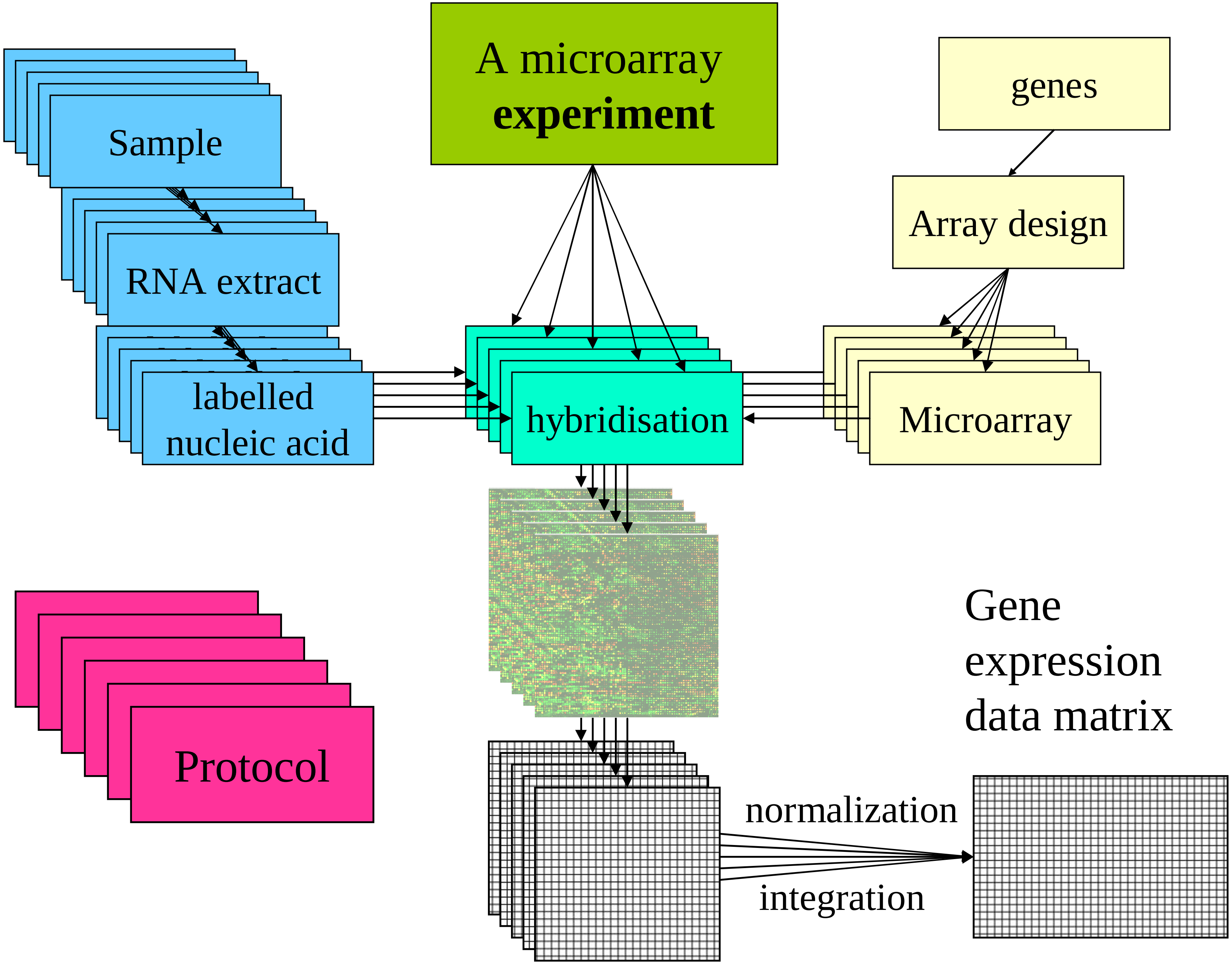
# Introduction to Microarray Data Analysis and Gene Networks

Alvis Brazma

European Bioinformatics Institute

# Content on today’s lecture

- What are microarrays measuring?
  - From fluorescence intensities to transcript abundance
- What is microarray data normalisation
- Look at some real experiments
  - Experimental design, experimental factors, replicates
- Open an account in Expression Profiler
- Load some data, normalise some data



A microarray **experiment**

Sample

RNA extract

labelled nucleic acid

hybridisation

genes

Array design

Microarray

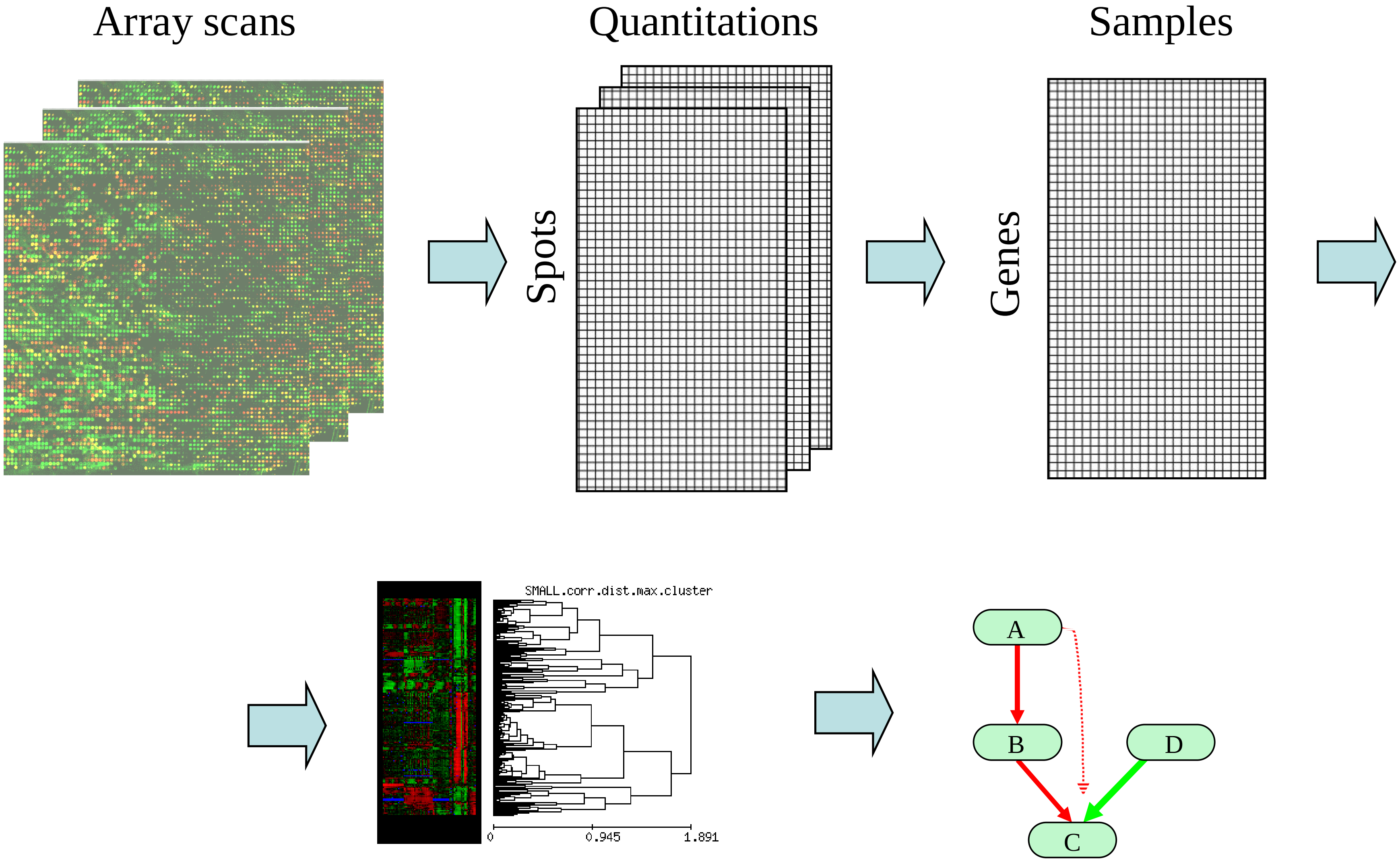
Protocol

Gene expression data matrix

normalization

integration

# Steps in microarray data processing



Array scans

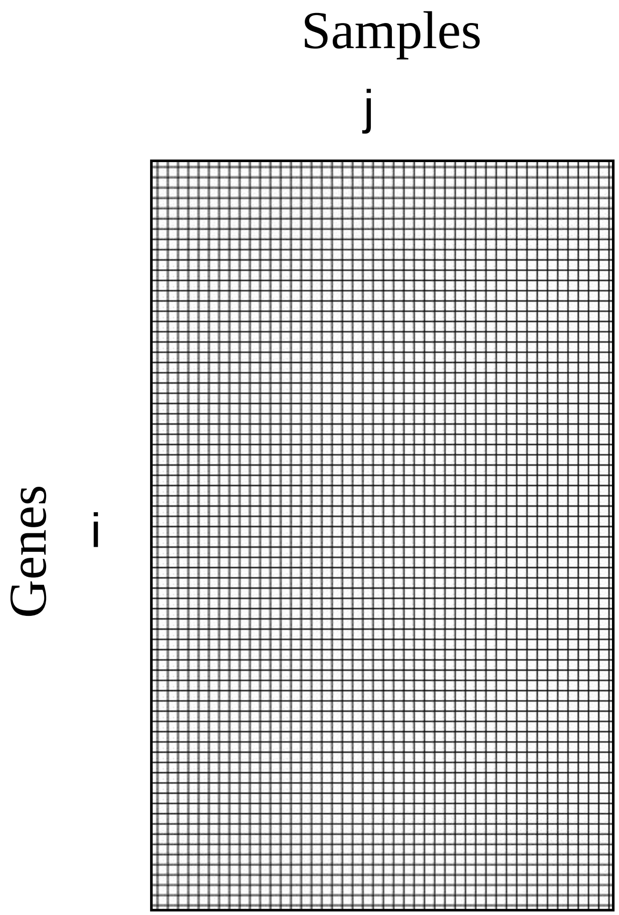
Quantitations

Samples

Spots

Genes

# The goal of data normalisation - Gene Expression Data Matrix



X(i,j) – amount of the RNA of the i-th gene in the j-th sample

# What are we measuring?

$$\text{Florescence Intensity} = X \times a \times b \times e + n + o$$

- X - is amount of RNA – this is what we are interested in
- a – hybridisation (sample and array (batch)) effect – particular experiment dependent
- b - sequence effect (probe efficiency) – i.e., what are the hybridisation properties of the particular DNA molecule – particular gene dependent
- e – multiplicative error
- n - non-specific binding and cross-hyb
- o - optical effects (from scanner)

# What are we measuring?

$$\text{Florescence Intensity} = X \times a \times b \times e \; + \; n \; + \; o$$

- X - is amount of RNA – this is what we are interested in
- a – hybridisation (sample and array (batch)) effect – particular experiment dependent
- b - sequence effect (probe efficiency) – i.e., what are the hybridisation properties of the particular DNA molecule – particular gene dependent
- e – multiplicative error
- n - non-specific binding and cross-hyb
- o - optical effects (from scanner)
  - Assume that Intensity is already adjusted for these two or they are negigible

# What are we measuring?

$$\text{Florescence Intensity} = X \times a \times b \times e$$

- X - is amount of RNA – this is what we are interested in
- a – hybridisation effect – particular experiment dependent
- b - sequence effect (probe efficiency) – i.e., what are the hybridisation properties of the particular DNA molecule – particular gene dependent
- e – multiplicative error

# What are we measuring?

Florescence Intensity(i,j) = X(i,j) × a(j) × b(i) × e

- X(i,j) - amount of RNA for i-th gene in j-th sample
- a(j) – hybridisation effect – depends on the particular hybridisation j, but does not depend on gene i
- b(i) - sequence effect (probe efficiency) – depends on the particular sequence i, but does not depend on sample j
- e – multiplicative error – strictly speaking e(i,j)

# What are we measuring?

$$\text{Intensity}(i,j) = X(i,j) \times a(j) \times b(i) \times e$$

$$\log(\text{Intensity}(i,j)) =$$

$$= \log(X(i,j)) + \log(a(j)) + \log(b(i)) + \varepsilon =$$

$$= Y(i,j)$$

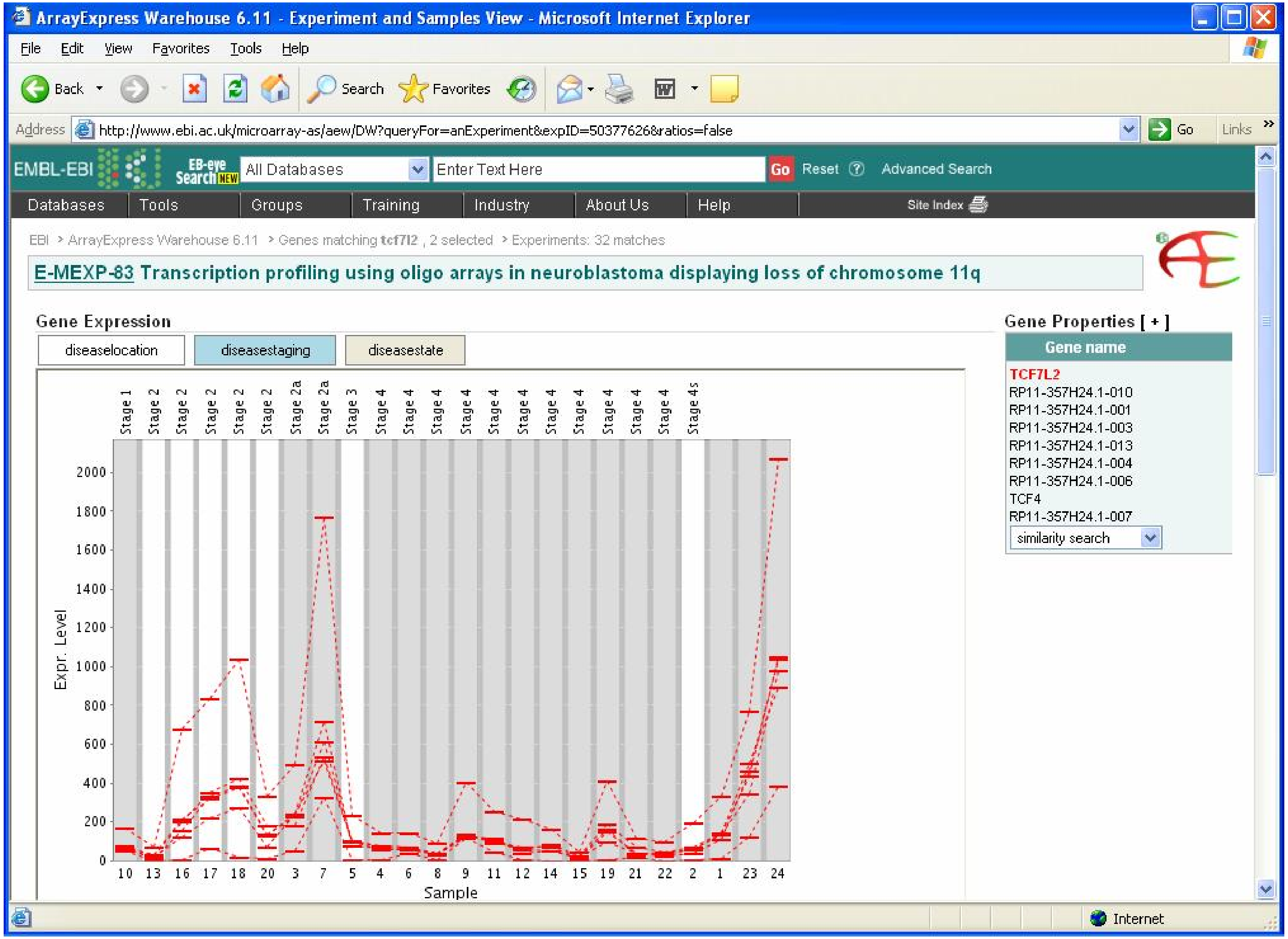
$$\log(X(i,j)) = Y(i,j) - (\log(a(j)) + \log(b(i)) + \varepsilon)$$

# What are we interested in?

$$\log(X(i,j)) = Y(i,j) - (\log(a(j)) + \log(b(i)) + \varepsilon)$$

- X(i,j) – the amount of RNA of ith gene in jth sample – the ‘expression level’
- All we need to do is to assess hybridisation effect a(j) and sequence effect b(i)

# Sequence effects are large



ArrayExpress Warehouse 6.11 - Experiment and Samples View - Microsoft Internet Explorer

Address: http://www.ebi.ac.uk/microarray-as/aew/DW?queryFor=anExperiment&expID=50377626&ratios=false

EBI > ArrayExpress Warehouse 6.11 > Genes matching **tcf7l2** , 2 selected > Experiments: 32 matches

E-MEXP-83 Transcription profiling using oligo arrays in neuroblastoma displaying loss of chromosome 11q

Gene Properties [ + ]

| Gene name |
|---|
| TCF7L2 |
| RP11-357H24.1-010 |
| RP11-357H24.1-001 |
| RP11-357H24.1-003 |
| RP11-357H24.1-013 |
| RP11-357H24.1-004 |
| RP11-357H24.1-006 |
| TCF4 |
| RP11-357H24.1-007 |

# Relative measures

- For the same gene i in two different samples j and h

$$Y(i,j)- Y(i,h) = \log(ratio) =$$

$$= \log(X(i,j)) + \log(a(j)) + \log(b(i)) + \varepsilon_1$$
$$- \quad (\log( X(i,h)) + \log(a(h)) + \log(b(i)) + \varepsilon_2) =$$

$$= \log(X(i,j)/ X(i,h)) + \log(a(j)/ a(h)) + \partial$$

- $\partial = \varepsilon_1 - \varepsilon_2$, error
- All we need to do is to estimate a(i)/ a(h)– the relative hybridisation effect

# Normalisation

$$Y(i,j) - Y(i,h) = \log(\text{ratio}) =$$

$$= \log(X(i,j)/ X(i,h)) + \log(a(i)/ a(h)) + \partial$$

- All we need to do is to estimate the relative sample effect a(i)/ a(h)
- this is known as **normalisation**

# One or two channel arrays?

- Two channel arrays estimate the ratio directly
- Depends on the question
  - If the question is only to compare two conditions (e.g., disease vs. normal), then this may be the right experiment
  - If there are more than two conditions to compare, it becomes more complicated to interpret the data
  - If more than two conditions are used, a 'reference' design if often used
- One channel experiments more easily to reuse the data
- Two channel arrays are cheaper – why?

# Relative measures – a catch

- For the same gene i in two different samples j and h

$$Y(i,j) - Y(i,h) = \log(\text{ratio}) =$$

$$= \log(X(i,j)) + \log(a(j)) + \log(b(i)) + \varepsilon_1$$
$$- \quad (\log( X(i,h)) + \log(a(h)) + \log(b(i)) + \varepsilon_2) =$$

$$= \log(X(i,j)/ X(i,h)) + \log(a(j)/ a(h)) + \partial$$

- For the derivation to be valid we need that b(j) indeed does not depend on the particular array used in the hybridisation j – arrays need to be very standardised

# Normalisation

$$\log(X(i,j)/ X(i,h)) \approx Y(i,j)- Y(i,h) - \log(a(j)/ a(h))$$

- All we need to do is to estimate the relative hybridisation effect a(j)/ a(h)
- *How to do this?*

# Normalisation

- Estimating the relative hybridisation effect a(j)/ a(h)
- *How to do this?* – By making assumptions

- Possibility 1 – as a(j) and a(h) are the same for all genes, if we knew the true expression ratio for one gene, we could make the estimate
- Possibility 2 – by using estimates of the average expression levels for all genes

# Normalisation

- Possibility 1 – as a(j) and a(h) are the same for all genes, if we knew the true expression ratio for one gene, we could make the estimate

$$\log(a(j)/ a(h)) = Y(*,j)- Y(*,h) - \log(X(*,j)/ X(*,h)) +$$

  - ‘House keeping genes’ – the genes that do not change the expression (in the particular experiment)
  - External controls
- Drawback – depends on our trust in a small number of genes

# Normalisation

- Possibility 2 – by using estimates of the average expression levels for all genes
  - The simplest version – assumption that the total (or average) expression does not change, i.e.,

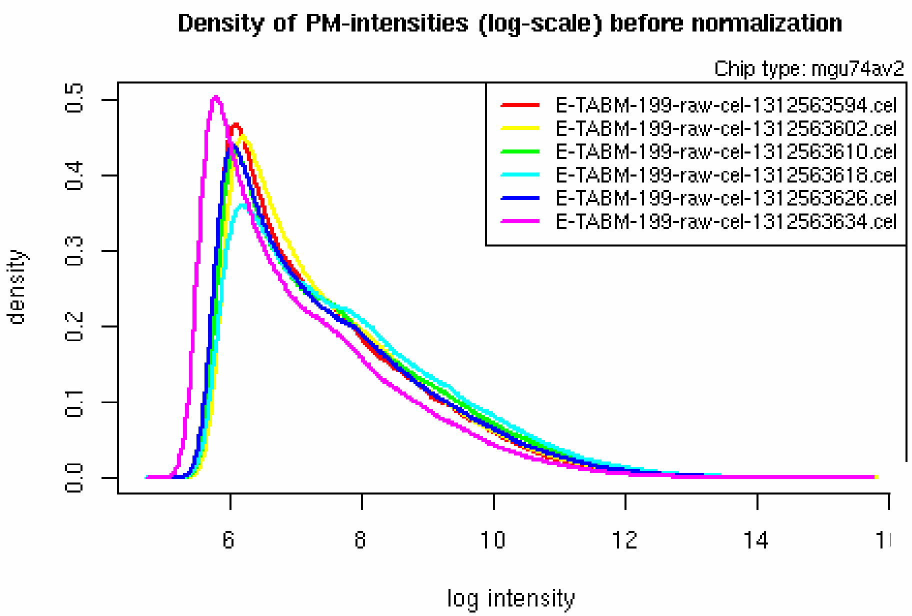
$$\Sigma_i X(i,j) = \Sigma_i X(i,h)$$

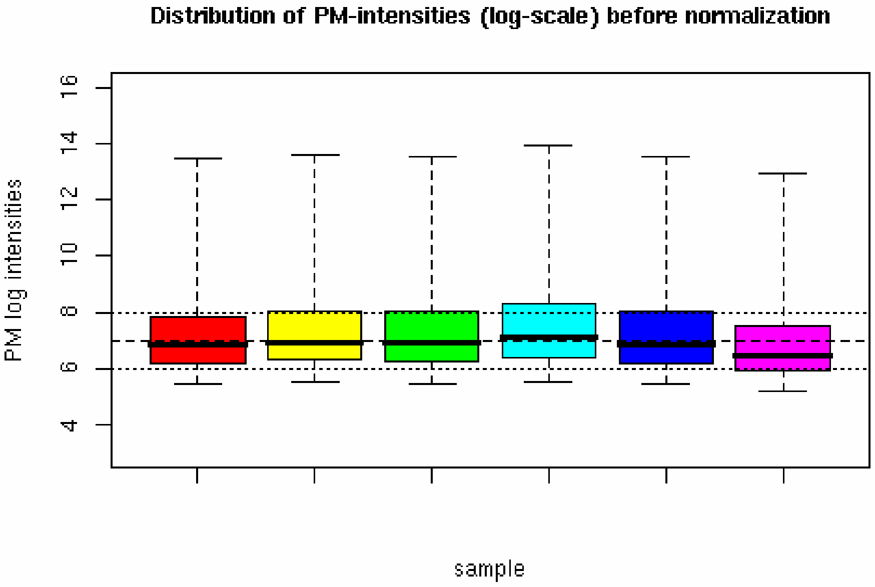
  - This is known as ‘total signal’ normalisation
  - The drawback – the error $\varepsilon$ depends on the expression level
- Possibility 2a - assume that most genes do not change their expression

# Normalisation

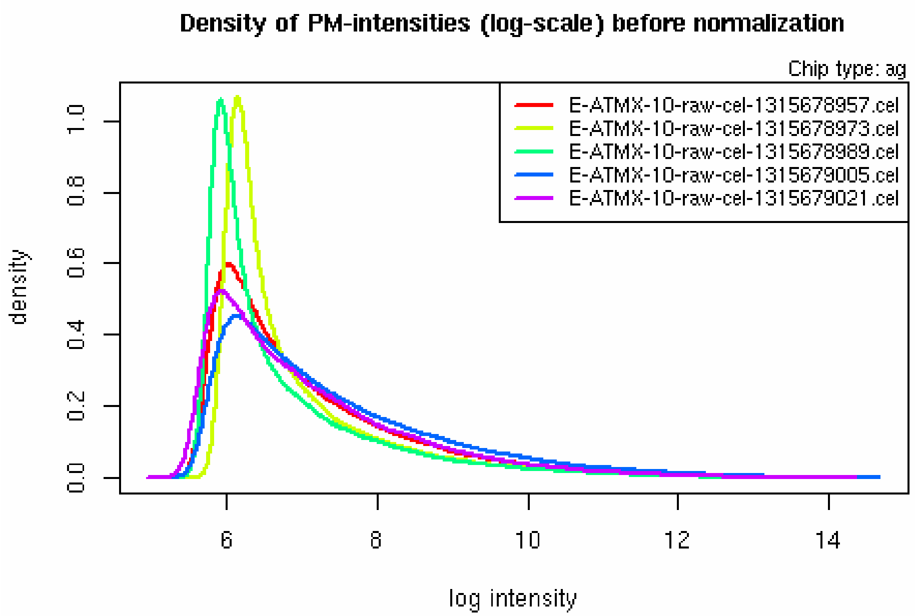
- Possibility 2a - assume that most genes do not change their expression – a quantile normalisation
- Note – we are moving away from the assumption that the normalisation factor is the same for all genes – although it won’t be sequence specific, it will be intensity (or expression level) specific
- Based on the assumption that the intensity distributions are the same in both hybridisations
- Can be easily generalised on more than two arrays

# Probe intensity distributions in 6 Affymetrix mouse (mgu74av2) arrays

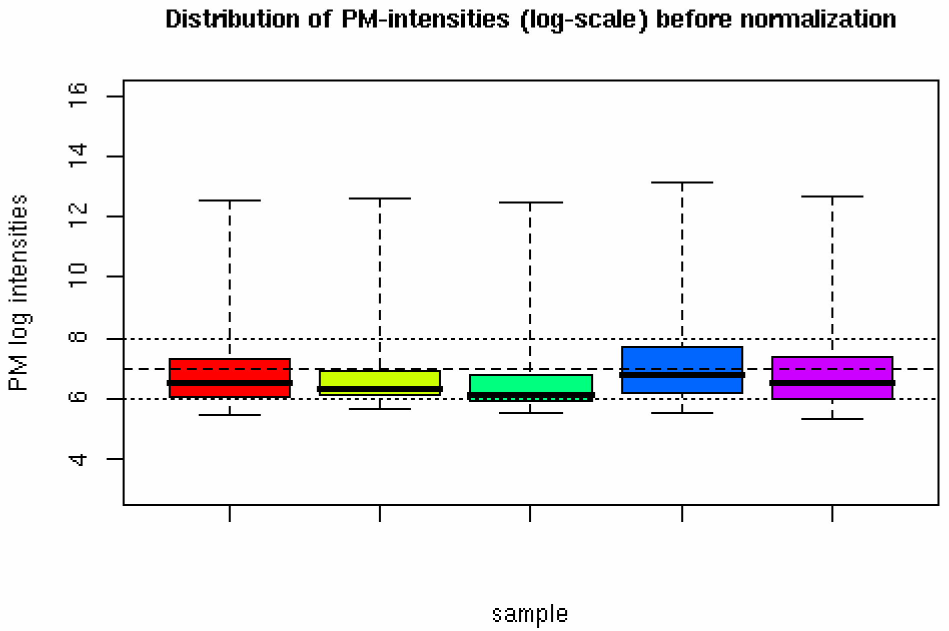




The assumption – the true expression value distributions are the same in all arrays



Density of PM-intensities (log-scale) before normalization

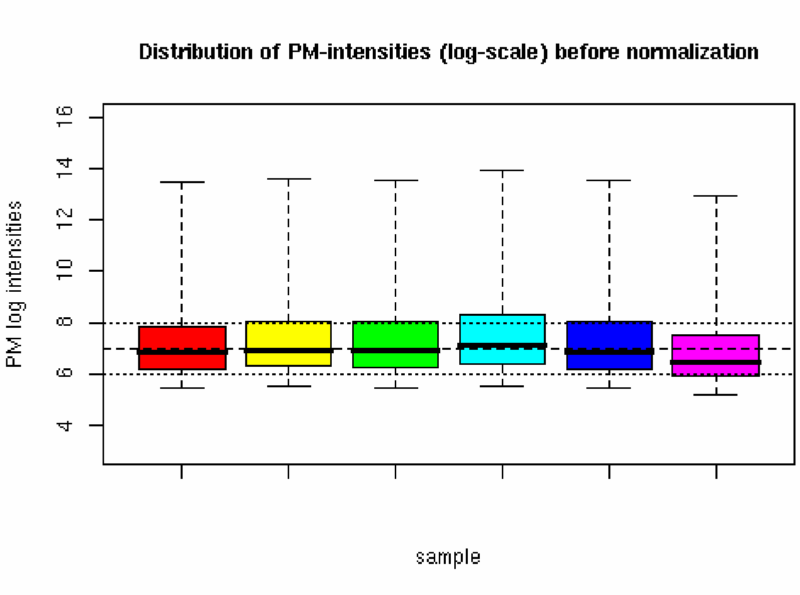


Distribution of PM-intensities (log-scale) before normalization

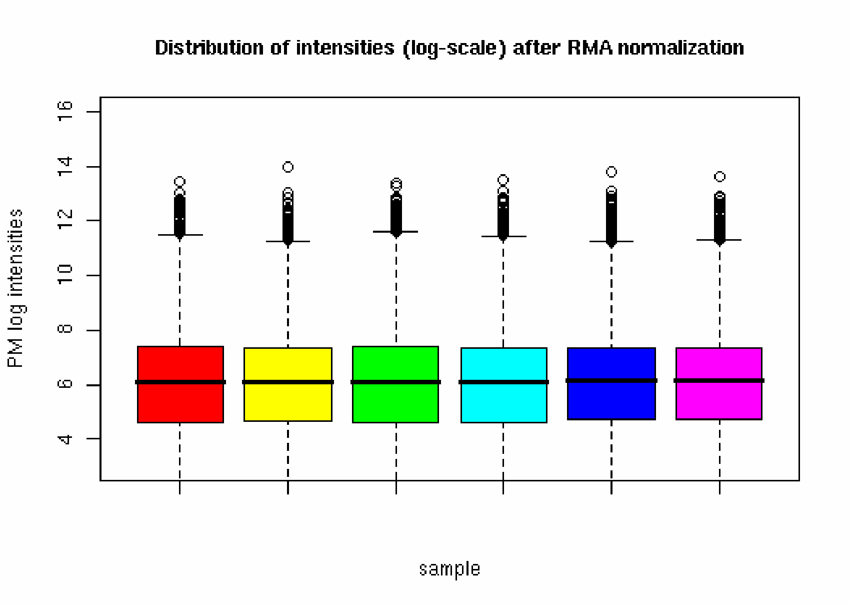
# **R**obust **M**ulti-array **A**verage (RMA) normalisation

- Order each column of data (i.e. the points from each array) from highest to lowest expression value
- Calculate the mean of the highest expression value in each column
- Replace each highest value in the original array by that mean value
- Repeat the procedure using the second-highest value in each column, and continue until all values have been replaced by their respective means

# Before and after RMA normalisation



Before



After

# RMA normalisation – steps from intensities to (pseudo) expression levels

1. Subtract the background intensity from each intensity value (if this has not already been done), in a way that ensures that all expression values are positive.
2. Take the log to base 2 of each expression value.
3. Normalise the log data as follows:
   a) Order each column of data (i.e. the points from each array) from highest to lowest expression value
   b) Calculate the mean of the highest expression value in each column
   c) Replace each highest value in the original array by that mean value
   d) Repeat the procedure using the second-highest value in each column, and continue until all values have been replaced by their respective means
4. The obtained ‘expression values’ will be gene specific

# Normalisation methods for Affymetrix arrays

- RMA
- GCRMA – the same as RMA, but additionally using sequence properties (what is the proportion of G or C vs A or T in the sequence) to account better for the sequence factor b(i,j), and thus making
- MAS5 – the original Affymetrix method, uses mismatch probe signal as a background