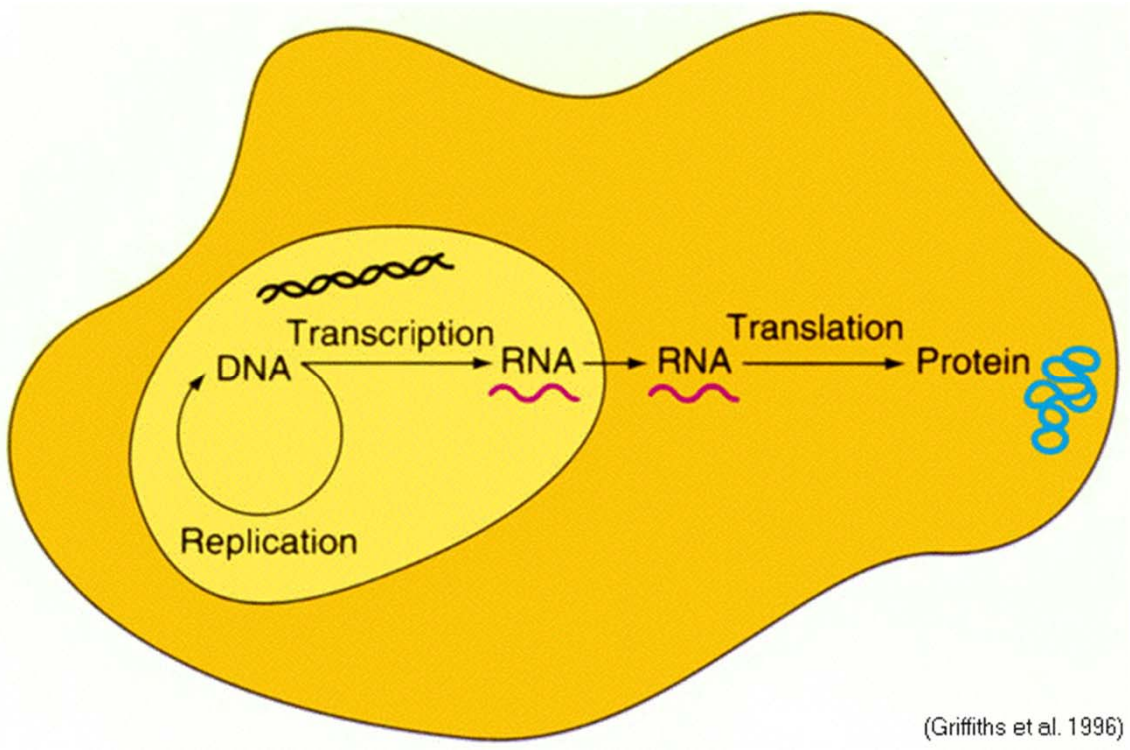
第6-1章: Microarray and Expression Measurements

- Introduction to Microarray
- Expression measurements

Transcriptome



Gene Expression

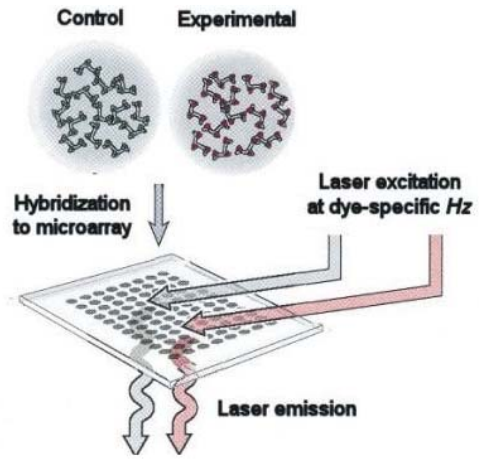
- Each cell contains a complete set of DNA.
- Only a fraction of these are used (or “expressed”) in any particular cell at any given time. For example, genes specific for erythroid cells, such as the hemoglobin genes, are not expressed in brain cells.

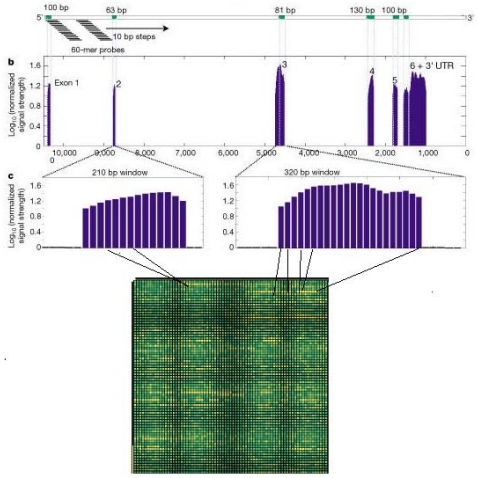
What is a DNA Microarray?

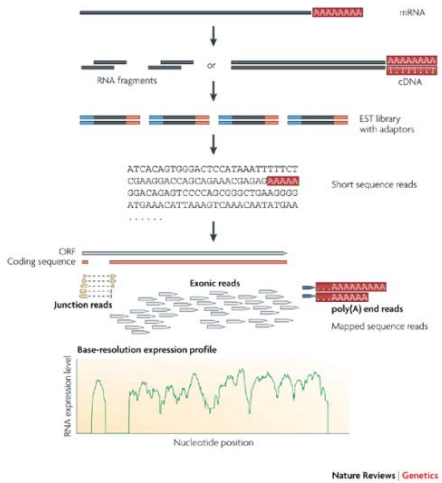
- Also known as DNA Chip
- Allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression)
- Transcription?
  - Process of copying of DNA into messenger RNA (mRNA)
  - Environment dependent!
- Microarray detects mRNA, or rather the more stable cDNA

The Evolution of Transcriptomics

Hybridization-based







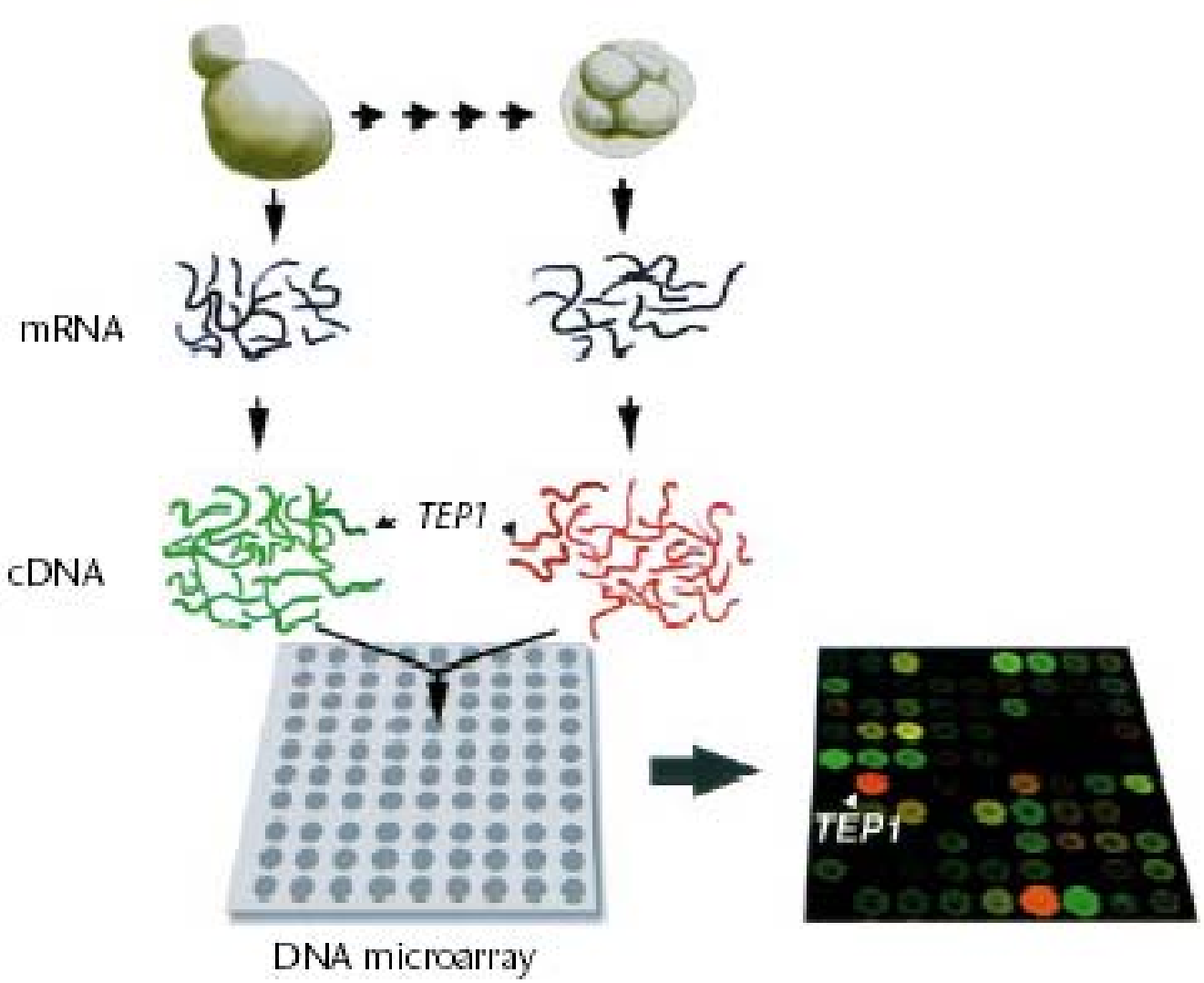
RNA-seq is still a technology under active development

1995 P. Brown, et. al.
Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

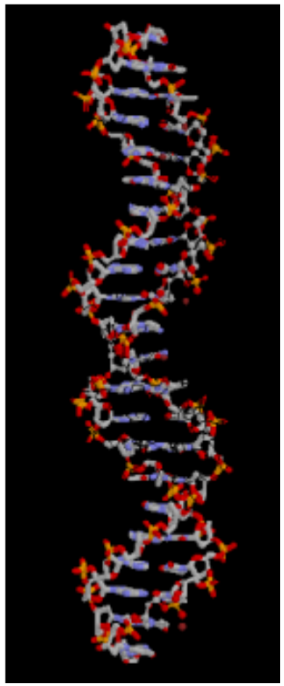
2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

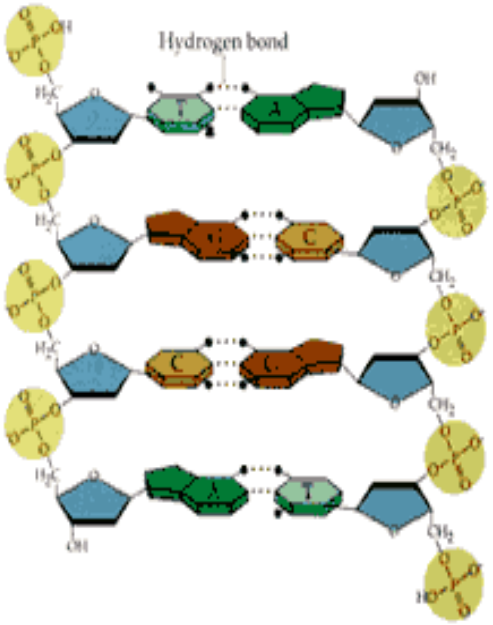
cDNA Microarray

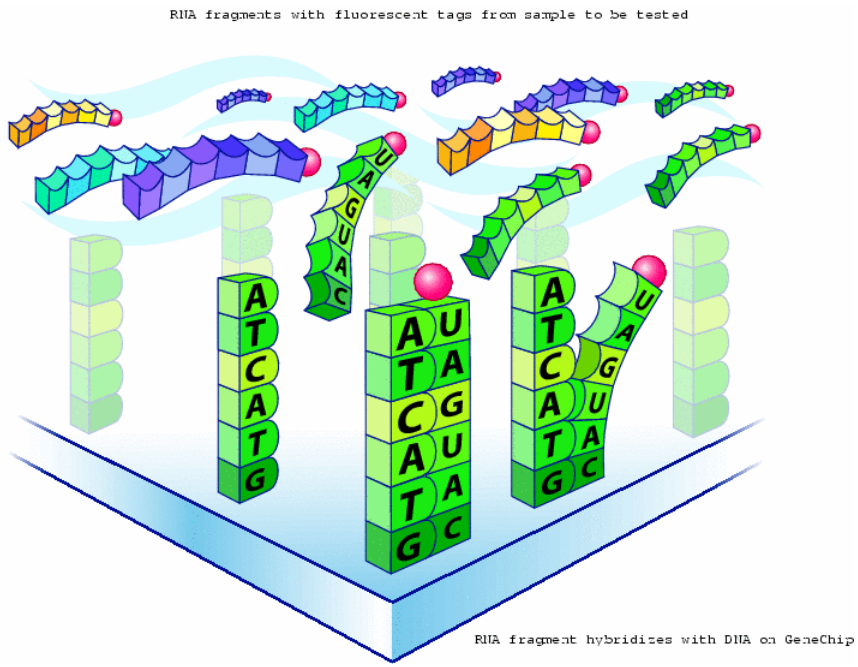


Brown & Botstein, 1999

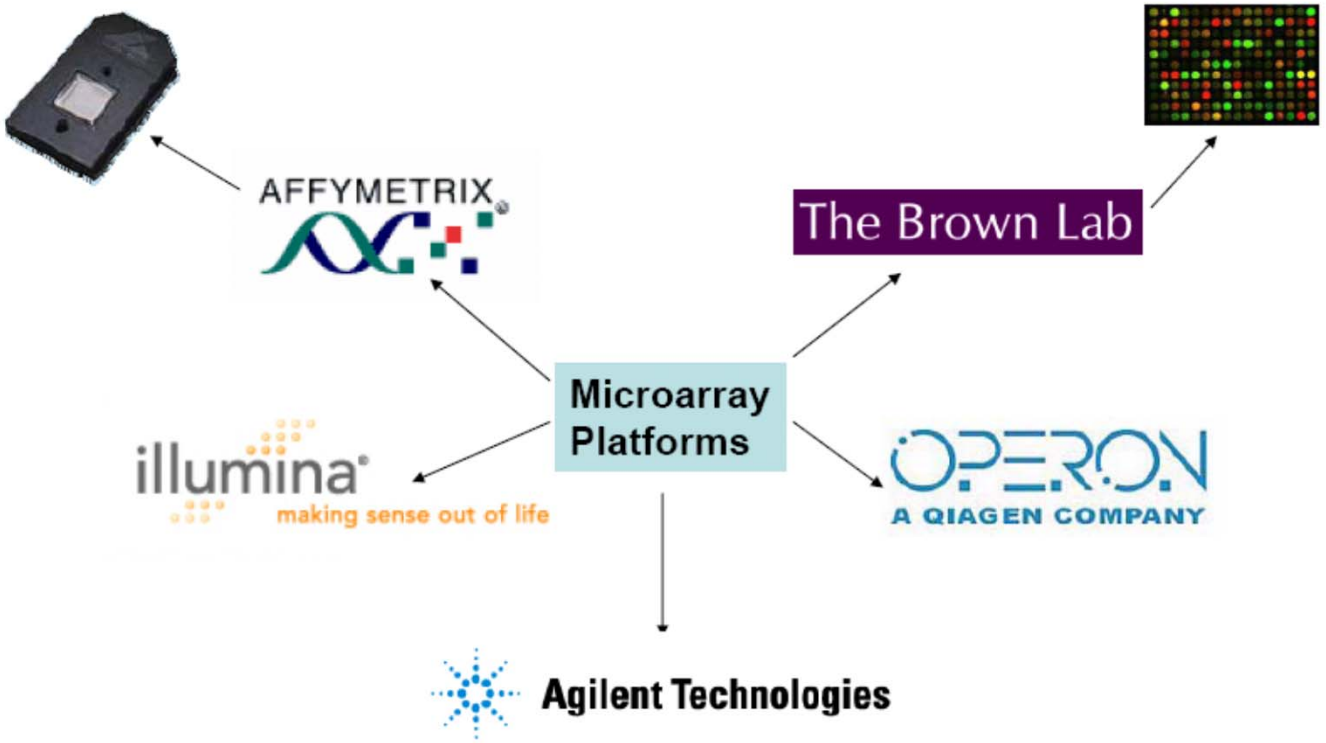
杂交机制(hybridization)->生物芯片



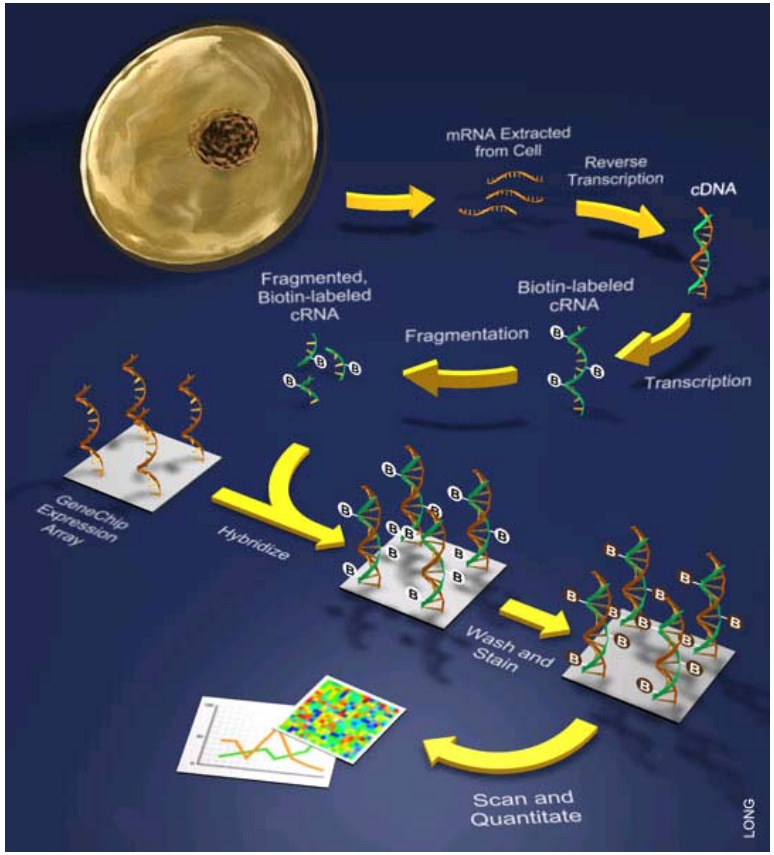


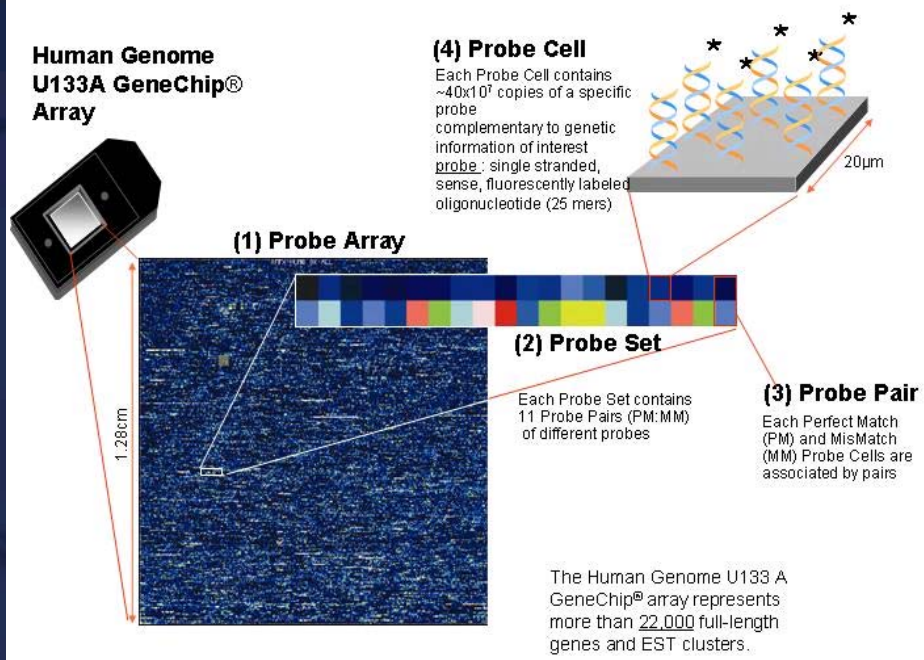


生物学芯片


AFFYMETRIX
The Brown Lab
Microarray Platforms
illumina
making sense out of life
OPERON
A QIAGEN COMPANY
Agilent Technologies

Affymetrix 表达芯片





Glossary

- Probe: gene specific oligonucleotides
- 20 - 25 bases in each probe, i.e., 20-25 mer
- Perfect match (PM)
- Miss match (MM)
- Probe pair: a (PM, MM) pair
- Probe-pair set: a collection of probe-pairs (usually 20) represent each gene
- “20 digit barcode”

GeneChip® Expression Array Design

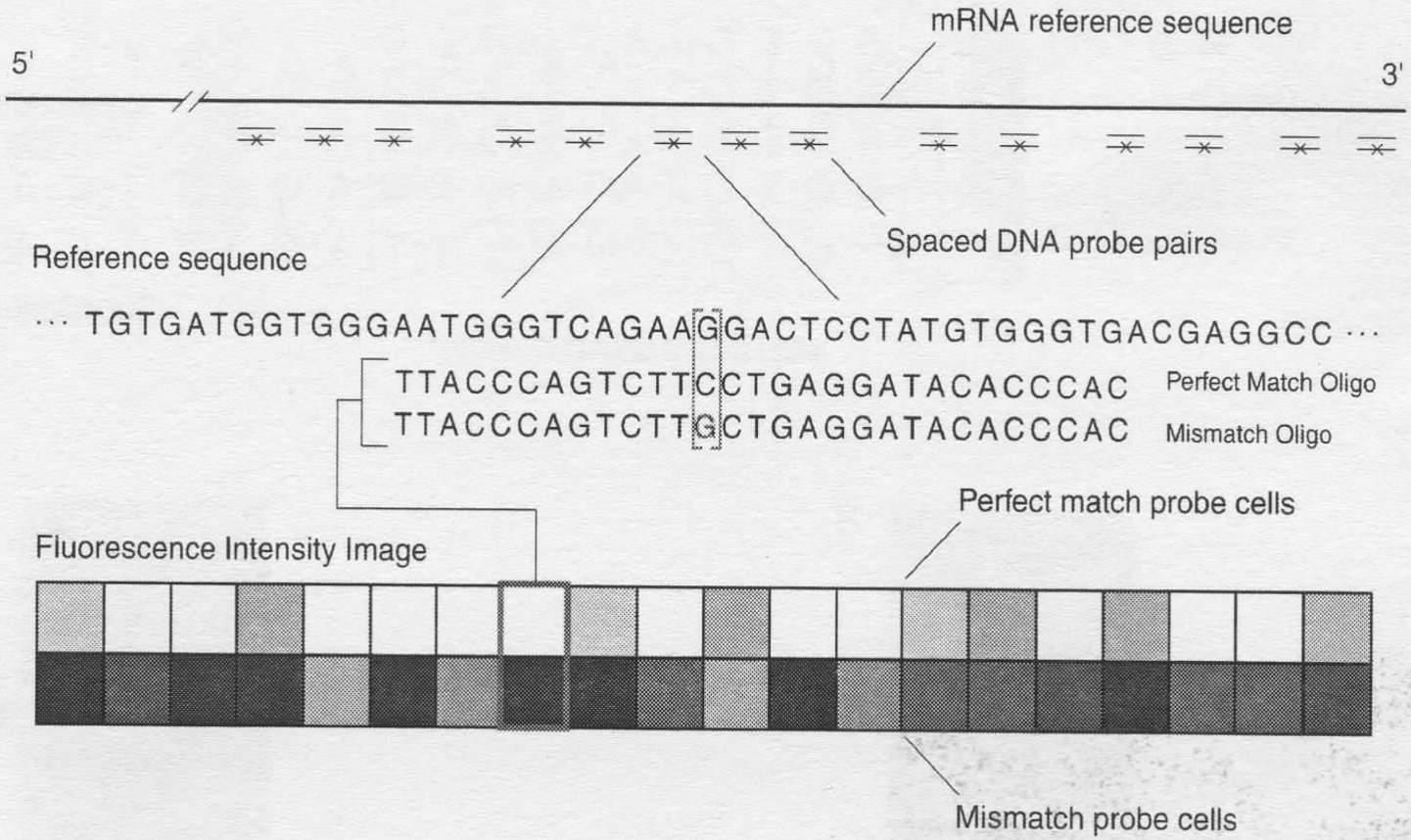
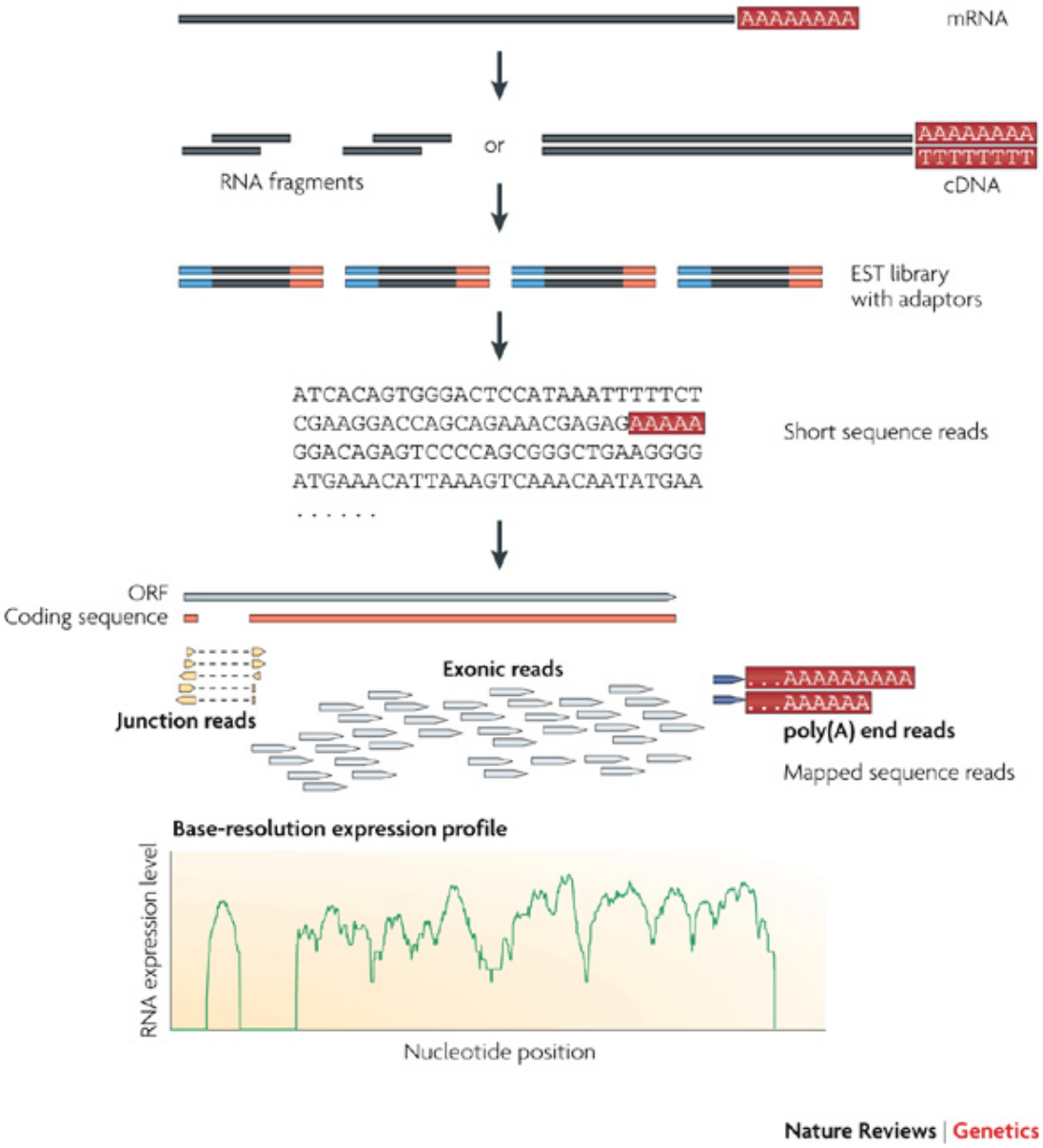


Figure 1-3 Expression tiling strategy



How RNA-seq works

Sample preparation

Next generation sequencing (NGS)

Data analysis:
- ✓Mapping reads
- ✓Visualization (Gbrowser)
- ✓De novo assembly
- ✓Quantification

Figure from Wang et. al, **RNA-Seq: a revolutionary tool for transcriptomics,** Nat. Rev. Genetics 10, 57-63, 2009)**.**

FPKM (RPKM): Expression Values

- **F**ragments ~~Reads~~ **P**er **K**ilobase of exon model per **M**illion mapped fragments
- Nat Methods. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Mortazavi A et al.

$$FPKM = 10^9 \times \frac{C}{NL}$$

C= the number of reads mapped onto the gene's exons
N= total number of reads in the experiment
L= the sum of the exons in base pairs.

Part I: Expression Measurement

1. MEBI
2. PDNN

References

- Cheng Li, Wing H. Wong. Model-based analysis of oligonucleotide arrays Expression index computation and outlier detection. PNAS 98:31-36, 2001.
- Zhang L, Miles MF, Aldape KD. A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnology 21(7):818-821, 2003.

Expression Measurement

- Affymetrix average approach
- Model Based Expression Index approach (Li & Wong)
- Robust Multi-Array approach (Irizarry & Speed)
- Position dependent nearest neighbor model (Zhang et al)

Data and Notation

- Probe intensity in chip i, probe j, and gene n

$$PM_{ijn}, MM_{ijn}$$

- i=1,…I (ranging from 1 to hundreds)
- j=1,…, J (usually 16 or 20)
- n=1,…, N (between 8,000 to 12,000)

Affymetrix Average Approach

- Affymetrix's Genechip@ software use Avg.diff

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

where A is a set of suitable pairs chosen by the software

Affymetrix's MAS5.0

- Affymetrix's new analysis approach

$$signal = TukeyBiweight\{\log(PM_j - \hat{MM}_j)\}$$

where MM* a version of MM that is never bigger than PM

dChip---MBEI
Model Base Expression Index

- Li-Wong full

$$\begin{cases} PM_{ij} = v_j + \alpha_j\theta_i + \phi_j\theta_i + \epsilon \\ MM_{ij} = v_j + \alpha_j\theta_i + \phi_j\theta_i + \epsilon \\ \epsilon \sim N(0, \xi^2) \end{cases}$$

with identifiablity constraint

$$\sum_j \phi_j^2 = J$$

dChip---MBEI
Model-based Expression Index

- Li-Wong reduced

$$\begin{cases} y_{ij} = PM_{ij} - MM_{ij} = \phi_j \theta_i + \epsilon \\ \epsilon \sim N(0, \sigma^2), \quad \sigma^2 = 2\xi^2 \end{cases}$$

with identifiability constraint

$$\sum_j \phi_j^2 = J$$

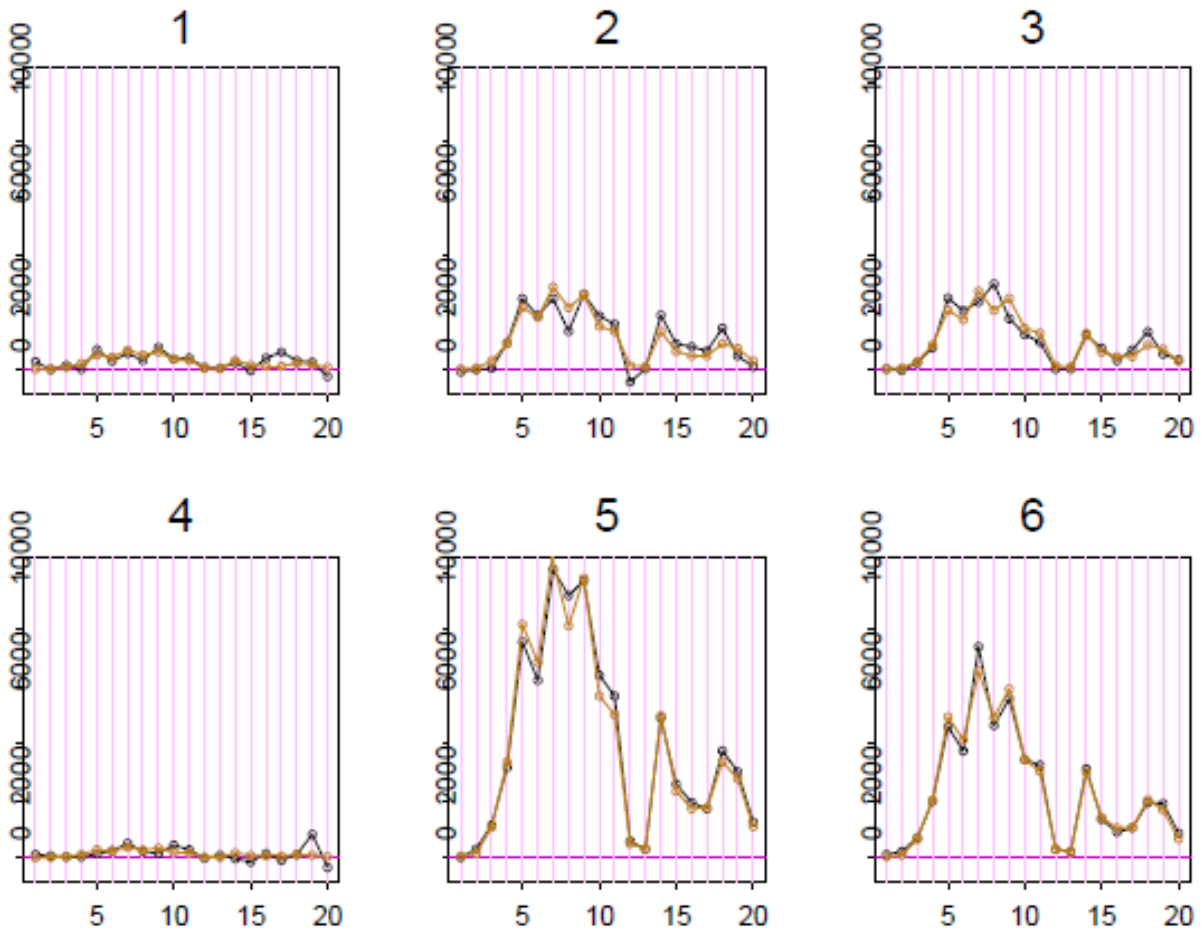


Figure 1.2. Black curves are the PM-MM difference data of gene A in the first 6 arrays. Light curves are the fitted values to model (2).

PDNN Model

- The model is named Position Dependent Nearest Neighbour (PDNN)

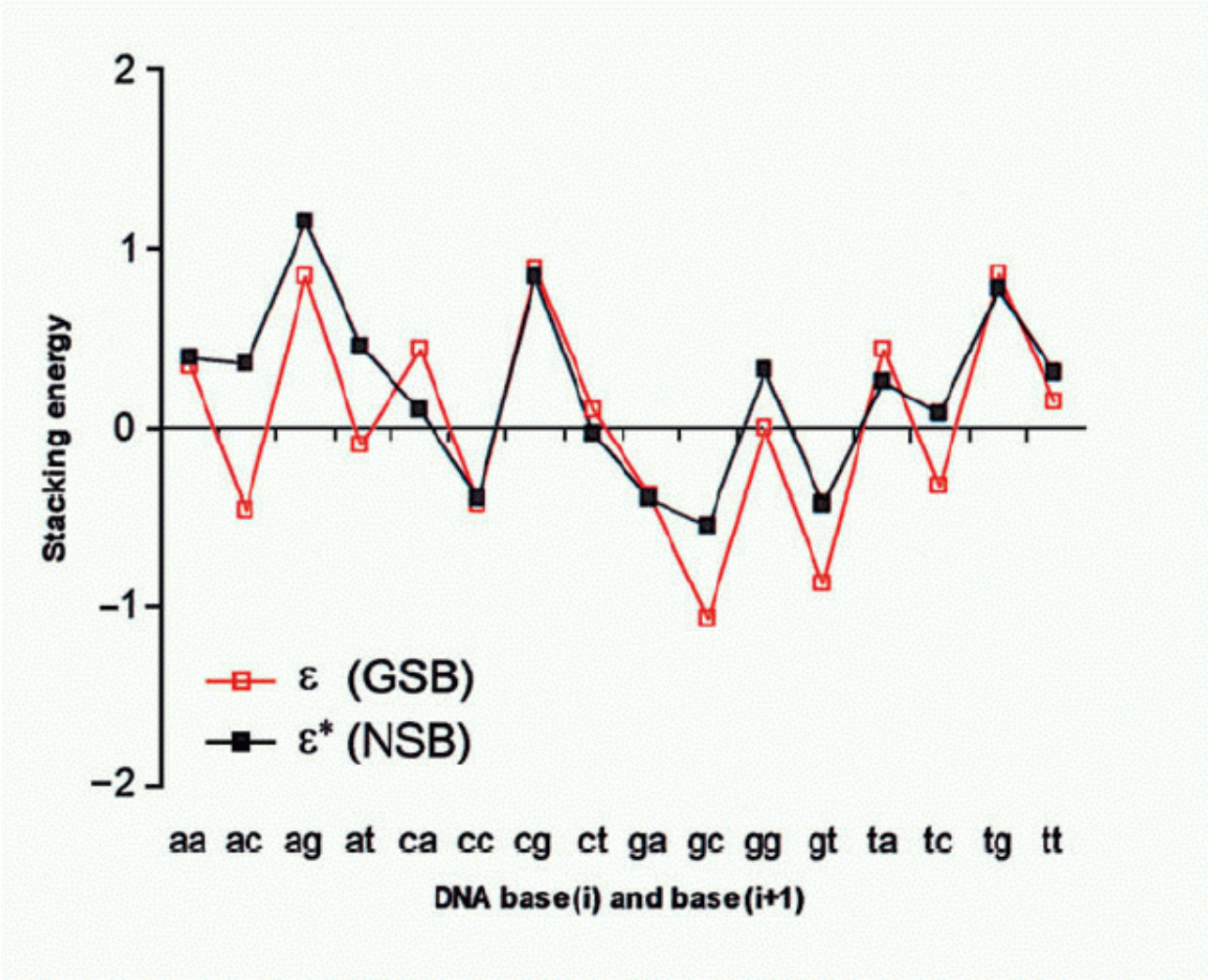
- The hybridization is characterized by a energy, and the energy is approximated by pair-wised interaction between nearest neighbours.

PDNN Model

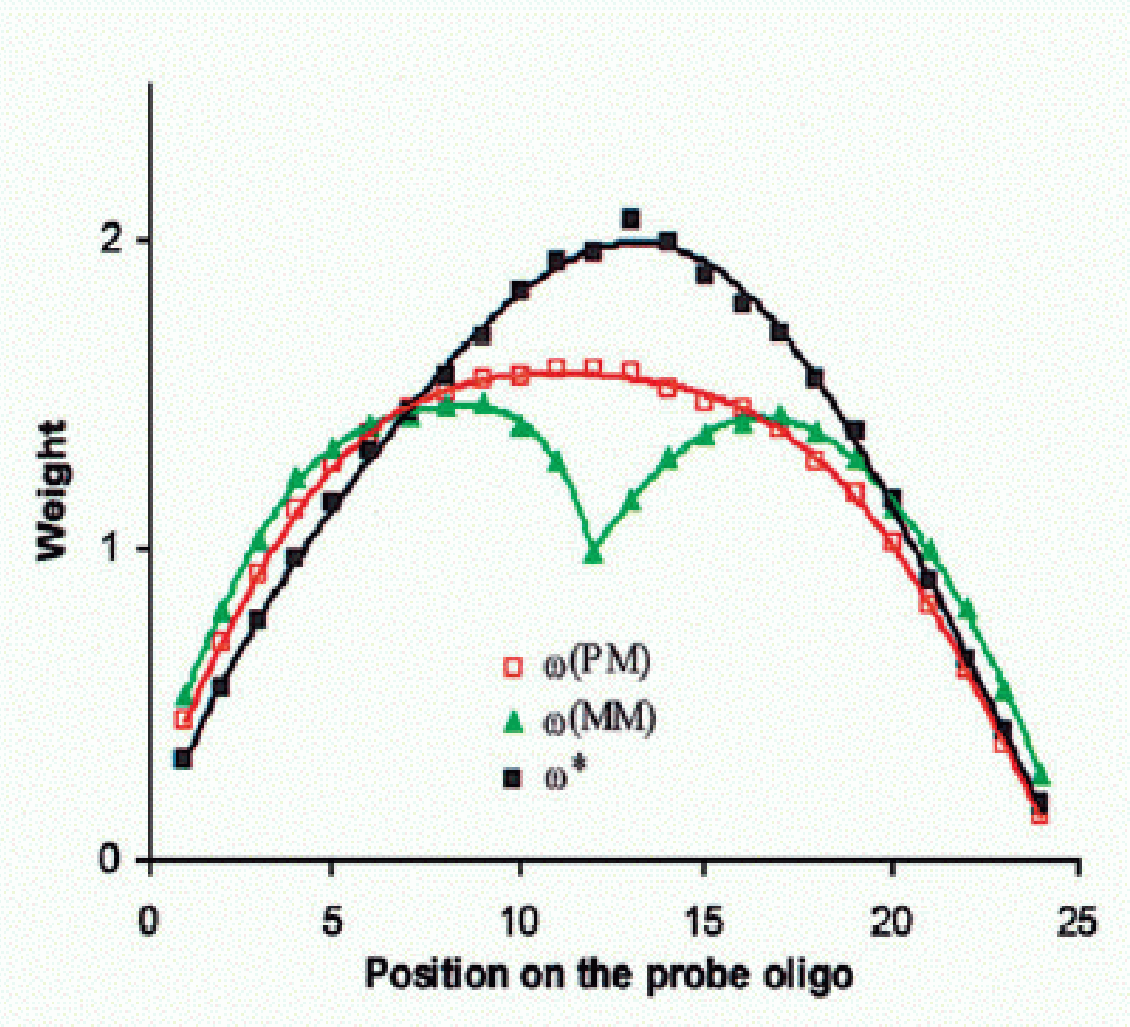
- Position Dependent Nearest Neighbor Model

$$\begin{cases} E = \sum_{i=1} 24\omega_i \lambda(b_i, b_{i+1}) \\ I = \dfrac{N}{1 + \exp(E)} + b + \epsilon \end{cases}$$

Stack Energy



Positional Weights



Extension : Generalized PDNN Model in SNP Array

$$
\begin{cases}
E(S^P, S^T) = \sum_{i=1}^{24} \omega_i \lambda(S_i^P, S_{i+1}^P) \\
E_1(S^P, S^T) = \left\{ \sum_{l=1, l\neq 12+j, 13+j}^{24} \theta_l^j \lambda(S_l^P, S_{l+1}^P) \right\} \\
\qquad + \kappa^j \delta(S_{12+j}^P S_{13+j}^P S_{14+j}^P, S_{12+j}^T S_{13+j}^T S_{14+j}^T) \\
E_2(S^P, S^T) = E_1^{S^P, S^T} + \xi^j(S_{12+j}^P S_{13+j}^P S_{14+j}^P, S_{12+j}^T S_{13+j}^T S_{14+j}^T)
\end{cases}
$$

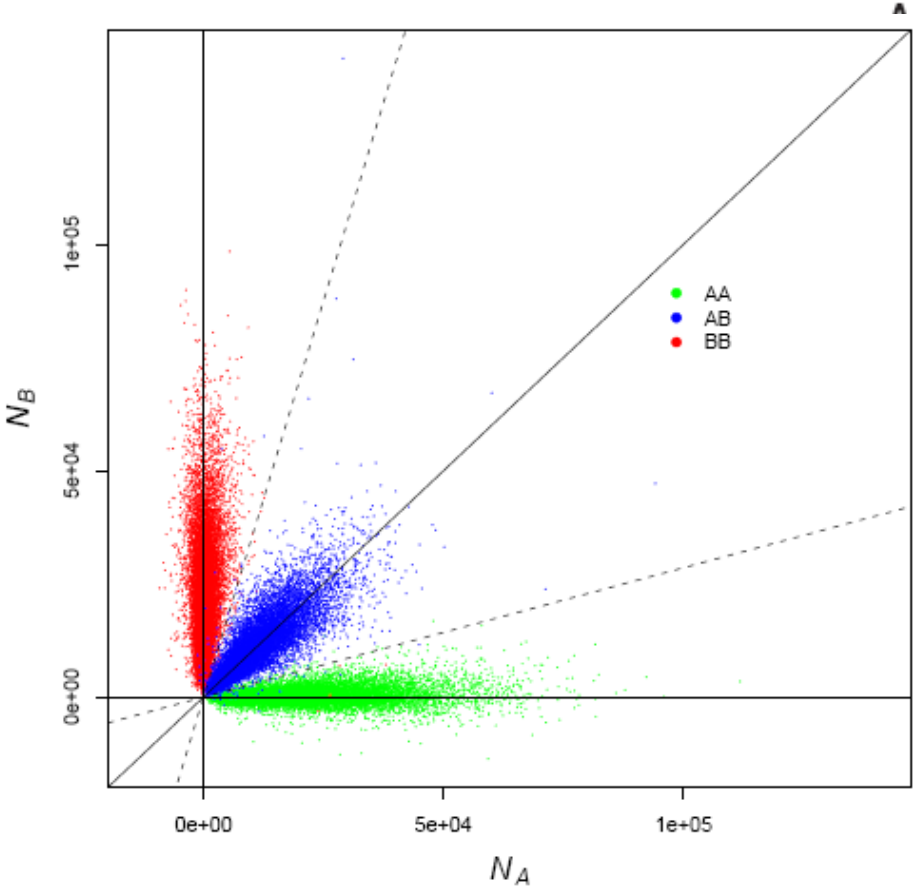
Wan L, Sun KL, Ding Q, Cui YH, Li M, Wen YL, Elston R, Qian MP and Fu WJ. Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Research, 37(17):e117.(2009)

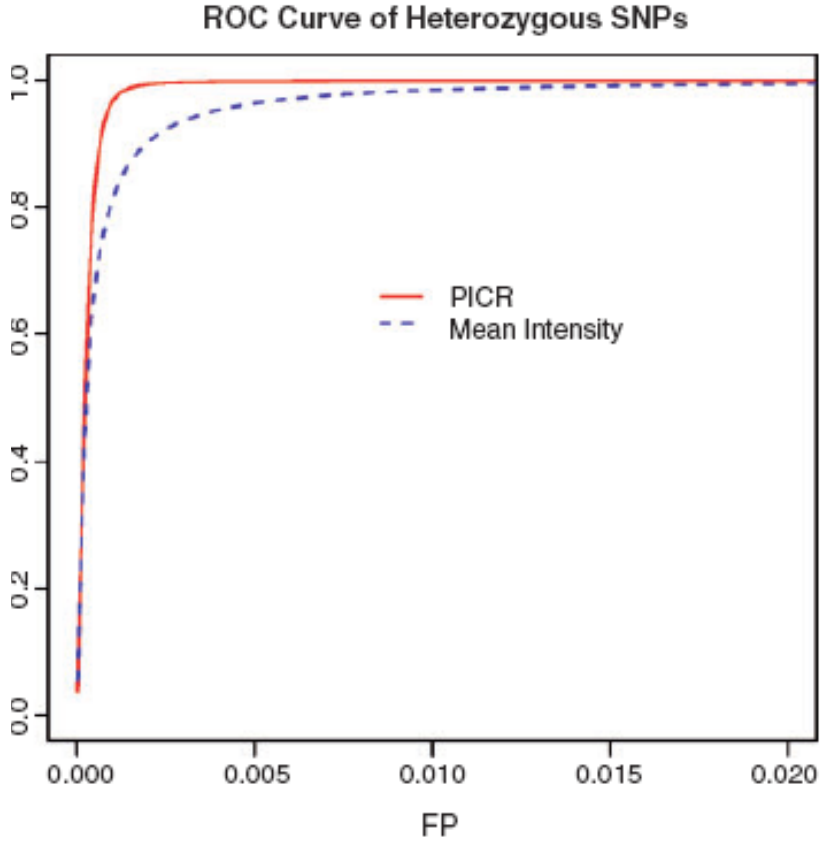
Extension: Generalized PDNN Model in SNP Array

$$
\begin{cases}
\phi(x) = \dfrac{1}{1+\exp(x)} \\
I_{PA,ks} = N_A \phi(E(S^{PA,ks}, S^{TA})) + N_B \phi(E_1(S^{PA,ks}, S^{TB})) \\
\qquad + b_{PA,ks} + \epsilon_{PA,ks} \\
I_{PB,ks} = N_A \phi(E_1(S^{PB,ks}, S^{TA})) + N_B \phi(E(S^{PB,ks}, S^{TB})) \\
\qquad + b_{PB,ks} + \epsilon_{PB,ks} \\
I_{MA,ks} = N_A \phi(E_1(S^{MA,ks}, S^{TA})) + N_B \phi(E_{t_k}(S^{MA,ks}, S^{TB})) \\
\qquad + b_{MA,ks} + \epsilon_{MA,ks} \\
I_{MB,ks} = N_A \phi(E_{t_k}(S^{MB,ks}, S^{TA})) + N_B \phi(E_1(S^{MB,ks}, S^{TB})) \\
\qquad + b_{MB,ks} + \epsilon_{MB,ks}
\end{cases}
$$

Wan L, Sun KL, Ding Q, Cui YH, Li M, Wen YL, Elston R, Qian MP and Fu WJ. Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Research, 37(17):e117.(2009)

More Accurate Genotyping





Wan L, Sun KL, Ding Q, Cui YH, Li M, Wen YL, Elston R, Qian MP and Fu WJ. Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Research, 37(17):e117.(2009)