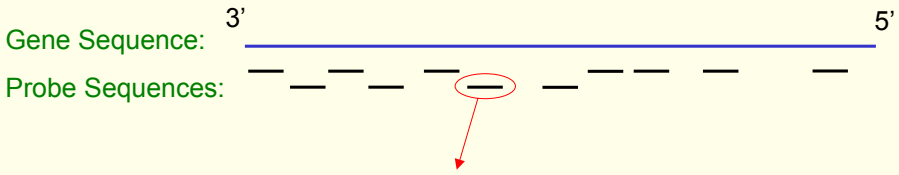
OLIGONUCLEOTIDE MICROARRAYS
(GeneChips)



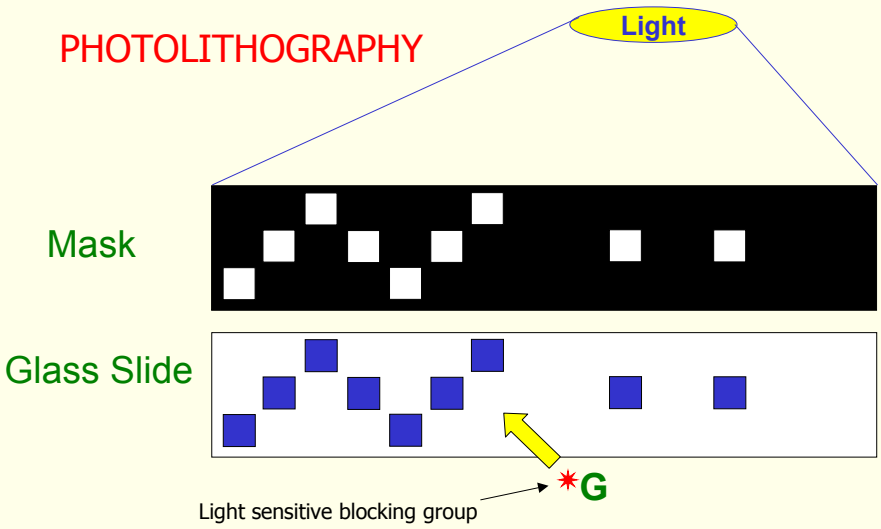
Perfect match: A-C-T-G-T-T-T-A-C-G-C-T-C-A-G-T-C-G-G-G-T-C-A-A-T

Mismatch: A-C-T-G-T-T-T-A-C-G-C-T-A-A-G-T-C-G-G-G-T-C-A-A-T

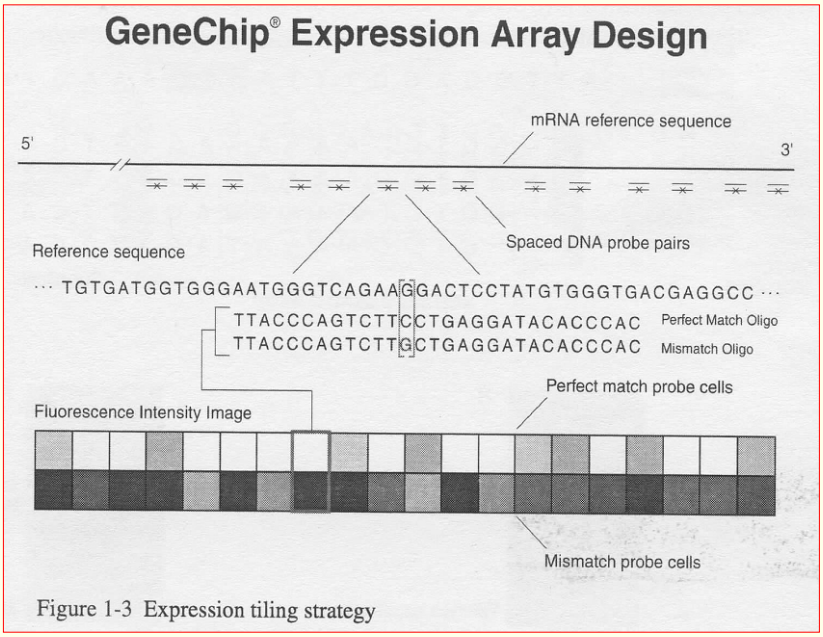
Probe set: 11 to 20 probe pairs (PM & MM)
to interrogate each gene

There may be 5,000-20,000 probe sets per chip

PHOTOLITHOGRAPHY



Repeat the process with different masks (for T, A and C)
until the entire chip contains a single nucleotide.



TERMINOLOGY

Probe: DNA sequence immobilized on the solid/array

Target: DNA or RNA sequence from the sample being interrogated

"Low-level Analysis": feature extraction, normalization, and computation of expression indexes

"High-level Analysis": self-organizing maps (SOM) (Tamayo et al., 1999); two-way clustering (Alon et al., 1999); hypothesis testing; etc.

Expression Index: (gene *g*)

$$e_g = \underbrace{\frac{1}{J}\sum_{j=1}^{J}(\mathrm{PM}_{gj} - \mathrm{MM}_{gj})}_{\text{"average difference" or "signal"}}$$

➡ Alternative: robust average

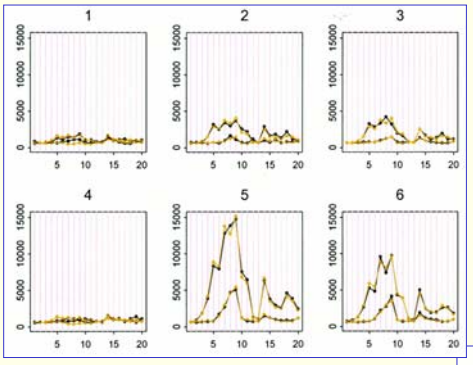
ANALYSIS OF MICROARRAY (OLIGO) DATA

- Large data sets
- Many levels of variation at different stages of the experiments
- Large differences among different probes used to interrogate the same gene (even after MM correction)

Example: (Li and Wong, 2001)

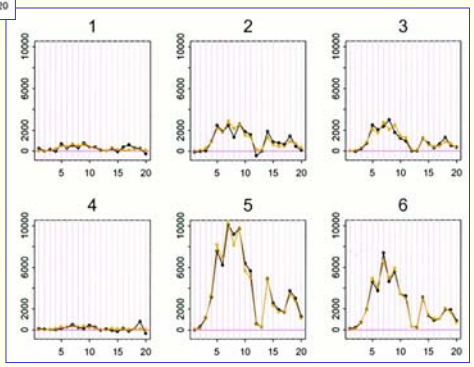
- 21 arrays (250,000+ probes features; 7129 probe sets)

- Results from 6 arrays; probe set # 6457 (gene A)
- Variation due to probe effects is larger than the variation due to arrays
- Probe-specific biases, however, are highly reproducible and predictable
- Probe-level statistical model accounting for probe-specific effects and model-based gene expression indexes: Software package DNA-Chip Analyser (dChip) at www.dchip.org



Legend: Black curves are the PM and MM data of gene A in the first six arrays. Light curves are the fitted values of model (1). Probe pairs are labeled 1 to 20 on the x-axis.

Legend: Black curves are the PM-MM difference data of gene A in the first six arrays. Light curves as the fitted values of model(3).



(Li and Wong, 2001)

STATISTICAL MODEL FOR A PROBE SET

- I samples in the experiment; raw data: $2 \times I \times 20$ for each gene
  - PM & MM
  - samples
  - probes
- θ_i : Model-based expression index (MBEI) for the gene in sample *i*

Model:

$$\begin{cases} MM_{ij} = \nu_j + \theta_i \alpha_j + \varepsilon \\ PM_{ij} = \nu_j + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon = \nu_j + \theta_i (\alpha_j + \phi_j) + \varepsilon \end{cases}$$

- ν_j: Baseline response of the probe pair due to nonspecific hybridization
- α_j: Rate of increase of MM response (α_j and $\phi_j > 0$)
- ϕ_j: Additional rate of increase of PM response
- ε: Random error

Assumptions:

- Intensity value of a probe increases linearly with θ_i
- Different rates of increase for different probes
- PM intensity increases at a higher rate than the MM intensity

MODEL FOR PM-MM DIFFERENCES (y_{ij})

$$y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij} \qquad (1)$$

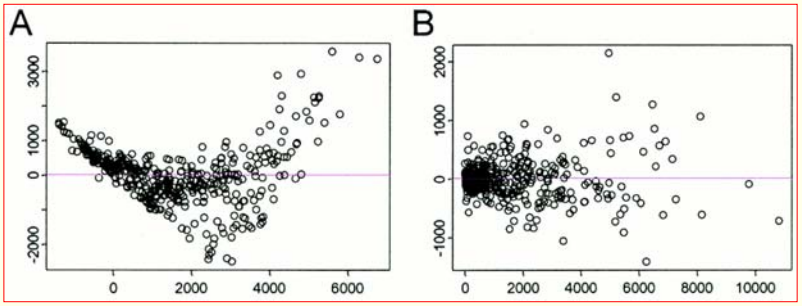
$$\begin{cases} \sum \phi_j^2 = J \quad \text{(model identifiability)} \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

- Choice of analysing y_{ij} or the pair M_{ij} and PM_{ij}
  - current preference

Alternatively: Standard additive model

$$y_{ij} = PM_{ij} - MM_{ij} = \mu + \theta_i + \phi_j + \varepsilon_{ij} \qquad (2)$$

Plot of residuals (y-axis) vs. fitted value (x-axis) for (A) additive model (2) and (B) multiplicative model (1).



$$R^2_{(1)} = 98.08\% \quad \text{and} \quad \hat{\sigma}^2_{(1)} = 1075$$

$$R^2_{(2)} = 87.85\% \quad \text{and} \quad \hat{\sigma}^2_{(2)} = 2705$$

Source: Li and Wong (2001)

CONDITIONAL MEAN AND STANDARD ERROR

➡ For a single array: $y_{ij} = \mathrm{PM}_{ij} - \mathrm{MM}_{ij} = \theta_i \phi_j + \varepsilon_{ij}$

but with ϕ_j replaced by $\tilde{\phi}_j$, learned from a large number of arrays

➡ Least Squares estimates for θ:

$$\tilde{\theta} = \frac{\sum_j y_j \phi_j}{\sum_j \phi_j^2} = \frac{1}{J}\sum_j y_j \phi_j$$

$$\begin{cases} E[\tilde{\theta}] = \theta \\ Var[\tilde{\theta}] = \sigma^2 / J \end{cases} \qquad (3)$$

➡ Approximate Standard Error for $\tilde{\theta}$: $SE[\tilde{\theta}] = \sqrt{\hat{\sigma}^2 / J}$

$$\hat{\sigma}^2 = (\text{fitted} - \text{observed})^2 / (J-1)$$

➡ Similarly, $SE[\tilde{\phi}]$ can be calculated regarding θ's as fixed.

used for outlier detection and probe selection

Probe Selection; Outlier and Artifact Detection

- Extension of equation (3) to model the response of a probe set to all genes in the sample:

$$y_{ij} = \theta_i^{(1)}\phi_j^{(1)} + \theta_i^{(2)}\phi_j^{(2)} + \ldots + \theta_i^{(n)}\phi_j^{(n)} + \varepsilon_{ij} = \sum_{k=1}^{n} \theta_i^{(k)}\phi_j^{(k)} + \varepsilon_{ij}$$

expression level of gene *k* in array *i*

sensitivity of probe *j* to gene *k*

- Ideally: $\phi_j^{(k')} = 0$ for gene k ($k \neq k$) → Sensitivity
- Cross-hybridization: It is expected that most cross-hybridizations have expression patterns different from that of the target gene; also, different probes in a probe set to cross-hybridize to different nontarget genes (Li et al., 2003)
- Discussion: Long and short oligos arrays

STANDARD OUTLIER ANALYSIS
(Wodicka et al., 1997)

① Compute mean and standard deviation of PM-MM for each probe set, in each array

② Outlier: probe pair with more than 3 standard deviations from the mean

- Drawback: probe with a large response might be the most informative…

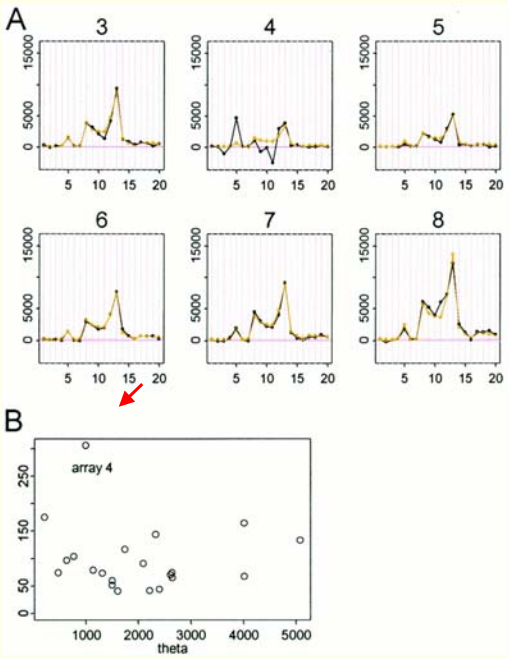
ALTERNATIVE

- "Probe response pattern": 20 ϕ values for a particular probe set
- Hypothesis: 20 PM-MM should follow this pattern (scaled by θ: target gene's expression index)
- Hypothesis: Use the (conditional) standard error of $\tilde{\theta}$ to assess agreement

Legend: (A) 6 arrays of probe set 1248; (B) Plot of standard error (SE, y-axis) vs θ. The probe pattern (black curve) of array 4 is inconsistent with other arrays, leading to unsatisfactory fitted curve (light) and large standard errors of θ_4.

Note: Similar approach may be used to identify problematic probes, using the conditional standard errors of the estimated ϕ_j's (see next figure)

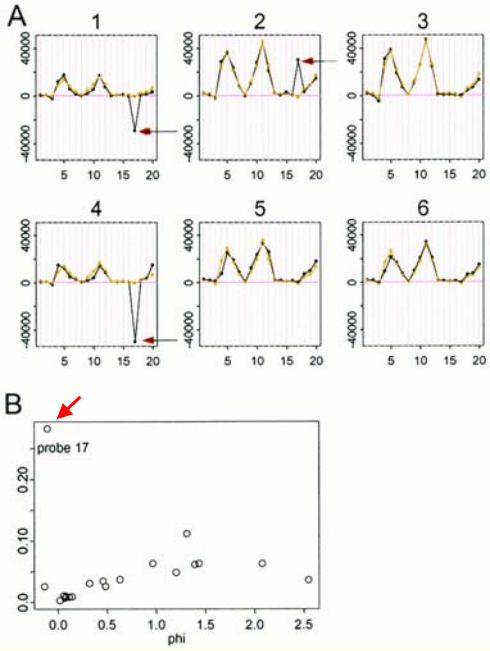
(Li and Wong, 2001)



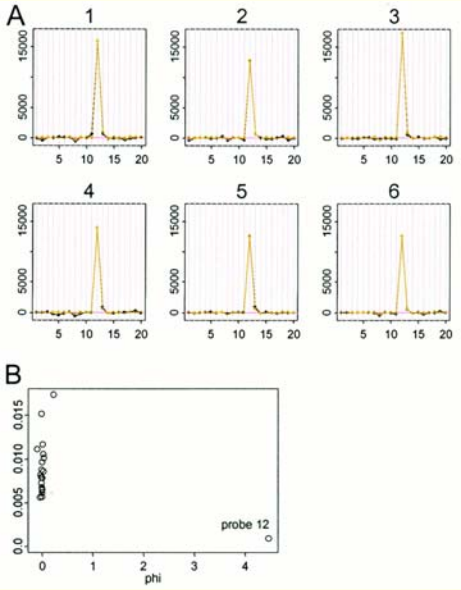
Legend: (A) The probe 17 of probe set 1222 is not concordant with other probes (black arrows); (B) Probe 17 is numerically identified by the large standard error of ϕ_{17}.

Note: To identify single outliers (e.g. y_{ij}), look at the residuals…

(Li and Wong, 2001)



Other undesirable artifacts: Cross-hybridization



Legend: (A) Probe set 3562 has a single high-leverage probe 12, and the fitted light curves almost coincide with the black data curve; (B) ϕ_{12} is large compared to other ϕ's close-to-zero value.

Note: It is expected that more than just one of the 20 probe pairs respond (at various sensitivities) if the target gene exists in the samples.

(Li and Wong, 2001)

ITERATIVE PROCEDURE

① Fit the model to the data table (I arrays by J probes) of each probe set

② Exclude array-outliers: $SE[\tilde{\theta}] > 3 \times \text{Median } SE[\tilde{\theta}]$

③ Fit the model again (data table with fewer rows)

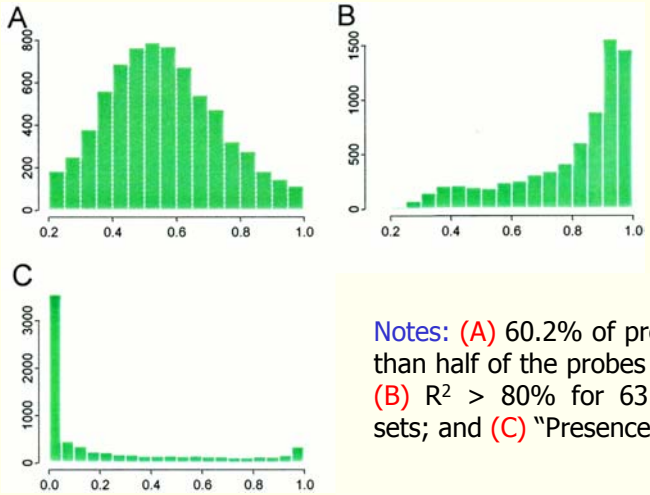
④ Exclude probe-outliers: standard error and magnitude of ϕ's; $\phi < 0$

⑤ Fit the model to the final data table (with fewer columns)

- Note: After probe-outliers are excluded, evaluate all array for outliers again, and compare to the previous set of array-outliers to see if any change occurs. Repeat until the set of probe-outlier and array-outlier does not change any more (in general 5-10 iterations). During this process, identify and exclude single outliers as well.

MODEL-FITTING SUMMARY

Legend: Histograms of (A) Percent of probes used; (B) Explained energy (R^2); and (C) Presence percentage for all 7129 probe sets.



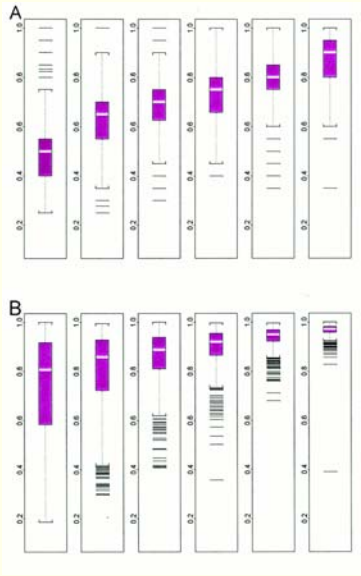
(Li and Wong, 2001)

Notes: (A) 60.2% of probe sets had more than half of the probes in the final model; (B) R^2 > 80% for 63.3% of the probe sets; and (C) "Presence" (GeneChip).

Legend: Boxplots of (A) probe usage and of (B) R^2, stratified by presence percentage (the number of presences of a gene in the 21 arrays and the subpopulation size for the 6 boxplots are: 0-3, 4365; 4-7, 817; 8-11, 567; 12-15, 520; 16-19, 518; and 20-21, 342).

When presence percentage is high, the excluded probes tend to be cross-hybridizing probes; when presence percentage is low, PM-MM differences fluctuating around zero may result in many negative ϕ estimates and exclusion of the corresponding probes. As more arrays enter the database, we may reuse these probes if they respond positively to target expression. The more arrays in which a target gene is present, the higher the R^2.

Comments: (A) High presence percentage leads to high probe usage; (B) When gene is present in many arrays, R^2 of the corresponding probe tends to be high.



Stability of Probe-sensitivity Indexes Across Tissues

- Suppose an experiment with samples corresponding to different treatments or conditions.
- Ideally: probe-sensitivity index (ϕ) should be independent of the treatment… But, cross-hybridization affinity to nontarget genes…
- Nevertheless: cross-hybridization of nontarget genes to only a few probes of a probe set. Also, cross- hybridization expression levels do not correlate with the target gene

---

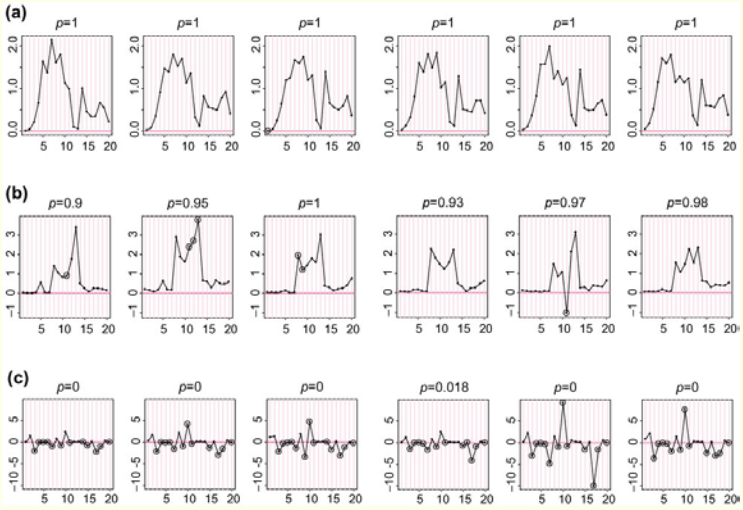
Example:

- Six sets of arrays:
  - 21 leukemia
  - 20 prostate cancer
  - 17 brain tumor
  - 55 cancer
  - 58 brain
  - 55 lung tumor
- Probe sets:

  6457 (a), 1248 (b), and 6571 (c).

Note: (b) Probe 11 in array set 5 is likely to be cross-hybridization (MM > PM $\Rightarrow$ ϕ < 0); Probes 8 and 13: different relative responses in array sets 1 and 4 (reasons: probes differentially cross-hybridized in different array sets, or same probe in different batches of arrays may systematically behave differently)

Legend: Estimated ϕ's (axis y) for three probe sets: (a), (b), and (c) in six array sets. "p" indicates the proportion of arrays called "Present" for the target gene in the array set. Large circles represent identified "probe-outliers" by negativity or large standard error of ϕ.



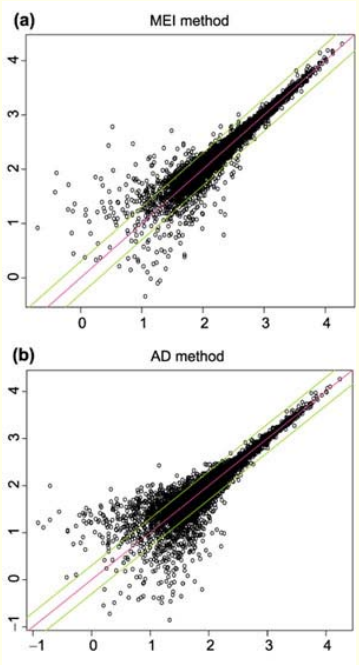
(Li and Wong, 2001)

Agreement of Expression Values of Replicated Arrays

- Array set 5 (Hakak et al., 2001): 29 pairs of replicated arrays (same mRNA sample split, amplified, labeled, and hybridized to two different arrays.

Legend: Log (base 10) expression indexes of single pair of replicate arrays for the MBEI (a) and the AD (b), methods. The center line is y = x and the flanking lines indicate the difference of a factor of two.

Comment: Fold change > 2 for low expressed genes….



CONFIDENCE INTERVAL FOR FOLD CHANGE

- After obtaining expression indexes (using DA or MBEI), fold change can be calculated between two arrays for every gene and be used to identify differentially expressed genes.
- Usually, low or negative expressions are truncated to a small number before calculating fold changes.
- The standard error from MBEI can be used to construct confidence intervals for fold changes.
- Suppose: $\hat{\theta}_1 \sim N(\theta_1, \delta_1^2)$ and $\hat{\theta}_2 \sim N(\theta_2, \delta_2^2)$

  where: θ_1, θ_2 : real (unknown) expression levels in the two samples
  $\hat{\theta}_1, \hat{\theta}_2$: model-based estimates
  δ_1, δ_2 : model-based standard errors

- Let: $r = \theta_1 / \theta_2$ be the real fold change.

➡ Then inference on r can be based on: (Wallace, 1998)

$$Q = \frac{(\hat{\theta}_1 - r\hat{\theta}_2)^2}{\delta_1^2 + \delta_2^2 r^2} \sim \chi^2_{1df}$$

Using expression levels and associated standard errors to determine approximate confidence interval of fold change.

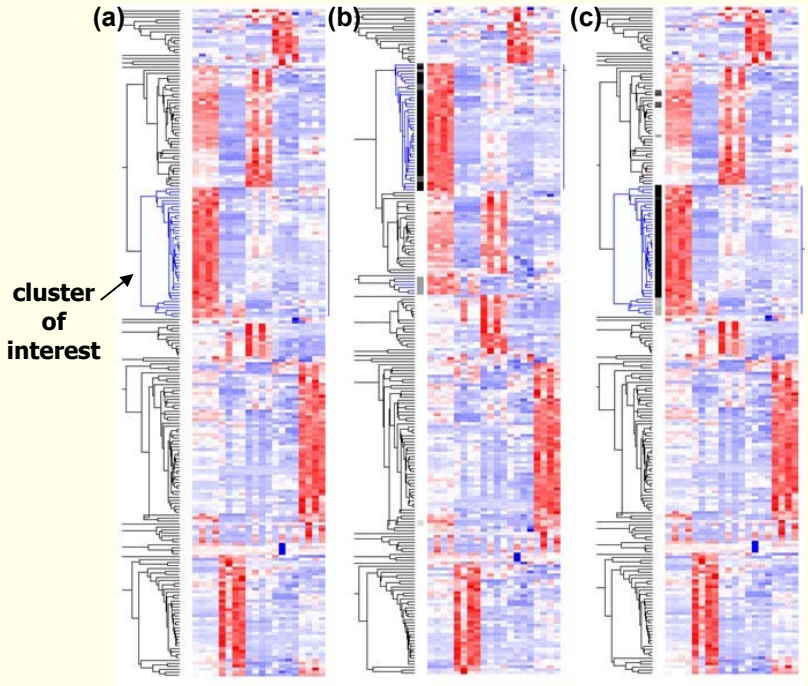
Gene	MBEI 1	SE 1	MBEI 2	SE 2	Fold Change	Lower CB	Upper CB
1	859.6	41.7	347.5	36.1	2.47	2.07	3.03
2	405.7	31.2	164.0	44.2	2.47	1.67	4.49
3	283.9	28.5	114.7	18.5	2.48	1.84	3.48
4	46.0	64.2	18.6	84.5	2.48	0	∞
5	225.2	57.5	90.9	36.2	2.48	1.18	7.49
6	247.0	50.6	99.7	19.5	2.48	1.51	4.02
7	50.0	21.5	20.2	22.6	2.48	0.49	∞
8	276.5	18.7	111.4	36.1	2.48	1.59	5.35
9	436.0	33.0	175.4	21.1	2.49	1.99	3.19
10	75.7	17.7	30.4	18.0	2.49	1.07	86.17
11	80.7	25.3	32.4	17.0	2.49	0.96	18.18
12	181.5	42.5	72.9	28.2	2.49	1.25	7.12
13	1122.2	99.2	449.9	63.3	2.49	1.92	3.35
14	168.2	40.6	67.4	30.3	2.49	1.18	9.82

A PRIME ON CLUSTERING ANALYSIS

- Clustering analysis of microarray data: two genes that are co-regulated (transcriptional level) present correlated expression values across samples.
- Clustering algorithm use these correlations (or the monotone transformation of correlations) to cluster genes.

➡ Example:

- Selected 225 genes: "Presence" proportion > 0.5 and CV > 70% across the 20 samples in array set 2.
- Standardize gene's expression values ~ (0, 1)
- Blue and red represent lower and higher expression levels, respectively
- Suppose we are interested in the gene branch colored in blue
- Panel (b): clustering after a particular resampling
- Panel (c): after resampling 30 times (vertical gray-scale bar denotes the reliability of each gene belonging to the original cluster)



FINAL REMARKS

➡ DNA-Chip Analyser (dChip): www.dchip.org

- normalization
- calculation of MBEI
- confidence intervals of fold changes
- hierarchical clustering (with resampling)

➡ Different expression index computation methods are constantly being proposed and compared; as more validation data are available the advantages and drawbacks of each of them are better assessed.

REFERENCES

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

Li, C. and W. H. Wong. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. 2(8), research 0032.1-0032.11.

Li C., G. C. Tseng, W. H. Wong (2003) Model-based analysis of oligonucleotide arrays and issues in cDNA microarray analysis, in Statistical Analysis of Gene Expression Microarray Data, T. Speed Ed., Chapman & Hall/CRC, Boca Raton, FL, p. 1-34.

Tamayo, P. et al. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. PNAS 96, 2907-2912.

Alon, U. et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 96, 6745-6750.

Hakak, Y. et al. (2001) Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. PNAS 98, 4746-4751.

Wodicka, L. et al. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat. Biotechnol. 15, 1359-1367.

Wallace, D. (1998) The Behrens-Fisher and Fieller-Creasy Problems, in Lecture Notes in Statistics 1, R.A. Fisher: An appreciation, Fienberg S. E. and D.V. Hinkley, Eds., New York: Springer – Verlag, 119-147.