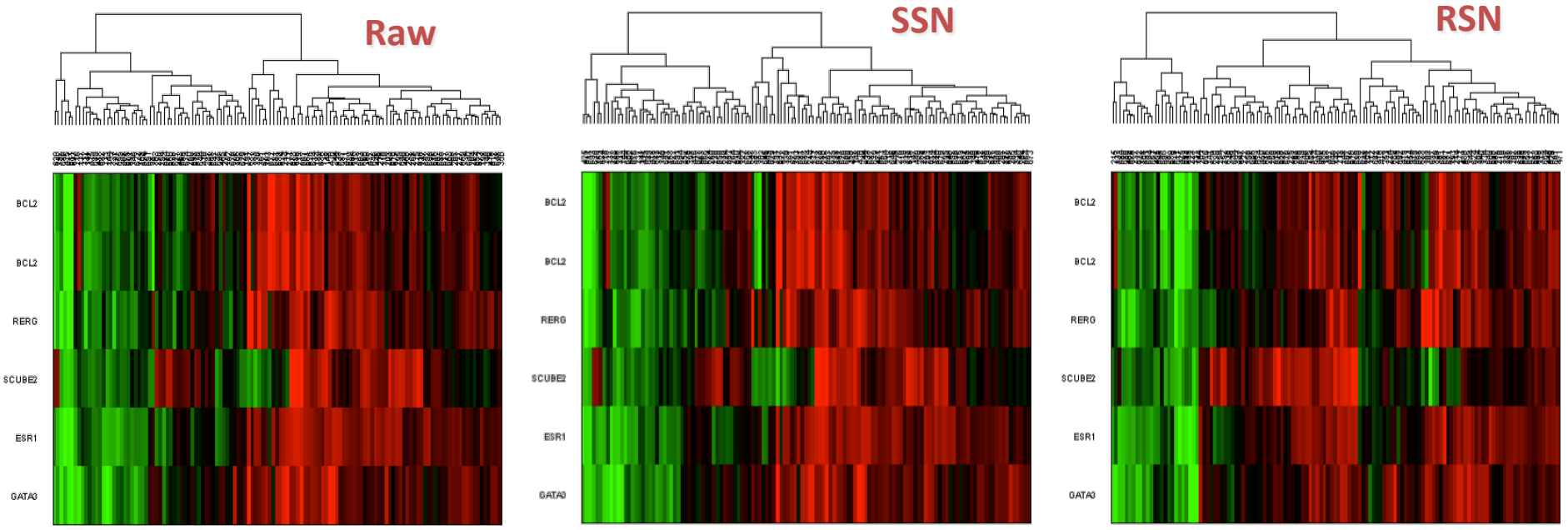


# Robust Spline Normalization

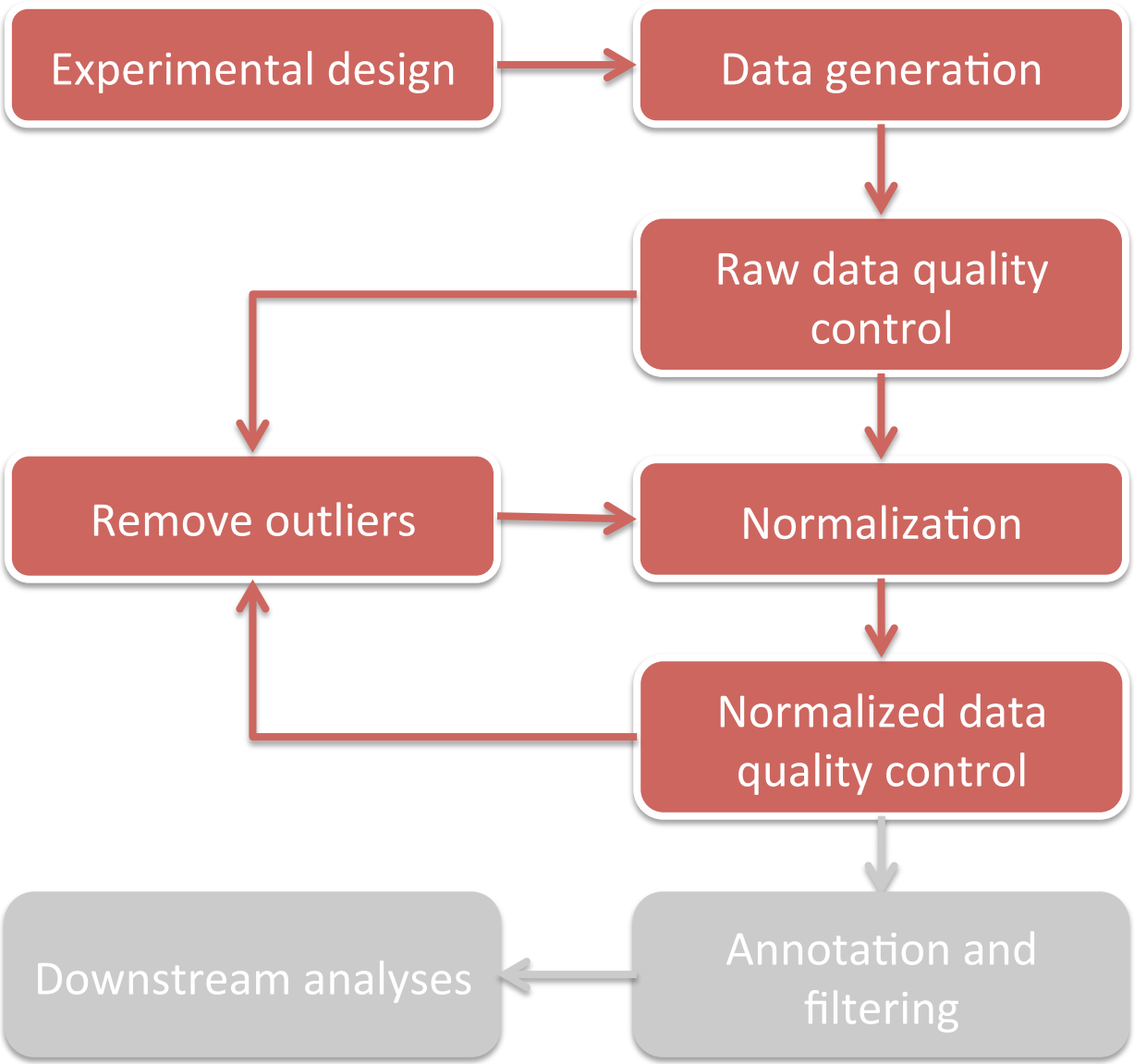


# How to choose?

- Schmid R et al. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. BMC Genomics. 2010 Jun 2;11:349
- Diagnostic plots (e.g. MA-plot, PCA)
- Search for known patterns in the data
- e.g. Expression of ESR1 gene and correlated genes in breast cancer

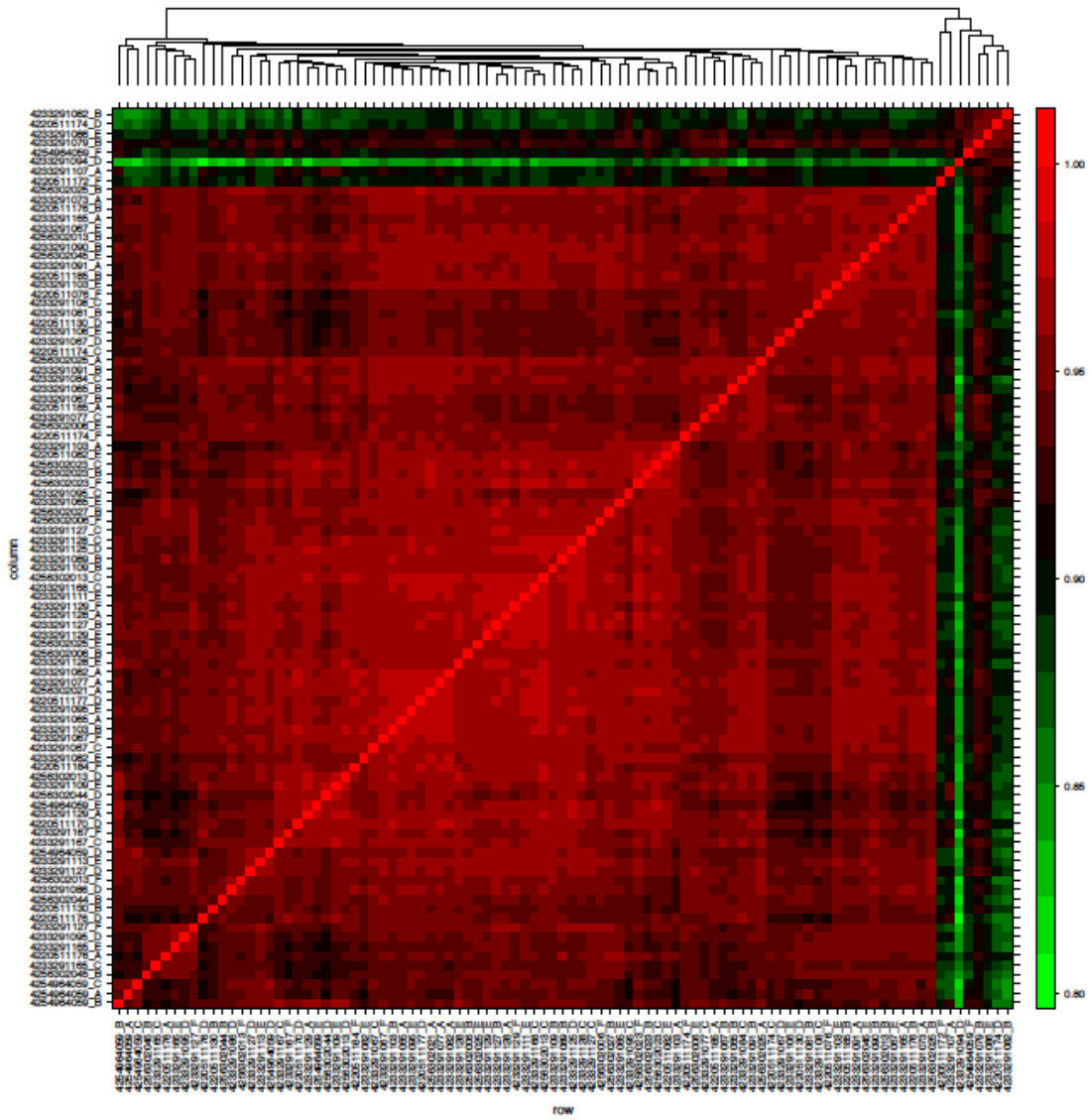


# The workflow



# Normalized data quality control

Sample-sample correlation after normalization

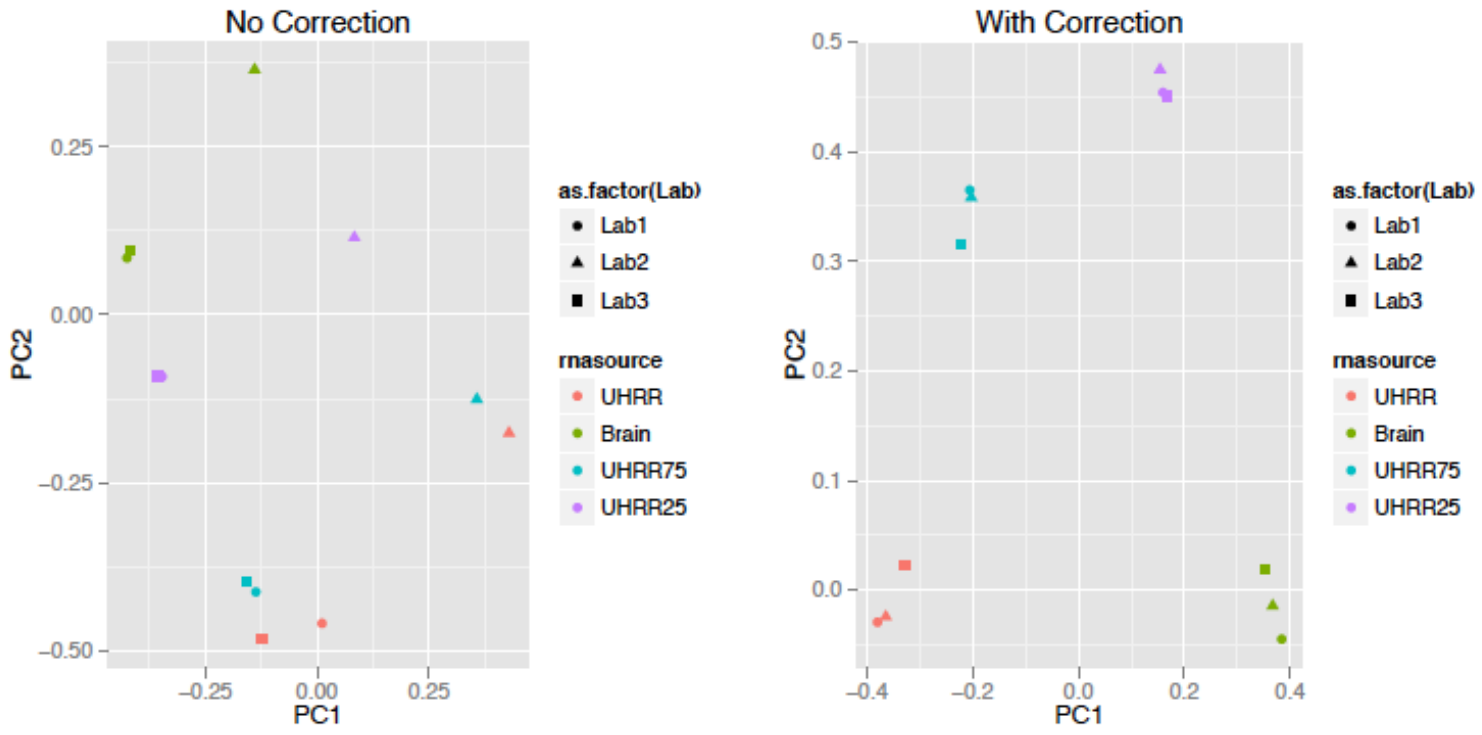


# Normalized data quality control

Normalized and summarized Illumina data can be treated similar to other microarray data (rows as genes, columns as arrays)

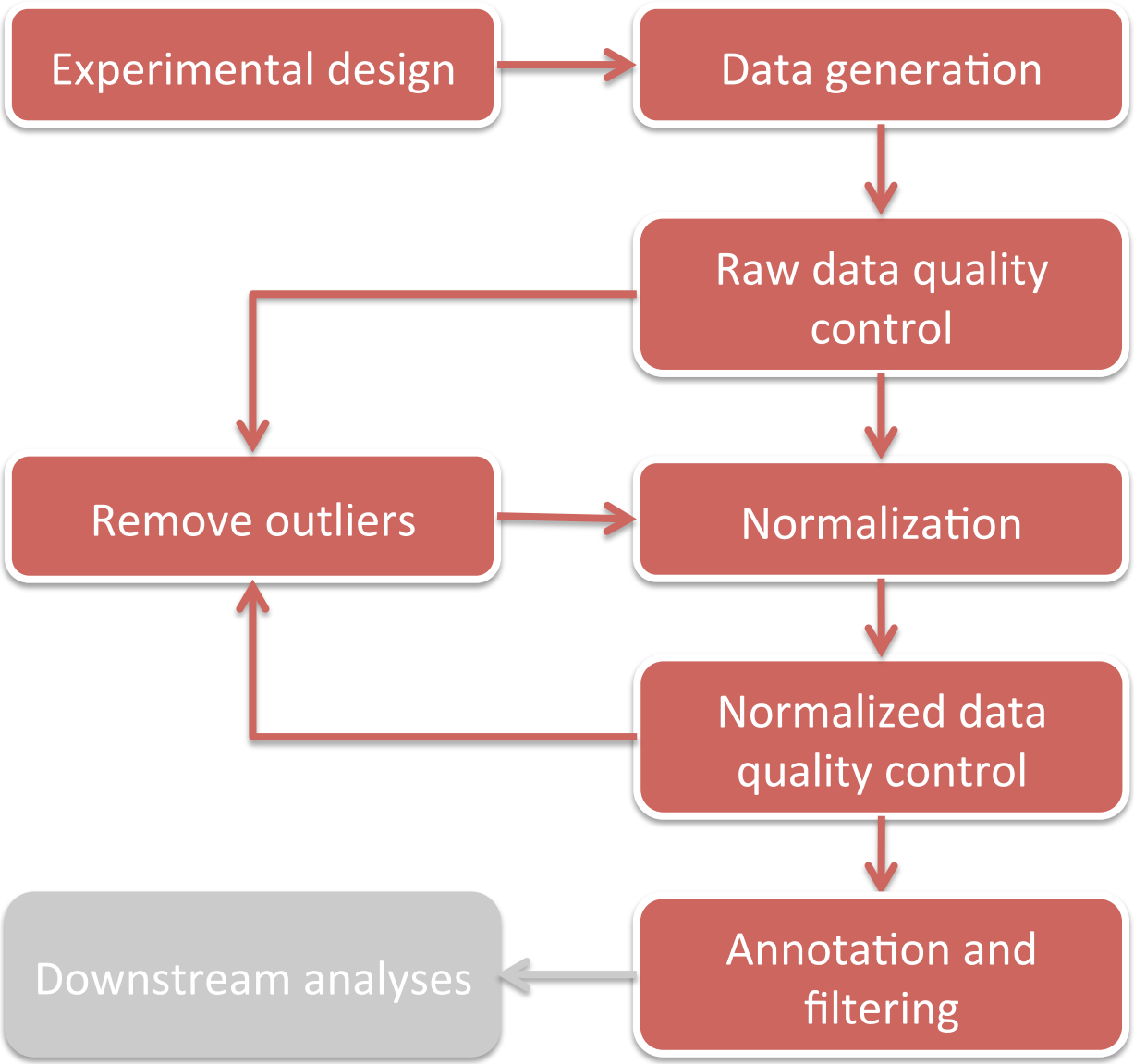
Diagnostic plot can highlight the presence of outliers or batch effects

**Example**: Experiment with Brain and Reference RNA hybridised at different labs




Several methods available, e.g. COMBAT (Johnson WE et al. Biostatistics. 2007)

# The workflow



# Probe annotation

- By default, each probe is assigned a numeric Code (ArrayAddress) if analysed at the bead-level, or a manufacturer identifier (ILMN_)

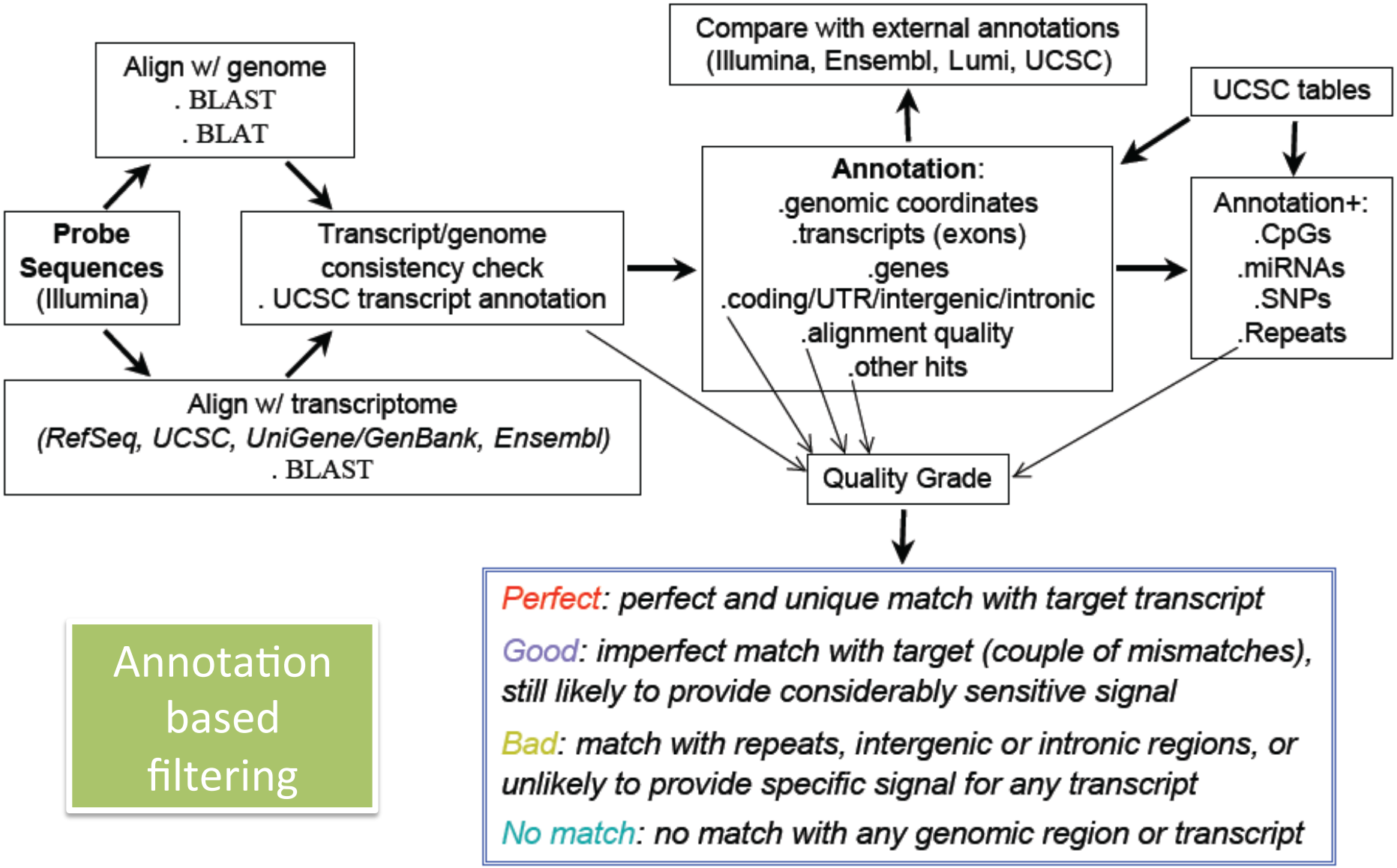
- The IDs need to be converted so that we could recognise our favourite genes in the data e.g. TP53, BRCA1 or relate the findings to a biological function

- Annotation can change over time (changes in transcript annotation, e.g. RefSeq/Ensembl database)

- Several factors can affect probe quality: probe uniqueness, position within the transcript, alternative splicing, presence of SNPs

- Original annotation provided by the manufacturer might be suboptimal (*Barbosa-Morais NL et al. NAR 2010*)

# Probe annotation




Annotation based filtering

Barbosa-Morais NL et al. NAR 2010

# Probe filtering

- A significant proportion of probes has signals not significantly different from background
- In Illumina microarrays a detection p-value is computed for each probe
- A significant p-value indicate that the probe has a signal significantly higher than the background (estimated through the negative controls)
- Probes not detected in most samples can be filtered out (~1/3 removed)
- Other parameters can be used, e.g. interquantile range (IQR)

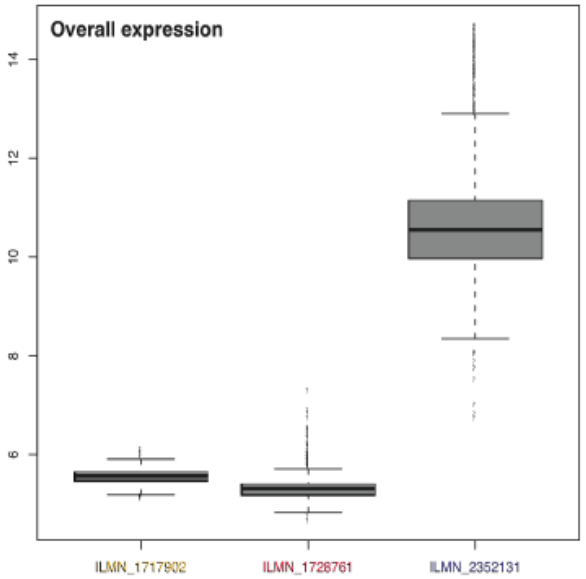
# Probe filtering

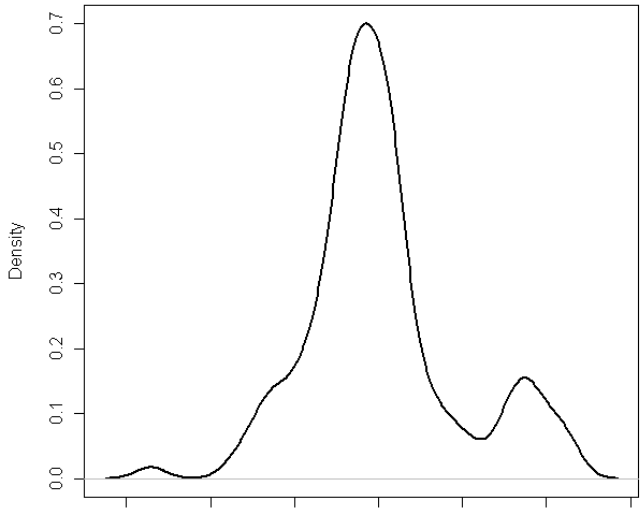
- Filtering helps to reduce the multiple testing problem

- Example: 50000 probes, differential analysis between two groups of interest. At a significance level of $p<0.01$, 500 (50000 x 0.01) probes would be differentially expressed by chance!

# Multiple probes per gene

- For many genes, more than one probe is present
- However, not all probes perform well
- Example:
  - ERBB2 gene, 3 probes available
  - Amplified in a subset of breast cancer ➔ bimodal signal expected
  - Only one of the three probes give a reliable signal



# Gene level data

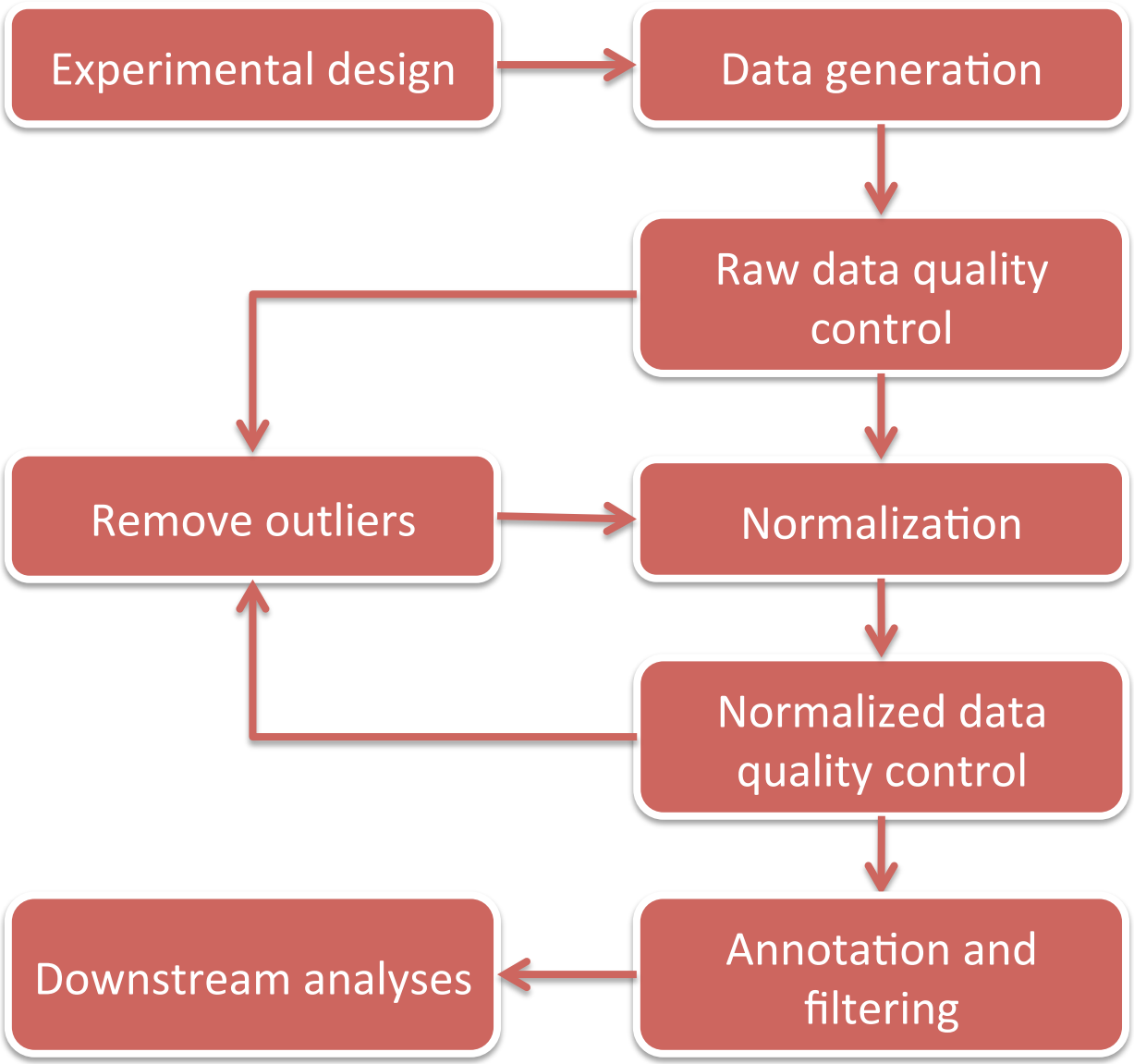
**Do not just average signals from different probes**

(one of the possible outputs in GenomeStudio)

**Do:**

- Take into account annotation
- Select the best performing probe (e.g. highest detection rate, highest IQR)
- Keep data at probe level (if you can)

# The workflow



# Downstream analyses

The analyses to answer your biological question will fall in one of three major categories:

- **Class discovery**
  - When classes are unknown (e.g. discovery of subtypes in a tumour type)
  - Approaches: hierarchical clustering, K-means clustering, PCA…
- **Class comparison**
  - When the groups or classes are known and we want to identify genes (or pathways) associated with them (e.g. treated/untreated cells)
  - Approaches: t-test, linear models, pathway analysis…
- **Class prediction**
  - When we want to identify a set of genes able to accurately predict the classes of interest in independent data
  - Approaches: PAM, SVM, survival analysis…

# Take home messages

- Microarray technology allowed the rise of -omics studies
- Well developed technology
- Plenty of tools for the analysis
- Expertise and experience of the bioinformatician are still crucial
- Computationally less demanding than RNA-seq
- The workflow for the analysis of Illumina gene expression data is partially valid for other Illumina microarrays, other commercial platforms and (to some extent) for sequencing based approaches

# References

Cairns JM, Dunning MJ, Ritchie ME, Russell R, Lynch AG. BASH: a tool for managing BeadArray spatial artefacts. Bioinformatics. 2008 Dec 15;24(24):2921-2

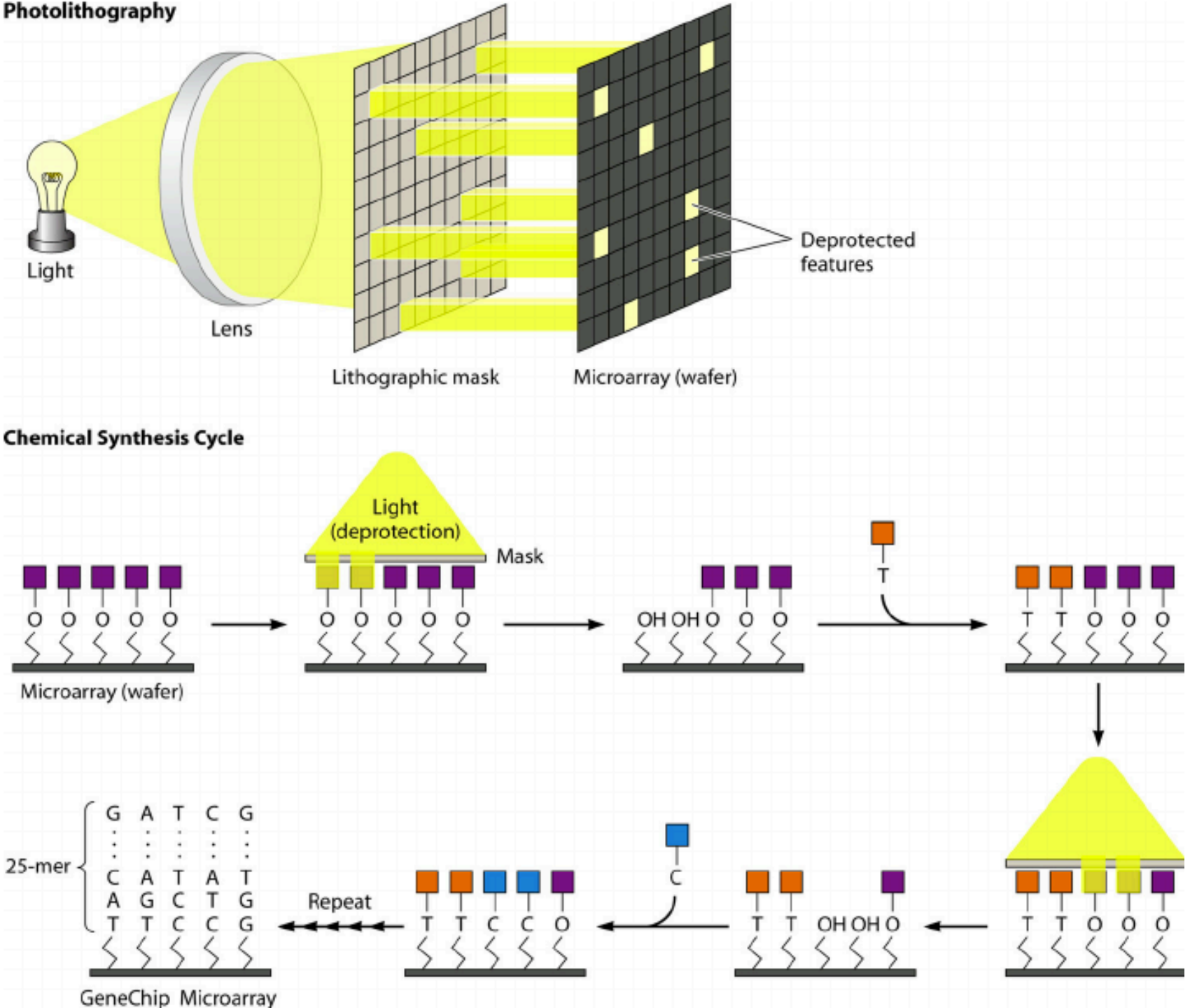
Schmid R, Baum P, Ittrich C, Fundel-Clemens K, Huber W, Brors B, Eils R, Weith A, Mennerich D, Quast K. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. BMC Genomics. 2010 Jun 2;11:349

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):733-9

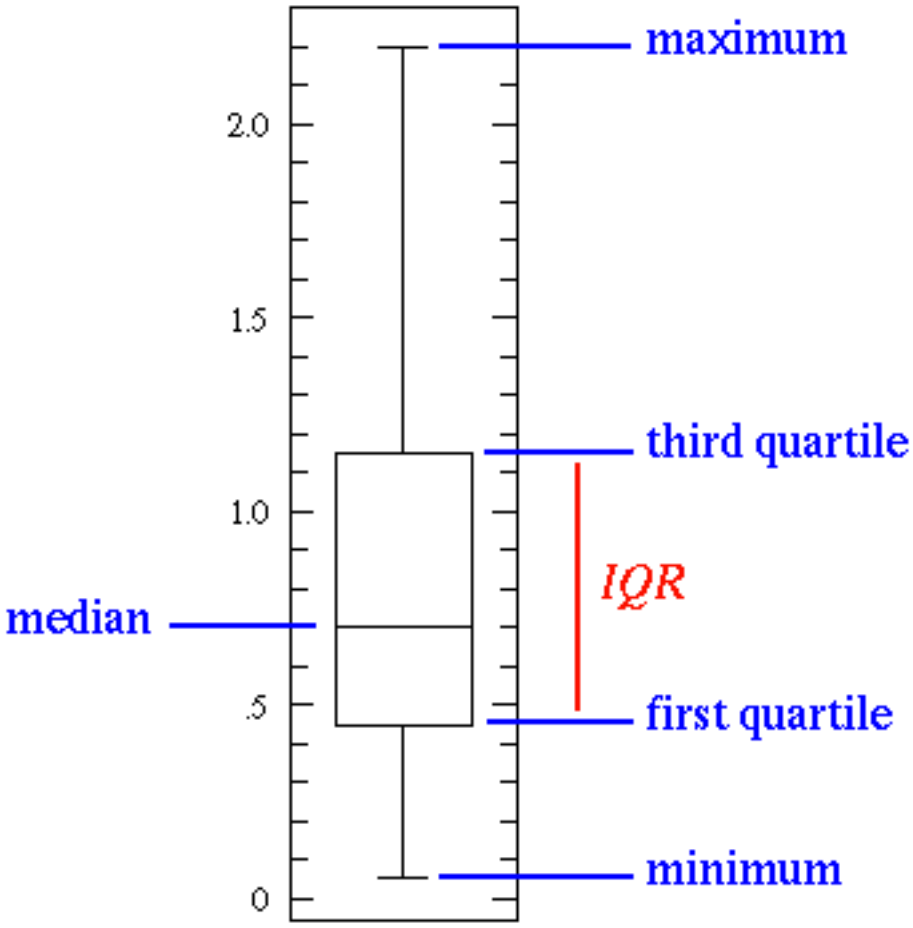
Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118-27

Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, Lynch AG, Tavaré S. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res. 2010 Jan;38(3):e17

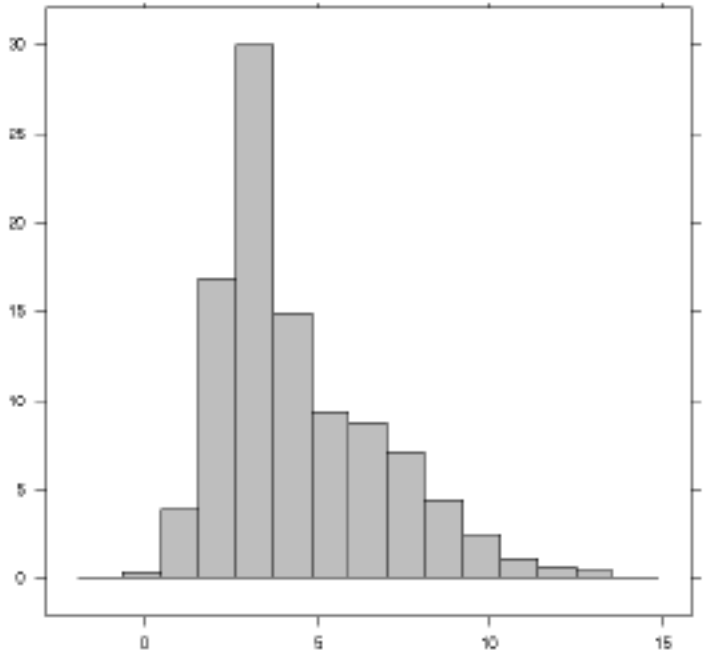
# Supplementary slides


Photolithography
Light
Lens
Lithographic mask
Microarray (wafer)
Deprotected features
Chemical Synthesis Cycle
Microarray (wafer)
Light (deprotection)
Mask
T
C
Repeat
25-mer
GeneChip Microarray

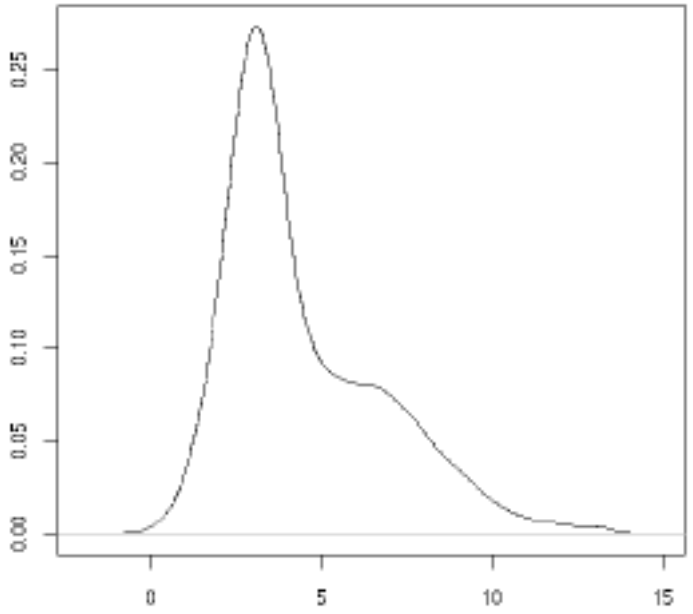
# Box-and-whisker plot

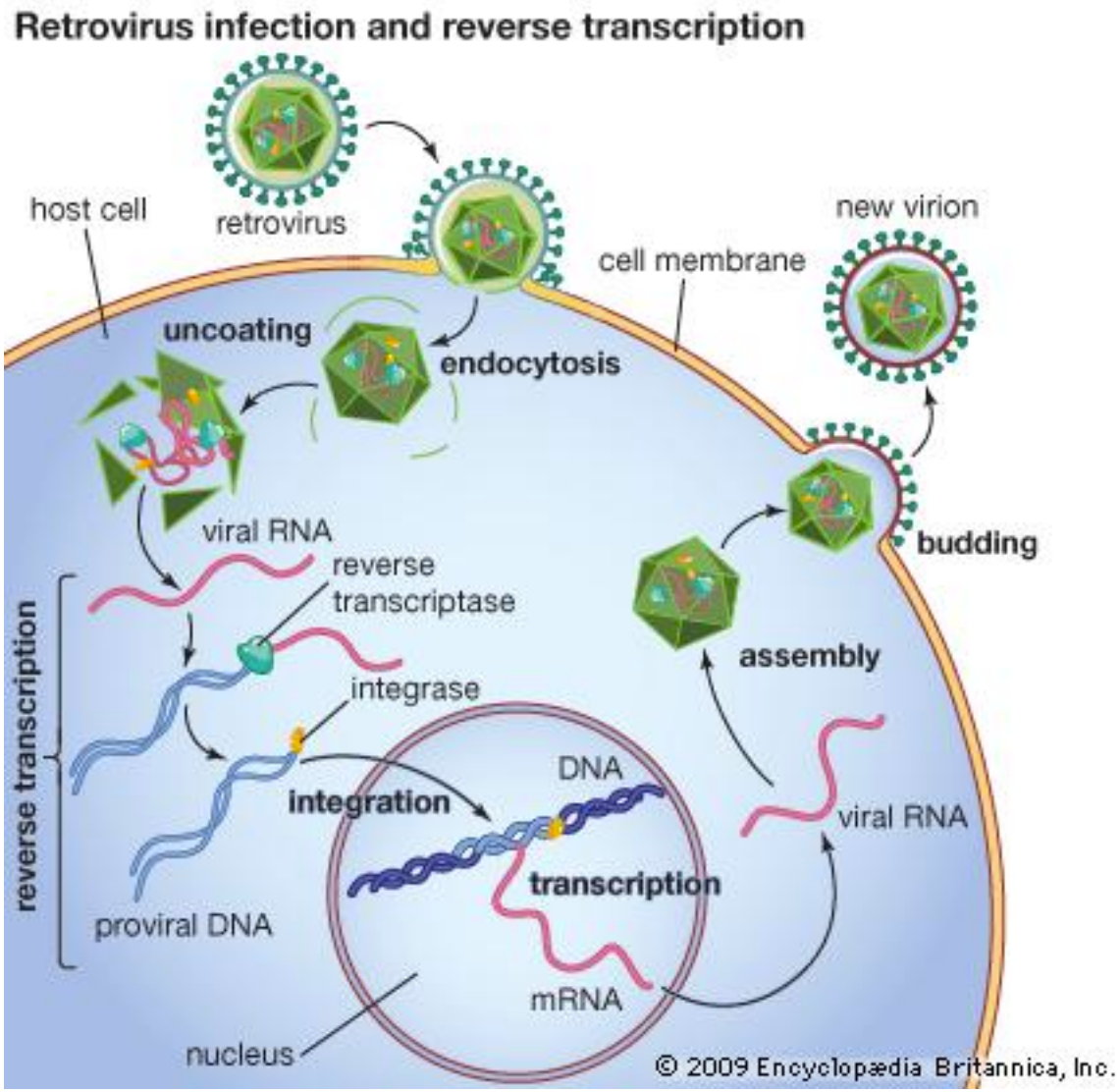


Histogram:
a graphical display where the data is grouped into ranges and then plotted as bars




Density plot: graph of a continuous probability distribution






Reverse transcriptase is also a fundamental component of a laboratory technology known as reverse transcription-polymerase chain reaction (RT-PCR), a powerful tool used in research