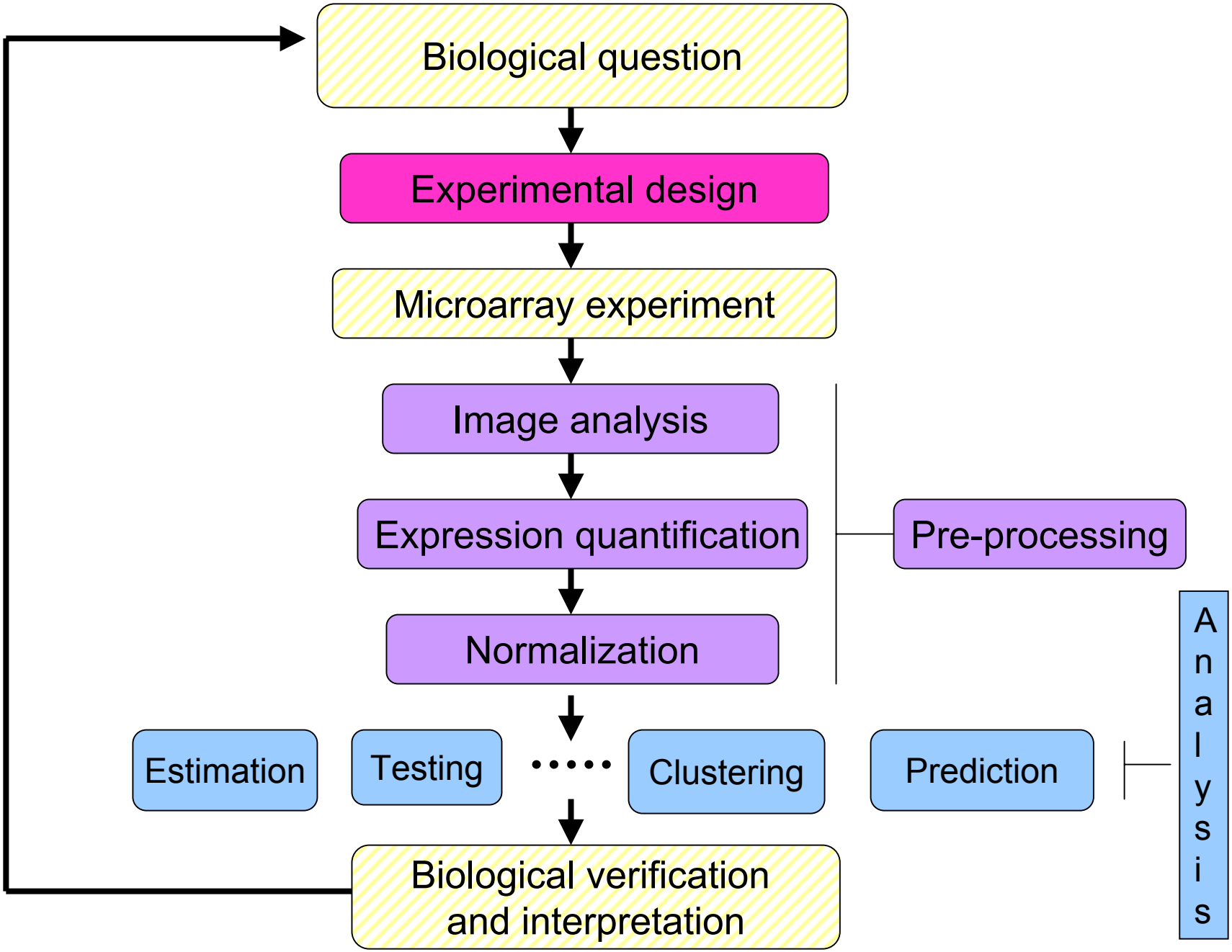
DNA Microarray Data
Oligonucleotide Arrays

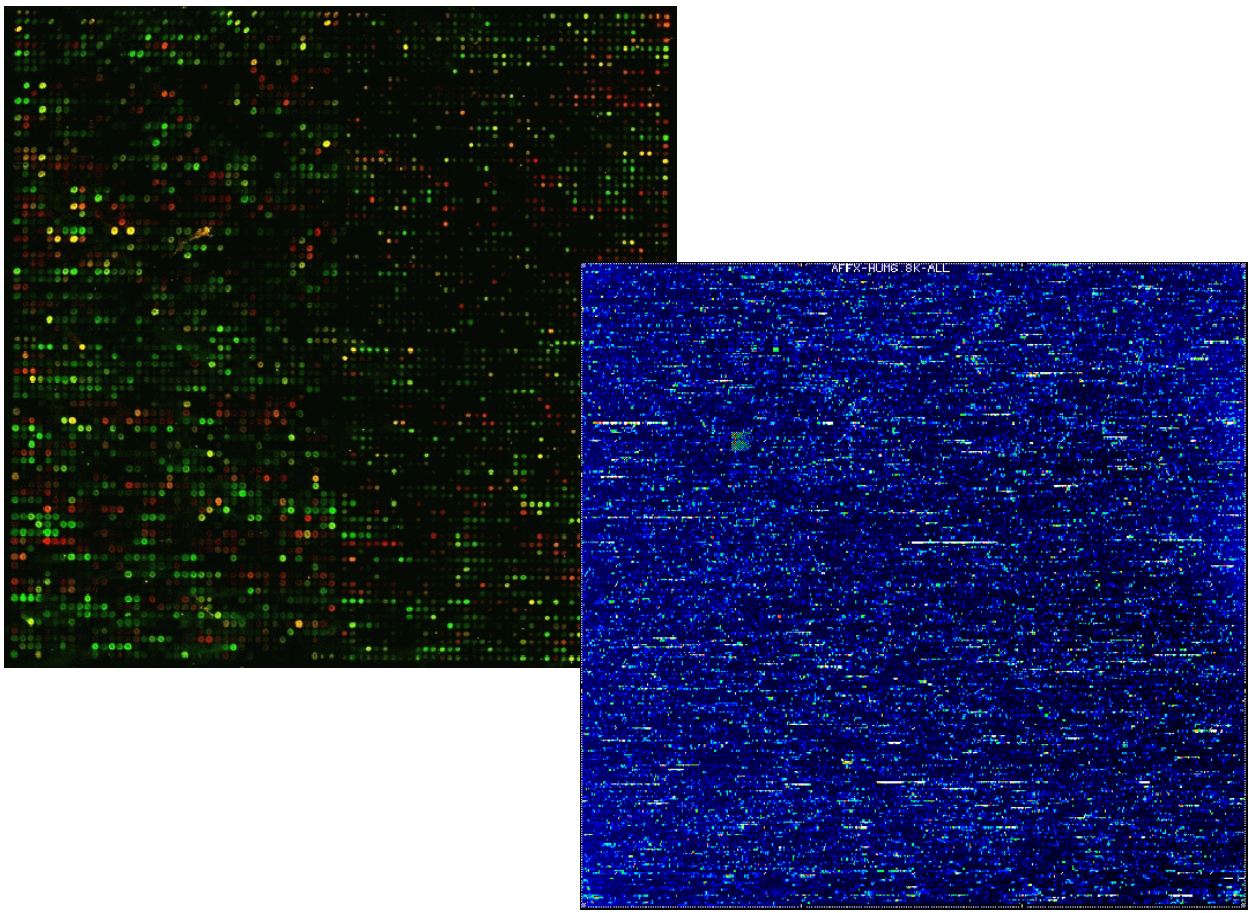
Sandrine Dudoit, Robert Gentleman, Rafael Irizarry, and Yee Hwa Yang

Bioconductor Short Course

2003


Biological question
Experimental design
Microarray experiment
Image analysis
Expression quantification
Normalization
Pre-processing
Estimation
Testing
.....
Clustering
Prediction
Analysis
Biological verification
and interpretation

DNA microarrays



DNA microarrays

DNA microarrays rely on the hybridization properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells.

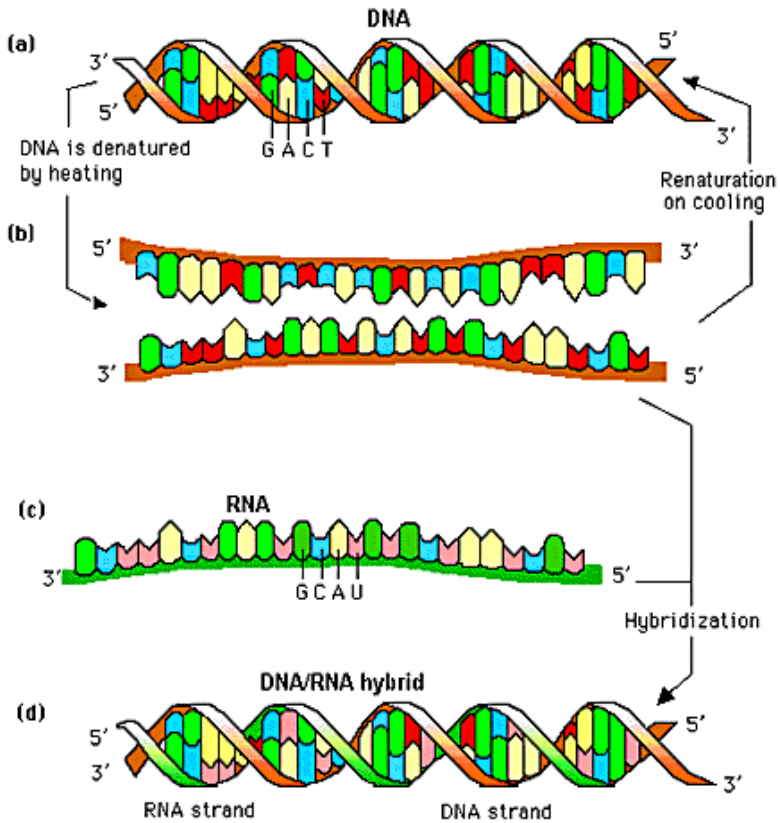
The ancestor of cDNA microarrays: the Northern blot.

Hybridization

- Hybridization refers to the annealing of two nucleic acid strands following the base-pairing rules.

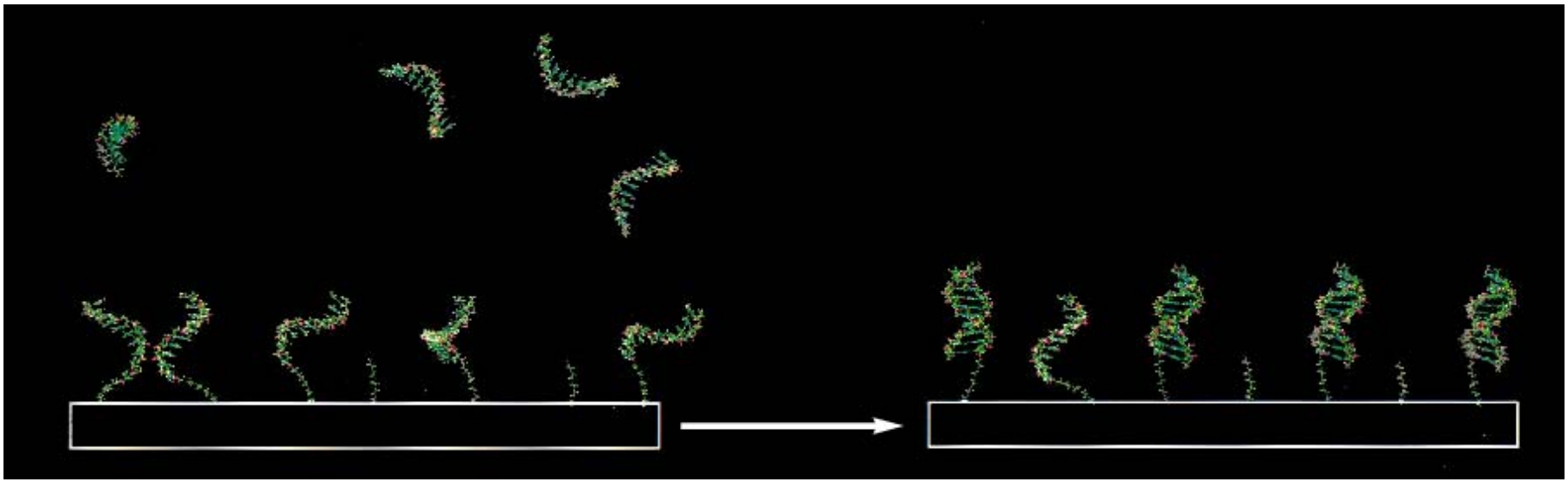
- Nucleic acid strands in a duplex can be separated, or denatured, by heating to destroy the hydrogen bonds.

Hybridization



Nucleic Acid Hybridization

Hybridization



Gene expression assays

The main types of gene expression assays:

- Serial analysis of gene expression (SAGE);
- Short oligonucleotide arrays (Affymetrix);
- Long oligonucleotide arrays (Agilent Inkjet);
- Fibre optic arrays (Illumina);
- Spotted cDNA arrays (Brown/Botstein).

Applications of microarrays

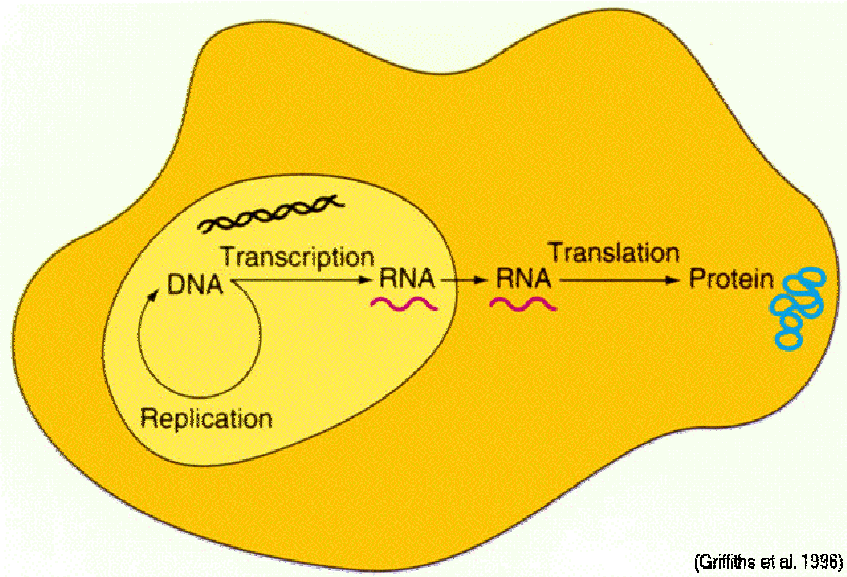
- Measuring transcript abundance (cDNA arrays);
- Genotyping;
- Estimating DNA copy number (CGH);
- Determining identity by descent (GMS);
- Measuring mRNA decay rates;
- Identifying protein binding sites;
- Determining sub-cellular localization of gene products;
- …

Applications of microarrays

- **Cancer research:** Molecular characterization of tumors on a genomic scale

  → more reliable diagnosis and effective treatment of cancer.
- **Immunology:** Study of host genomic responses to bacterial infections.
- ...

Transcriptome



- mRNA or transcript levels sensitively reflect the state of a cell.
- Measuring protein levels (translation) would be more direct but more difficult.

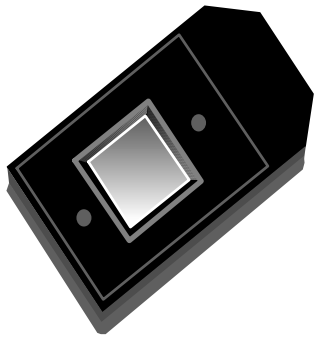
Transcriptome

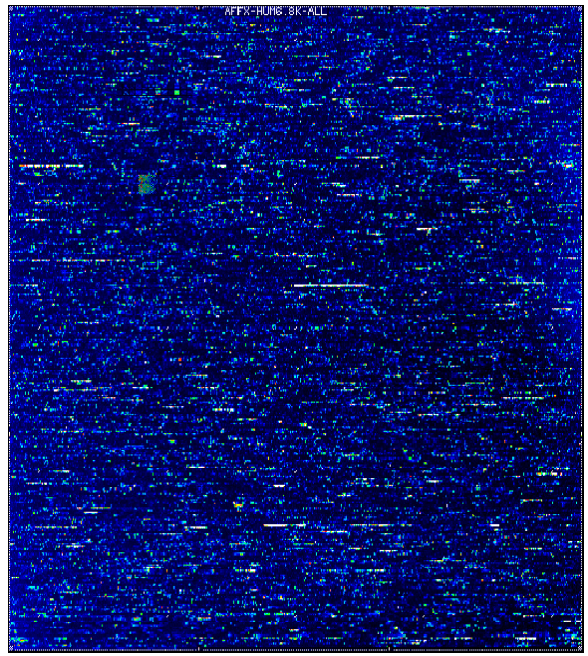
- The transcriptome reflects
  - Tissue source: cell type, organ.
  - Tissue activity and state:
    - Stage of development, growth, death.
    - Cell cycle.
    - Disease vs. healthy.
    - Response to therapy, stress.

Applications of microarrays

- Compare mRNA (transcript) levels in different types of cells, i.e., vary
  - Tissue: liver vs. brain;
  - Treatment: drugs A, B, and C;
  - State: tumor vs. non-tumor, development;
  - Organism: different yeast strains;
  - Timepoint;
  - etc.

Oligonucleotide chips





Terminology

- Each gene or portion of a gene is represented by 11 to 20 oligonucleotides of 25 base-pairs.

- Probe: an oligonucleotide of 25 base-pairs, i.e., a 25-mer.
- Perfect match (PM): A 25-mer complementary to a reference sequence of interest (e.g., part of a gene).
- Mismatch (MM): same as PM but with a single homomeric base change for the middle (13^{th}) base (transversion purine <-> pyrimidine, G <->C, A <->T) .
- Probe-pair: a (PM,MM) pair.
- Probe-pair set: a collection of probe-pairs (11 to 20) related to a common gene or fraction of a gene.
- Affy ID: an identifier for a probe-pair set.
- The purpose of the MM probe design is to measure non-specific binding and background noise.

Probe-pair set

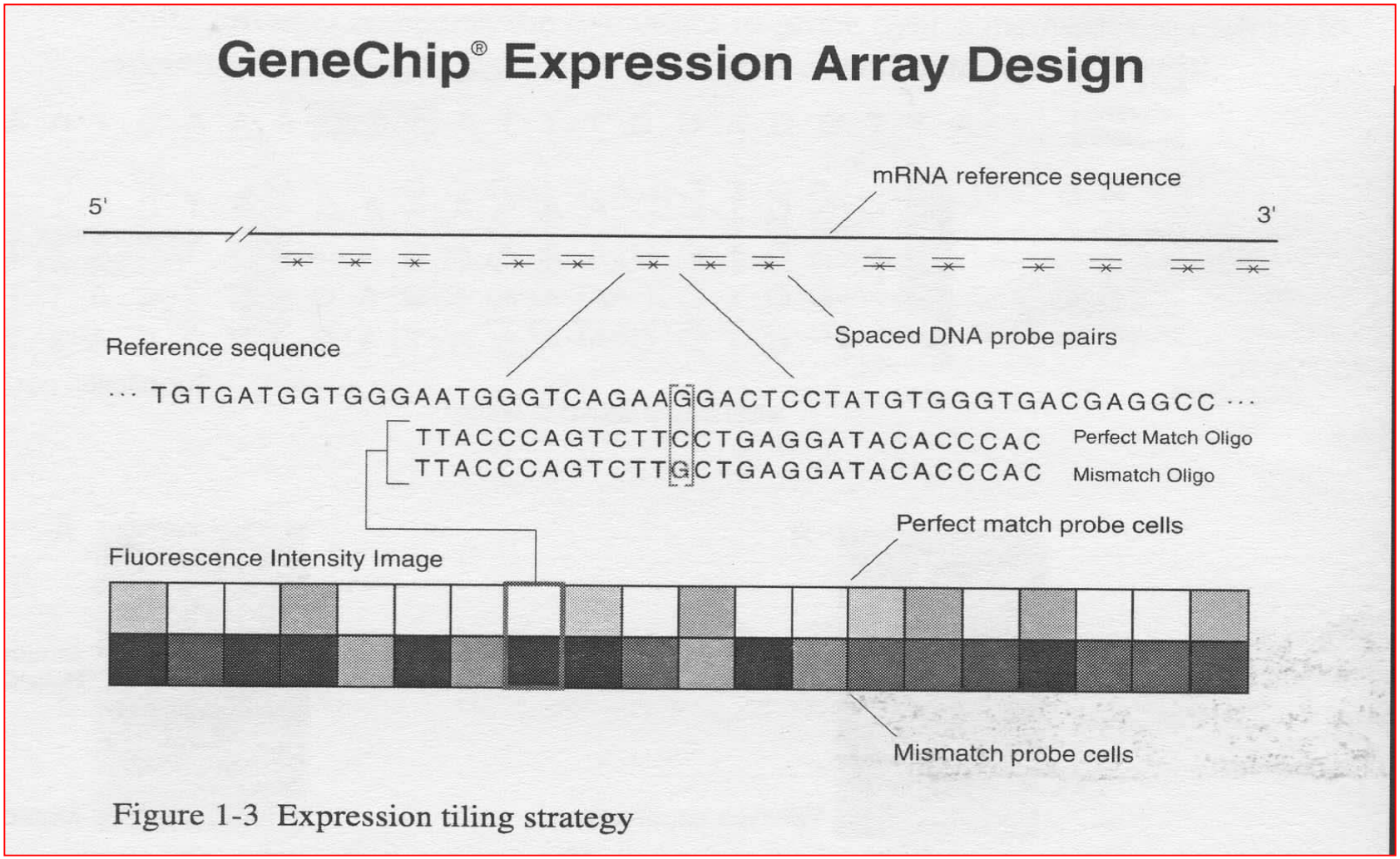
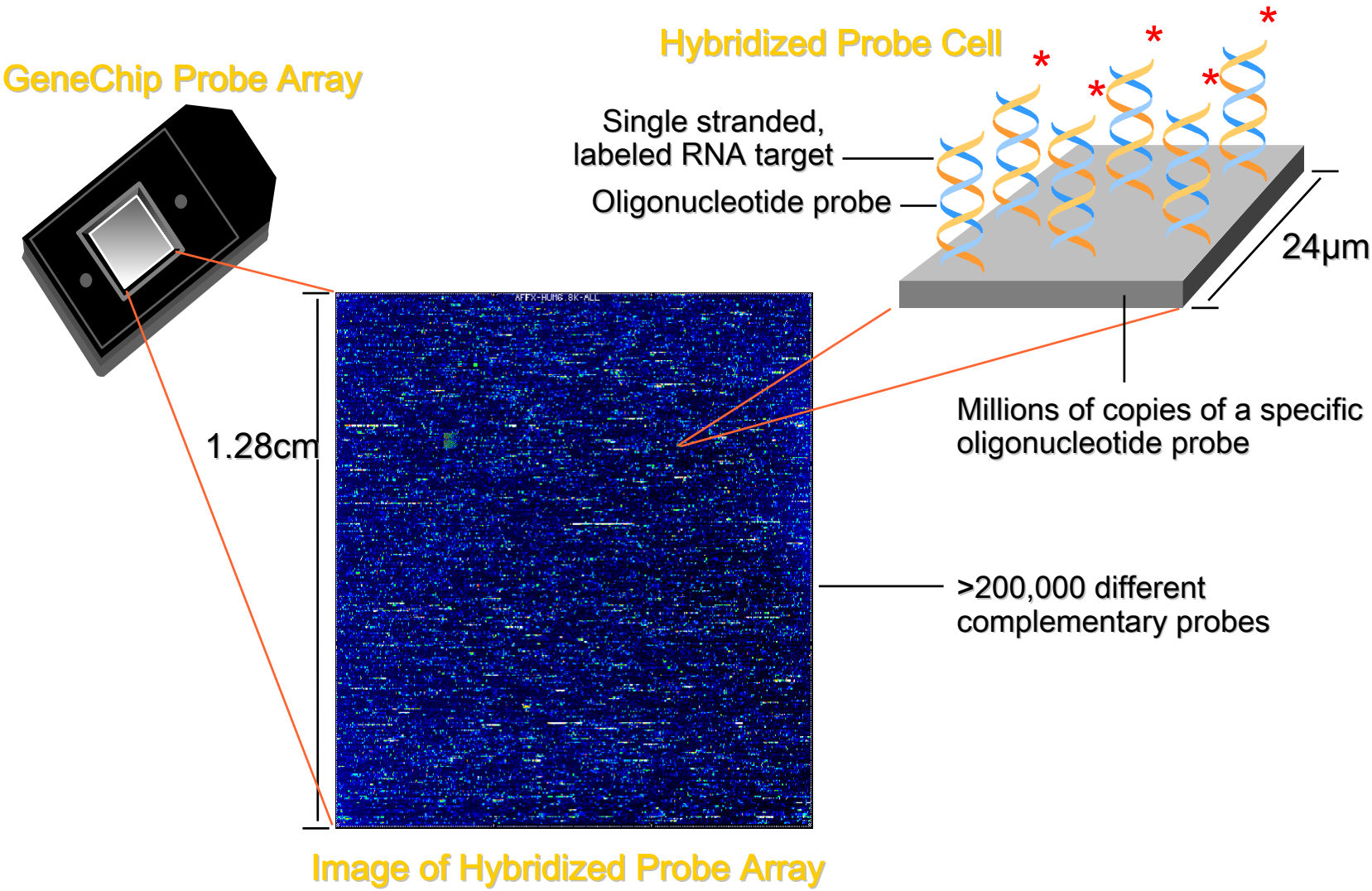


Figure 1-3 Expression tiling strategy

Spotted vs. Affymetrix arrays

Spotted arrays	Affymetrix arrays
One probe per gene	11 – 20 probe-pairs per gene
Probes of varying length	Probes are 25-mers
Two target samples per array	One target sample per array

Oligonucleotide chips

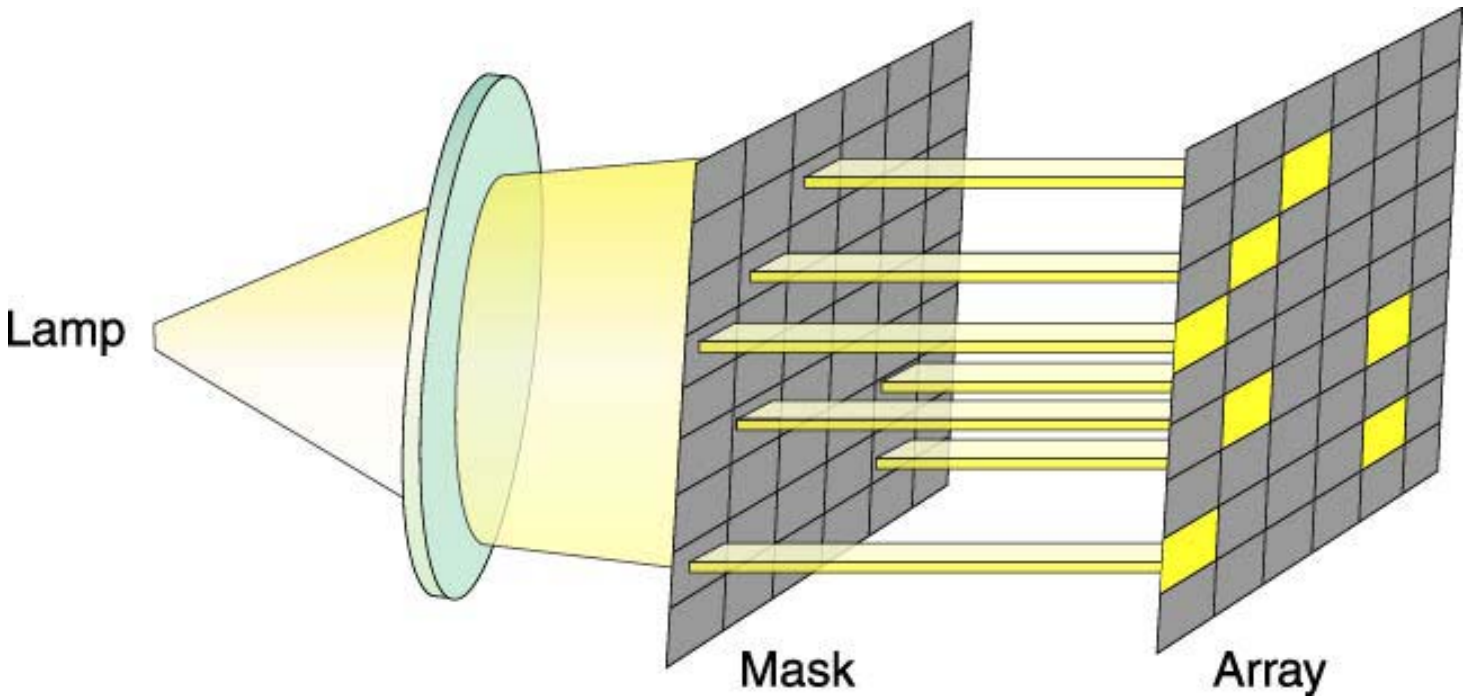


Compliments of D. Gerhold

Oligonucleotide chips

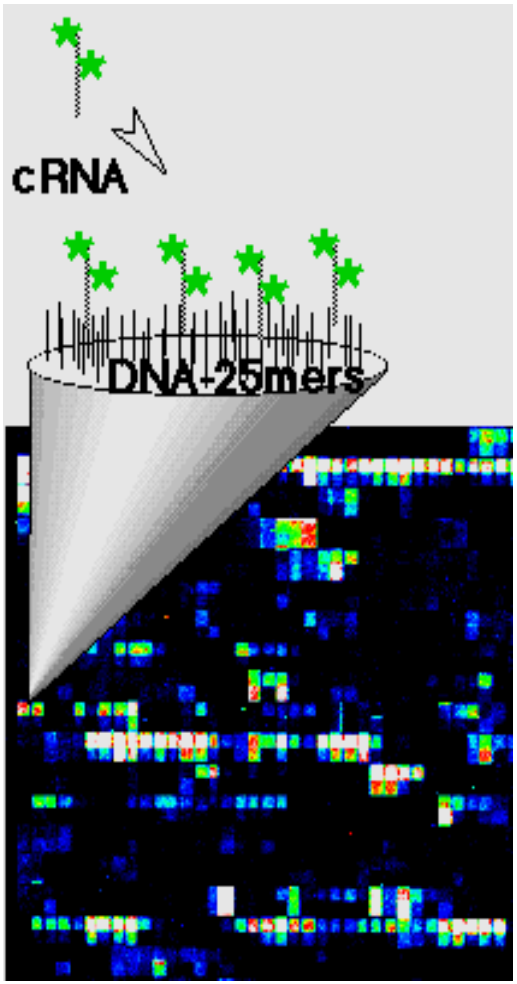
- The probes are synthesized *in situ*, using combinatorial chemistry and photolithography.
- Probe cells are square-shaped features on the chip containing millions of copies of a single 25-mer probe. Sides are 18-50 microns.

Oligonucleotide chips



The manufacturing of GeneChip® probe arrays is a combination of photolithography and combinational chemistry.

Image analysis



- About 100 pixels per probe cell.
- These intensities are combined to form one number representing the expression level for the probe cell oligo.
- → CEL file with PM or MM intensity for each cell.

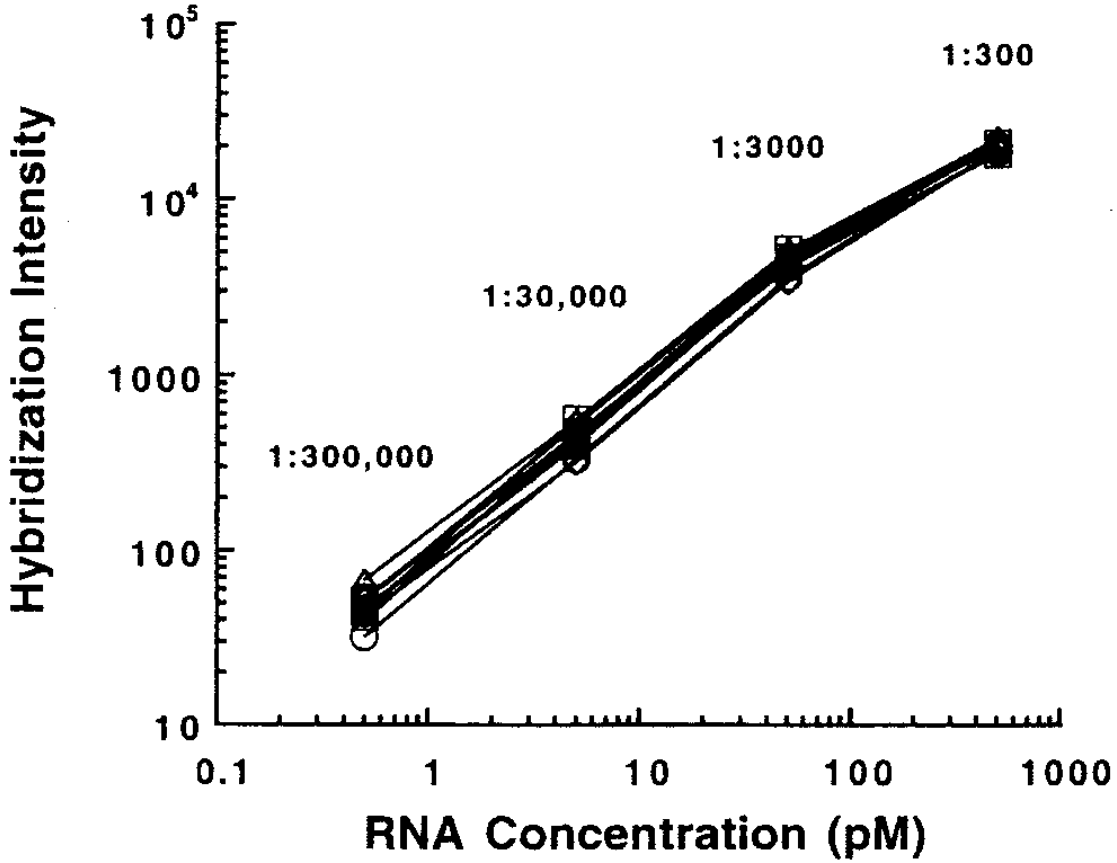
Expression measures

- Many expression measures are based on differences of **PM-MM**.
- The intention is to correct for background and non-specific binding.
- E.g. *MarrayArray Suite*® (MAS) v. 4.0 uses Average Difference Intensity (ADI) or

  AvDiff = average of PM-MM.
- Problem: MM may also measure signal.
- More on this in lecture *Pre-processing DNA Microarray Data.*

What is the evidence?

Lockhart et. al. Nature Biotechnology 14 (1996)



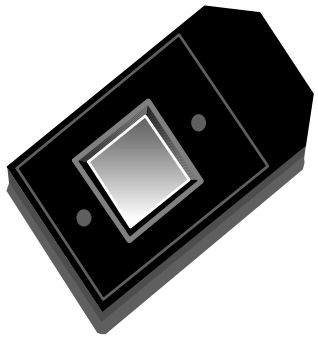
Integration of experimental and biological metadata

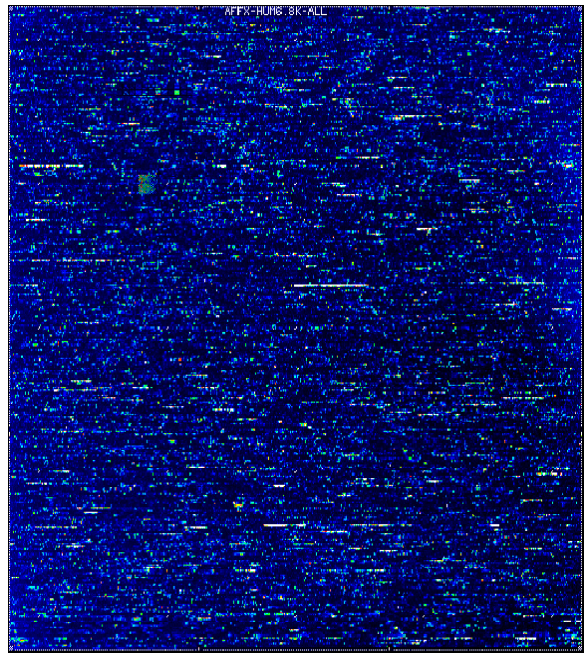
- Expression, sequence, structure, annotation, literature.
- Integration will depend on our using a common language and will rely on database methodology as well as statistical analyses.
- This area is largely unexplored.

Pre-processing

- Affymetrix oligonucleotide chips
  - Image analysis;
  - Background adjustment
  - Normalization;
  - Expression measures.

Pre-processing: Oligonucleotide chips





Affymetrix files

- Main software from Affymetrix company *MicroArray Suite* - *MAS*, now version 5.
- **`DAT`** file: Image file, ~10^7 pixels, ~50 MB.
- **`CEL`** file: Cell intensity file, probe level PM and MM values.
- **`CDF`** file: Chip Description File. Describes which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs).

Image analysis

- Raw data, DAT image files ➔ CEL files
- Each probe cell: 10x10 pixels.
- Gridding: estimate location of probe cell centers.
- Signal:
  - Remove outer 36 pixels ➔ 8x8 pixels.
  - The probe cell signal, PM or MM, is the 75th percentile of the 8x8 pixel values.
- Background: Average of the lowest 2% probe cell values is taken as the background value and subtracted.
- Compute also quality measures.

Data and notation

- PM_{ijg}, MM_{ijg} = Intensity for perfect match and mismatch probe in cell *j* for gene *g* in chip *i*.
  - *i = 1,…, n* -- from one to hundreds of chips;
  - *j = 1,…, J* -- usually 11 or 20 probe pairs
  - *g = 1,…, G* -- between 8,000 and 20,000 probe sets.
- Task: summarize for each probe set the probe level data, i.e., 20 PM and MM pairs, into a single expression measure.
- Expression measures may then be compared within or between chips for detecting differential expression.

Expression measures MAS 4.0

- GeneChip® MAS 4.0 software uses AvDiff

$$AvDiff = \frac{1}{|\mathrm{A}|} \sum_{j \in \mathrm{A}} (PM_j - MM_j)$$

where A is a set of "suitable" pairs, e.g., pairs with $d_j = PM_j - MM_j$ within 3 SDs of the average of $d_{(2)}, \ldots, d_{(J-1)}$.

- Log-ratio version is also used: average of log(PM/MM).

Expression measures MAS 5.0

- GeneChip® MAS 5.0 software uses Signal

$$signal = \text{Tukey Biweight}\{\log(PM_j - MM_j^*)\}$$

with MM * a new version of MM that is never larger than PM.

- If MM < PM, MM* = MM.
- If MM >= PM,
  - SB = Tukey Biweight (log(PM)-log(MM)) (log-ratio).
  - log(MM*) = log(PM)-log(max(SB, +ve)).
- Tukey Biweight: B(x) = (1 – (x/c)^2)^2 if |x|<c, 0 ow.

Expression measures
Li & Wong

- Li & Wong (2001) fit a model for each probe set, i.e., gene

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \ \varepsilon_{ij} \propto N(0, \sigma^2)$$

  where
  - θ_i: model based expression index (MBEI),
  - ϕ_j: probe sensitivity index.
- Maximun likelihood estimate of MBEI is used as expression measure for the gene in chip *i*.
- Need at least 10 or 20 chips.
- Current version default works with PMs only.

Expression measures

- Many expression measures are based on PM-MM, with the intention of correcting for non-specific binding and background noise.
- Problems:
  - MMs are PMs for some genes,
  - removing the middle base does not make a difference for some probes .
  - Subtracting MM adds variance. Especially at low end.
- Why not simply average PM or log PM? Not good enough, still need to adjust for background.
- Also need to normalize.

Expression measures RMA

Irizarry et al. (2003).

1. Estimate background BG and use only background-corrected PM: $\log_2$(PM-BG).
2. Probe level normalization of $\log_2$(PM-BG) for suitable set of chips.
3. Robust Multi-array Average, RMA, of $\log_2$(PM-BG).

RMA background

More refined background estimation

- Model observed PM as the sum of a signal intensity SG and a background intensity BG

$$PM = SG + BG,$$

where it is assumed that SG is *Exponential* (α), BG is *Normal* (μ, σ^2), and SG and BG are independent.

- Background adjusted PM values are then E(SG|PM).

Quantile normalization

- Probe level quantile normalization (Bolstad et al., 2002).
- Co-normalize probe level intensities, e.g. PM-BG or just PM or MM, for *n* chips by averaging each quantile across chips.
- Assumption: same probe level intensity distribution across chips.
- No need to choose a baseline or work in a pairwise manner.
- Deals with non-linearity.

Curve-fitting normalization

- Astrand (2003), Bolstad et al. (2003). Generalization of M vs. A robust local regression normalization for cDNA arrays.
- For *n* chips, regress orthonormal contrasts of probe level statistics on the average of the statistics across chips.

RMA expression measures

- Robust regression method to estimate expression measure and SE from PM-BG values.
- Assume additive model

$$\log_2(PM_{ij} - BG) = a_i + b_j + \varepsilon_{ij}$$

- Estimate RMA = a_i for chip *i* using robust method, such as median polish (fit iteratively, successively removing row and column medians, and accumulating the terms, until the process stabilizes).
- Fine with *n*=2 or more chips.

Summary

- Don’t subtract MM.
- “Background correct” PM. Even global background improves on probe-specific MM.
- Take logs: probe effect is additive on log scale.
- PMs need to be normalized (e.g. quantile normalization).
- RMA is arguably the best summary in terms of bias, variance, and model fit. Comparison study in Irizarry et al. (2003).

`affy`: Pre-processing Affymetrix data

- Basic classes and methods for probe-level data.
- Widgets for data input.
- Diagnostic plots: 2D spatial images, boxplots, MA-plots, etc.
- Background estimation.
- Probe-level normalization: quantile and curve-fitting normalization (Bolstad et al., 2002).
- Expression measures: MAS 4.0 AvDiff, MAS 5.0 Signal, MBEI (Li & Wong, 2001), RMA (Irizarry et al., 2003).
- Three main functions: **`ReadAffy`**, **`expresso`**, **`rma`**.

Combining data across slides

Data on *G* genes for *n* hybridizations

→ *G x n* genes-by-arrays data matrix

Arrays

Genes		Array1	Array2	Array3	Array4	Array5	...
	Gene1	0.46	0.30	0.80	1.51	0.90	...
	Gene2	-0.10	0.49	0.24	0.06	0.46	...
	Gene3	0.15	0.74	0.04	0.10	0.20	...
	Gene4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	Gene5	-0.06	1.06	1.35	1.09	-1.09	...
	...	...	...	...	...	...	

M = $\log_2$(Red intensity / Green intensity)
expression measure, e.g, RMA

Probe Level Data

- recently Affymetrix has made probe level data available for all their chips
- this provides us with several additional data analytic opportunities
- methodology and software is being developed (but we would certainly appreciate collaborations)

Some Definitions

- **cDNA**: a DNA molecule that is complementary to a mRNA
- in some sense cDNAs reflect the transcriptome (set of transcribed genes)
- **EST**: expressed sequence tag. A sequenced piece of cDNA
- a full length cDNA defines the structure of a transcript; an EST merely indicates an association between a sequence and a gene

Labeling mRNA

- most microarray technologies rely on labeling mRNAs of interest
- Affymetrix arrays use a process that produces biotinylated amplified RNA (aRNA)
- most procedures rely on the poly-A tail that is attached to the 3' end of mRNA to stabilize it

Labeling mRNA

- since the technologies rely on the poly-A tail, it seems that they can be mislead by internal poly-A sequences
- it is also important to realize that most of the data are produced relying on certain aspects of nucleic acid binding that could potentially be checked by computer modeling (computational biochemistry)

Probe Level Data

- recall that the probe level intensities reflect the binding of labeled mRNA to the 25mers that are fixed on the surface of the chip
- sets of 11 or more 25mers are selected from each mRNA of interest (*probe set*)
- the intensities across the probe set are processed (averaged) to produce an estimate of the expression of the gene

Probe Level Data

- perhaps the largest challenge is the one that exists due to cross-hybridization
- for our purposes cross-hybridization means that a mRNA other than the one intended binds to a particular probe
- cross-hybridization has the potential to greatly alter estimated expression and adequate adjustment could greatly improve the performance of Affymetrix chips

Cross-hybridization

- the **transcriptome** for an organism or tissue is the entire set of transcripts and their relative levels under defined conditions
- if we know the transcriptome then we can explore the potential for cross-hybridization
- we can define the notions of sensitivity and specificity with respect to a transcriptome

Sensitivity and Specificity

- recall that a probeset is a collection of probes (25mers) that are labeled as coming from a specific gene
- for each *probe set* we define the **sensitivity** as the proportion of the probes whose sequences are actually contained in the specific gene
- for *a probe* we define the **specificity** of the probe to be one divided by the number of transcripts that contain the 25mer

Sensitivity and Specificity

- a good probe set will be one with high sensitivity and where all probes have high specificity
- updating or refining methods such as RMA, MAS or Li-Wong to accommodate sensitivity and specificity should be helpful

Other Uses

- two other uses for probe level data
- adjustment for GC content
- identification (and possibly adjustment) of degraded mRNA

Probe Level Data

- we saw that A and T have **2** hydrogen bonds while C and G have **3**
- this means that C and G form stronger bonds than A and T
- we might then expect to see higher intensity values for probes that are GC rich due to increased binding
- again this could be accounted for in the analysis

Probe Level Data

- as we have mentioned one cannot directly compare mRNA abundance using Affymetrix expression values
- if one probe set has an estimated expression level that is twice that of another probe set that does not mean that there is twice as much mRNA
- we can compare samples, within probes

Comparisons

- between sample comparisons are meaningful
- note that GC content is constant across samples (within genes) and so not relevant *under ideal conditions*
- but not all samples have the same sets of genes expressed and cross-hybridization can profoundly affect the outcome
- methods for adjusting due to GC are relevant and important

RNA degradation

- the purpose of attaching a poly-A tail to the mRNA prior to export from the nucleus is to stabilize the mRNA
- degradation tends to be from the 5' end towards the 3' end
- not all mRNAs will degrade at the same rate
- not all samples will be exposed to the same conditions and hence may have different levels of degradation

mRNA Degradation

- again we have many reasons to be interested in this phenomenon and to develop tools that will let us easily detect it
- cross-hybridization and GC content can add to the problem
- much more sophisticated probe level analyses seem warranted

Combining Data

- everyone agrees on the importance of joint normalization of samples prior to analysis
- for all algorithms currently available this means that you cannot combine data from Hu6800 chips with Hgu95 chips or with U133 chips (or indeed Mgu74 chips)
- by using probe level data there is some chance of extracting subsets of common probes (or nearly matching) and using those as the basis for normalization