Analysis of Affymetrix GeneChips

By Christina Curtis (updates by Benilton Carvalho)

Overview

- **Overview of Affymetrix GeneChip technology**
  - Preliminaries and terminology
  - GeneChip manufacturing and expression array design
  - Sample preparation

- **Low-level analysis & handling probe-level data**
  - Array quality assessment
  - Background correction
  - Normalisation
  - Computing gene expression measures

- **High-level analysis**
  - Detecting differential expression
  - Correcting for multiple testing
  - Downstream analysis: functional enrichment, machine learning

- **Other GeneChip arrays**

RESEARCH ARTICLE

Expression monitoring by hybridization to high-density oligonucleotide arrays

David J. Lockhart*, Helin Dong, Michael C. Byrne[1], Maximillian T. Follettie[1], Michael V. Gallo[1], Mark S. Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi[1], Heidi Horton[1], and Eugene L. Brown[1]

Nat Biotechnol, 1996, 14(13):1675–1680

review

© 1999 Nature America Inc. · http://genetics.nature.com

High density synthetic oligonucleotide arrays

Robert J. Lipshutz, Stephen P.A. Fodor, Thomas R. Gingeras & David J. Lockhart

Affymetrix, Inc. 3380 Central Expressway, Santa Clara, California 95051, USA. e-mail: rob_lipshutz@affymetrix.com

Nat Genetics supplement, 1999, 21:20–24

To date, more than 10,000 scientific articles using or reviewing Affymetrix technology, perhaps the most widely used commercially available platform

Affymetrix GeneChip Arrays

- Employ short, 25 nt oligonucleotide **probes** that are surface immobilized on a quartz wafer

- Genomic sequences of interest are represented by a **probe set** comprised of 11-20 different 25-mers
  - Allows for high sensitivity & reproducibility

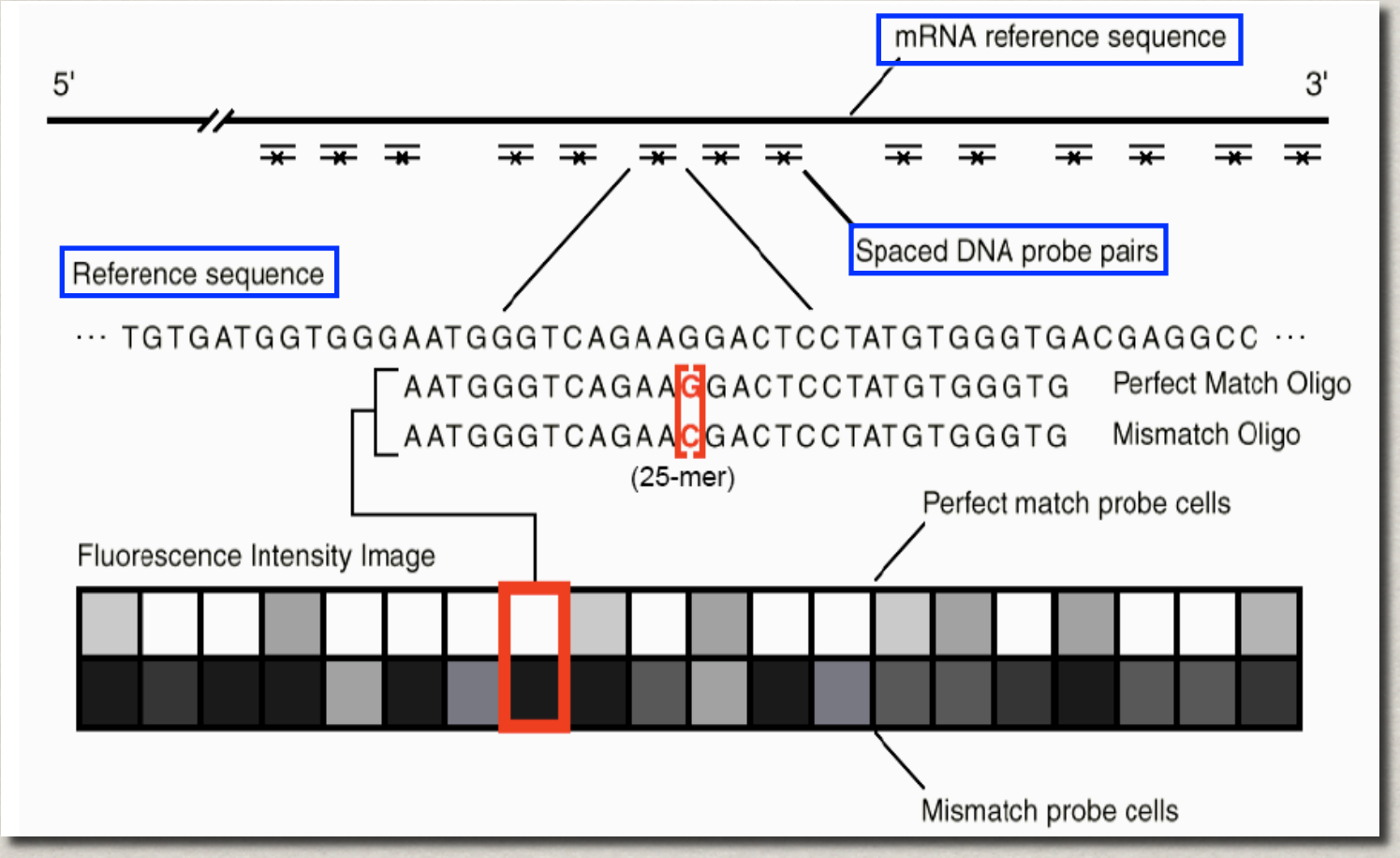
  - **Probe** = individual 25-mer sequence
  - **Probe set** = set of 25-mer sequences corresponding to a particular gene or EST
  - **Target** = labeled nucleic acids derived from sample of interest

Perfect Match vs Mis-Match Probes

- ✦In an *attempt* to measure non-specific binding, a paired set of **Mis-Match** sequences are included on *many** GeneChip arrays
- ✦**Perfect Match (PM)** probes = correct 25-mer sequence
- ✦**Mis-Match (MM)** probes = 25-mer sequence with middle (13th) nucleotide replaced by its complement
- ✦The two probes are referred to as a **probe-pair**

*Note that the new Exon arrays, Gene1.0 ST arrays, SNP 5.0/6.0 and Axiom arrays do not contain MM probes

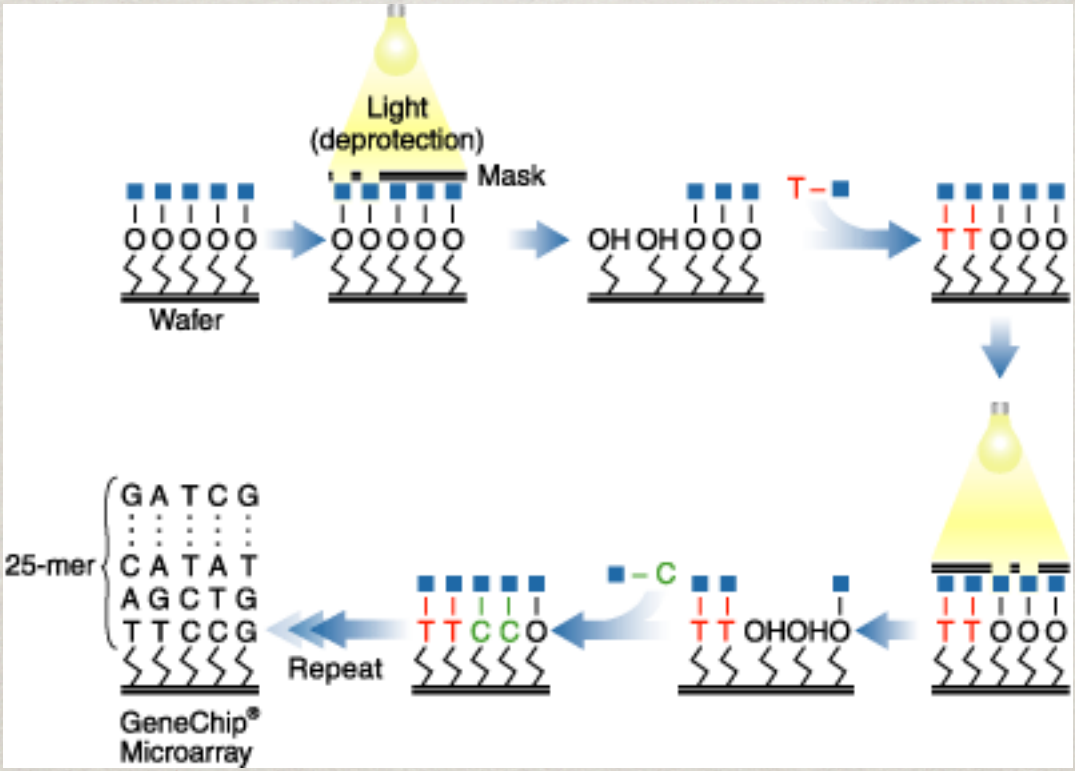
Affymetrix Probe design



Probe selection

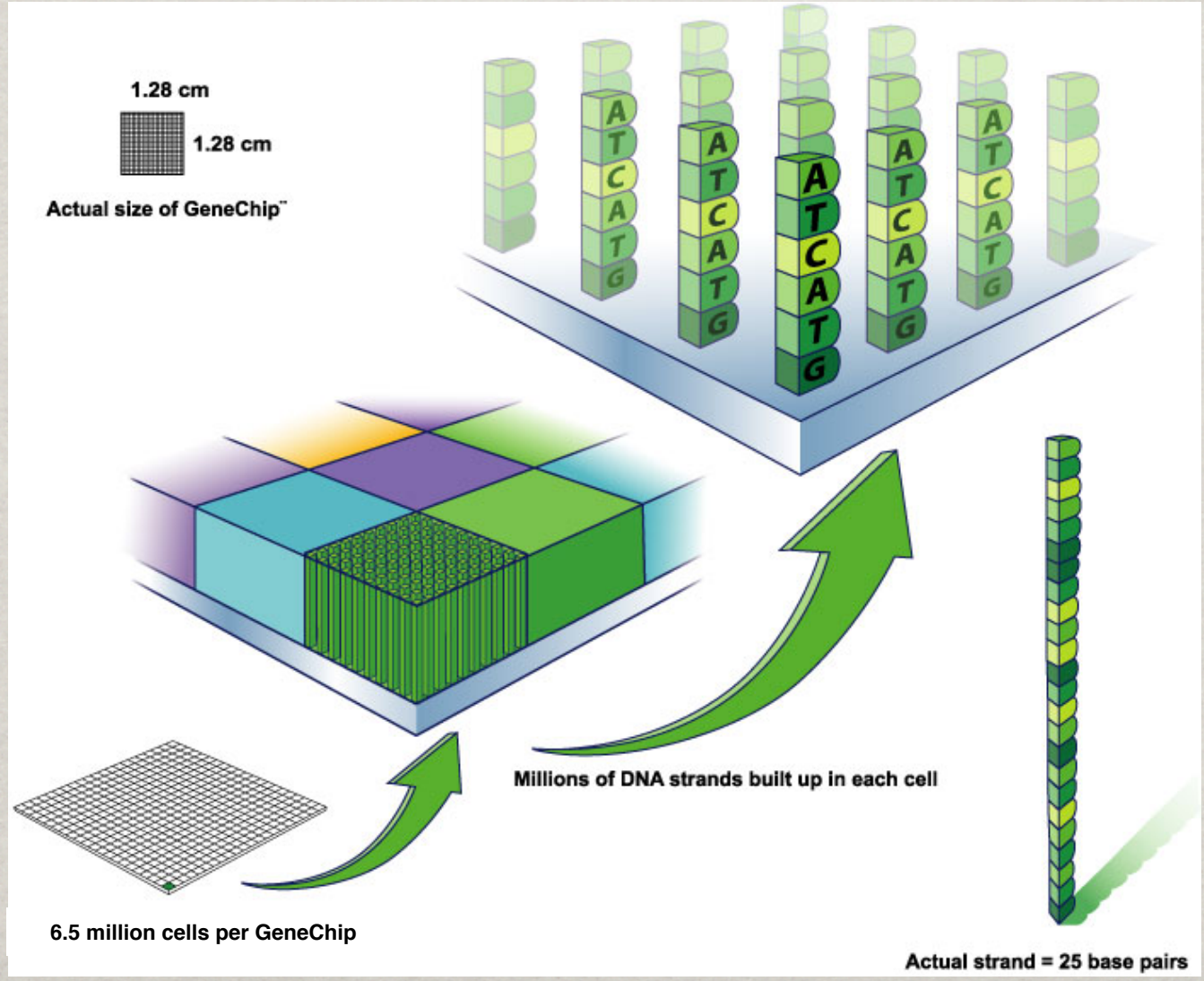
- ✦Dictated by intended use of the array
- ✦For expression arrays, probes representative of each transcript are selected based on a thermodynamic model of nucleic acid duplex formation to identify responsive probes
- ✦Other considerations for 3' expression arrays include:
  - ✦Minimize cross hybridization
  - ✦Uniqueness
  - ✦Spatial separation
  - ✦Restricted to first 600 bases proximal to polyadenylation site

Synthesis of affymetrix genechips

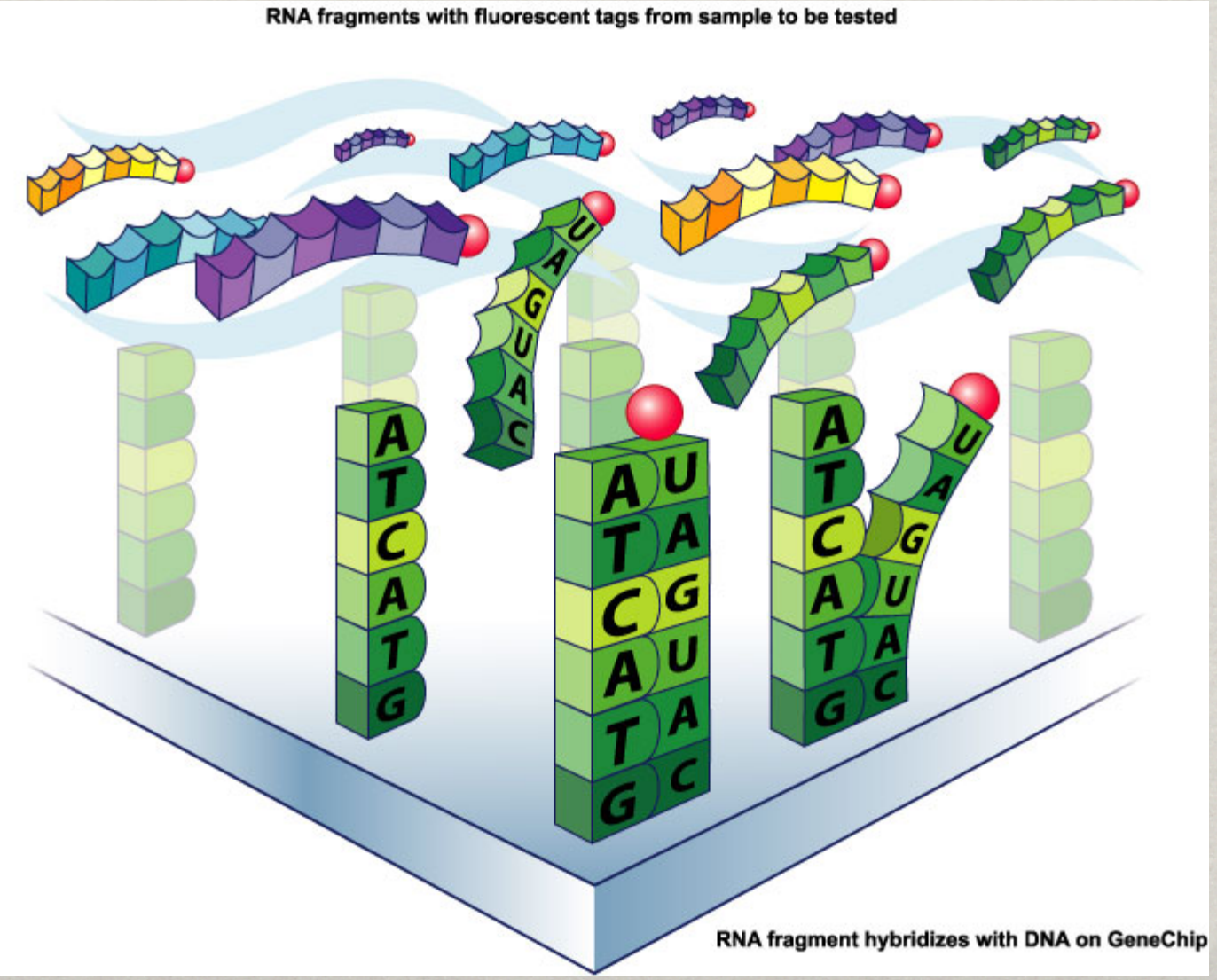


The synthesis of Affymetrix GeneChips involves a combination of photolithography and combinatorial chemistry. A series of masks are used to deprotect different locations and base by base, 25 nucleotide long probes are synthesized in parallel on the wafer.

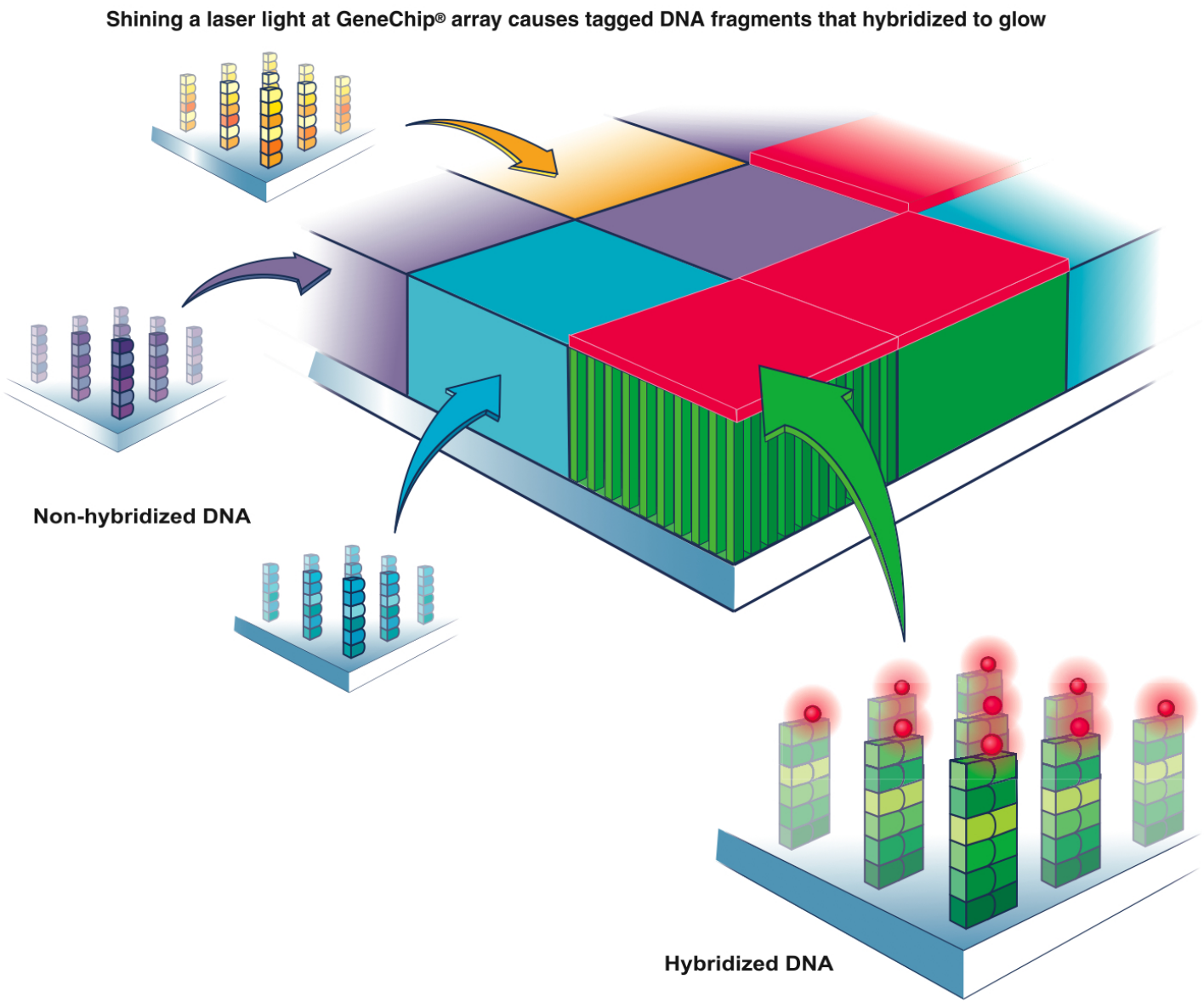
Probes built up systematically In Situ



Hybridization of RNA to probes



Scanning of Arrays

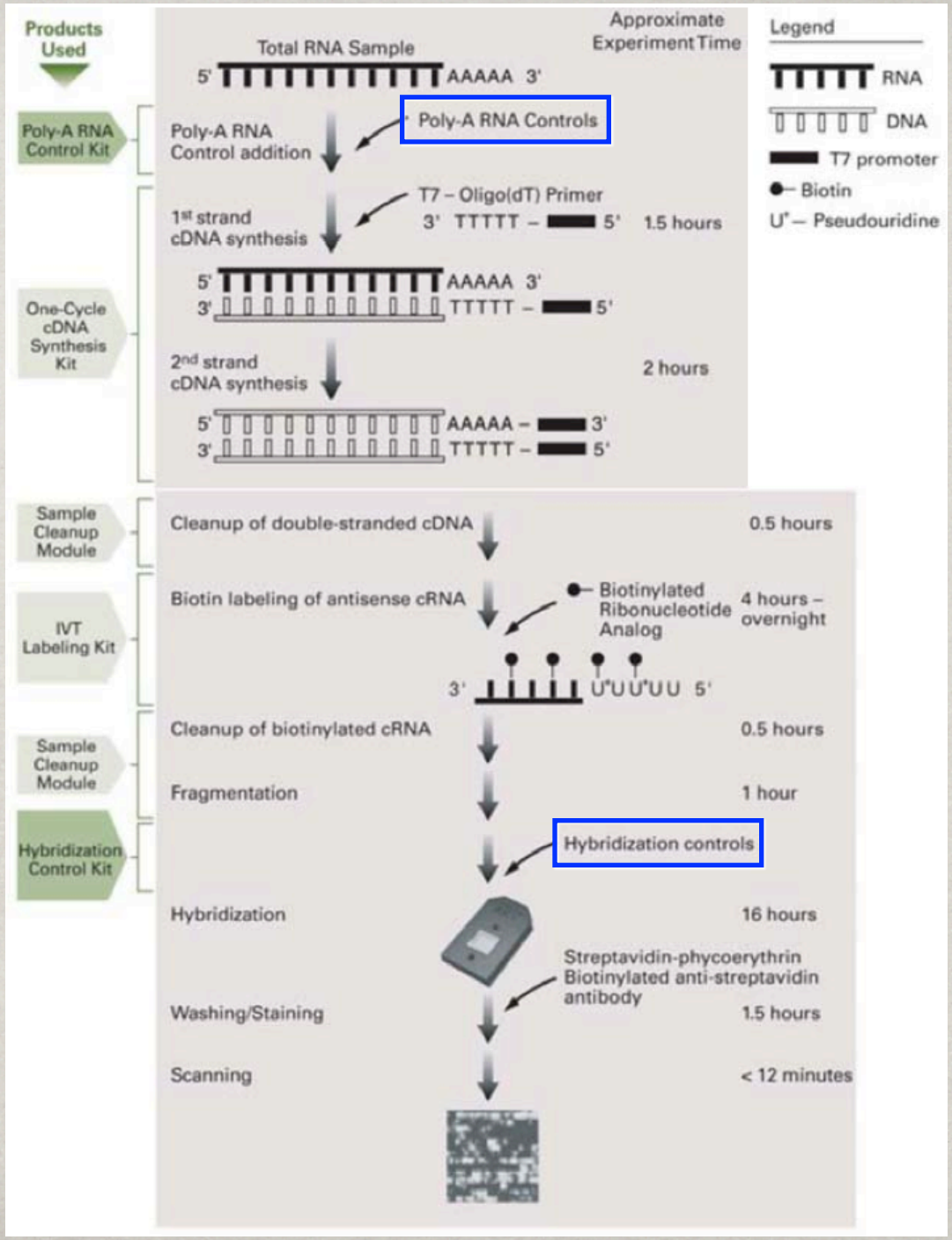


Sample Prep

Two types of controls:

Poly-A-RNA spike-in controls from *B. subtilis* - provide exogenous positive control to monitor labelling process

Hybridization controls are used to monitor the hyb procedure. Derive from *E.coli* and are added in the form of a pre-mixed control at staggered concentrations

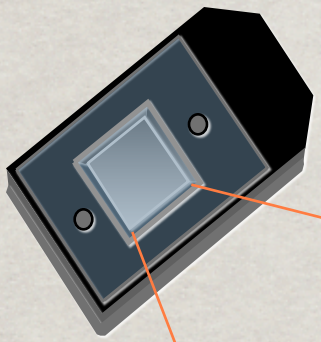


Equipment (Old Gen)

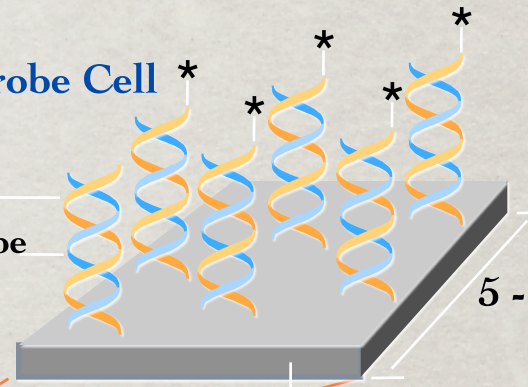
Equipment (New Gen)

Affymetrix GeneChip Arrays

GeneChip Probe Array



Hybridized Probe Cell



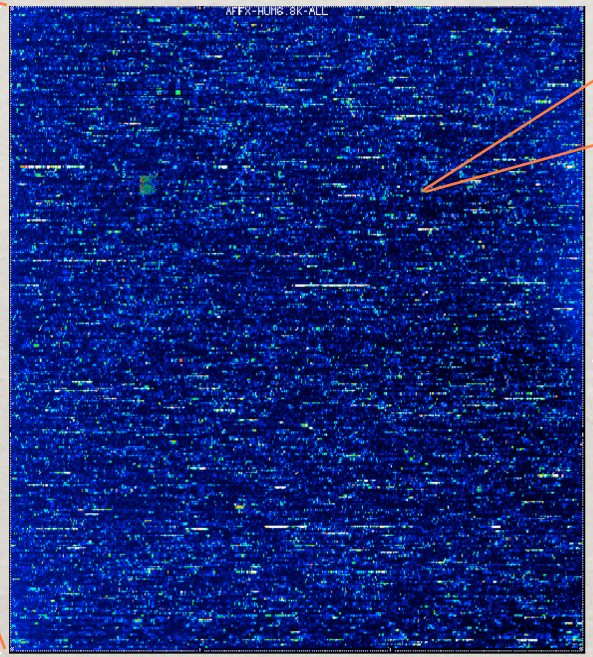
Single stranded, labeled cRNA target

Oligonucleotide probe

5 - 24µm

Millions of copies of a specific oligonucleotide probe

1.28cm



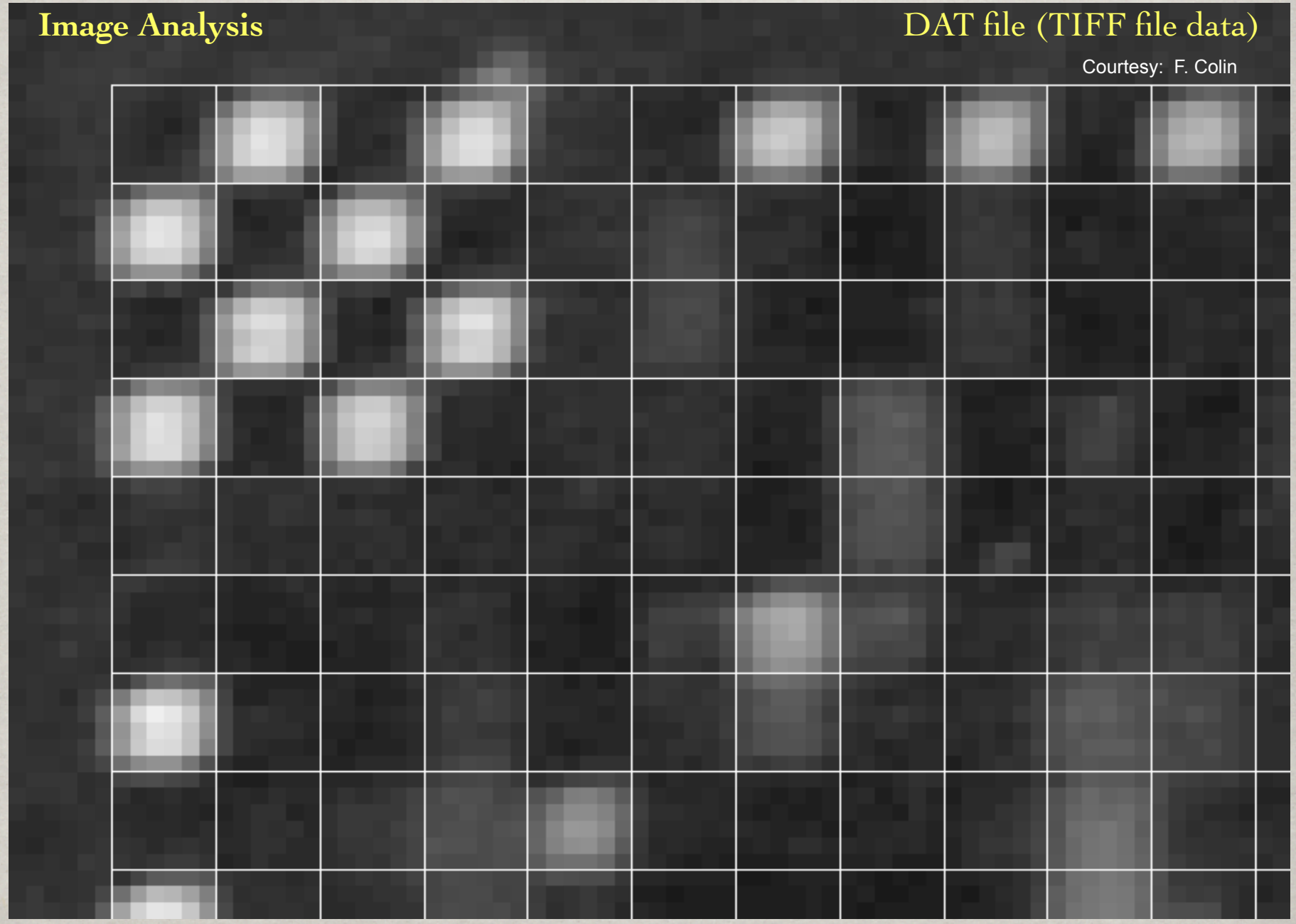
up to 6.4 million features/chip

Image of Hybridized Probe Array

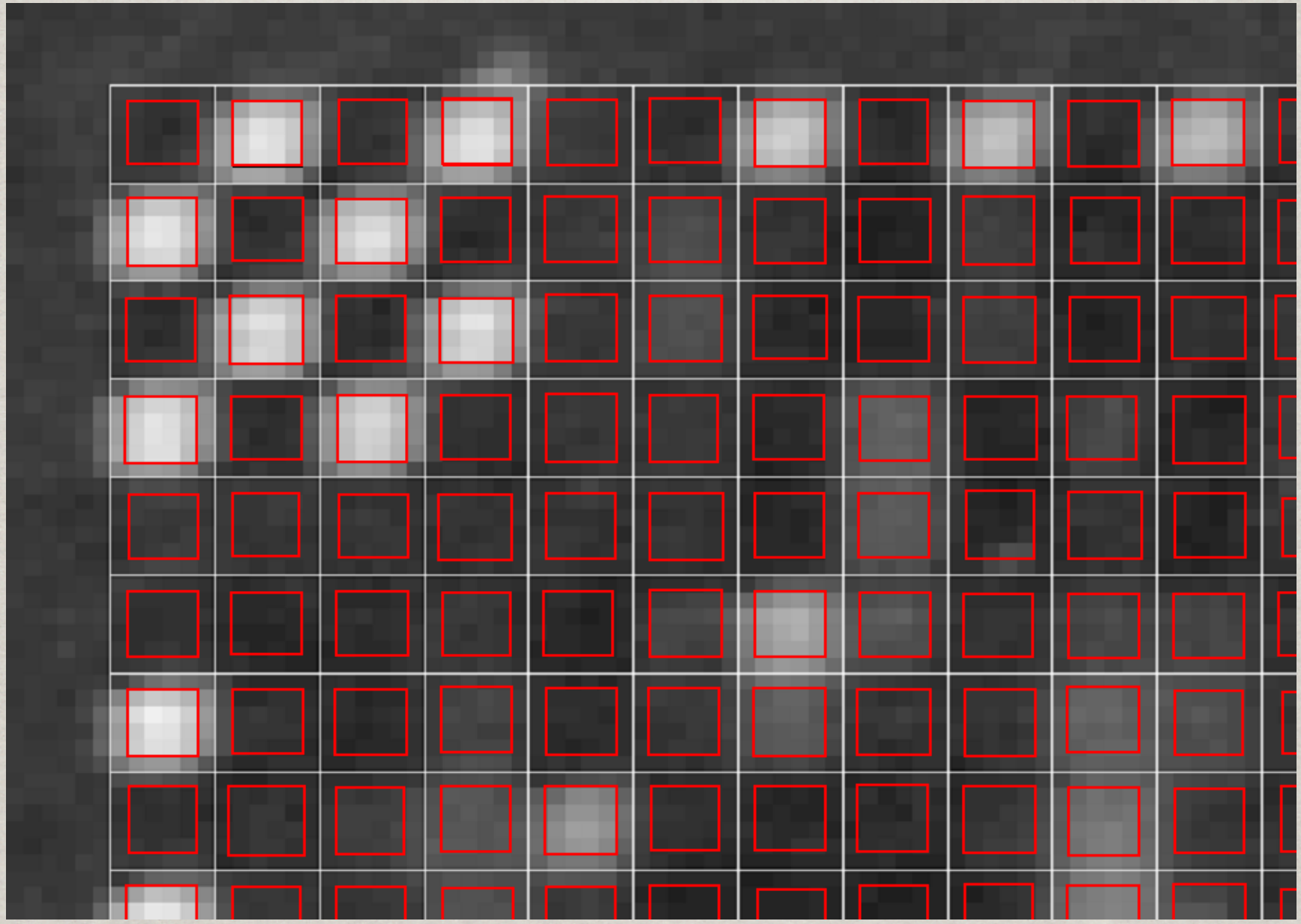
Compliments of D. Gerhold

GeneChip HG-U133A

Chip dat file – checkered board – close up w/ grid



Chip dat file – checkered board – close up pixel selection



Sample	Col1	Col2	Col3
Row1	123	234	345
Row2	456	567	678
Row3	789	890	901

Design XYZ	Col1	Col2	Col3
Row1	Set2	Set3	Set1
Row2	Set1	Set2	Set3
Row3	Set3	Set1	Set2

Design XYS

Index	Row	Col	Int
1	1	1	123
2	1	2	234
3	1	3	345
4	2	1	456
5	2	2	567
6	2	3	678
7	3	1	789
8	3	2	890
9	3	3	901

CEL File

Index	Row	Col	Set	Gene	Type
1	1	1	2	A	PM
2	1	2	3	B	PM
3	1	3	1	A	PM
4	2	1	1	A	MM
5	2	2	2	A	MM
6	2	3	3	B	MM
7	3	1	3	B	PM
8	3	2	1	A	MM
9	3	3	2	A	PM

Sample1	Col1	Col2	Col3
Row1	123	234	345
Row2	456	567	678
Row3	789	890	901

Sample2	Col1	Col2	Col3
Row1	012	123	234
Row2	345	456	567
Row3	678	789	890

Sample3	Col1	Col2	Col3
Row1	0123	1234	2345
Row2	3456	4567	5678
Row3	6789	7890	8901

Index	Row	Col	Int
1	1	1	123
2	1	2	234
3	1	3	345
4	2	1	456
5	2	2	567
6	2	3	678
7	3	1	789
8	3	2	890
9	3	3	901

Index	Row	Col	Int
1	1	1	012
2	1	2	123
3	1	3	234
4	2	1	345
5	2	2	456
6	2	3	567
7	3	1	678
8	3	2	789
9	3	3	890

Index	Row	Col	Int
1	1	1	0123
2	1	2	1234
3	1	3	2345
4	2	1	3456
5	2	2	4567
6	2	3	5678
7	3	1	6789
8	3	2	7890
9	3	3	8901

Representation With BioConductor

Index	Sample1	Sample2	Sample3
1	123	012	0123
2	234	123	1234
3	345	234	2345
4	456	345	3456
5	567	456	4567
6	678	567	5678
7	789	678	6789
8	890	789	7890
9	901	890	8901

Index	Row	Col	Set	Gene
1	1	1	2	A
2	1	2	3	B
3	1	3	1	A
4	2	1	1	A
5	2	2	2	A
6	2	3	3	B
7	3	1	3	B
8	3	2	1	A
9	3	3	2	A

Design XYS

CEL and CDF files

- Image analysis of each scanned GeneChip produces a **CEL** file which stores the mean intensity & SD for each square probe cell
  - Each hybridization has a unique CEL file
- The **CDF** (Chip Description File)specifies the probe and the probe set that belongs to each cell; implies it is necessary to decode the **CEL** file
  - Each chip type has a unique CDF file
- CDF information packages can be downloaded from BioConductor for standard Affymetrix arrays

Practical

Exercise 1

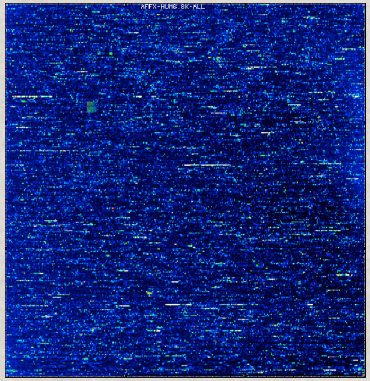
Low-level Analysis

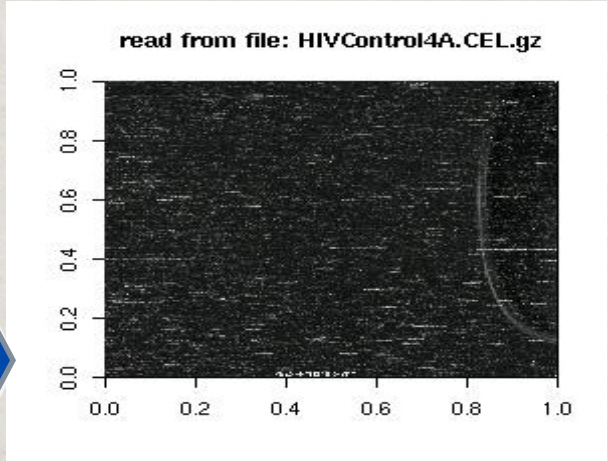
- ✦Low-level analysis (aka pre-processing) is the initial and possibly most important step in microarray data analysis
- ✦Deals with probe-level intensities rather than expression values
- ✦Involves:
  - ✦Data exploration
  - ✦Quality assessment
  - ✦Background correction, normalisation, summarization
- ✦Goal is to obtain more accurate gene expression values

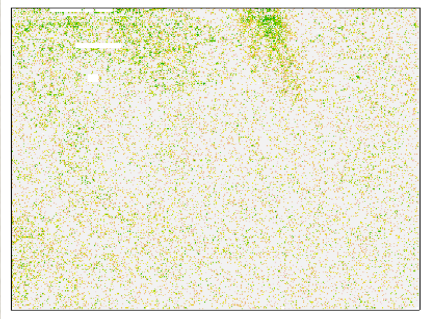
Practical

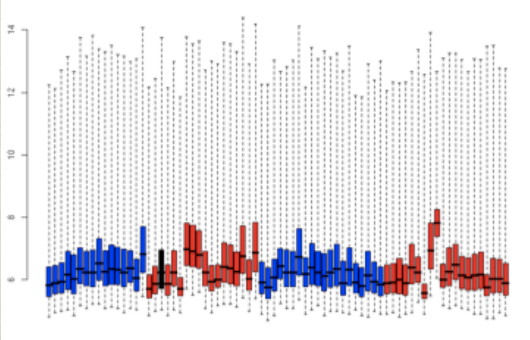
Exercises 2-4

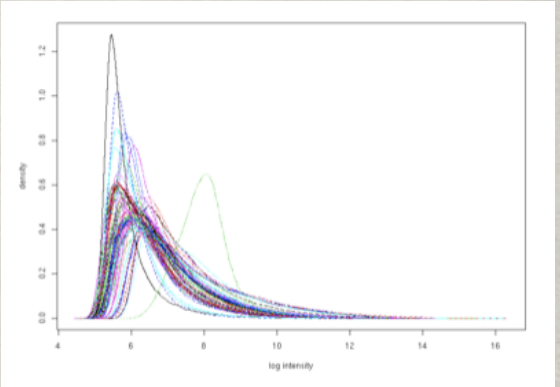
Quality assessment



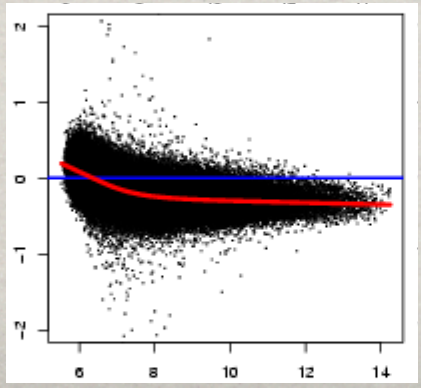


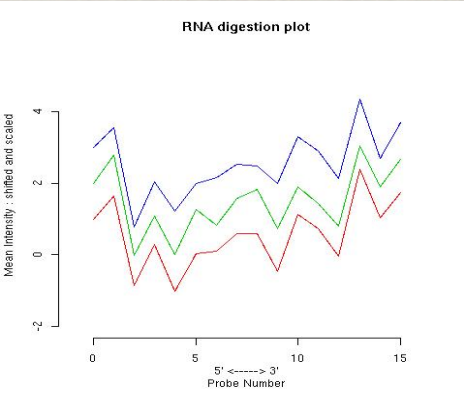




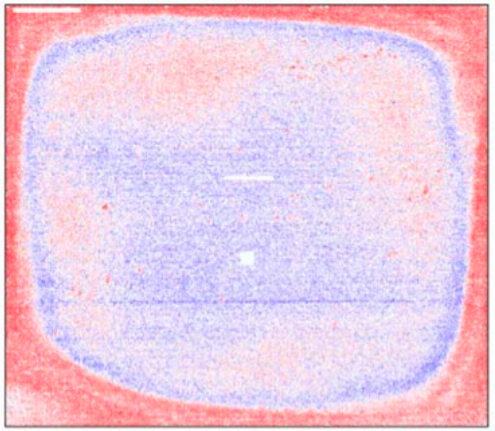


- chip pseudo images
- boxplots
- density plots
- MA plots
- RNA degradation plots



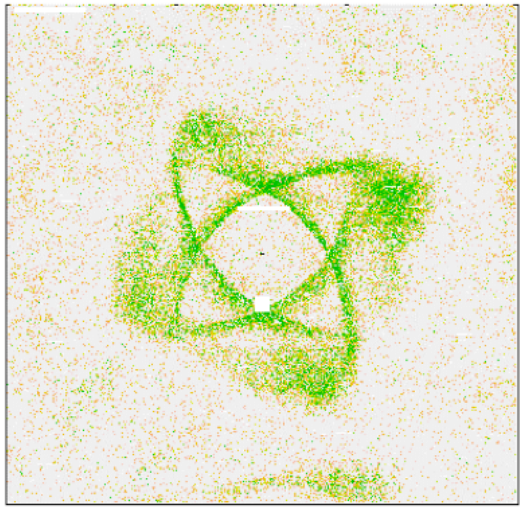


Chip Pseudo-images From PLM Fit

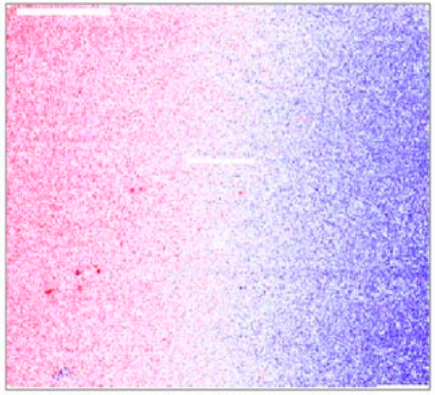


"Ring of Fire"

"Crop Circles"



"Tricolor"

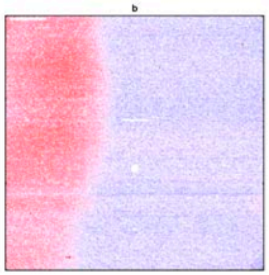


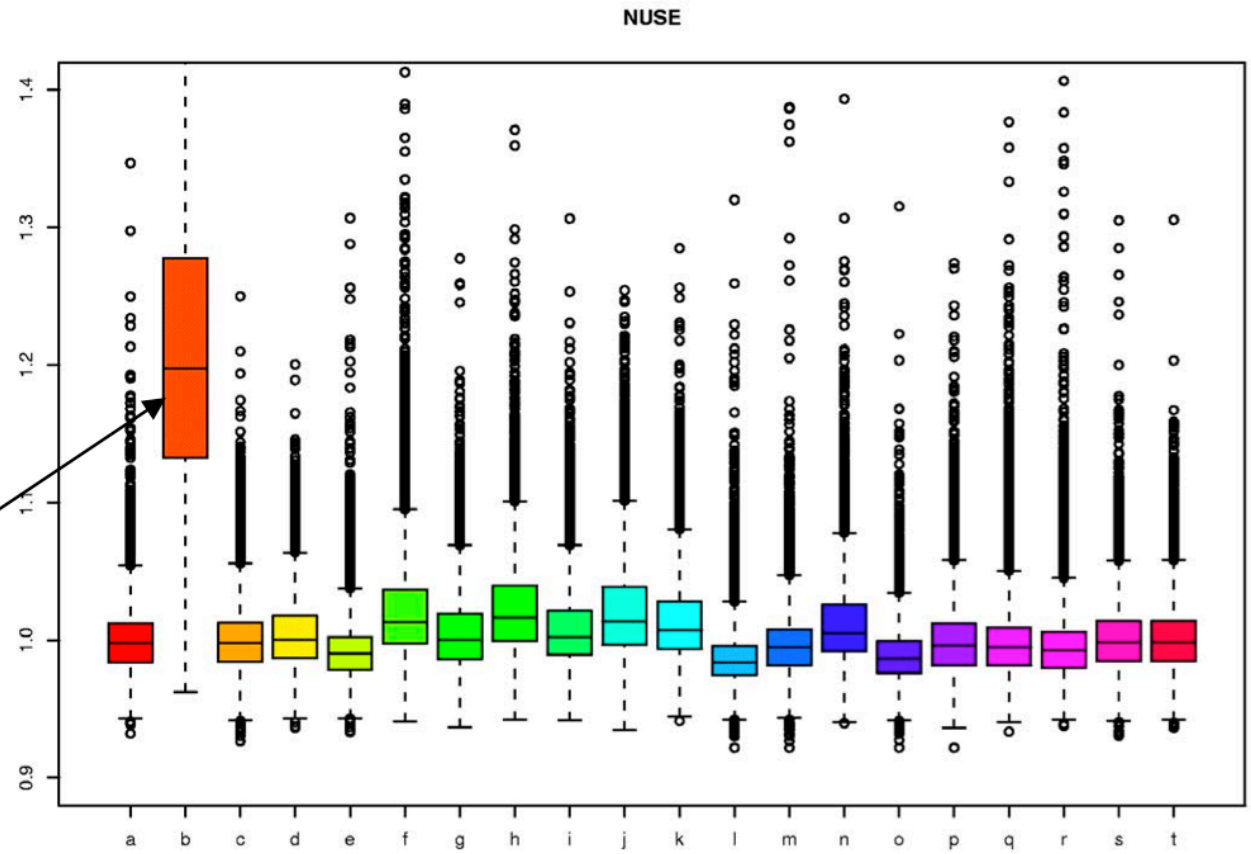
http://PLMImageGallery.bmbolstad.com

Slide from Ben Bolstad

Detecting outlier arrays

Normalized
Unscaled
Standard
Errors





Slide from Ben Bolstad

Practical

Exercise 5

Pre-processing steps

Computing expression values for each probe set: a 3-step scheme

- ✦Background correction
- ✦Normalisation
- ✦Probe set summarization

Background correction

- ✦In addition to noise associated with optical detection system there is also non-specific physico-chemical reactions between labeled targets and microarray surface
- ✦Therefore, observed signal intensities contain contributions from both specific and non-specific signals
- ✦Requires a proper treatment of background signal

Affymetrix's MAS 5

Background correction

$$y_j = PM_j - MM_j^*$$

where MM_j^* is chosen so that y_j is non-negative

Normalisation

Scale so that mean y_j is same for each chip

Probe Set Summary

$$\log(\text{Signal Intensity}) = \text{TukeyBiweight}(\log y_j)$$

j = probe

✦**PLIER (successor of MAS)**: Probe Logarithmic Intensity Error is a model-based signal estimator which benefits from multi-array analysis.

dChip / MBEI

(Model Based Expression Index)

Background correction

$$y_{ij} = PM_{ij} - MM_{ij}$$

Probe Set Summary

$$y_{ij} = \theta_i \varphi_j + \varepsilon_{ij}$$

dProbe set signal = θ_i

i = array, j = probe

Li and Wong (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *PNAS* 98, 31-36

RMA

Background correction

Normalisation

Probe Set Summarization

Irizarry, Bolstad, Collin, Cope, Hobbs, Speed (2003). Summaries of Affymetrix GeneChip probe level data. *NAR* 31(4):e15

✦**GCRMA** (GC content RMA)

Wu Z, Irizarry RA, Gentleman R, Murillo FM, and Spencer F (2005). A Model Based Background Adjustment for Oligonucleotide Expression Arrays. JASA 99, 909-917

✦**Others**: vsn, pdnn, NLFIT

RMA

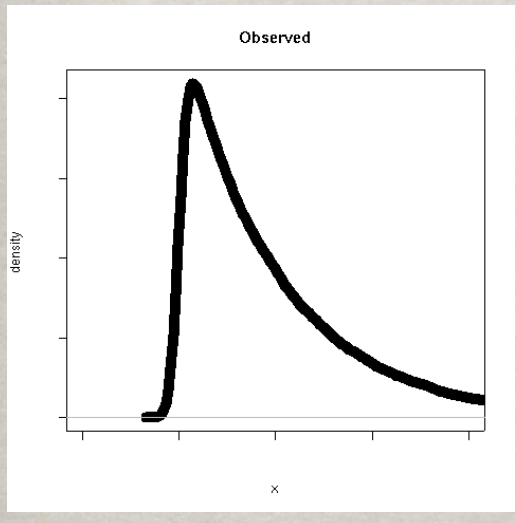
Probabilistic approach based on empirical dist of probe intensities: assume PM = S (exponential) + B (Normal); convolution method

Background correction

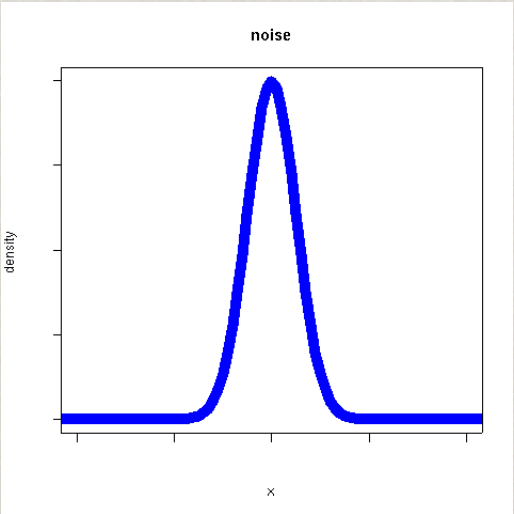
Ignore MM, fit model to PM

PM = Background + Signal

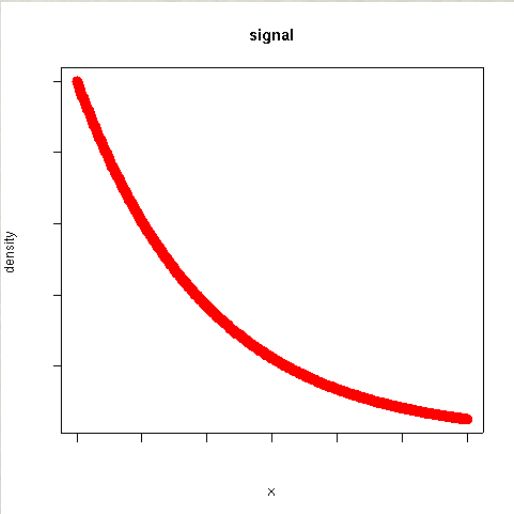
Normal(μ,σ^2) Exponential(α)



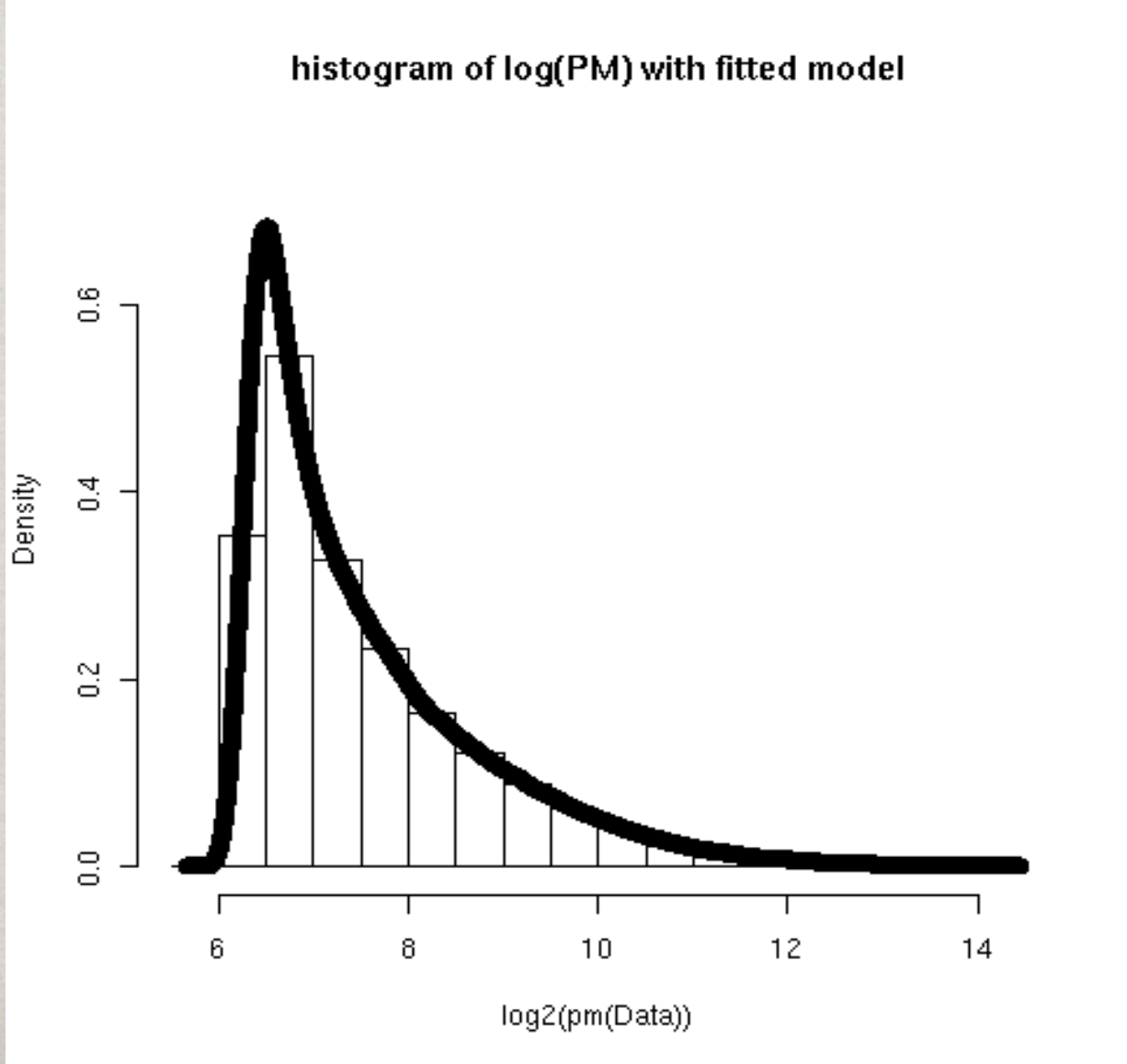
=



+



PM data on $\log_2$ scale: raw and fitted model



RMA: Normalisation

- ✦Force the empirical distribution of probe intensities to be the same for every chip in an experiment
- ✦The common distribution is obtained by averaging each quantile across chips

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Sort Each Column →

1	1	1
2	2	5
3	3	6
4	4	7
9	9	9

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Sort Each Column →

1	1	1
2	2	5
3	3	6
4	4	7
9	9	9

Average Each Row →

1
3
4
5
9

Target Distribution

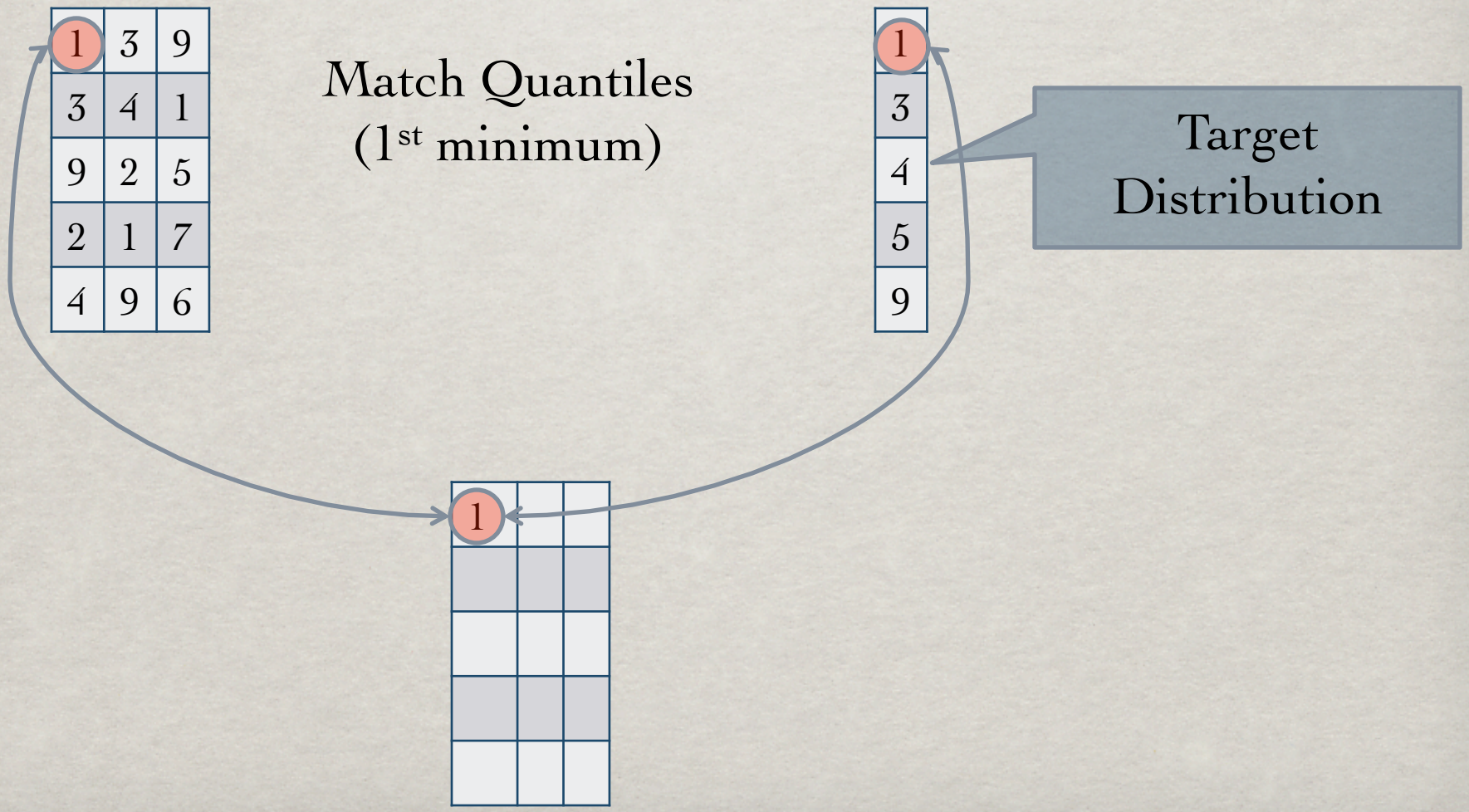
Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

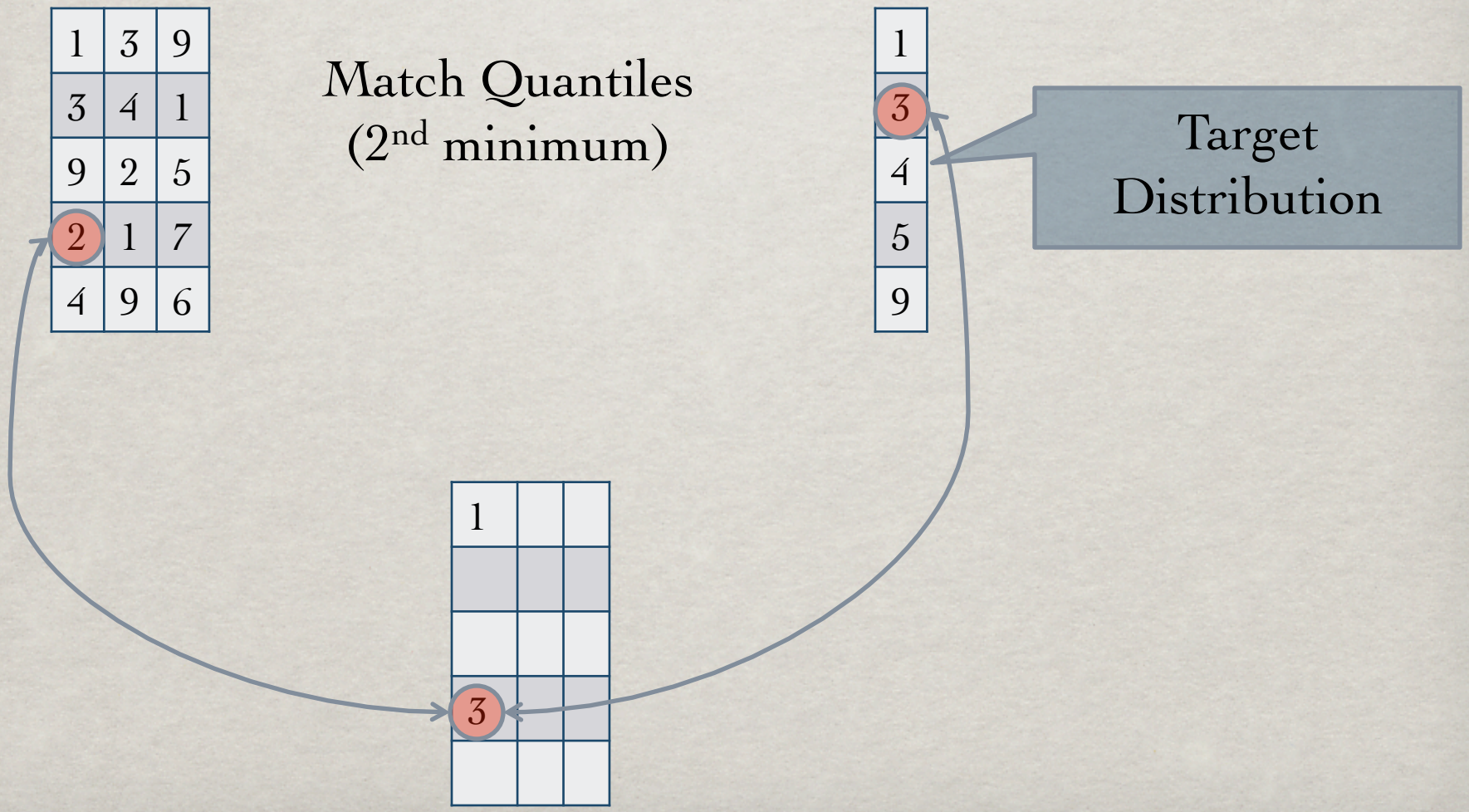
1
3
4
5
9

Target Distribution

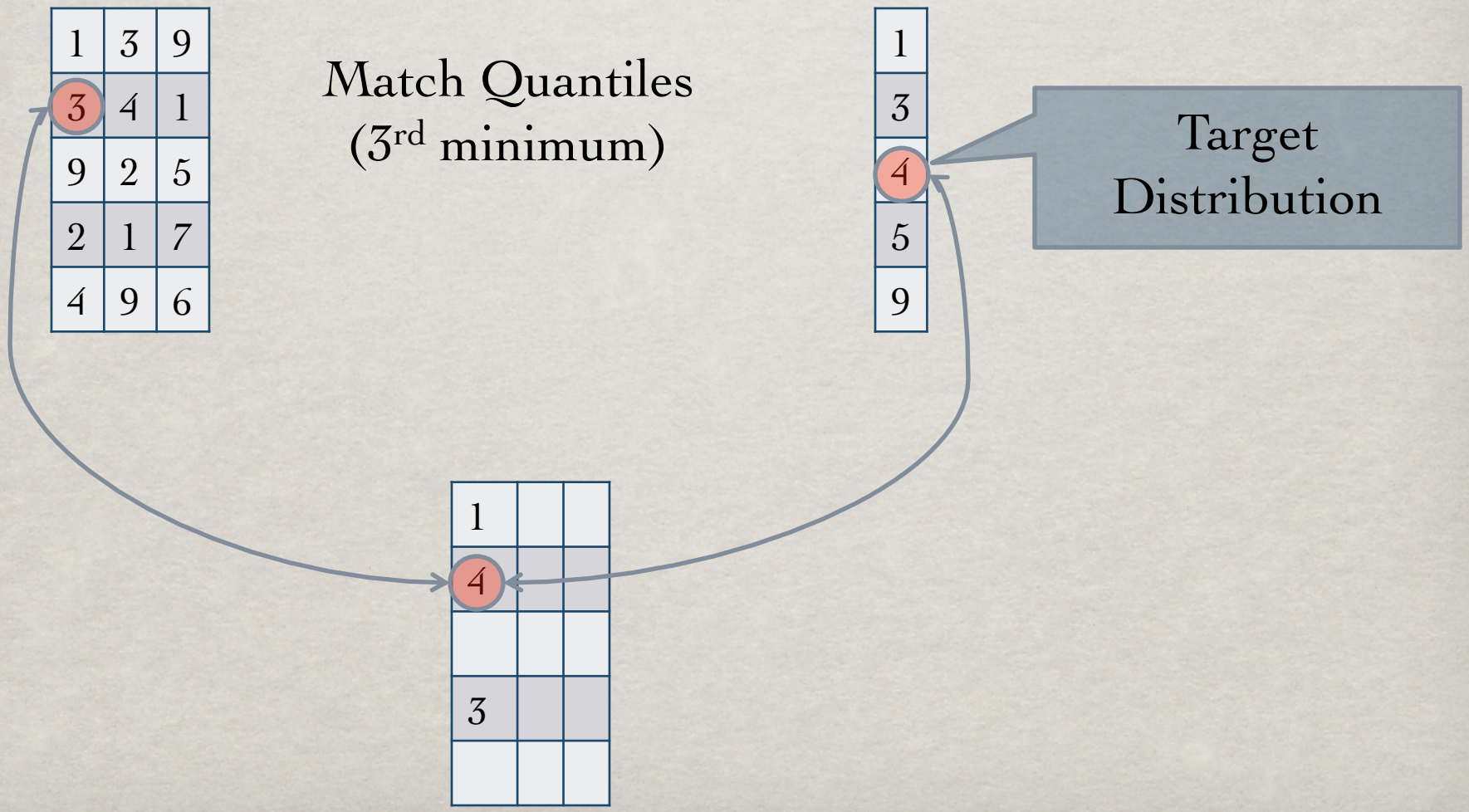
Quantile Normalisation



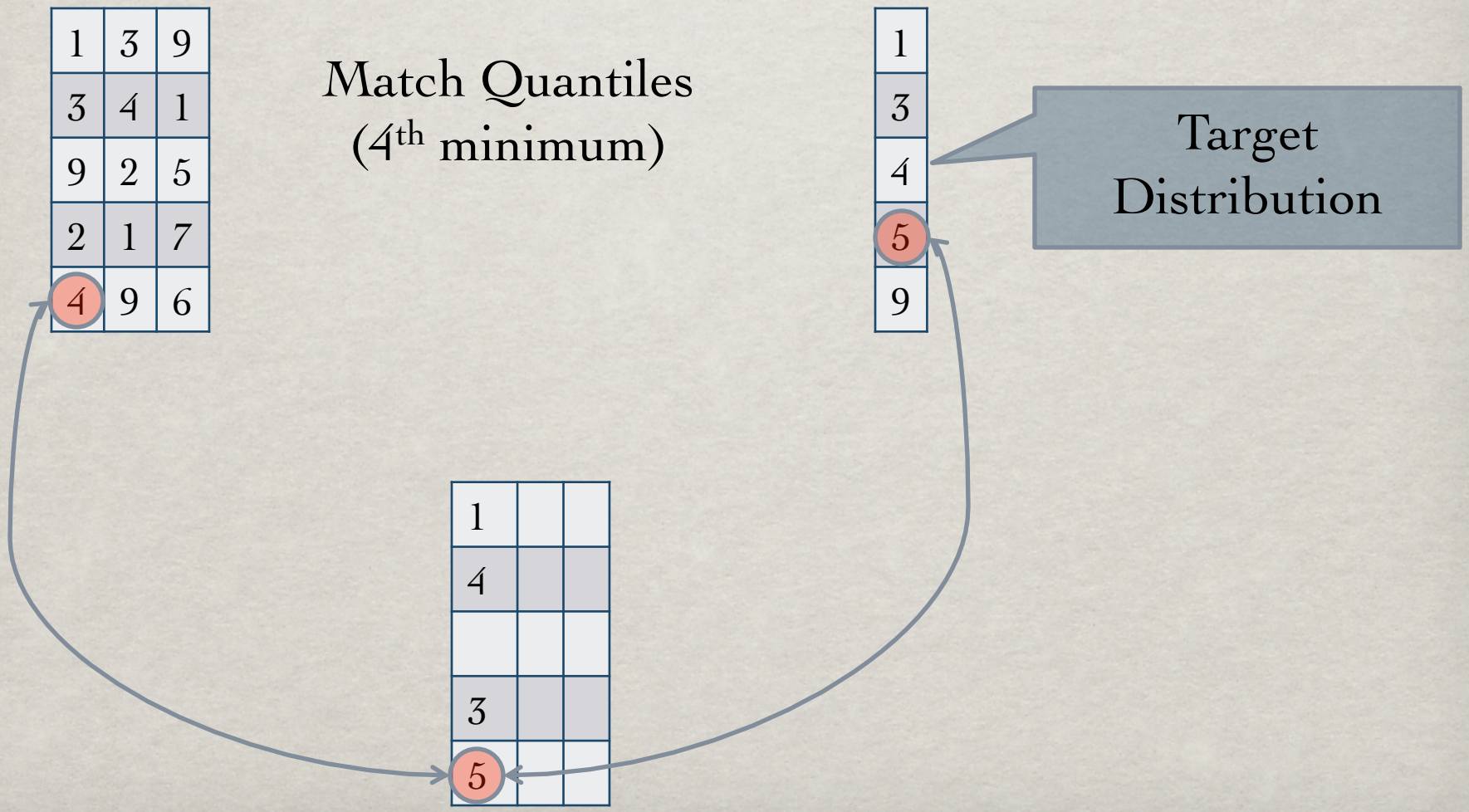
Quantile Normalisation



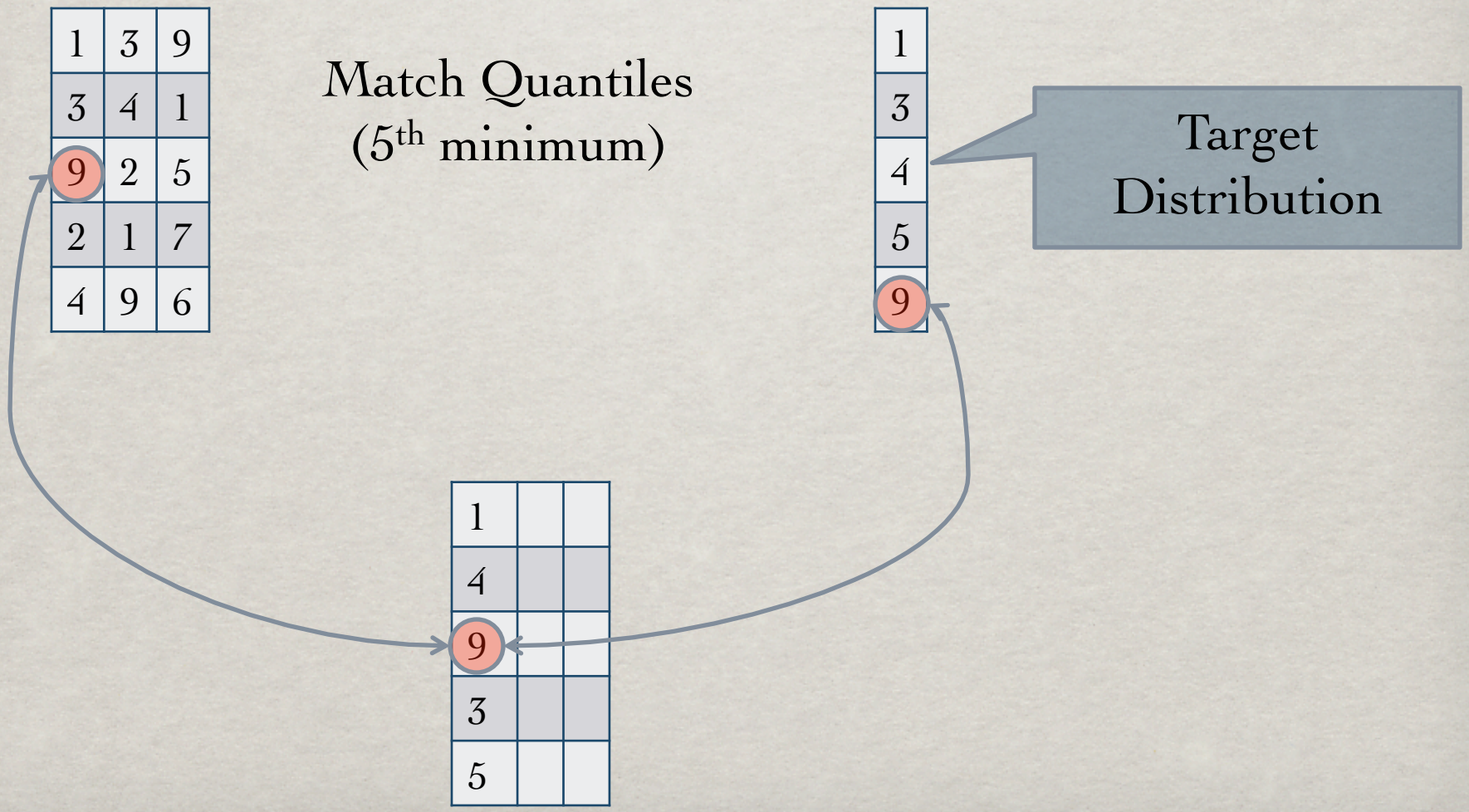
Quantile Normalisation



Quantile Normalisation



Quantile Normalisation



Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Match Quantiles
(repeat for Sample 2)

1
3
4
5
9

Target Distribution

1	4	
4	5	
9	3	
3	1	
5	9	

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Match Quantiles
(repeat for Sample 3)

1
3
4
5
9

Target Distribution

1	4	
4	5	
9	3	
3	1	
5	9	

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Match Quantiles
(repeat for Sample 3)
1st minimum

1
3
4
5
9

Target Distribution

1	4	
4	5	
9	3	
3	1	
5	9	

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Match Quantiles
(repeat for Sample 3)
2nd minimum

1
3
4
5
9

Target Distribution

1	4	
4	5	1
9	3	
3	1	
5	9	

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Match Quantiles
(repeat for Sample 3)
3rd minimum

1
3
4
5
9

Target Distribution

1	4	
4	5	1
9	3	3
3	1	
5	9	

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Match Quantiles
(repeat for Sample 3)
4th minimum

1
3
4
5
9

Target Distribution

1	4	
4	5	1
9	3	3
3	1	
5	9	4

Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

Match Quantiles
(repeat for Sample 3)
5th minimum

1
3
4
5
9

Target Distribution

1	4	
4	5	1
9	3	3
3	1	5
5	9	4

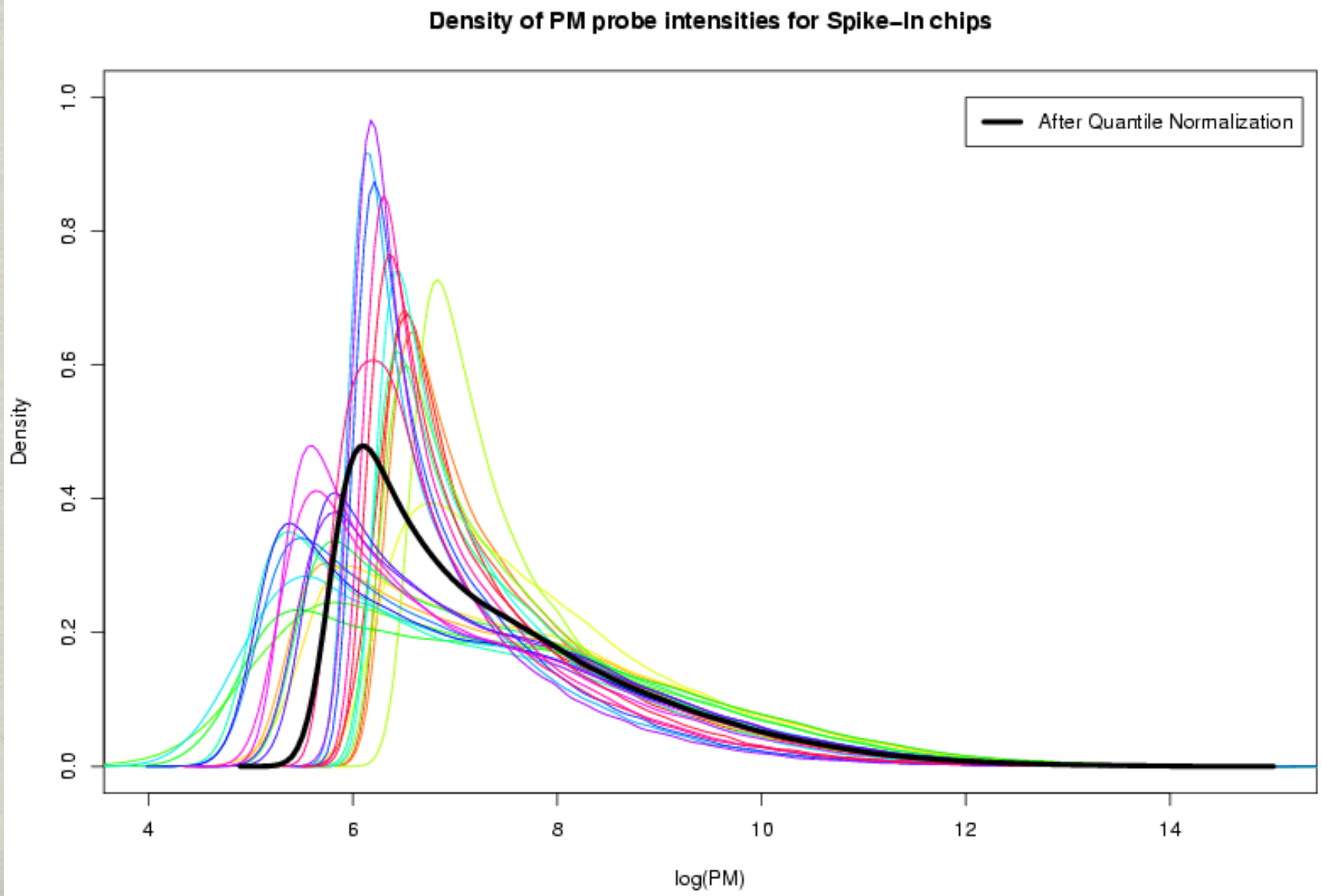
Quantile Normalisation

1	3	9
3	4	1
9	2	5
2	1	7
4	9	6

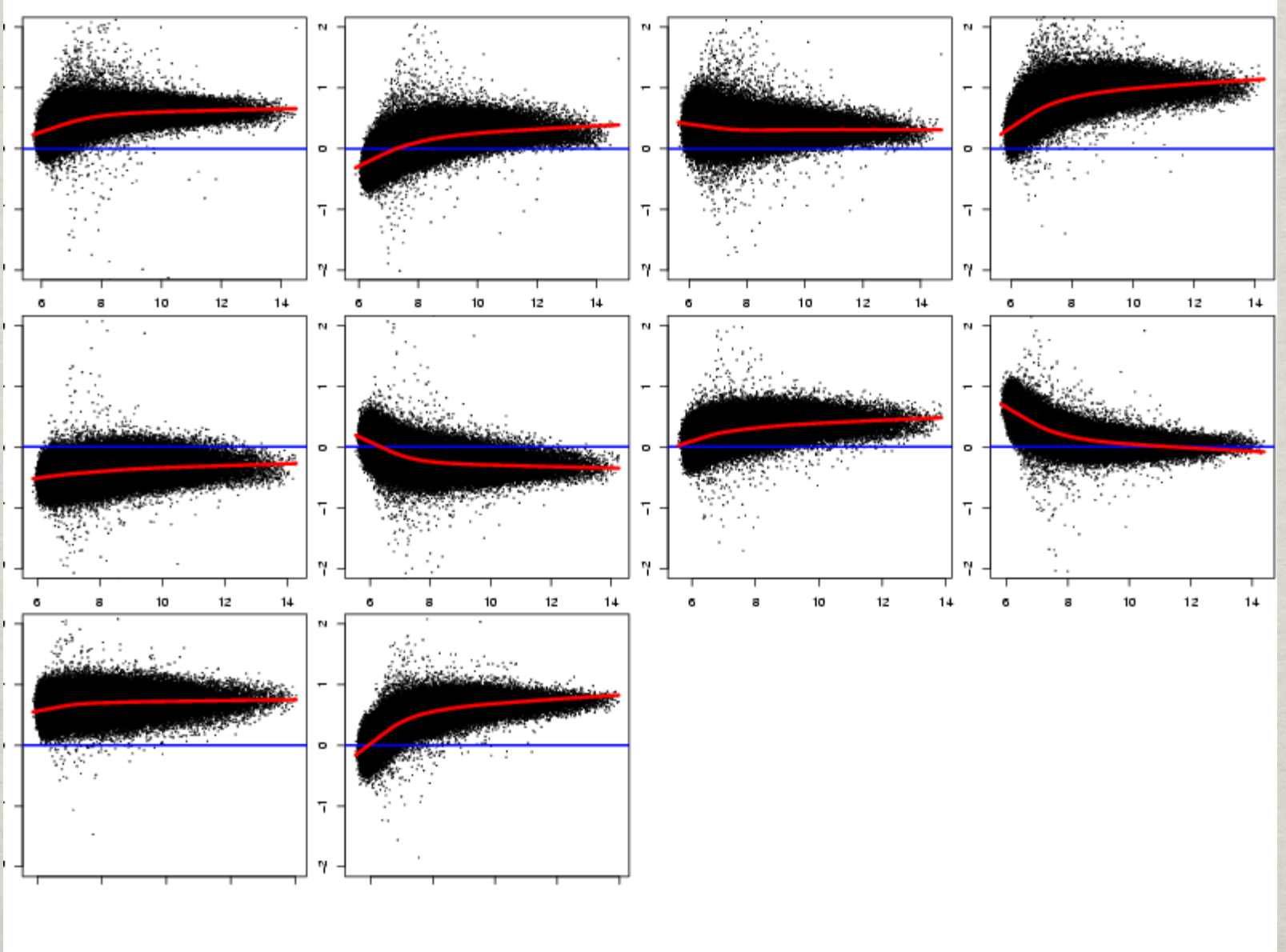
Normalised Data

1	4	9
4	5	1
9	3	3
3	1	5
5	9	4

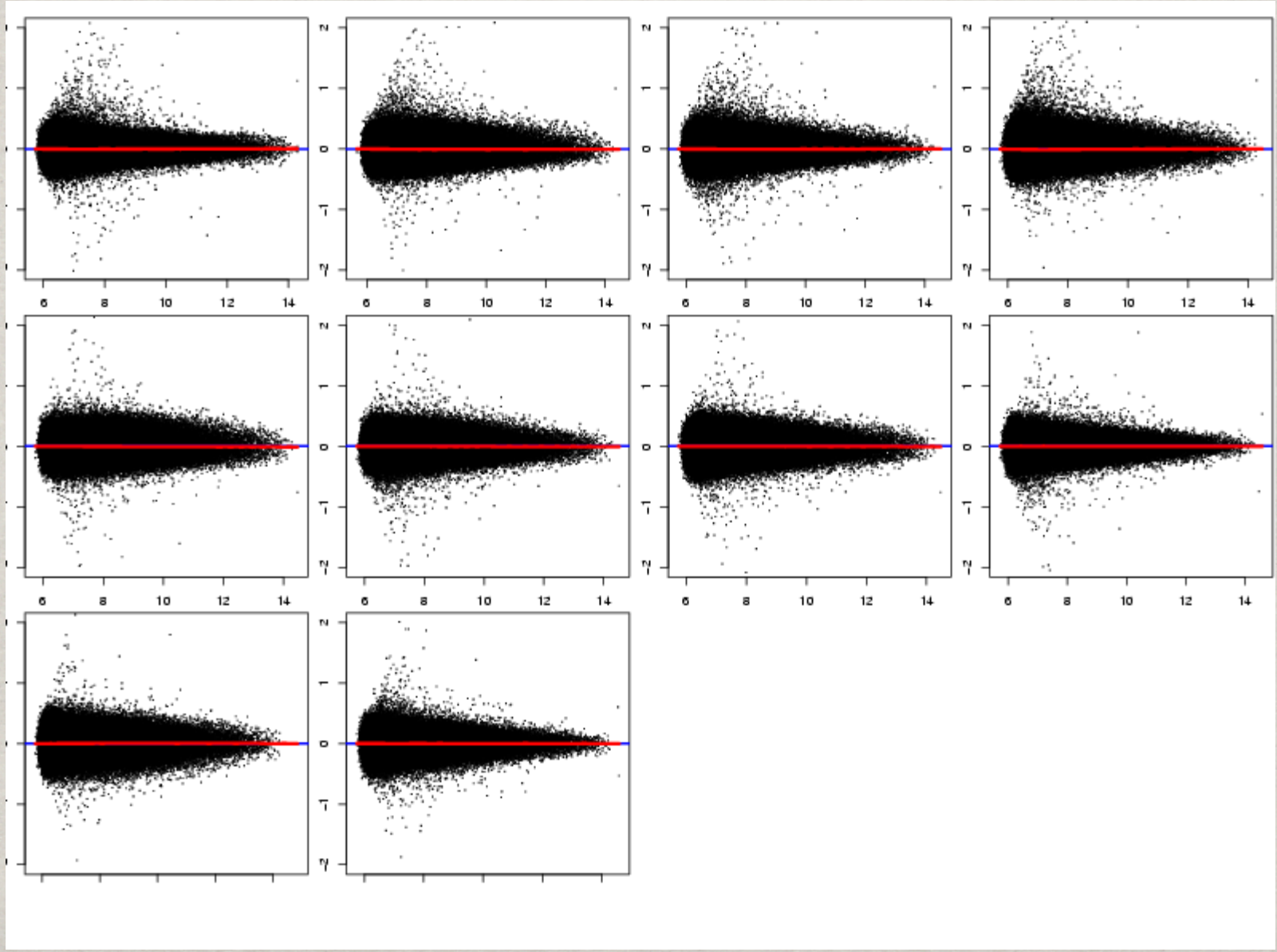
One distribution for all: the black curve



MvA plots of chip pairs: Pre-normalisation



MvA plots of chip pairs: Post-normalisation



RMA:

Probe Set Summarization

Robustly fit the two-way model for each probe set

$$\log_2 y_{ij} = a_i + b_j + e_{ij}$$

i = array, j = probe

Fit may be by median polish (quick) or by M-estimation (slower but yields standard errors and good quality diagnostic tools)

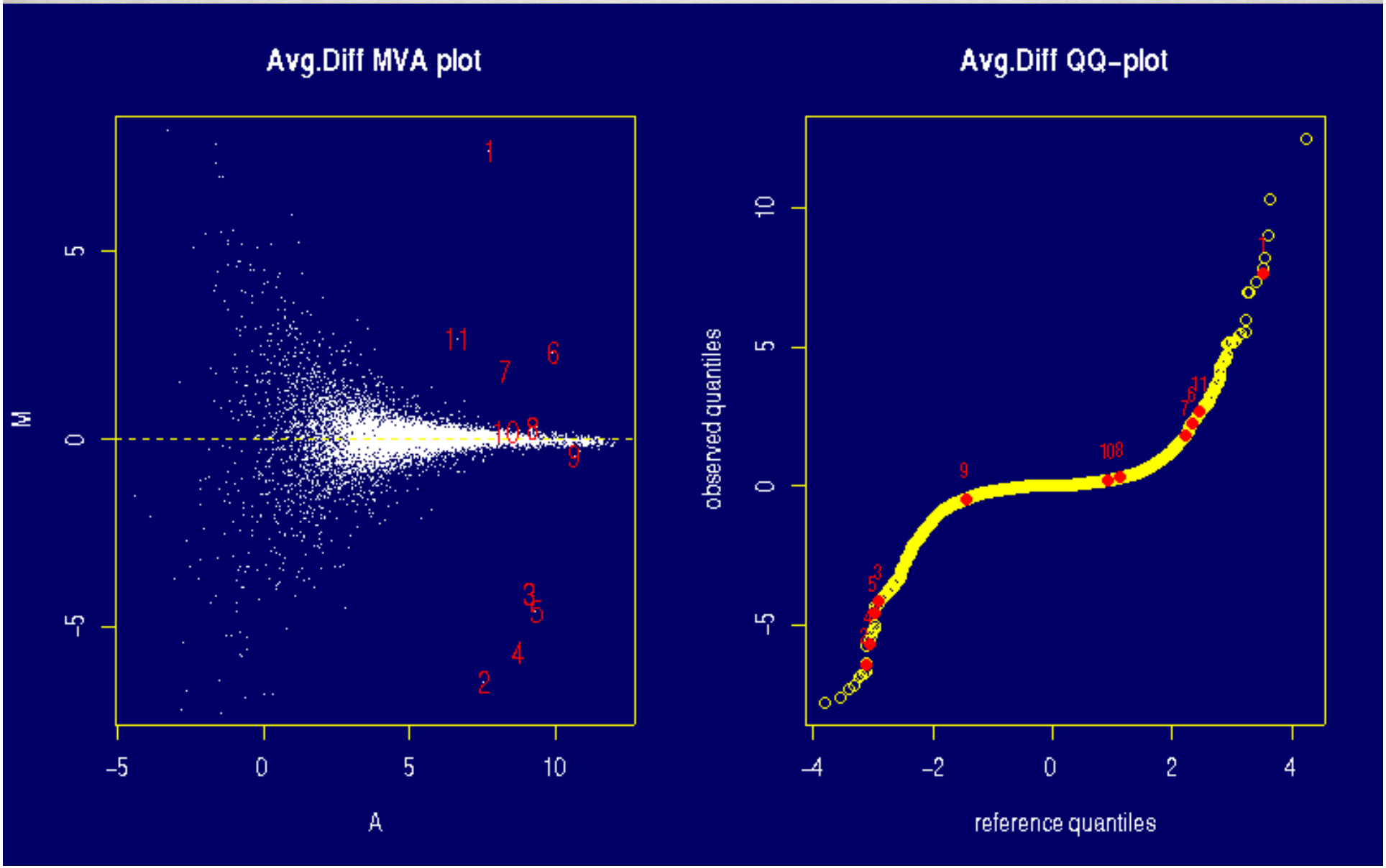
Practical

Exercise 6

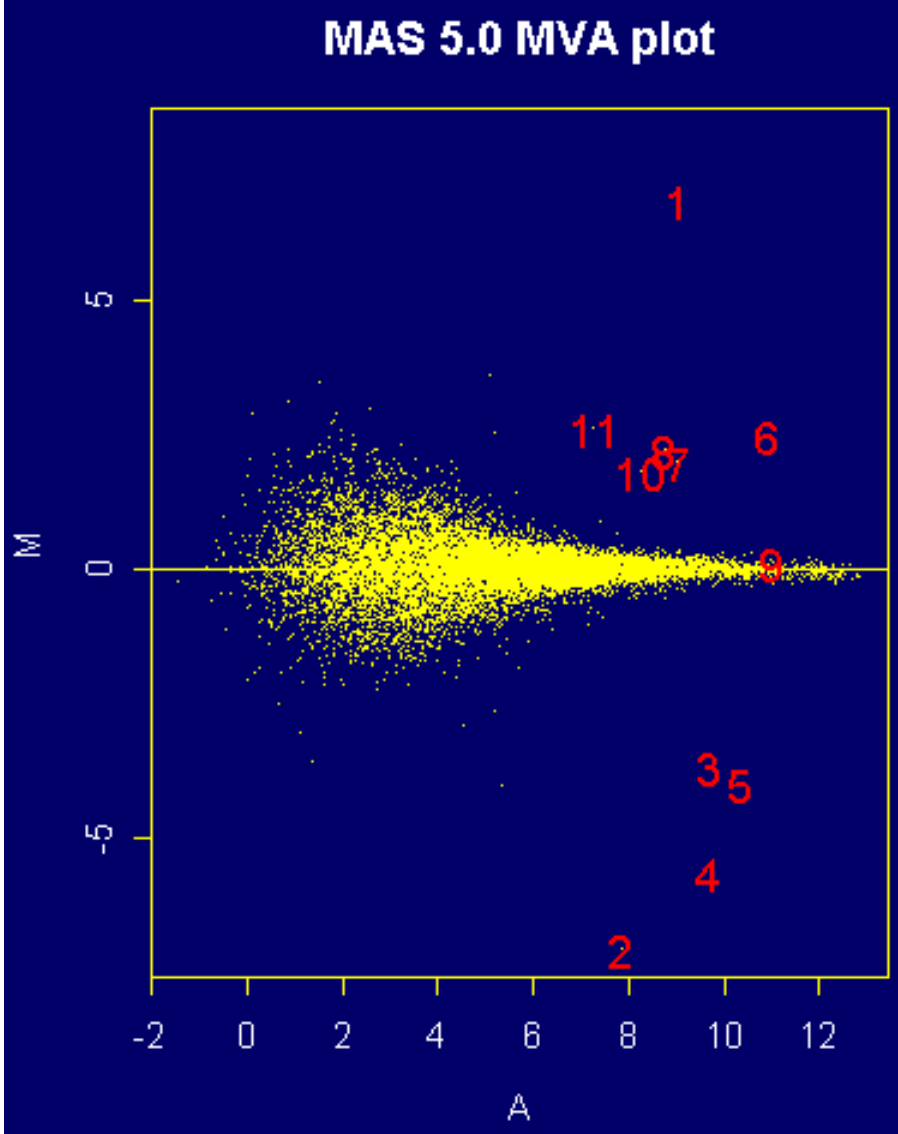
A measure of Truth

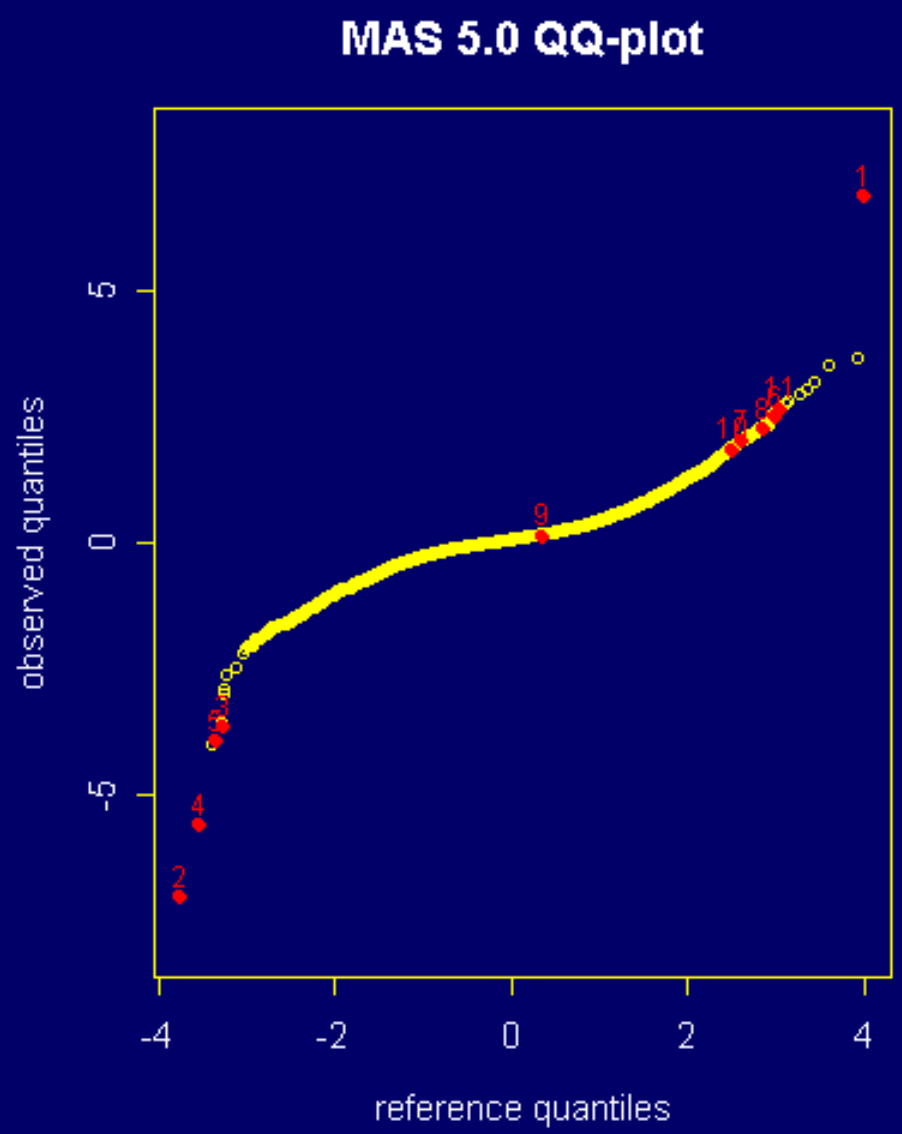
- ✦The availability of the Affymetrix & Genelogic spike in studies has led to significant efforts in exploring the relationship between concentration and microarray signal
- ✦Extremely useful for assessing bias/variance
- ✦Ideally expression values are both precise (low variance) and accurate (low bias), but there is always a tradeoff!

Assessing Differential expression

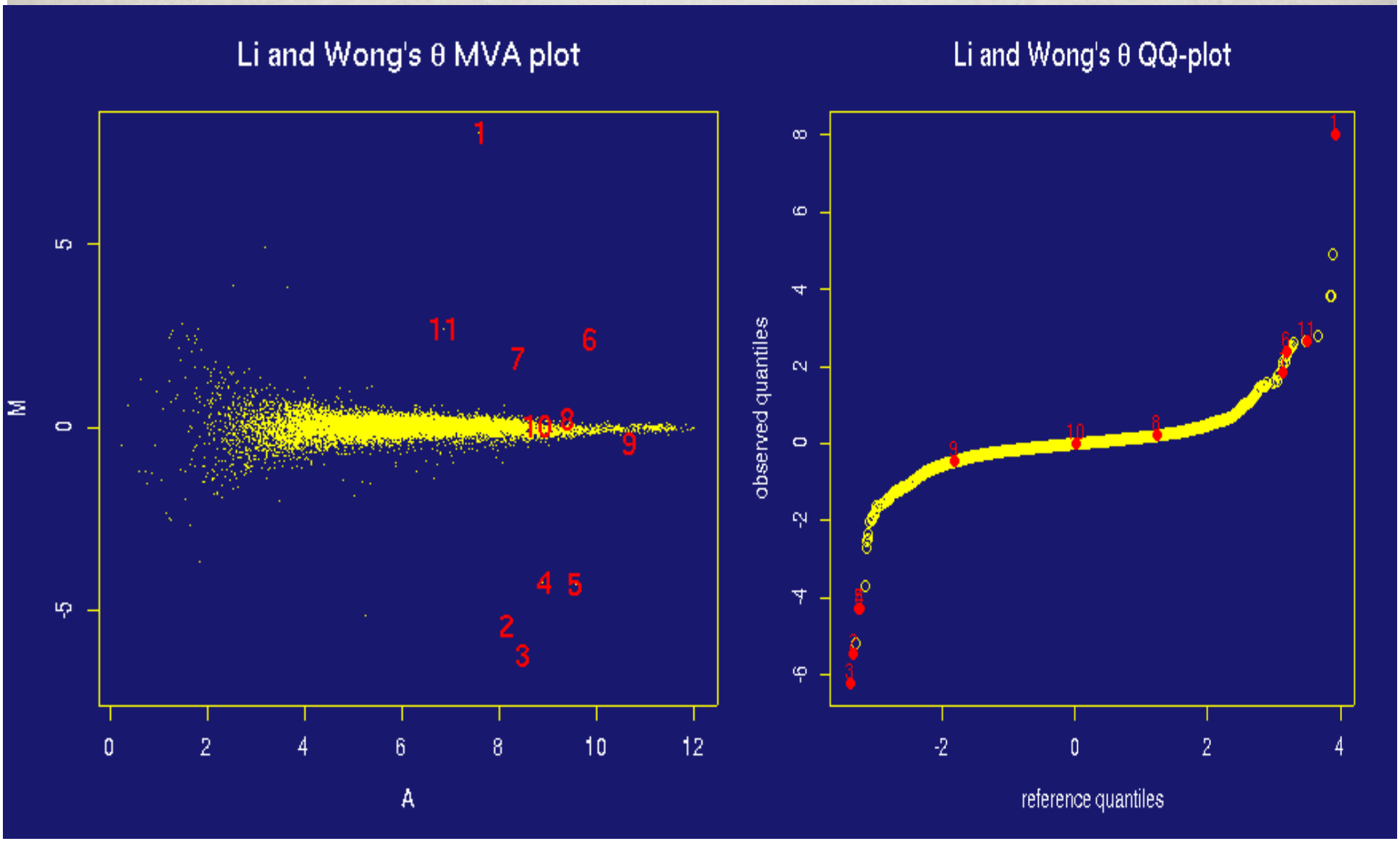


Assessing Differential expression

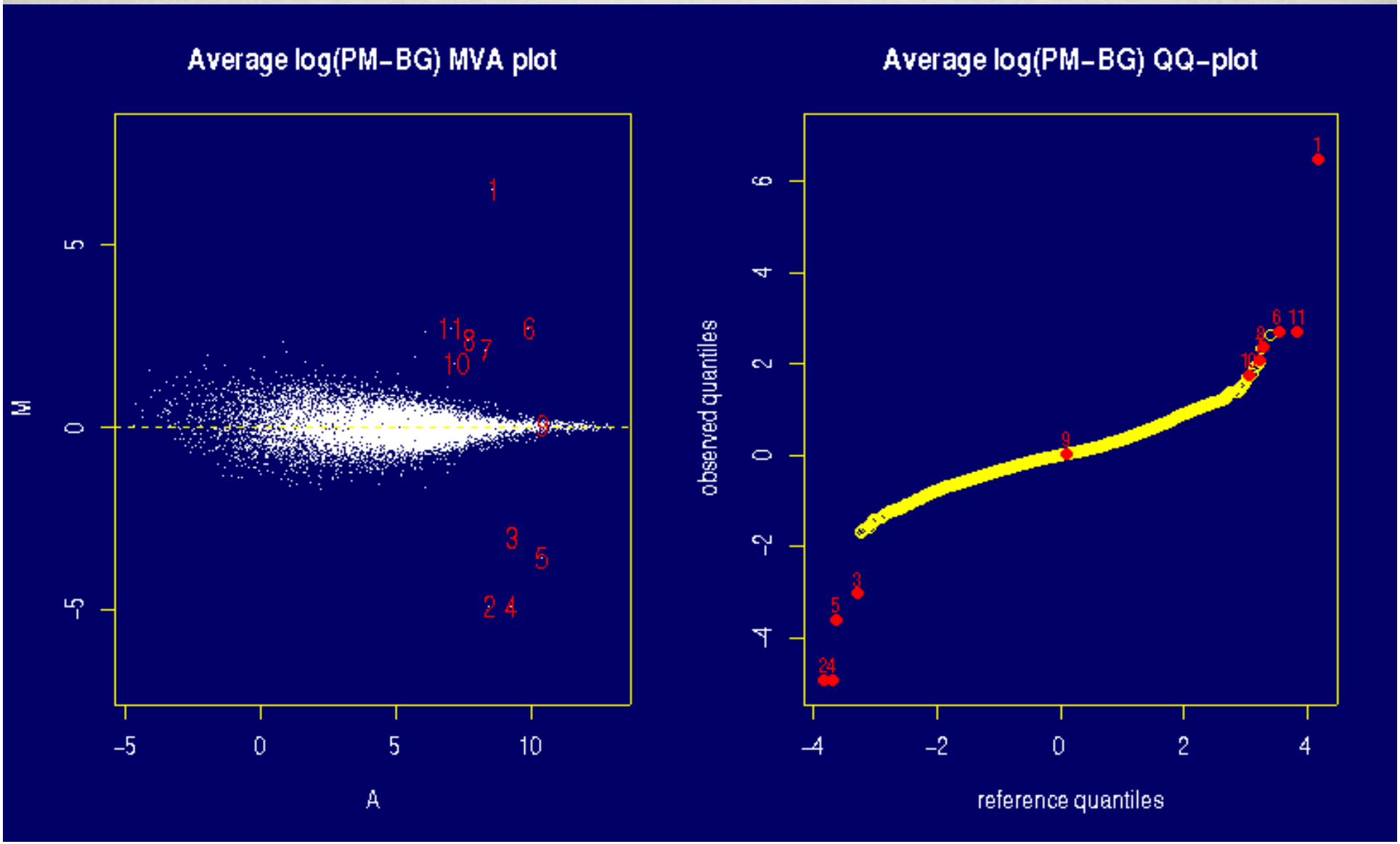




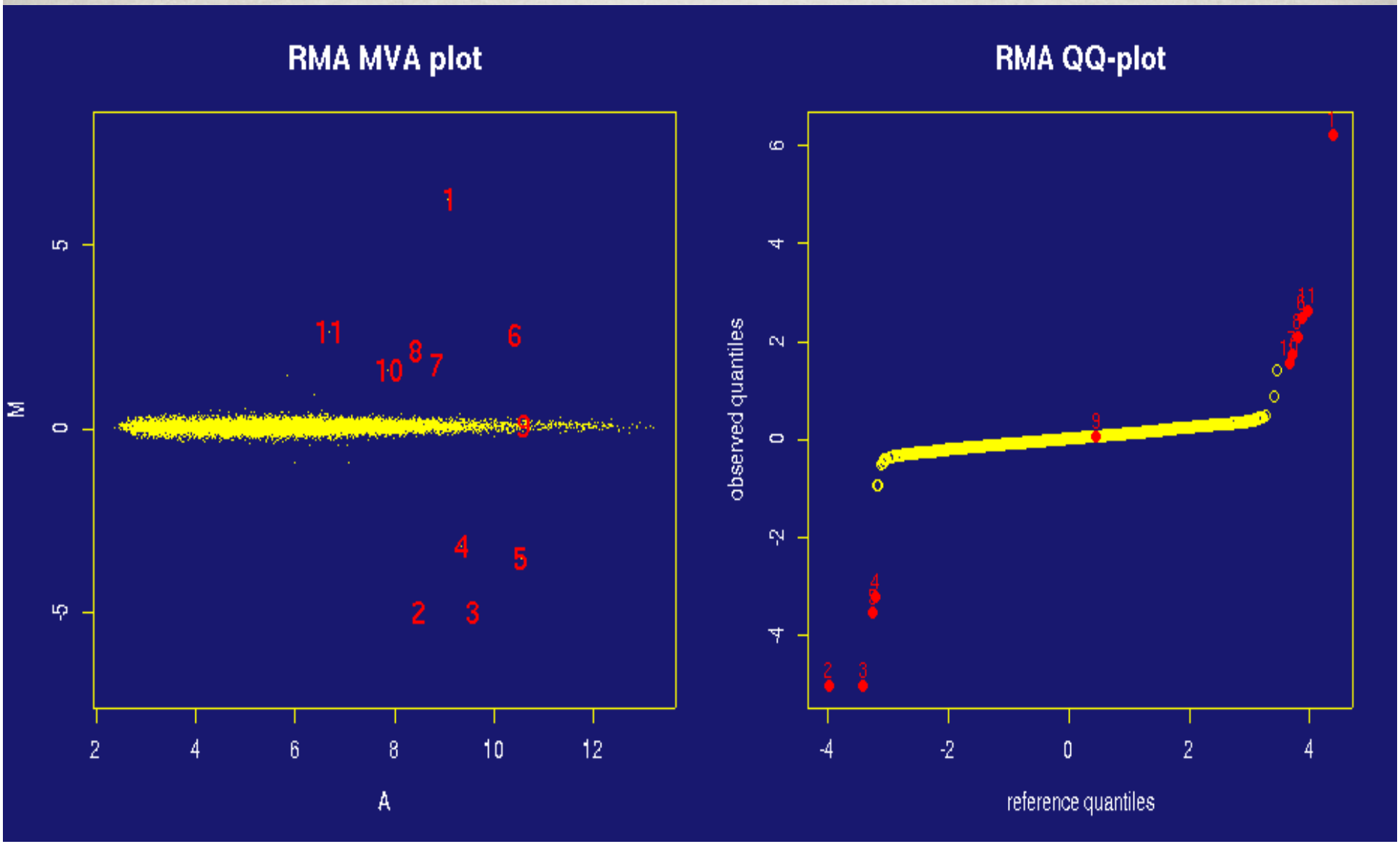
Assessing Differential expression



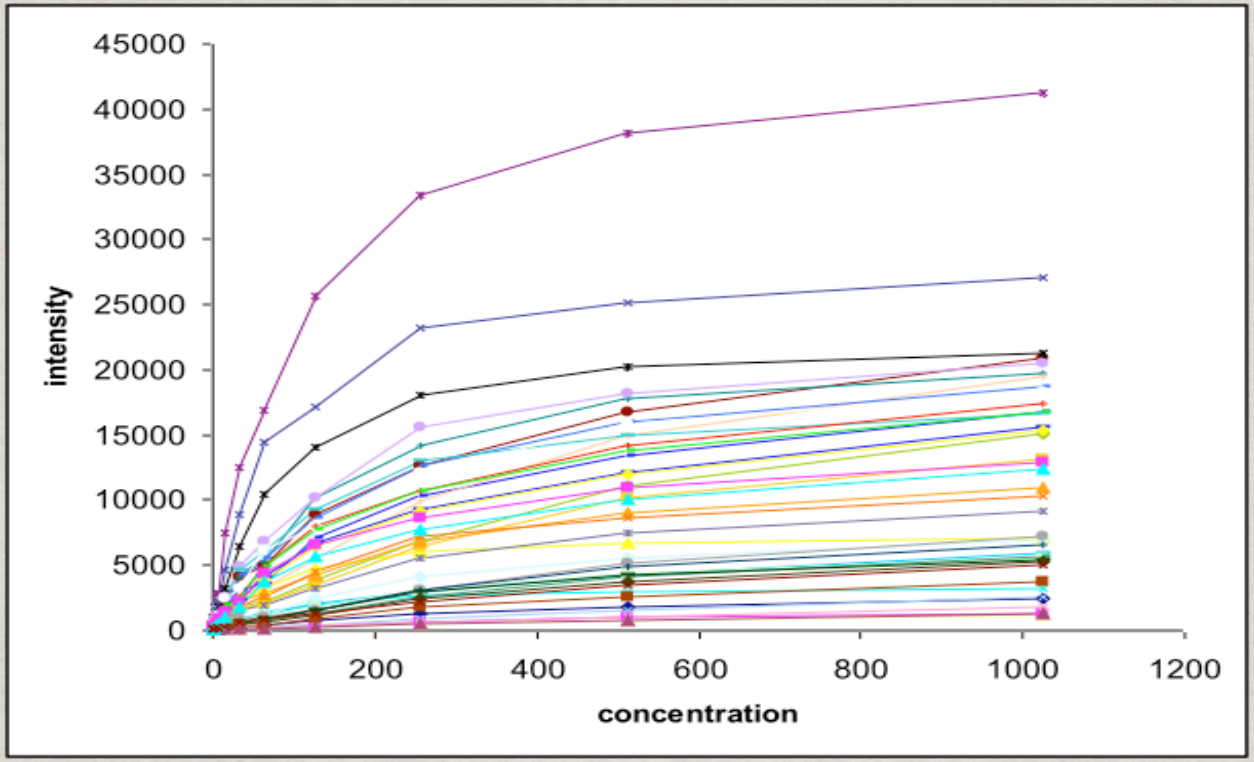
Assessing Differential expression



Assessing Differential expression



The Saturation Intensity Mystery



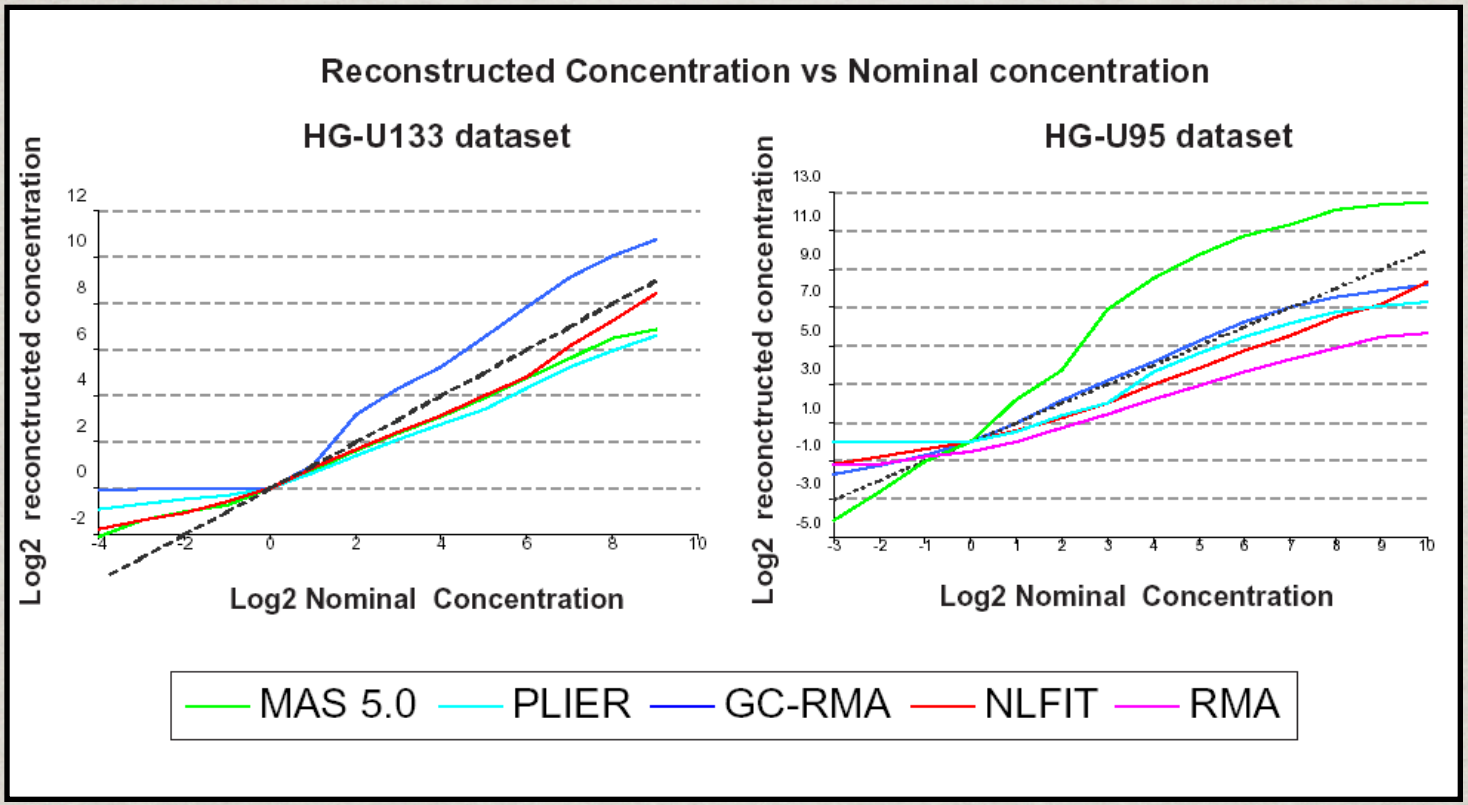
Skvortsov D, et al NAR. 2007

Adsorption isotherms exhibit different saturation intensities for Affymetrix probes

Visual examination of the signal response to concentration reveals a nonlinear relationship in transcript abundance and signal

Amongst different probes there is up to a 50x difference in saturation levels and variation within these levels correlates with probe sequence composition

Comparing Expression Measures



Abdueva D, et al NAR. 2006

Reconstructed concentration range versus nominal concentrations of HGU133 (left panel) and HG-U95 (right panel) spike-in datasets. Reconstructed curves were shifted vertically to meet zero at the 1 pM spike-in. The identity line is shown in dashed black.

Summary

RMA reduces variability without losing the ability to detect differential expression

Its not perfect, especially at detecting low and high abundance transcripts, but it performs well in terms of bias, variance and model fit & is available in BioConductor

See **Affycomp** for further comparisons

Probe annotation matters!

- ✦ Annotation by Affymetrix is the best available at the time the arrays are first release
- ✦Gene and transcript definitions constantly evolve and hence original annotations can be out-dated
- ✦Additionally, many genes are defined by multiple probe sets and the resultant expression values for the same gene can differ
- ✦Has motivated the remapping of probes to the latest genome annotation and new probe mapping definitions (several variations exist)
  - ✦ Custom CDFs are available, some of which are geared towards different databases of transcripts (Entrez Gene, RefSeq, Unigene)
- ✦ Differing annotations can dramatically alter the interpretation of your results

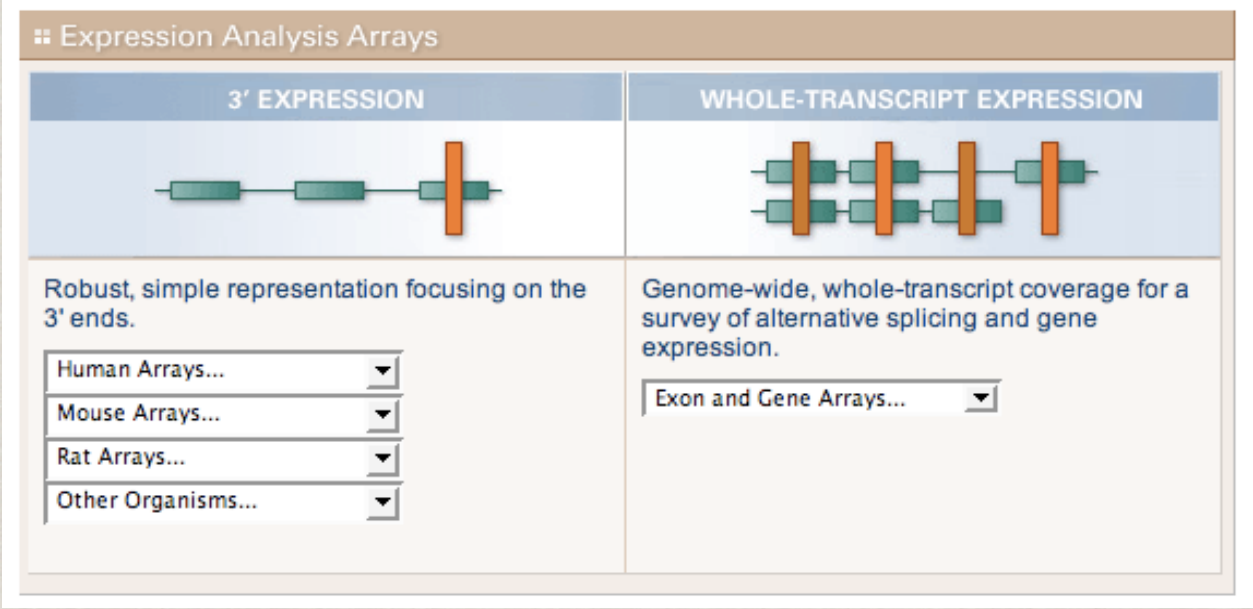
Alternate Probe annotations

- Dai, M., Wang, P., Boyd, D., Kostov, G., Athey, B., Jones, E., Bunney, W., Myers, R, Speed, T.P., Akil, H., Watson, S. and Meng F. (2005). Evolving gene/transcript definitions signicantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33:175.

- Lu, J., Lee, J.C., Salit, M.L., Cam, M.C. (2007). Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. *BMC Bioinformatics*, 8:108.

- Ferrari, F., Bortoluzzi, S., Coppe, A., Sirota, A., Safran, M., Shmoish, M and Ferrari, S.,Lancet, D., Danieli, G.A., Bicciato, S. (2007). Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*, 8:446.

Other GeneChip Arrays

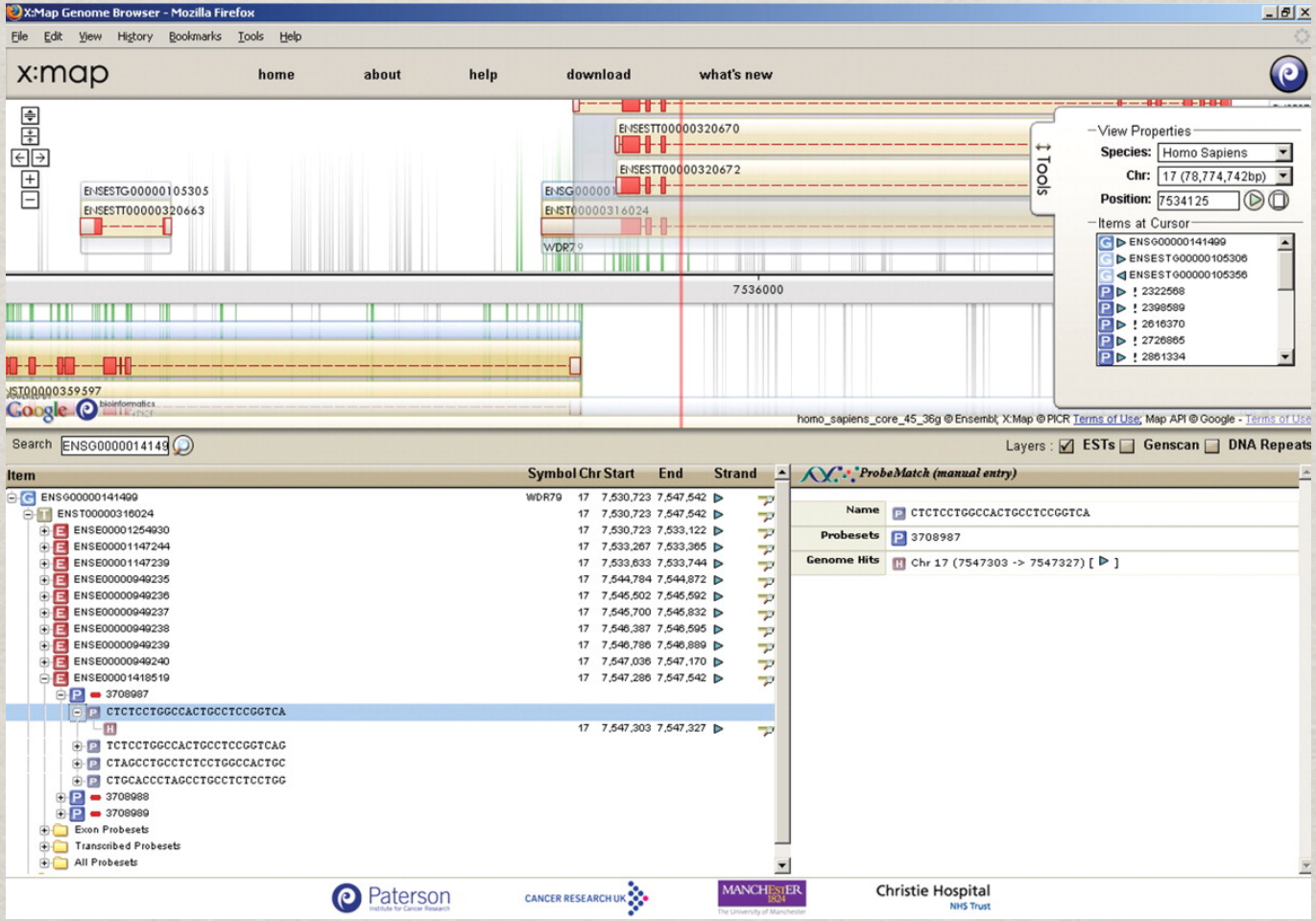


More than 50 chip types available, representing nearly 30 different organisms and with various applications including: Exon arrays, SNP arrays, and tiling arrays

Exon Arrays

- Provide an additional form of genomic information, namely, **alternative splicing**
- GeneChip Human Exon 1.0 ST (HuEx) array contains 5.4 million features grouped into 1.4 million probe sets, interrogates >1 million exon clusters
- Four probes per probe set (affects distribution of gene-level variance)
- Does not use paired MM probes
  - Surrogate background intensities can be derived from 1000 pooled probes with same GC content as PM probe
- Median # probes per RefSeq gene is 30-40 along transcript
- Designed to be as inclusive as possible at the exon level

X:MAP (tool for exon arrays)



http://xmap.picr.man.ac.uk

- Yates T, Okoniewski MJ, Miller CJ (2008). X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res.* 36:D780-6

Human Gene Array

- ✦Designed to measure expression over the full length of a transcript, like exon arrays
- ✦Primarily consists of a subset of HuEx probes, namely those that are well annotated
- ✦Could serve as an economical exon array for well annotated probes ... further study required

SNP Arrays

- ✦Genotyping and copy number (CN) analysis
- ✦SNP 5.0 contains of 8 probes/SNP, no MMs
  - ✦Contains > 500K SNPs & ~340K CNPs
  - ✦Median physical distance between SNPs is 1.8 kb
- ✦SNP 6.0 contains 6 probes/SNP, for alleles A and B, there are 3 reps of PMA and 3 of PMB
  - ✦Contains > 900K SNPs & > 900K CNPs
- ✦CNPs are unreplicated non-polymorphic probes for CN inference

BioConductor software

✦**3' Expression arrays**

✦affy

✦affyPLM

✦simpleaffy

✦affycomp

✦gcrma

✦plier

✦vsn

✦affylmGUI

✦aroma.affymetrix

✦annotation packages

Other flavors

✦XMap; Exon array-specific

✦oligo (All Affy/NimbleGen)

✦crlmm (Affy/Illumina SNP)

Acknowledgements

- ✦Christina Curtis
- ✦Matt Ritchie
- ✦Terry Speed
- ✦Rafael Irizarry
- ✦Jean Yang
- ✦Bridget Hobbs
- ✦Ben Bolstad
- ✦Francois Collin
- ✦Gordon Smyth

Practical

Exercises 7-11