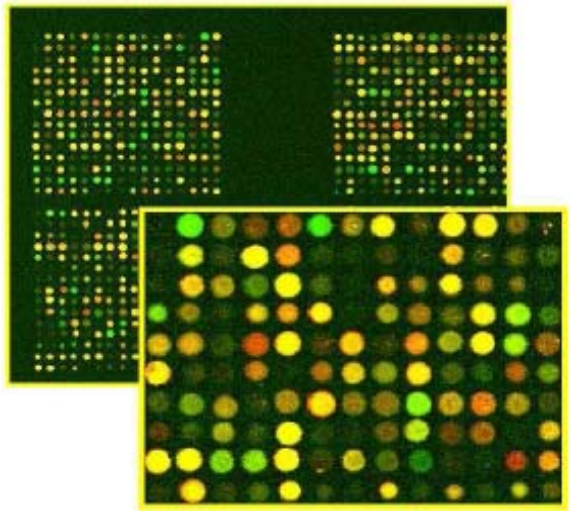
PART II: Genome Analysis

Chapter 7. DNA Microarrays
Chapter 8. DNA Analysis



7. DNA Microarrays

7. DNA Microarrays

7.7 Different combinations of the processing steps

7.8 Statistics

7.9 Gene Selection

7.10 Next Generation Sequencing

454 Sequencing http://www.roche.com/

Solexa Illumina http://www.illumina.com

Solid ™ System http://www3.appliedbiosystems.com/AB_Home/index.htm

7. DNA Microarrays
7.1 Motivation

High-density and high through-put method

- To monitor [mRNAs]
- Patterns of gene expression
- Genetic networks information
- Systematical analysis of both cell condition responses and states
- Medical applications

  - Diagnosis and Prognosis
    - Tumor sample → marray → GEPs: Kind of cancer and current status (Leukaemia, Schizophrenia, breast cancer, etc,…)

Treatments selection Drug dose adjustment Indicative genes

Techniques

Southern blot (1975) complementary nts sequences affinity on porous surfaces

↓

↓

Northern blot (Western blot)

↓

Immunoassays Antibody-antigen affinity

↓

↓

Microarrays complementary nts sequences affinity on glass array surfaces

http://www.affymetrix.com/

1. Expression Arrays

- ✓ Whole-Transcript Expression : Exon and Gene Arrays
  - Human-Mouse-Rat Exon 1.0 ST Array
  - Human-Mouse-Rat Gene 1.0 ST Array
- ✓ 3' Expression :
  - Human Genome U133 Plus 2.0: 61,000 probe sets: 47,000 transcripts + 45,000 human genes
  - Human Genome U133: 45,000 probe sets: 39,000 transcripts + 33,000 human genes
  - Human Genome U95 Set: 63,000 probe sets: 54,000 UniGene clusters
- ✓ 3' Array Plates:
  - Human-Mouse-Rat
- ✓ microRNA Expression:
  - microRNA Array

2. Genomic-DNA Analysis Arrays

- ✓ SNP Genotyping and CNV Analysis:
  - *Mapping and Genome-wide Human SNP arrays
- ✓ Targeted Genotyping
  - Universal 3,5,10,25,70 K Array (to avoid cross-hybridization)
- ✓ Resequencing
  - CustomSeq, Human Mito and SARS (Complete sequences)

3. Gene Regulation Arrays

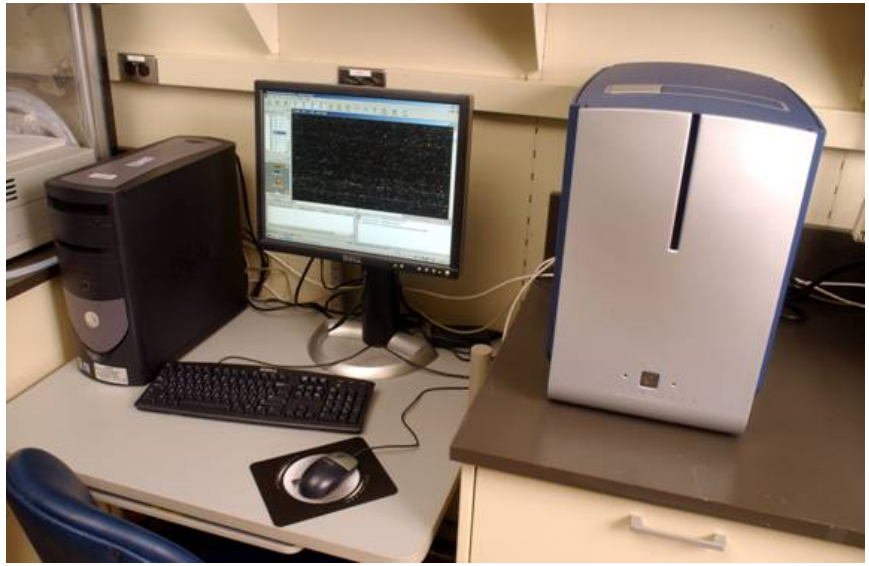
- ✓ CHIP-ON-CHIP
  - *Whole Genome Analysis:
    - Human-Mouse-Arabidopsis-Drosophila-C.Elegans-S.Pombe- S.Cerevisiae Tiling 1.0 and/or 2.0 Array Set ()
  - Promoters
- ✓ Transcript Mapping
  - Whole Genome Arrays (as *)
  - Promoters and ENCODE regions

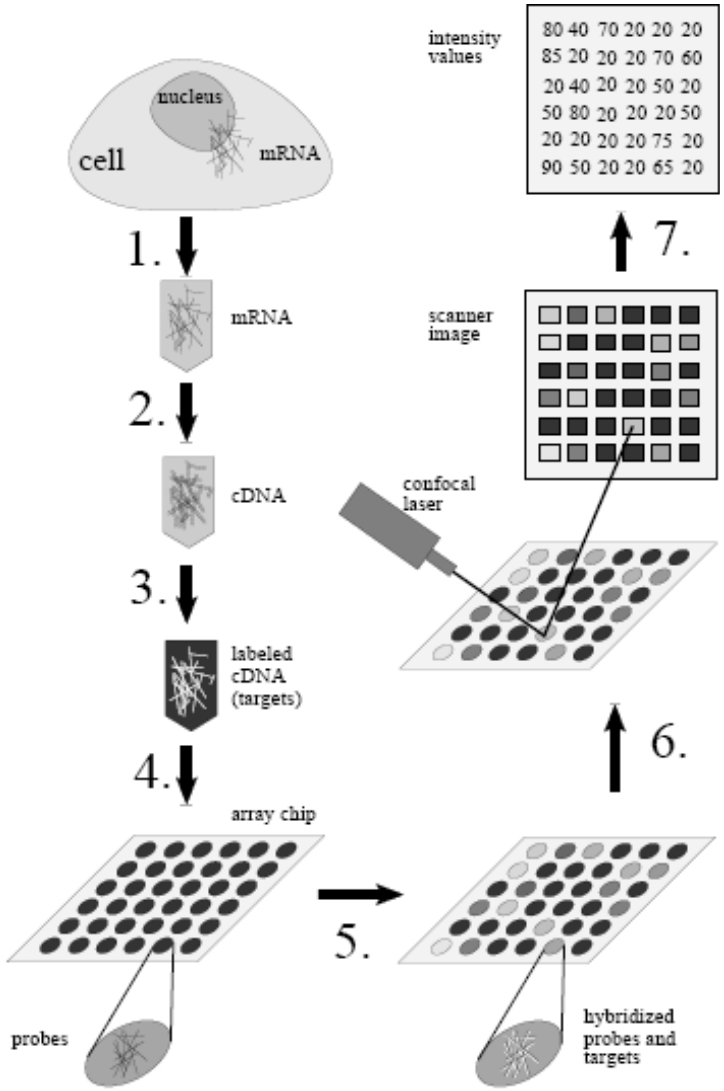
Affymetrix devices

Fluidics station

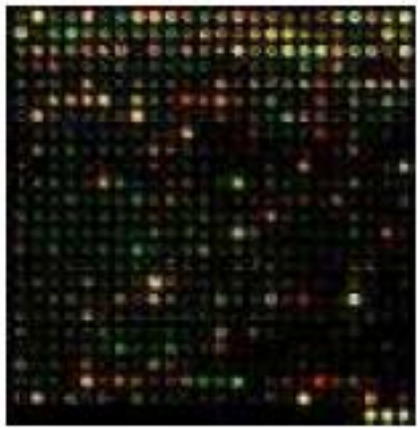
Wash / Stain

Scanner/Computer Station





1. mRNA extraction
2. Reverse transcription (OligodT + T7 and SP6 promoters-3' and 5' ends- + RNA polymerase)
3. Target cDNA (cRNA) marked with fluorescent marker
4. Probes in the chip
5. Hybridize cDNA (cRNA) targets against chip probes
6. Scan with confocal laser microscope
7. Strength of the fluorescent light is recorded
8. Intensity values to real numbers



Matrix entry

Value measurement

	Sample1	Sample2	Sample3	Sample i
Gene 1	mv_{11}			
Gene 2				
Gene 3		mv_{23}		
		...		
			...	
Gene j				mv ij

Oligonucleotide Arrays

Complementary short sequences or *probes* of 20-70 nts immobilized in the chips

One dye: Biotin

Spotted or cDNA arrays

Hundreds of complementary nucleotides for detecting mRNA

Two dyes: cys3 and cys5

When target is available: Expressed Gene

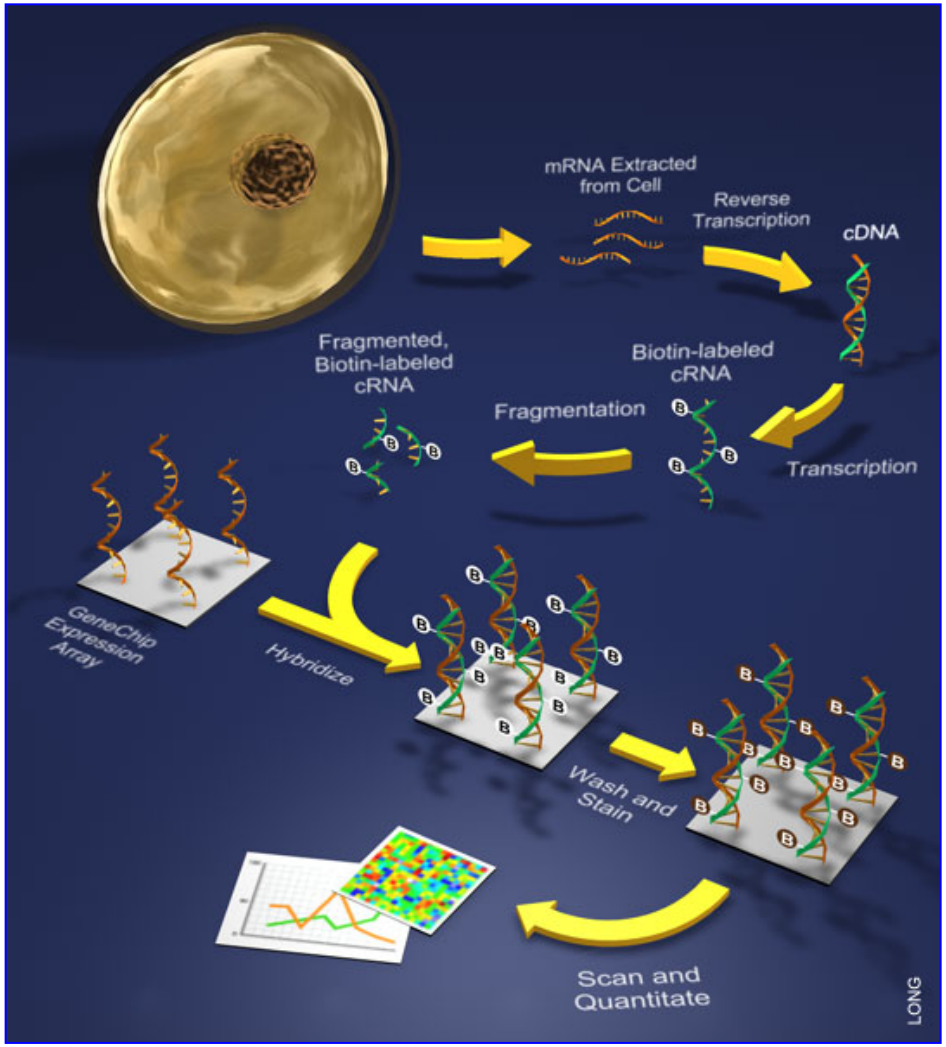
→ Hybridization: Probe + Target

Target: labeled/marked sequence from the sample to be analyzed (cDNA / cRNA)

Probe: Target complementary sequences immobilized in the arrays

7. DNA Microarrays
7.3 DNA Microarray Techniques

1 Oligonucleotide Arrays



DGC---→ PCR products : cDNA clones

96-Well plated format

||

Each PCR product: Electrophoresis

||

Plates -→→Pools 19 vials 1.5mL

||

Every Pool 1-4 Plates: Approx. 98-384 clones

|| + Biotin dye

$cDNA^{Bio}$ ($cRNA^{Bio}$) : Purity and [] control

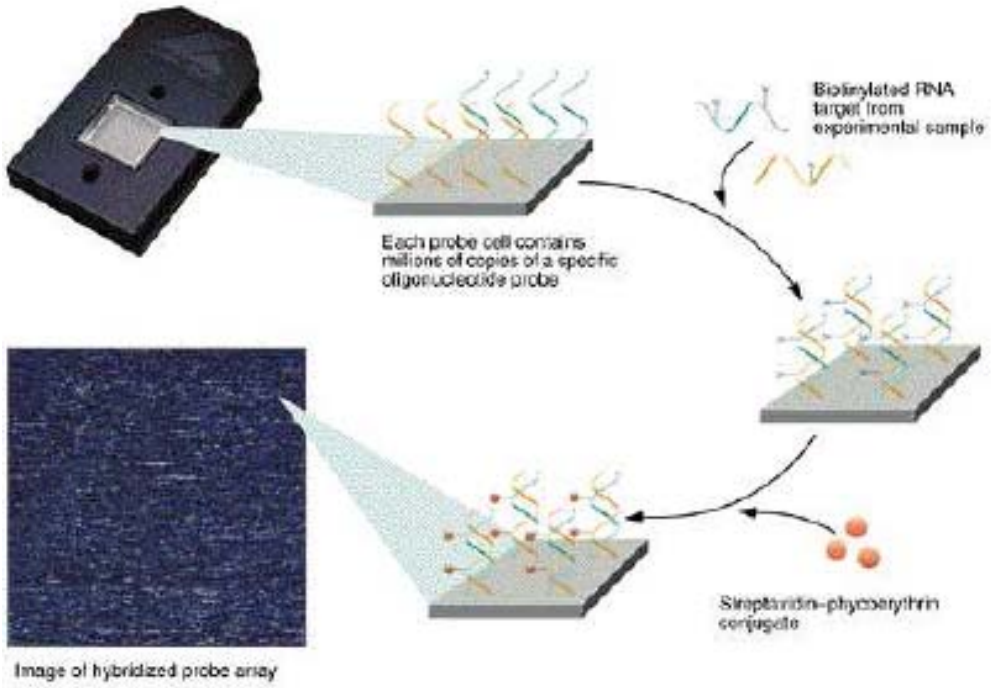
||

Hybridization Affymetrix GeneChip$^{mbox©}$

||

Confocal scanning and Quantification

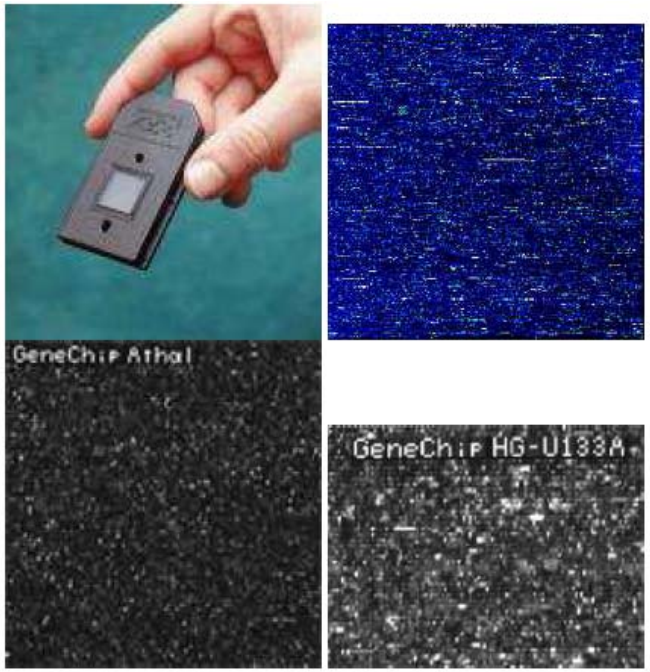
1 Oligonucleotide Arrays



Affymetrix technology

11-20 Oligonucleotides of 25 bp length

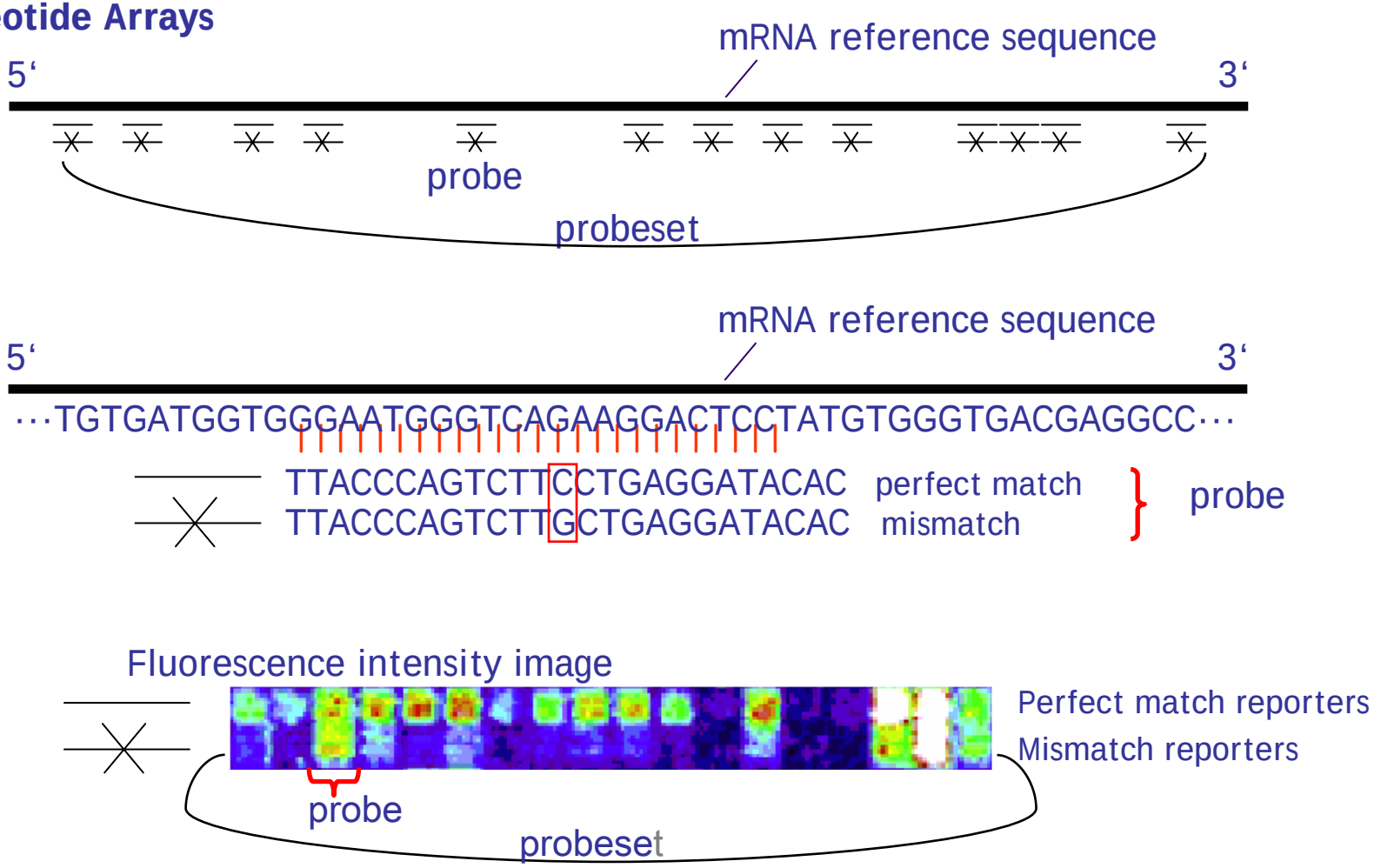
600 bp mRNA reference sequence (3')



GeneChip[mbox©]

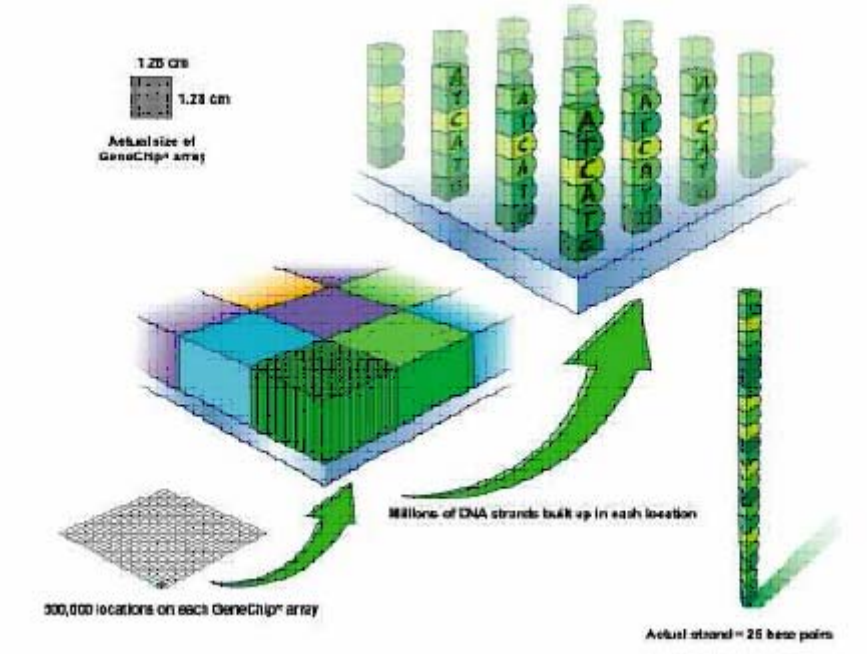
Images obtained

1 Oligonucleotide Arrays



Expression level of each RNA specie reported by a probeset distributed over the 600 bp

1 Oligonucleotide Arrays



PCR products or oligonucleotides generated from the genome public databases corresponding to those genes can be spotted onto the chip

The development of new solid supports and miniaturization permits that the genes are spotted at very high density: allows the parallel analysis of thousands of genes

1 Oligonucleotide Arrays

Each Target sequence → One probe Set : 11-20 pairs of oligonucleotides

Target Transcript for Human recA gene:

ctcagcttaagtcatggaattctagaggatgtatctcacaagtaggatcaag

```
ctcagcttaagtcatggaattctag                          PM1
ctcagcttaagtgatggaattctag                          MM1
 tcagcttaagtcatggaattctaga                         PM2
 tcagcttaagtct tggaattctaga                        PM2
                  attctagaggatgtatctcacaagt         PM3
                  attctagaggatctatctcacaagt         MM3
                        aggatgtatctcacaagtaggatca   PM4
                        aggatgtatctc t caagtaggatca MM4
```

PM: target mRNA measurement

MM: background measurement

A/P call PM-MM

> 0 transcript detected Present call

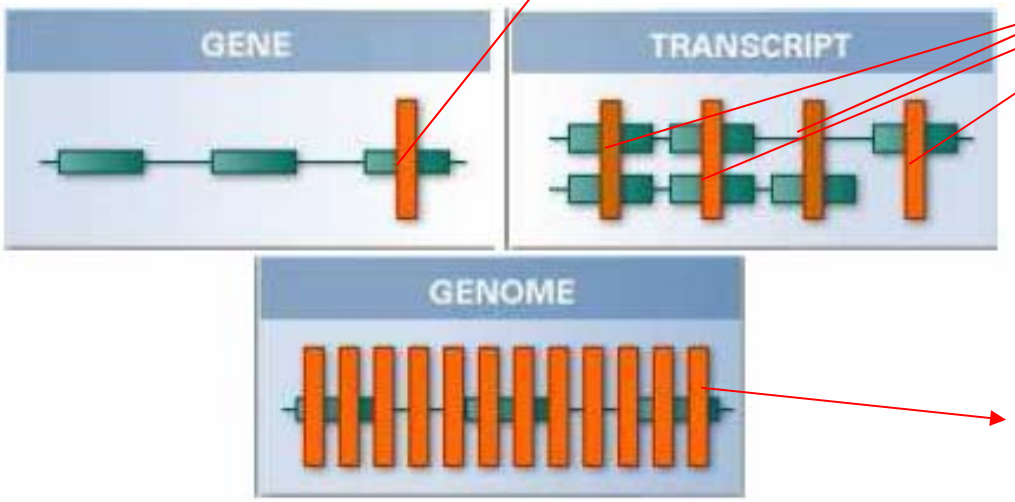
<0 transcript not detected Absent

Summarized to avoid the use of the information of each probe on the noise level information used in the I/NI call

1 Oligonucleotide Arrays

Affymetrix techniques

3'exon end marked by a probe set



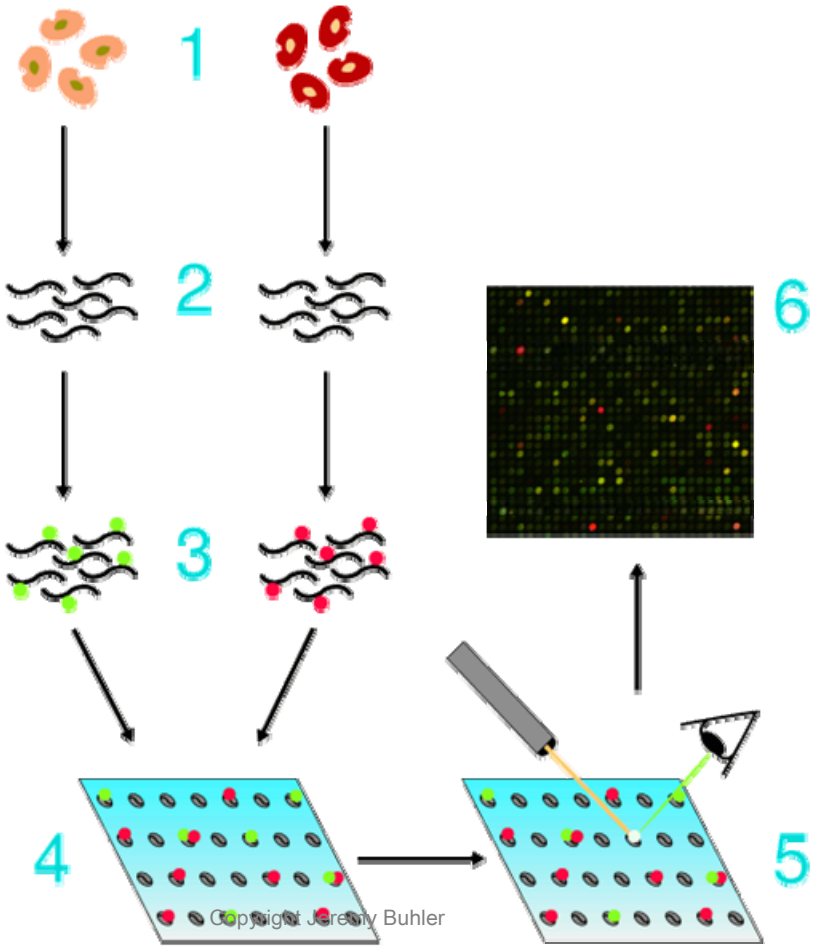
Probe location of an exon array marked by 3 to 4 probe set

Alternative Splicing detection

35bp shifting in a Tiling array leaving gaps of 10 bp

Non coding transcript detection

2 cDNA / spotted Arrays



Red-Green Microarray Technique

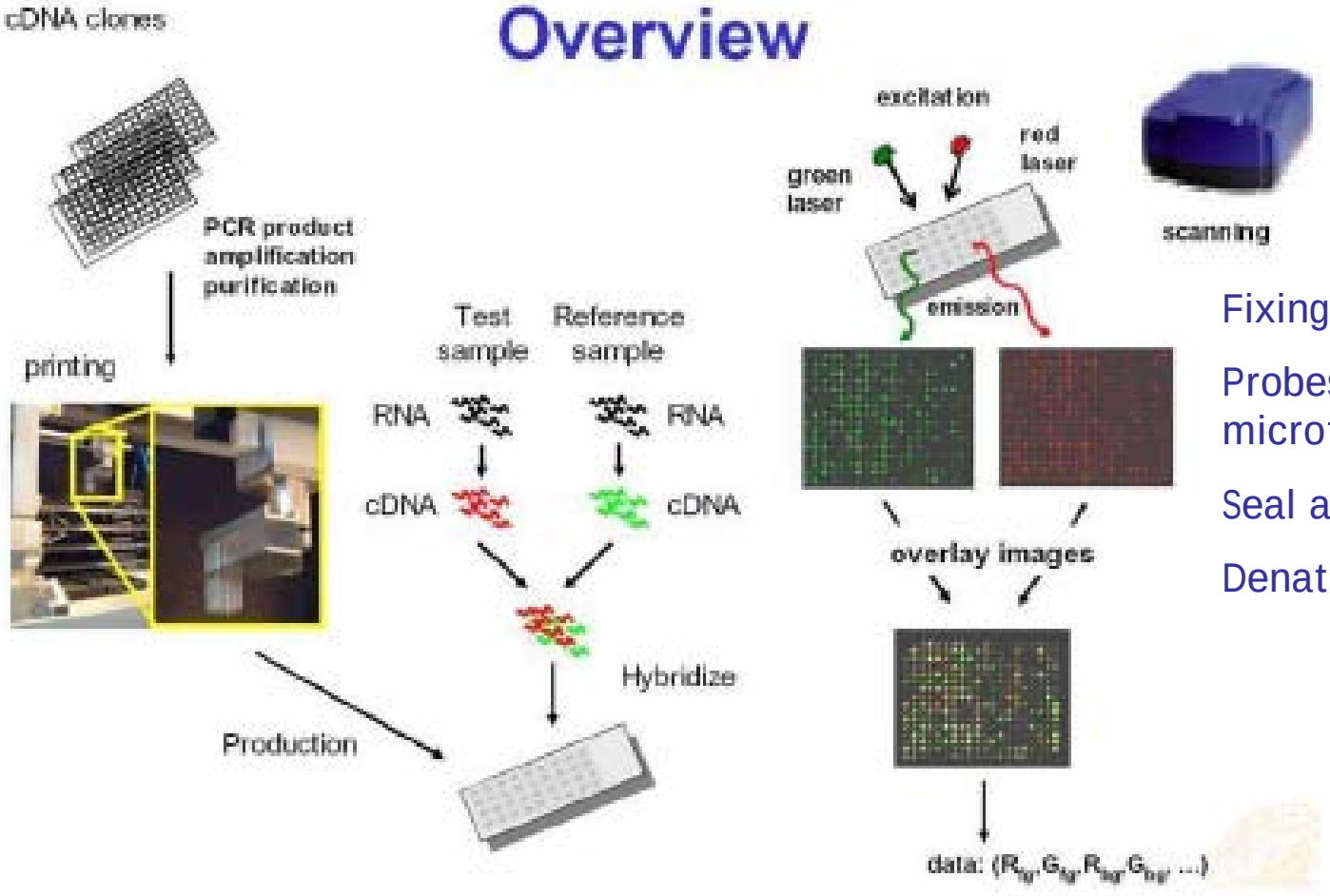
1. Select samples
2. Extract mRNA (PCR) and perform reverse transcription (cDNA)
3. Label with fluorescent marker (Cy3 red, Cy5 green)
4. Dyed lines mixed
5. Hybridize (chip immobilized probes and cDNA/cRNA targets bind)
6. Confocal microscope scanning with excitation to lead red and green emitting fluorescent

Segmentation and average or ratio of R/G (log-ratio) intensities are computed

Process 2X Control and sample

2 cDNA / spotted Arrays

Glass chip manufacture



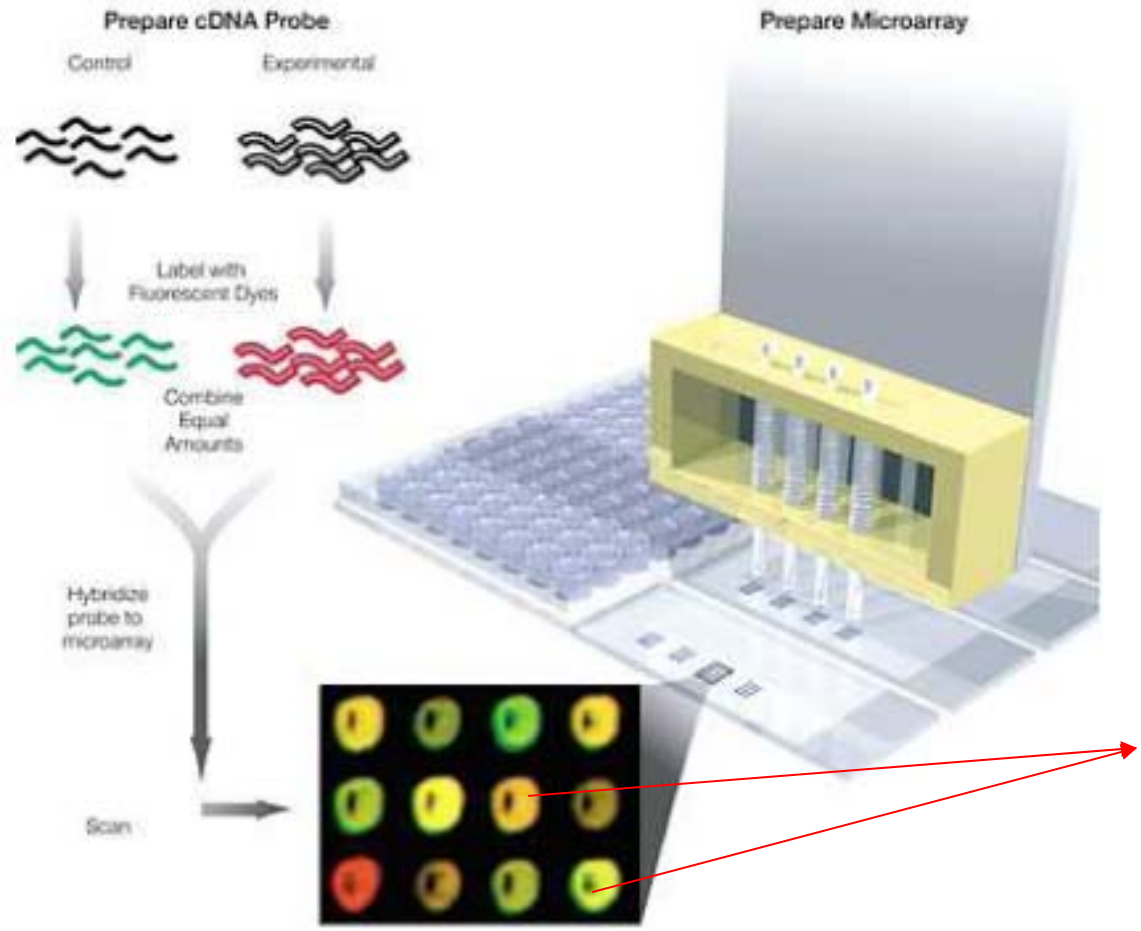
Fixing regions preparation

Probes synthesis and microtiter

Seal and deactivation

Denatured DNA SSPs

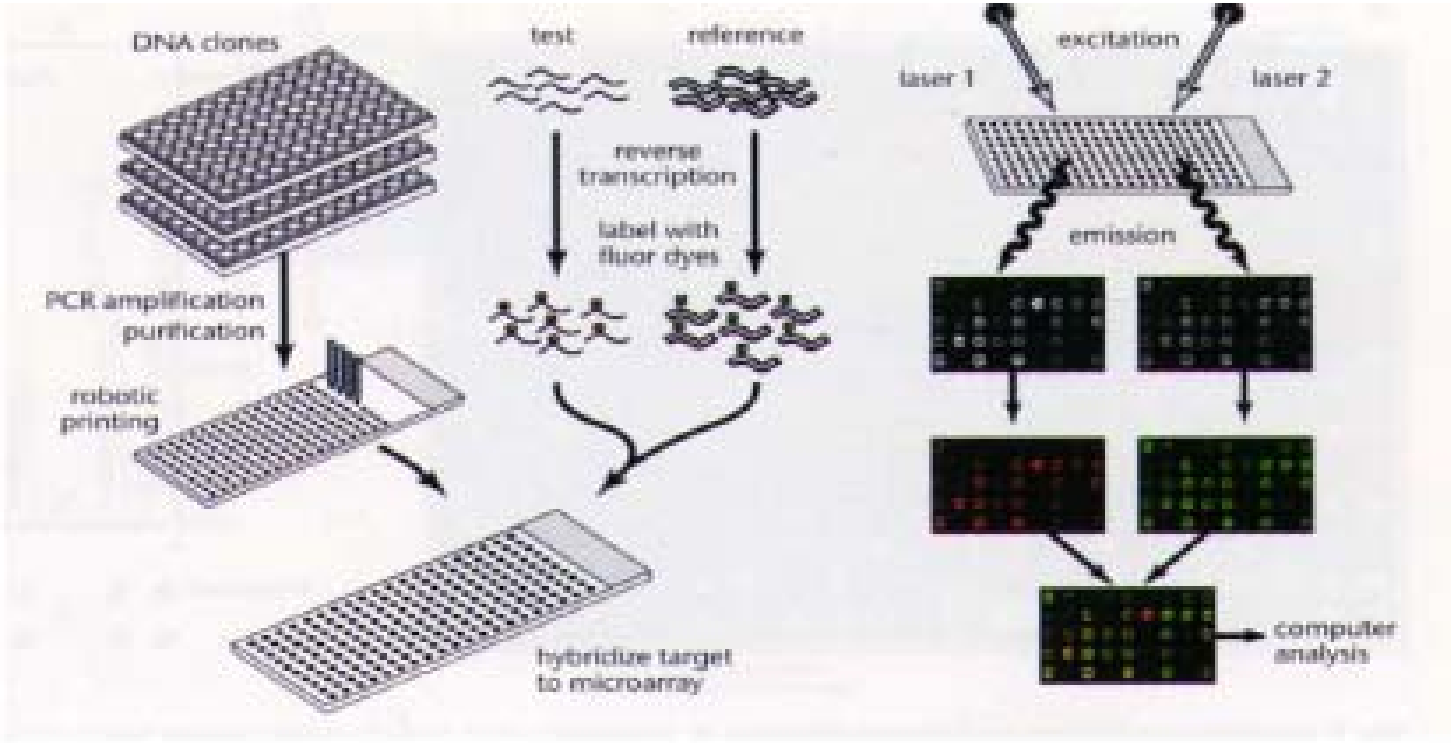
2 cDNA / spotted Arrays



Robot spotter brings small quantities of the probes onto a glass plate: fixed to glass

Signal: DEGs

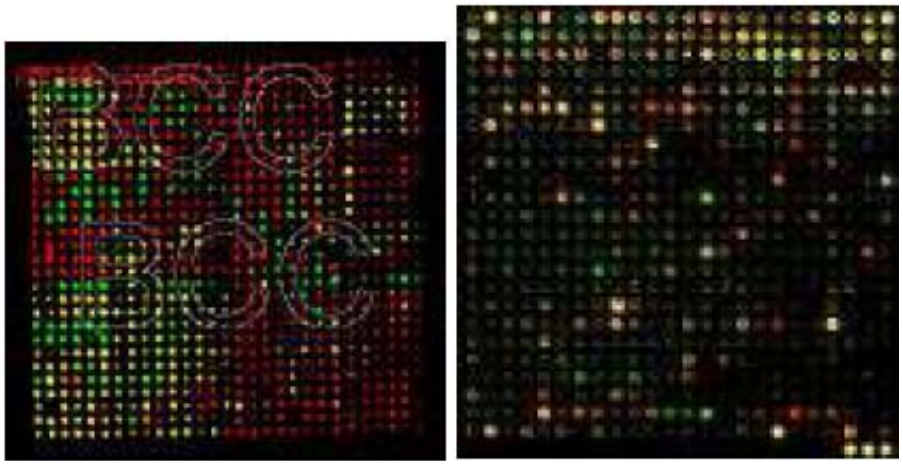
2 cDNA / spotted Arrays

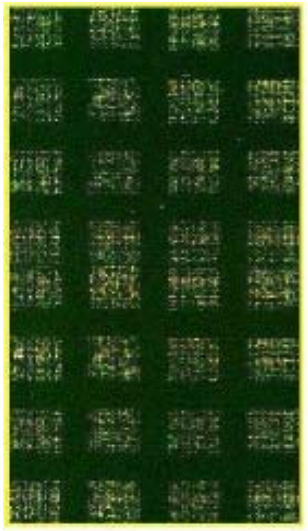


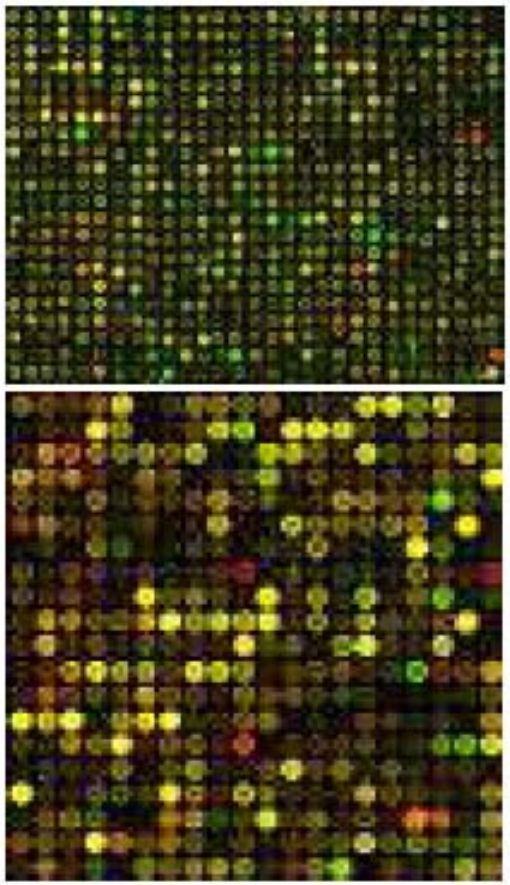
Log-ratio **R**/**G**

Intensities transformed to real numbers after segmentation of the stained location

2 cDNA / spotted Arrays







Scanner images of **R**-**G** spotted arrays

3 Other techniques

Serial Analysis of Gene Expression (SAGE): sequencing short ESTs

- Standard methods: tags or transcription segments of 100-300 bases
- SAGE: tags or transcription segments of 9-14 bases within a gene

Drawback: more than one gene sharing the same tag
No guarantee of whole gene transcription

- Presence of a tag → Transcription / Expressed Gene

Digital Micromirror Arrays: oligonucleotide DNA arrays reading put by a CCD camera

- Chips under lights that activate the probe and labeled target is added

3 Other techniques

Inkjet Arrays: Standard inkjet printing from Hewlett Packard

- Probes pre-synthesized on a glass slide can be printed or created nt by nt

Bead Arrays: Oligonucleotides attached on small glass beads

- Beads linked to a substrate are brought into the array. Localization and hybridization

Nanomechanical Cantilevers: Oligonucleotides probes attached to cantilevers silicon gold surfaces

- Binding of probes and targets detection by deflection angle of a laser beam

7. DNA Microarrays
7.4 Microarray Noise

Expression values → **Noise origin**

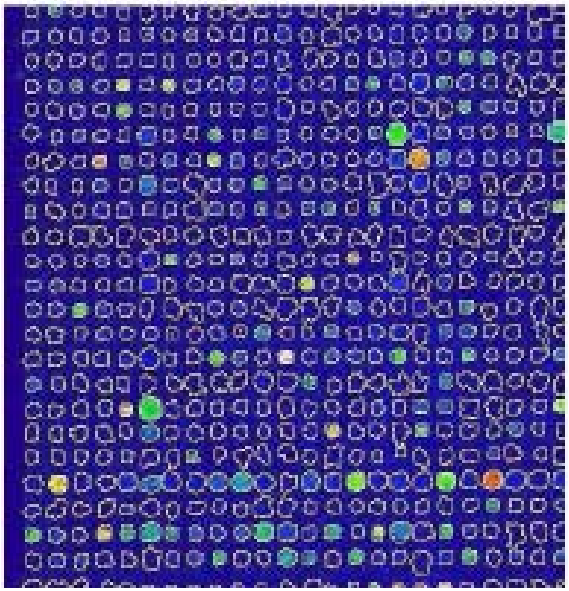
- Chip fabrication
- Microarray measurement technique
- mRNA extraction
- Reverse transcription
- Background intensity
- Non-uniform target labelling (multiple times bind, [dye], etc,..)
- Pipette errors
- Temperature fluctuations
- Hybridization efficiency
- Scanning deviations
- Biological variations (tissue samples vary in their RNA content)

Not Gaussian distribution: Log expression distribution of noise with heavy tails

7. DNA Microarrays
7.5 Image Analysis

First computational step with computer science tools to improve the results

Goal → Get intensity value for each spot



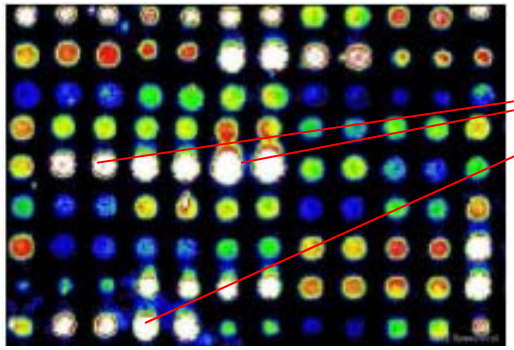
1. Addressing or “gridding”: Spot centers localization . Grids aligned to intensity peaks

2. Segmentation: Spots from background separation. Classification of pixels either as signals and bg

3. Intensity/Information extraction: from both spots and bg

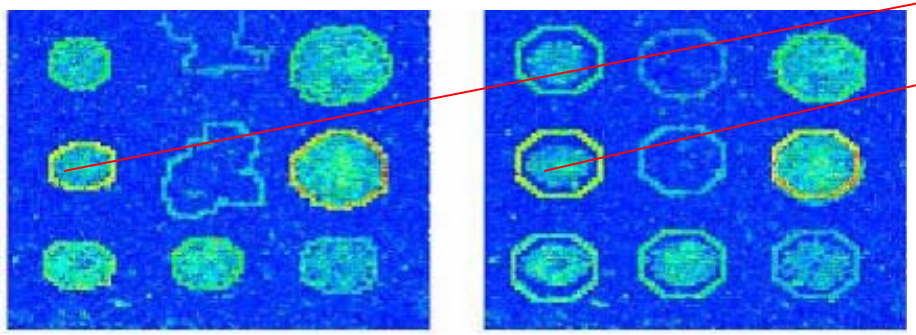
   - Background correction
   - Signals intensity pairs calculus
   - Quality measurement

Difficulties during “gridding” step: How to align intensity peaks



Spots with different size, shape and intensities that can overlap

Spots Segmentation



Growth until intensity is decreased

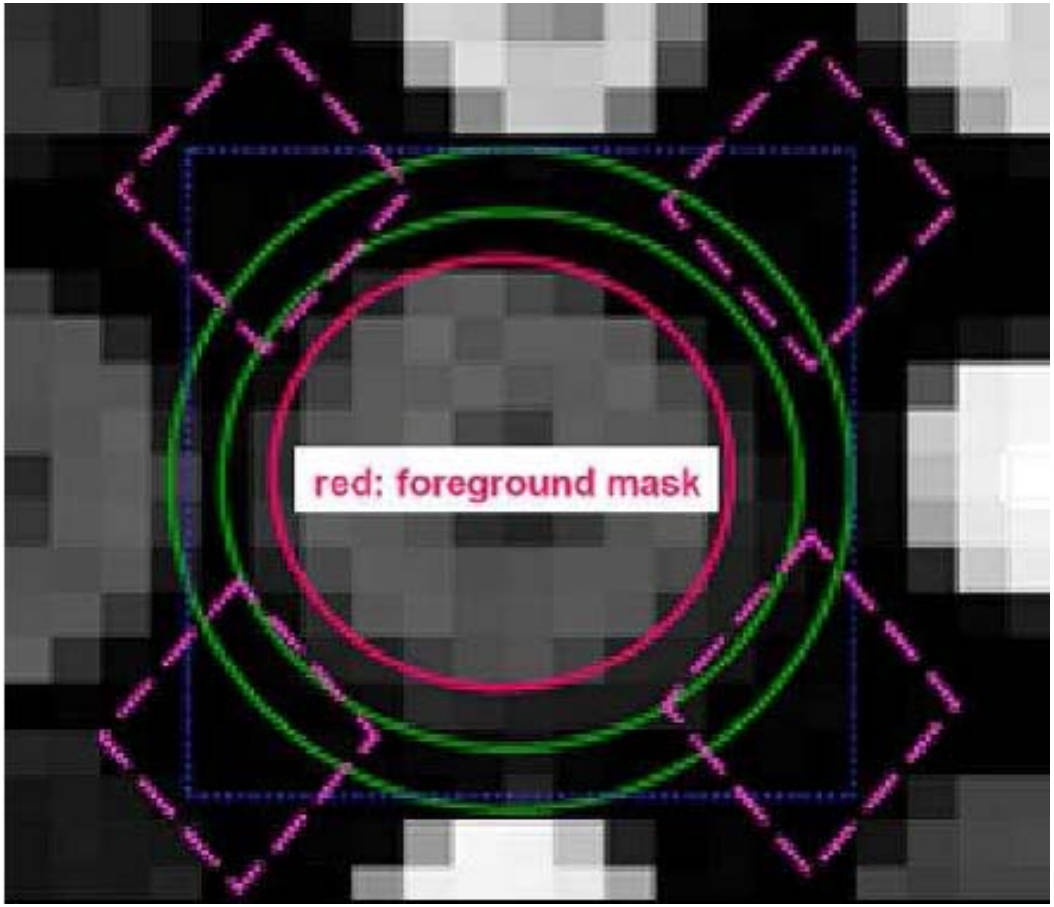
Not always matching btw spot-circle

Adaptive circles / radius adjustment

Seeded Region Growing Fixed Circle

7. DNA Microarrays
7.5 Image Analysis: Background correction

Spot Intensity = Surrounding environment intensity →



NOT probes are attached!!!

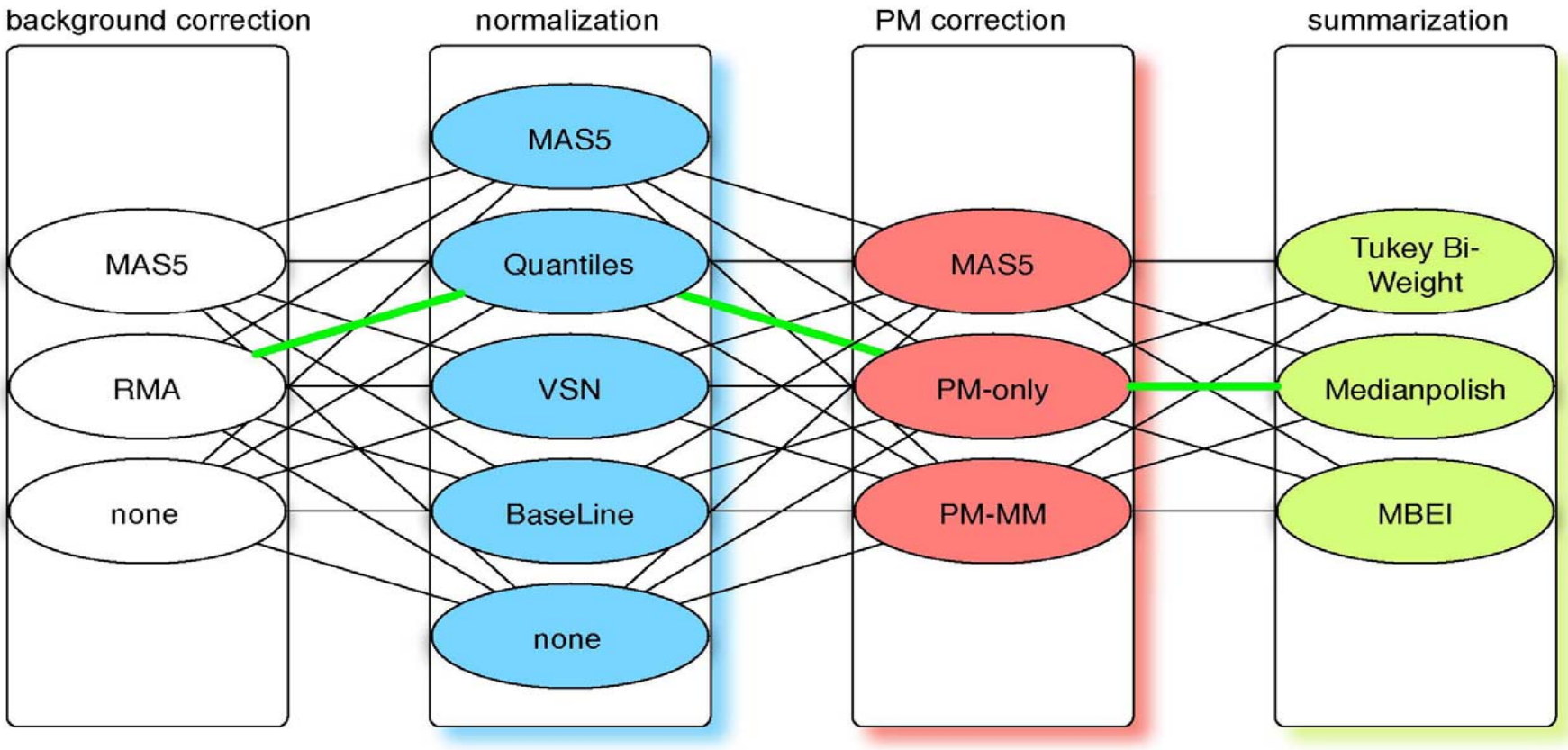
Spot Value = BG Value = Zero

BG should be extracted

Red: foreground mask

Pink: BG mask

Preprocessing Goal: extract signal **s***i* [**mRNA]** for each sample/chip *i*



BG correction
↓
Normalization
↓
PM correction
↓
Summarization
↓
Statistics
↓
Results Interpretation

Metrics used

Log_2 (PM/MM)

PM-MM

Log_2 √(PMxMM)

PM-MM/PM+MM

Metrics comparing the PM with MM signals perform better at distinguish signals due to correct hybridization than PM alone NO!!

1 Background Correction

Background Correction Techniques: Subtracting the BG from the signal

Affymetrix Microarray Suite (MAS5)

Robust Multi-array Average

Felix Naef

Background Correction

MAS5: Affymetrix Microarray Suite 5.0 [Aff.2001,Hubbel et al.,2002]
signals due to non specific bindings

Metric PM-MM/PM+MM ≃ Log_2 (PM/MM)
Array divided into 16 rectangular "zones"
Local background: the lowest 2% intensities in the "zones"

Local background subtracted from both PMs and MMs
PMs and MMs kept above a positive threshold

Background Correction

RMA: Robust Multi-array Average [Irrizary et al., 2003b,a,Bolstad et al., 2003]

Assumptions:

Signal density S is distributed exponentially $p_S(S) = \alpha\, e^{-\alpha x}$

Background density B distributed normally

$$p_B(B) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(B-\mu)^2/(2\sigma^2)}$$

BG: Only positives contributions exist== BG additive and positive
→truncated Gaussian

α : Estimated by the average distance of PM to their mean signal intensity value
μ: The mean of the MM values
σ^2: Averaged squared distance of MM values which are bellow the mean to the mean

Background Correction

Joint density

$pS,(S,O) = pS(S)pB(O-S)$

Ensure $O-S \geq 0$

Estimation:

$\check{S} = E(S \mid S + B)$ where S + B is the observed PM/MM

Felix Naef

The PM – MM < 50% are selected

Gaussian is fitted to estimated the mean of Background intensity

The small PMs differences -→→ NO signal ,, Background easily be extracted

Normalization Techniques

- Different arrays to be compared→ different conditions
  - » Different intensity levels

  - **Affymetrix- MAS5**
  - **Affymetrix- Baseline**
  - **Quantile normalization (RMA)**
  - **Invariant Difference Selection (IDS,[Schadt et al.,2001])**
  - **Cyclic loess**

Normalization Techniques

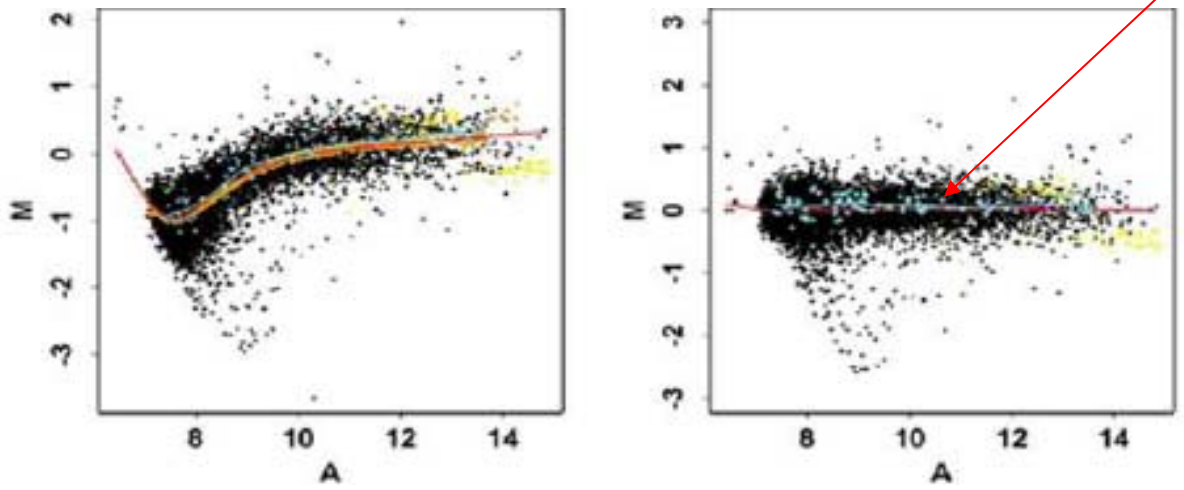
MvA plot:

- Shows the difference between chips
- Approximate the median M=0 equal intensity to avoid artifacts and intensity patterns

M= logp1 – logp2 Difference

A= 0.5 (logp1 + logp2) Intensity level

Desired



Normalization Techniques

Baseline: Affymetrix

Exclude highest and lowest 2% probes per array

Chose baseline array

The average intensities of all arrays to this baseline are globally scaled

Arrays are normalized to the median mean index

Normalization Techniques

IDS: Invariant Difference Selection ,[Schadt et al.,2001]

Find probe pairs with same order intensity → Invariant Probe Pairs (1)

↓

Same intensity difference PM-MM in an array and in a baseline array (median)

Likely NOT differentially expressed

$$R_i = \frac{(L\ (B_i + E_i) + H\ (2\,N - B_i - E_i))}{2\,N}$$

$$D_i = \frac{2\ |B_i - E_i|}{B_i + E_i},$$

1. The i-th difference is viewed as invariant

$D_i < R_i$ 2. GCVSS generalized cross validation to fit the relation of these genes

3. Final approximation to M = 0 in a MvsA plot

Normalization Techniques

Quantile: (RMA)

Goal→ distributions of probe intensities for each array in a set of arrays to be the same

Quantile → The fraction (or percent) of points below the given value [Wikipedia]

The PMs are sorted per array

Each sorted array is "aligned" one to another ("multiple alignment")

The median per column is computed and all values in a column are to the median

Forces each array/chip to have the same distribution of signal intensity

Quantile-quantile plot shows a straight diagonal when n data vectors have the same distribution

Assumes all probes in the array show constant expression level

Few expression values change with the conditions

Normalization Techniques

Transform the quantiles so that they all lie in the straight diagonal

$$\text{proj}_{\boldsymbol{d}}\boldsymbol{q}_k = \left(\frac{1}{n}\sum_{j=1}^{n} q_{kj}, \ldots, \frac{1}{n}\sum_{j=1}^{n} q_{kj}\right)$$

$$\boldsymbol{q}_k = (q_{k1}, \ldots, q_{kn})$$

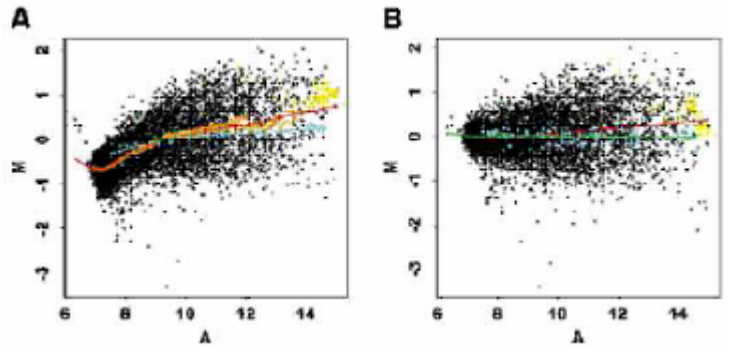
$$\boldsymbol{d} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$$

"We can give each array the same distribution by taking the mean quantile and substituting it as the value of the data item in the original dataset" B.M Bolstad

Normalization Techniques

Cyclic Loess: [Cleveland 1979, Cleveland and Devlin 1998]

- Local regression to fit data and readjust back to linear scale
- Nonlinear intensity dependent/dye dependent for cDNA **R**-**G** arrays
- Predicted loess value subtracted from the data to decrease the standard deviation and place the mean log ratio at 0
- Normalization for pairs of arrays
- Finally averaging for the resulting M and A values



Original data (curve fitted)

Data mapped to a linear scale

Curve used to map the intensity values back to linear scale

$M = \log p_1 - \log p_2$

$A = 0.5 (\log p_1 + \log p_2)$ Intensity level

Normalization Techniques

Cyclic Loess:

Weighting function: More weight to points whose response is being estimated

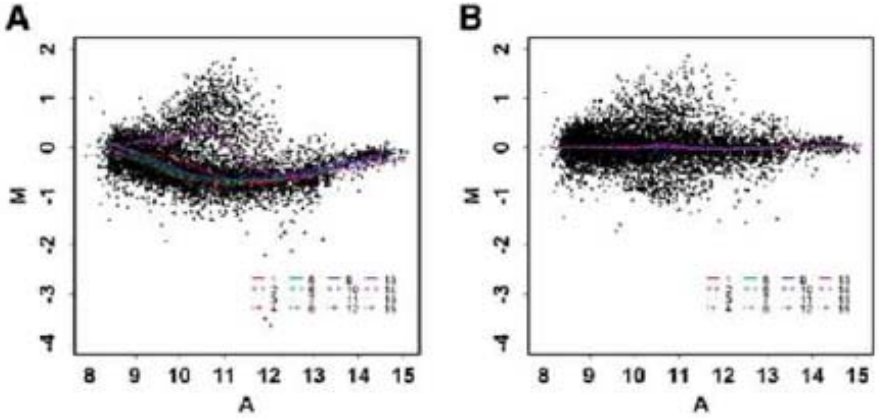
Less weight to points further away

At each point in the data set a polynomial is fit to a subset of data using the weighted least squares

Finished after regression function values are c

$$w(x) = \begin{cases} (1 - |x|^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases}$$

oints



Local linear quadratic model is obtained

Computationally very intensive

PM Correction

Combine PM and MM intensity correction to remove nonspecific signals contribution and to obtain one value for each probe pair

PM-MM

Cut-off at the value in which the MM exceeds its corresponding PM intensity

PM only

IM values **(MAS5)**

Ideal Mismatches: Negative values are avoided by estimating the nonspecific signals when MM > PM

$$\mathrm{IM}_i = \begin{cases} \mathrm{MM}_i & \text{for} \quad \mathrm{PM}_i > \mathrm{MM}_i \\ \exp(-\mathrm{SB})\,\mathrm{PM}_i & \text{for} \quad \mathrm{PM}_i \leq \mathrm{MM}_i \end{cases}$$

$$\mathrm{SB} = \frac{\tau}{1 + 0.1\ (\tau - \mathrm{SB}_1)}$$

$$\mathrm{SB}_1 = \mathrm{TB}(\log(\mathrm{PM}_j) - \log(\mathrm{MM}_j), 1 \leq j \leq N)$$

Biweight Specific BG for probe pair j in PS1

Log on base 2 exclusively

Estimate is probe-specific

Estimate is NOT probe-specific

PM Correction

TB : Tukey's biweight estimation→ PM – MM are computed

TB of x with parameters c (c = 5) and ε (ε = 0.0001) is computed as

$$
\begin{aligned}
m &= \text{median}(x) \\
s &= \text{median}(\{|x_i - m|\}) \\
u_i &= \frac{x_i - m}{c\ s + \epsilon} \\
w_i &= \left(1 - u_i^2\right)^2 \\
\text{TB}(x, c, \epsilon) &= \frac{\sum_i w_i\ x_i}{\sum_i w_i} .
\end{aligned}
$$

4 Summarization

Estimate [mRNA] by combining the multiple preprocessed probe intensities to a single expression value per probe set → expression level of each gene

Tukey-biweight (MAS5)

Tukey-biweight function of $\log_2(PM - IM)$

Arrays are normalized to the median (over the arrays) mean index

Median Polish (RMA)

Fit of an additive model by median polish

Can not handle negatives data
Not **A**pplicable

4 Summarization

MBEI: Model Based Expression Index [Li and Wong, 2001]

Least square fit the linear model

$$\mathrm{PM}_{ij} - \mathrm{MM}_{ij} = y_{ij} = \theta_i \phi_j + \epsilon_{ij}$$

θ_i: Expression Index
ϕ_j: Probe pattern

Parameter estimation Li-Wong algorithm

A $$\hat{\theta}_i = \frac{\sum_{j=1}^{J} y_{ij}\phi_j}{\sum_{j=1}^{J} \phi_j^2}$$

B $$\hat{\phi}_j = \frac{\sum_{i=1}^{I} y_{ij}\theta_i}{\sum_{i=1}^{I} \theta_i^2}$$

→ Derived from the squared error

$$R_{\mathrm{emp}} = \sum_{ij} (y_{ij} - \theta_i \phi_j)^2$$

Solved for θ_i results in A
Solved for Φ_i results in B

4 Summarization

FARMS: Factor Analysis for Robust Microarray Summarization [Hochreiter et al., 2006]

"Summarization method based on a factor analysis model for which a Bayesian Maximum a Posteriori method optimizes the model parameters under the assumption of Gaussian measurement noise" Hochreiter et al. (2006)

RNA concentration estimation directly from the model

Summarization problem→ linear model with Gaussian noise: Factor Analysis Model with one hidden factor = [mRNA]

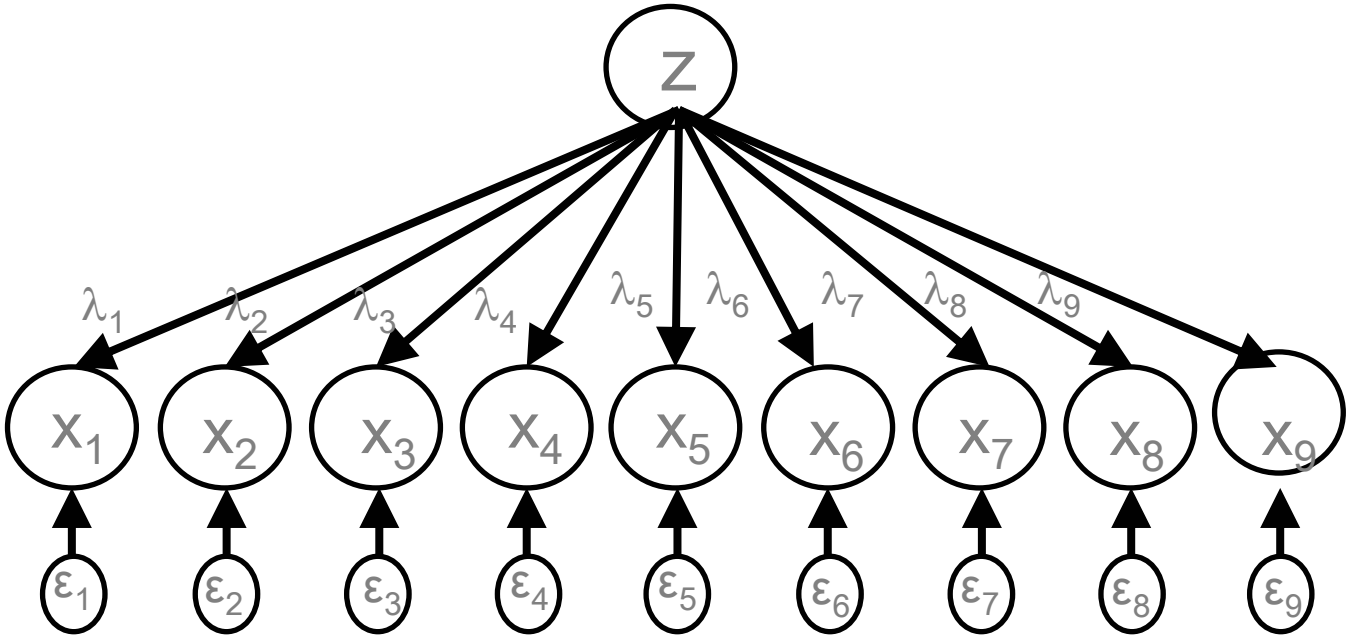
4 Summarization

Factor

loading matrix

observations

additive noise



z = variation in mRNA concentration or log [mRNA]

λ_i = sensitivity of log-PM_i

ε_i = measurement noise for log-PM_i

x_i = observed log-PM_i (mean normalized to zero)

4 Summarization

The model

$$x = \lambda z + \varepsilon$$

Generative model:

z: factor $\mathcal{N}(0, I)$ One dimensional standard G Distribution

ε: noise $\mathcal{N}(0, \Psi)$ with Ψ = diagonal noise covariance matrix

$\mathcal{N}(\mu, \Sigma)$ Multidimensional G distribution

λ: loading matrix

The observation vector x is Gaussian distributed: $x \sim \mathcal{N}(0, \lambda\lambda^T + \Psi)$

BIOINF

4 Summarization

Expectation-Maximization algorithm to estimate the model parameters

Ψ, I $\rightarrow$ EM-algorithm

$$\mathrm{E}-Step:$$

Compute $E\left[z|x_i\right]$ and $E\left[zz^T|x_i\right]$

$$\mathrm{M}-Step:$$

$$\Lambda^{new}=\left(\sum_{i=1}^{n}x_iE\left[z|x_i\right]^T\right)\left(\sum_{i=1}^{n}x_iE\left[zz^T|x_i\right]\right)^{-1}$$

$$\Psi^{new}=\frac{1}{n}diag\left\{\sum_{i=1}^{n}x_ix_i^T-\Lambda^{new}E\left[z|x_i\right]x_i^T\right\}$$

4 Summarization

High density oligonucleotide array data summarized at probe level

- Probe-level data to assess prob set quality
- Only a small subset of Probesets are retained : selected features

Spiked signals are unlikely due to the low observed variance in the data

Chip normally with more constant gene signals than variable signal

Negatives values are not plausible (↑[mRNA] ↓signal intensity) NO!!!!

Factor analysis model with deffault parameters

- No BG correction
- Normalization by Quantiles and Cyclic Loess
- PMs only