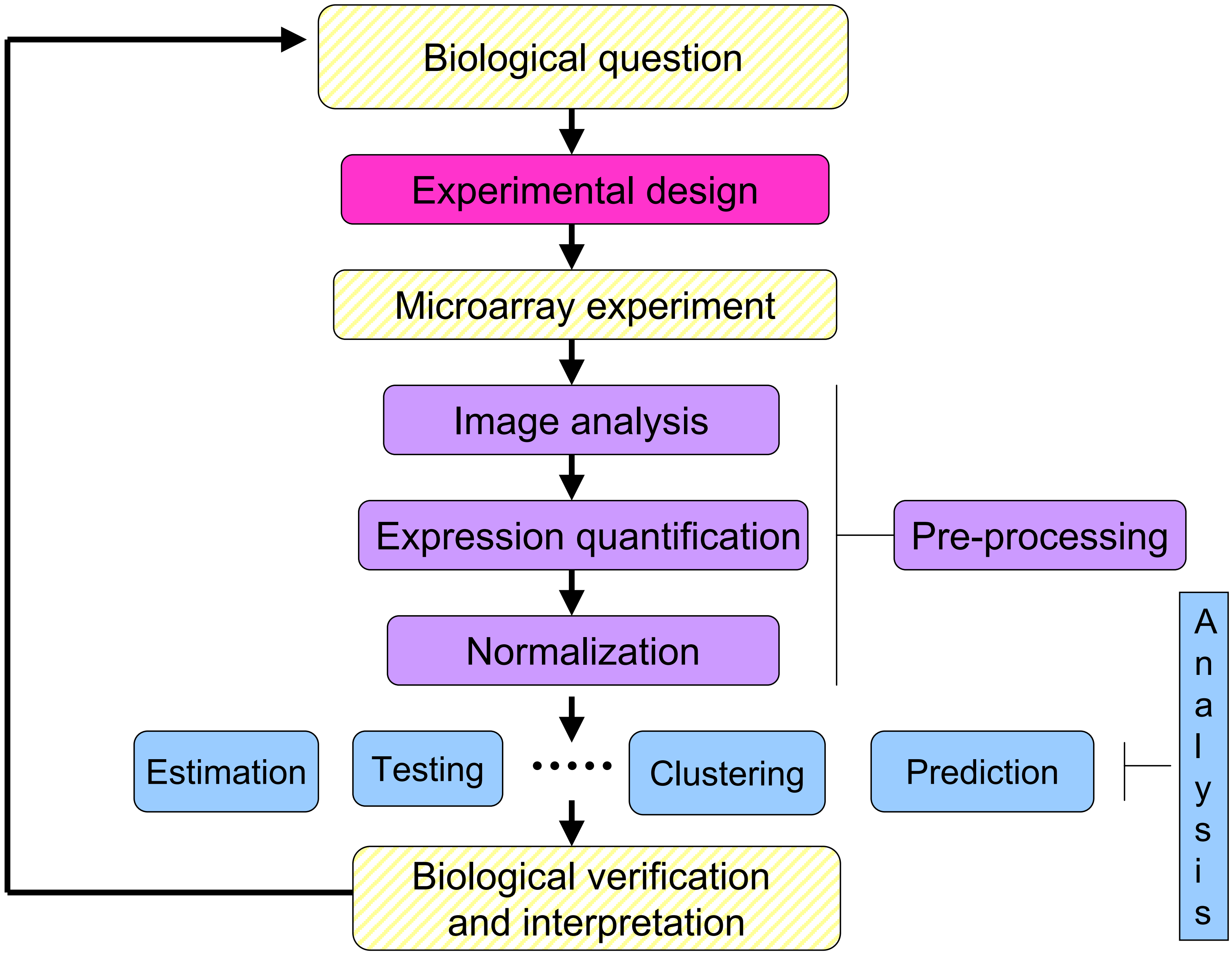
DNA Microarray Data
Oligonucleotide Arrays

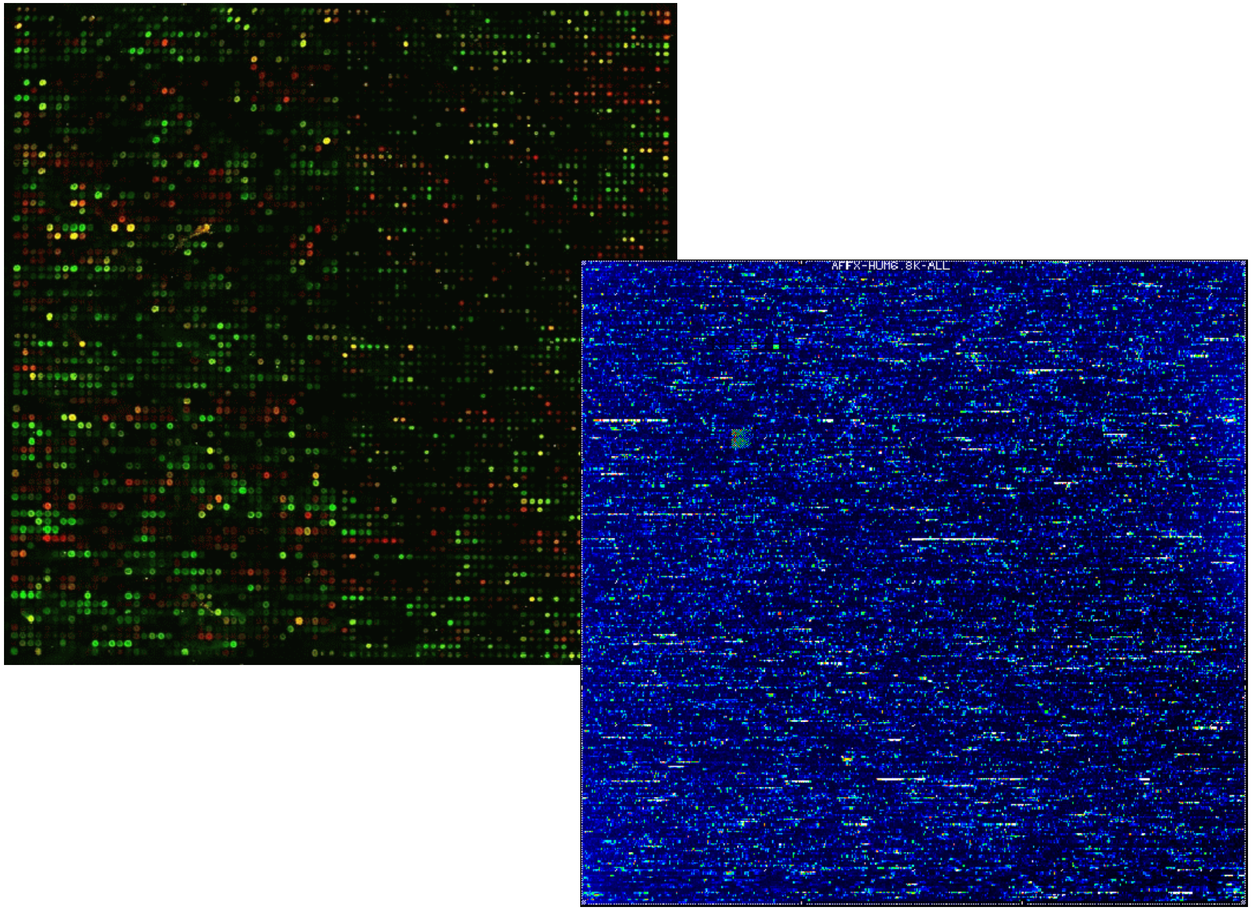
**Sandrine Dudoit, Robert Gentleman,
Rafael Irizarry, and Yee Hwa Yang**

Bioconductor Short Course

Winter 2002


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.

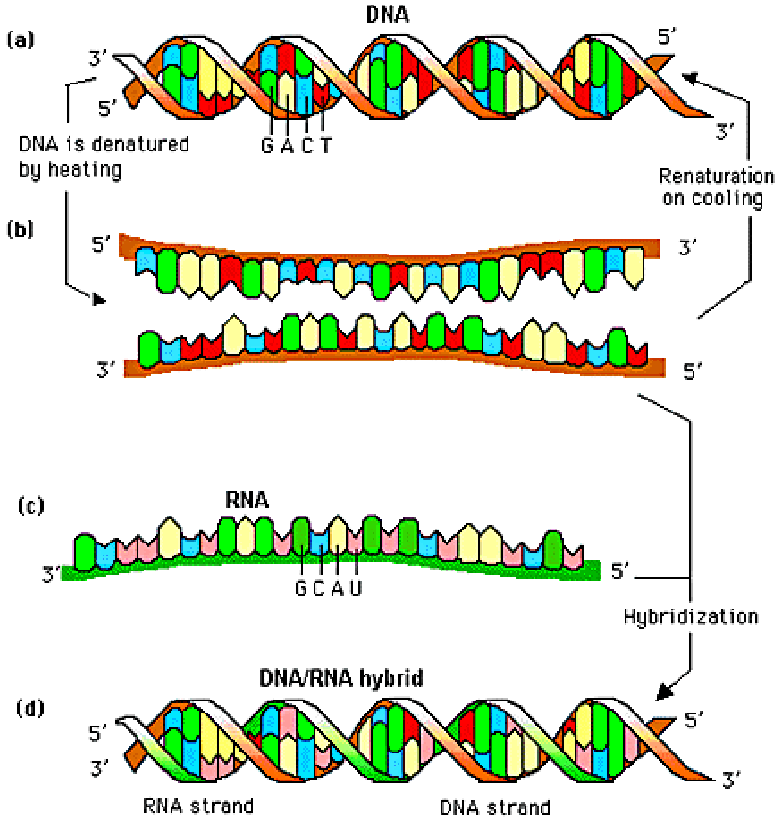
The ancestor of cDNA microarrays: the Northern blot.

Hybridization

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

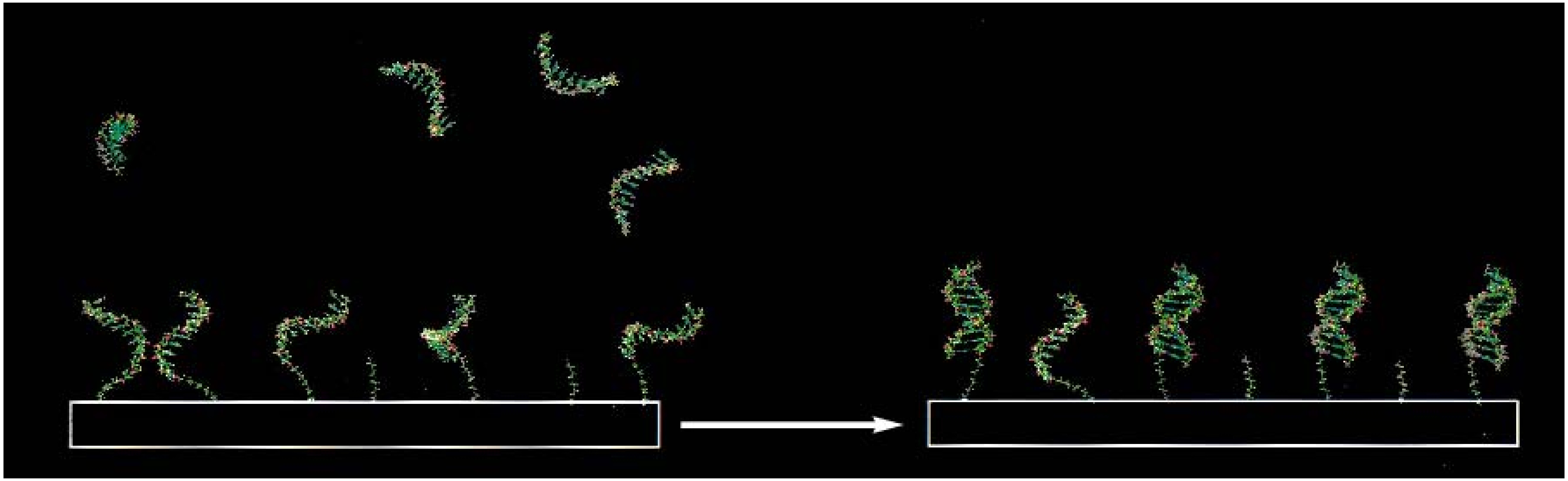
- Nucleic acid strands in a duplex can be separated, or denatured, by heating to destroy the hydrogen bonds.

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

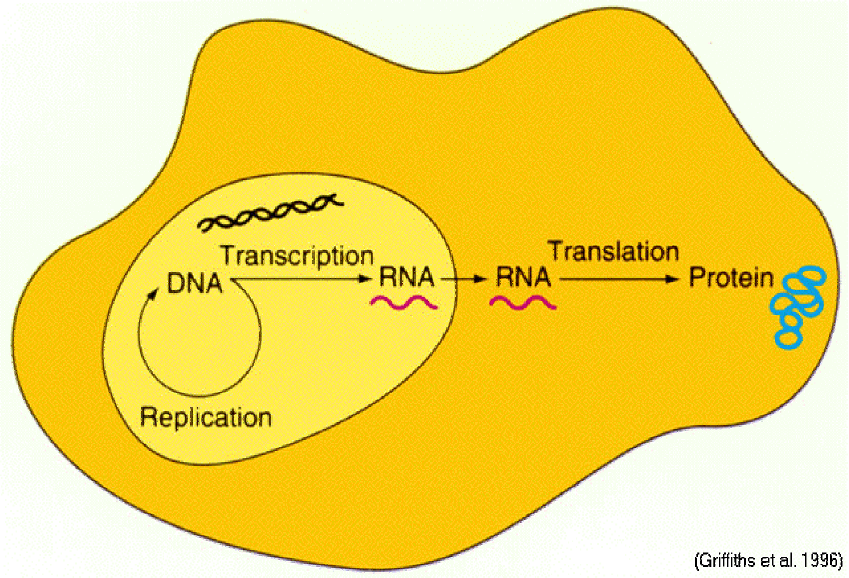
- Measuring transcript abundance (cDNA arrays);
- Genotyping;
- Estimating DNA copy number (CGH);
- Determining identity by descent (GMS);
- Measuring mRNA decay rates;
- Identifying protein binding sites;
- Determining sub-cellular localization of gene products;
- …

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



(Griffiths et al. 1996)

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

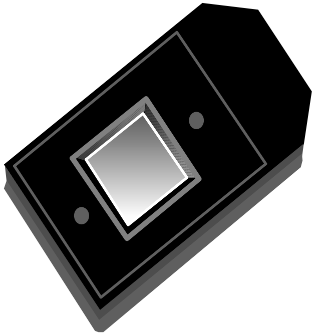
Transcriptome

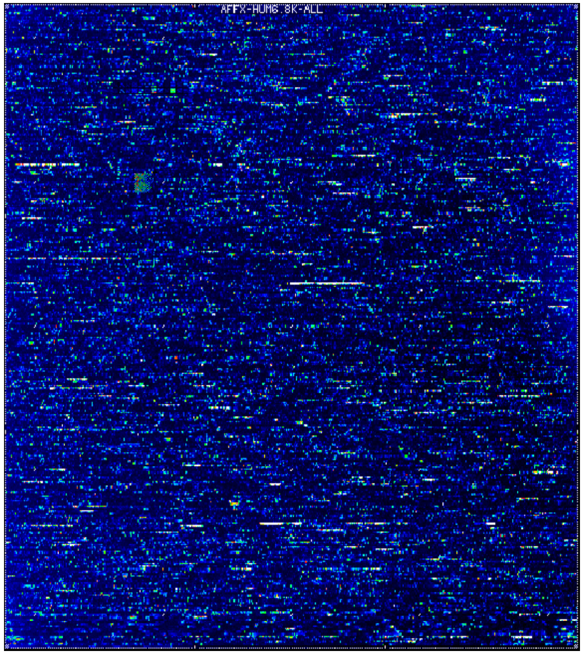
- The transcriptome reflects
  - Tissue source: cell type, organ.
  - Tissue activity and state:
    - Stage of development, growth, death.
    - Cell cycle.
    - Disease vs. healthy.
    - Response to therapy, stress.

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



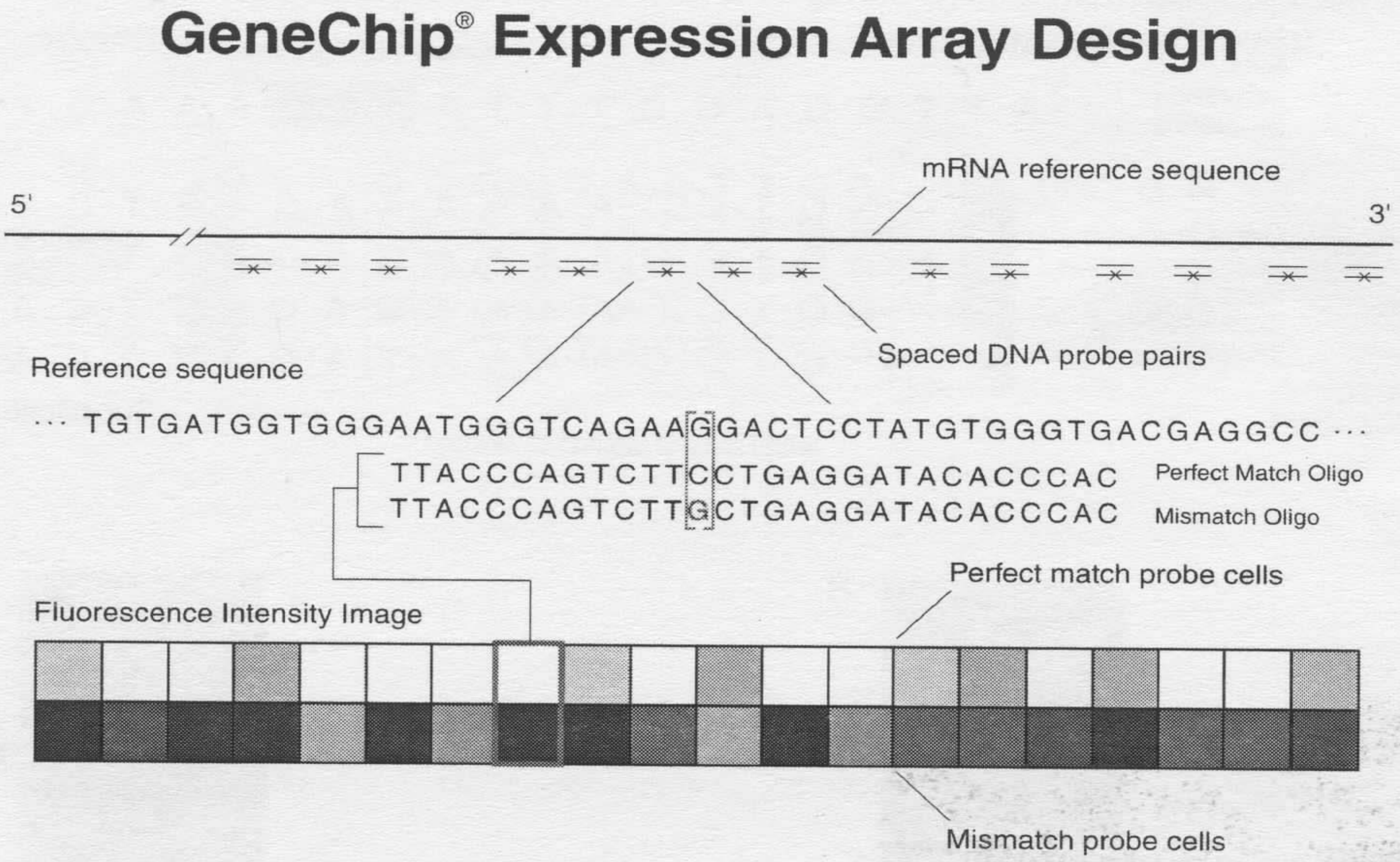


Terminology

- Each gene or portion of a gene is represented by 16 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 (13th) base (transversion purine <-> pyrimidine, G <->C, A <->T) .
- Probe-pair: a (PM,MM) pair.
- Probe-pair set: a collection of probe-pairs (16 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


GeneChip® Expression Array Design

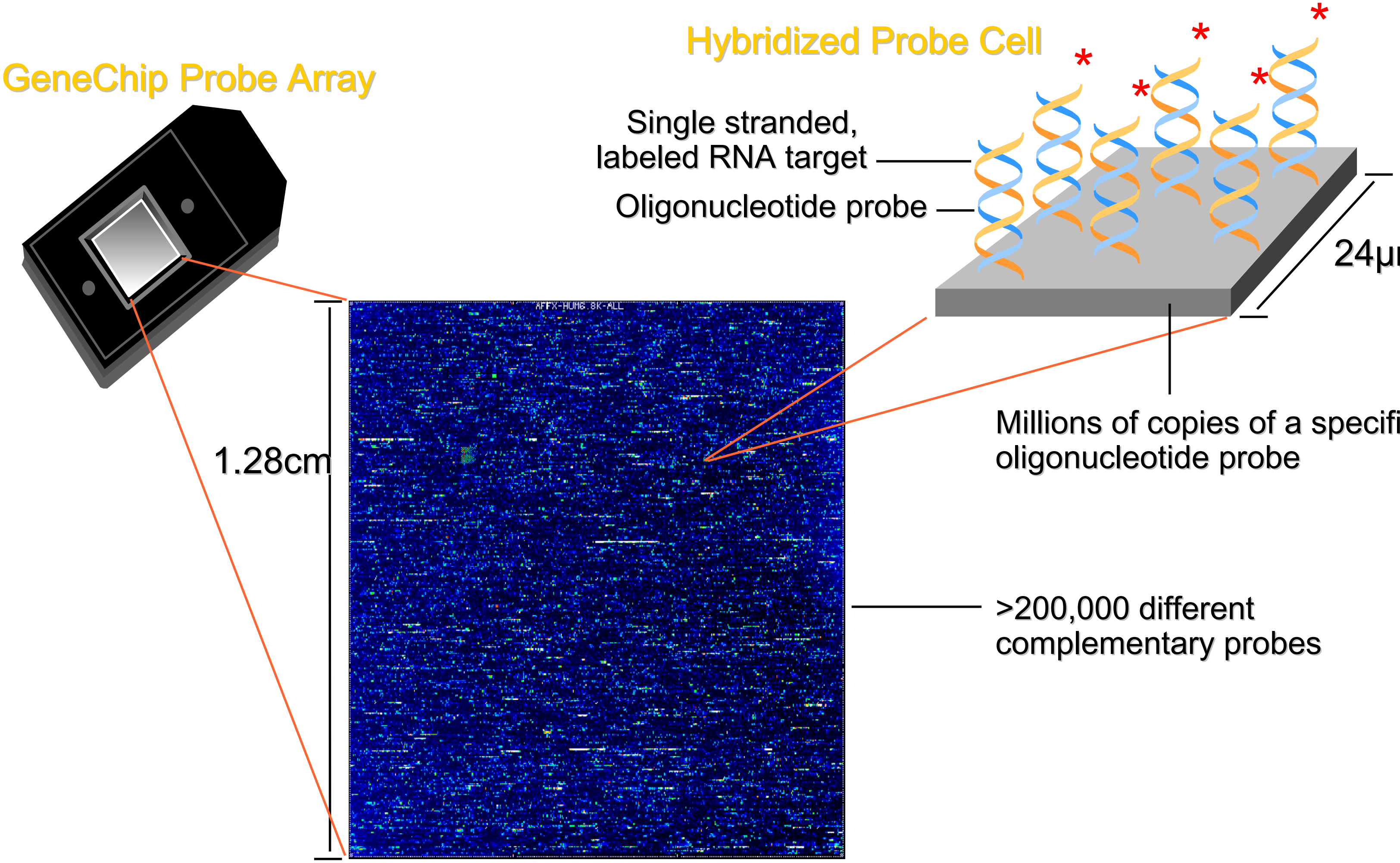
mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
TTACCCAGTCTTCCTGAGGATACACCCAC
Perfect Match Oligo
TTACCCAGTCTTGCTGAGGATACACCCAC
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 1-3 Expression tiling strategy

Spotted vs. Affymetrix arrays

Spotted arrays	Affymetrix arrays
One probe per gene	16 – 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



GeneChip Probe Array

Hybridized Probe Cell

Single stranded, labeled RNA target

Oligonucleotide probe

24µm

1.28cm

Millions of copies of a specific oligonucleotide probe

>200,000 different complementary probes

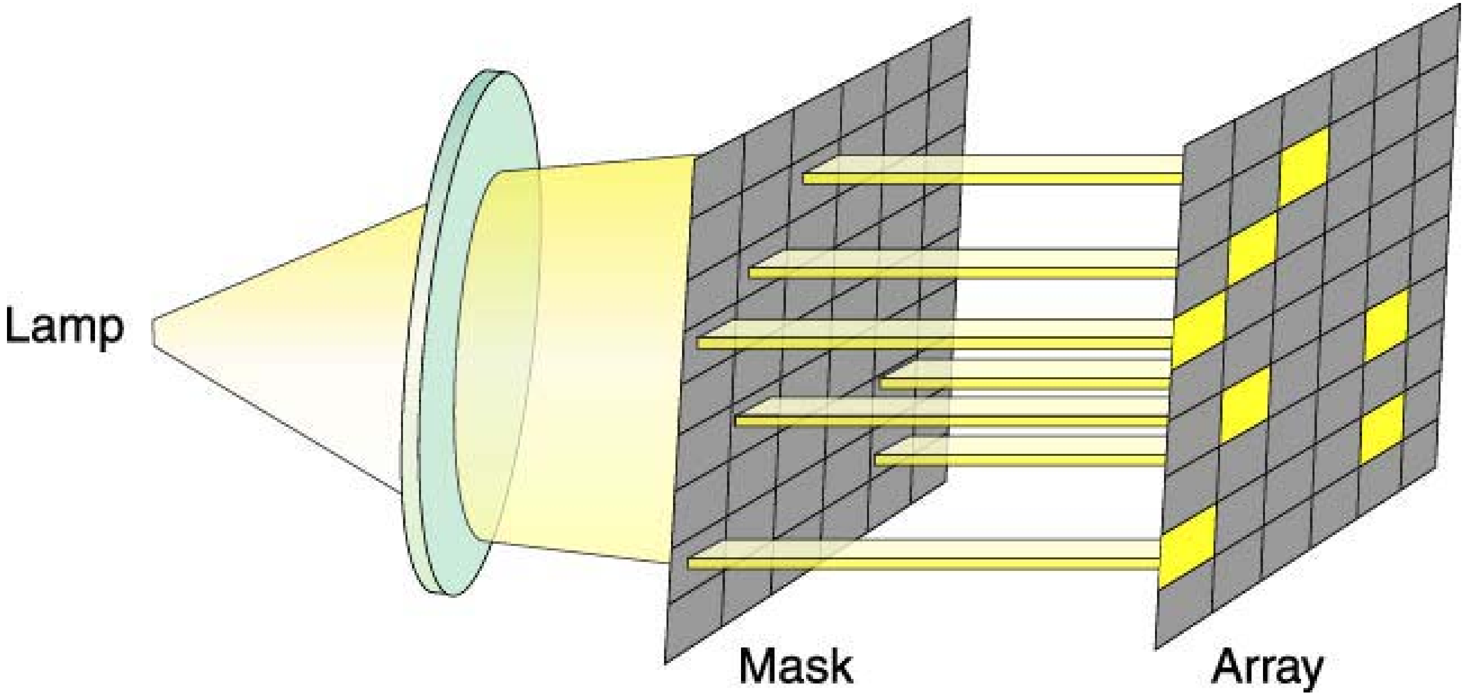
Image of Hybridized Probe Array

Compliments of D. Gerhold

Oligonucleotide chips

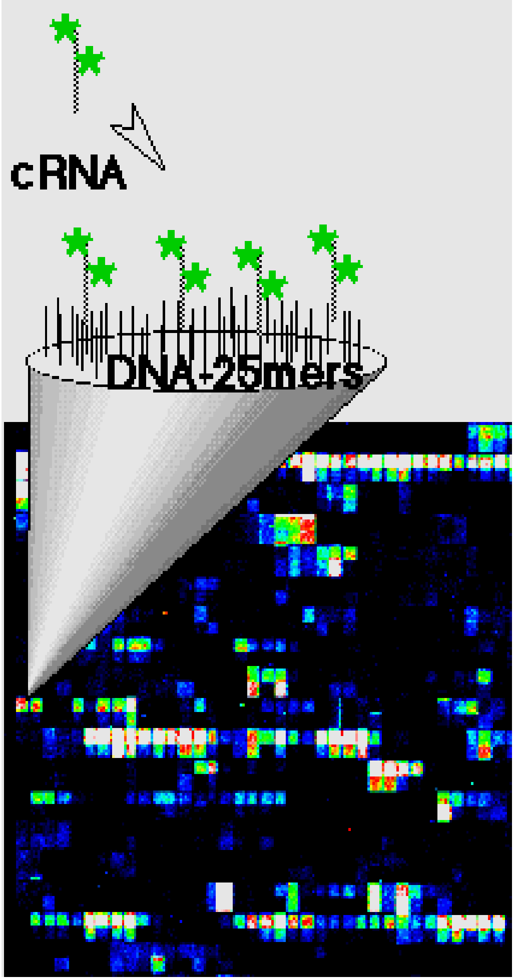
- The probes are synthesized *in situ*, using combinatorial chemistry and photolithography.
- Probe cells are square-shaped features on the chip containing millions of copies of a single 25-mer probe. Sides are 18-50 microns.

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.

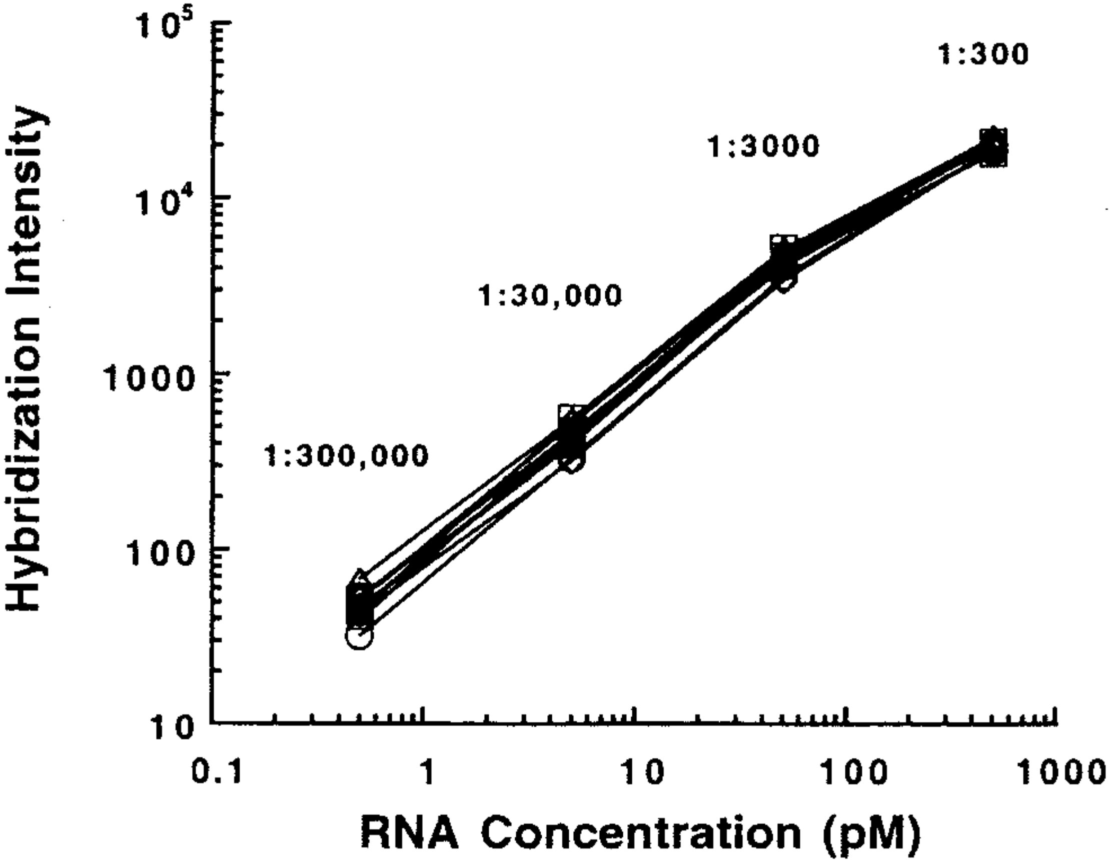
Expression measures

- Most 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)



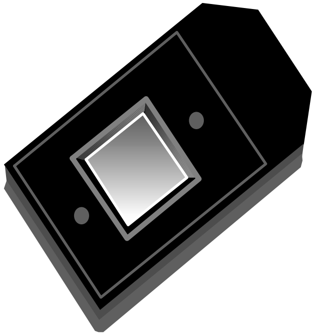
Integration of experimental and biological metadata

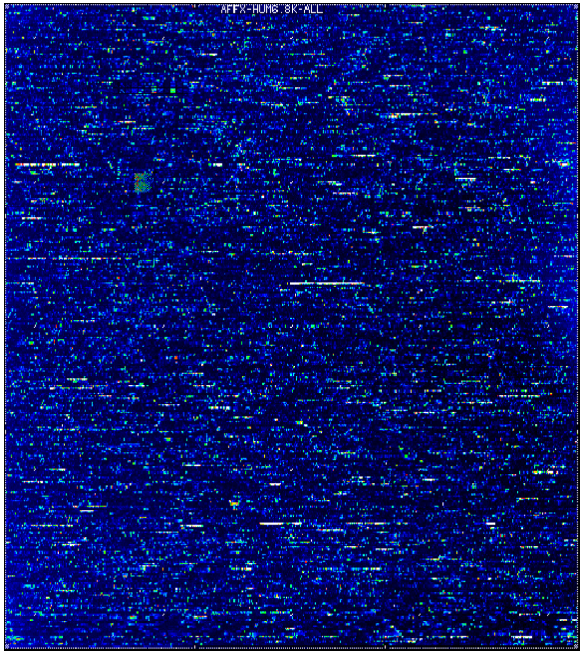
- Expression, sequence, structure, annotation, literature.
- Integration will depend on our using a common language and will rely on database methodology as well as statistical analyses.
- This area is largely unexplored.

Pre-processing

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

Pre-processing: Oligonucleotide chips





Terminology

- Each gene or portion of a gene is represented by 16 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 (13th) base (transversion purine <-> pyrimidine, G <->C, A <->T) .
- Probe-pair: a (PM,MM) pair.
- Probe-pair set: a collection of probe-pairs (16 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

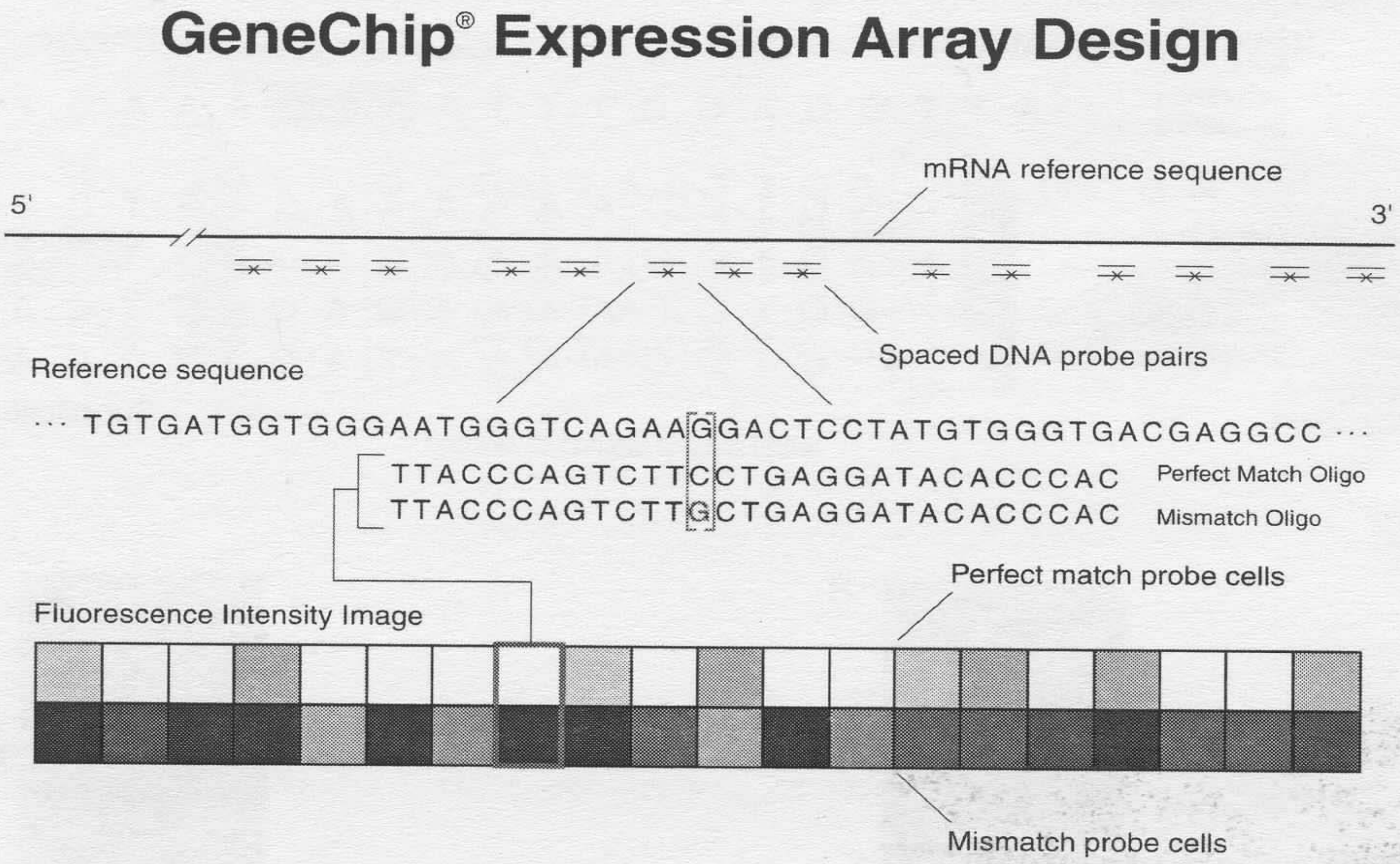

GeneChip® Expression Array Design
mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
TTACCCAGTCTTCCTGAGGATACACCCAC
Perfect Match Oligo
TTACCCAGTCTTGCTGAGGATACACCCAC
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 1-3 Expression tiling strategy

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*.
  - $i = 1,\ldots, n$ -- from one to hundreds of chips;
  - $j = 1,\ldots, J$ -- usually 16 or 20 probe pairs;
  - $g = 1,\ldots, G$ -- 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

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

  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 $d_{(2)}$, …, $d_{(J-1)}$.
- 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 works with PMs only.

Expression measures

- Most 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 .
- 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, I

Simple background estimation

- Estimate $\log_2$(BG) as the mode of the $\log_2$(MM) distribution for a given chip (kernel density estimate).

- Quick fix when PM <= BG: use half of the minimum of $\log_2$(PM-BG) for PM > BG over all chips and probes.

RMA background, II

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

- Bolstad et al. (2002). 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, I

Simple measure

$$\mathrm{RMA} = \frac{1}{|\mathrm{A}|} \sum_{j \in A} \log_2 (PM_j - BG_j)$$

with A a set of “suitable” pairs.

RMA expression measures, II

- 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 use 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).
- Two main functions: **`ReadAffy`**, **`express`**.

Combining data across slides

Data on *G* genes for *n* hybridizations

→ *G x n* genes-by-arrays data matrix

Arrays

		Array1	Array2	Array3	Array4	Array5	...
Genes	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