Day 2. cDNA microarray analysis

1. Basic techniques
   Clustering

2. Prediction of phenotype given cDNA pattern
   Partial Least Squares

3. Genetical genomics
   Heat shock proteins (rats)
   Whole genome (yeast)
   Combining expression and markers for gene detection

Reasons for success

Impressive, extremely powerful technology
Potentially very useful in Human genetics
Many data publicly available!
Some webpages

Stanford Microarray Database
http://genome.www5.stanford.edu/MicroArray/SMD/

Lymphoma/Leukemia Molecular Profiling Project

Gene Expression Omnibus
NIH / NCBI

EISEN's lab
Treeview

cDNA microarray principle

Prepare cDNA Probe

"Normal" Tumor

RT / PCR Label with Fluorescent Dyes

Combine Equal Amounts

Hybridize cDNA to Microarray

Microarray Technology

SCAN

NHGRI
A typical cDNA microarray data consists of the measurements of laser intensity, which are assumed to be proportional to the original amount of mRNA in the tissue, of the i-th individual / sample and the j-th gene, \{G_{ij}\}

Some questions that can be addressed by microarrays

- Is a gene expressed differentially in two or more treatments (tissues, time, disease status, etc)?
- How much different are several treatments / genes in terms of their expression profile?
- How does evolution affect gene expression?
- Phenotype prediction: disease status, disease subtype, survival time.
- What is the genetic basis in the variation of gene expression?
- Can expression data be useful to identify causal genes?
Learning techniques

**Unsupervised:** no information on outcome
- Clustering
- Principal components (PCA)
- Self Organizing Maps (SOM)

**Supervised:** information on outcome
- Linear Discriminant Analysis (LDA)
- Support Vector Machine (SVM)
- Neural networks (NN)
- Partial Least Squares (PLS)

---

Day 2. cDNA microarray analysis

1. Basic techniques
   - Clustering

2. Prediction of phenotype given cDNA pattern
   - Partial Least Squares

3. Genetical genomics
   - Heat shock proteins (rats)
   - Whole genome (yeast)
   - Combining expression and markers for gene detection
Unsupervised Learning

There is usually not a measure of 'success', as compared to the supervised methods.
⇒ Proliferation of approaches, as their validity is a matter of opinion.

Clustering techniques

The idea behind is to group genes that show a similar behavior, thus identifying patterns of gene expression

There exist dozens of variants that can be grouped in

• Hierarchical / Non hierarch. clustering
• Agglomerative / Divisive
• Self-organizing maps

Among others
**All ⇒ Definition of distance or ’proximity’**

Euclidean distance:

\[ d_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \]

Pearson’s correlation

\[ r_{xy} = \frac{\sum_{i=1}^{n}x_iy_i - \frac{\sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i}{n}}{\sqrt{\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}} \sqrt{\sum_{i=1}^{n}y_i^2 - \frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}} \]

**WARNING!**

- Results depend on distance chosen
- Difficult to justify any given distance measurement

---

**Hierarchical Clustering**

Unweighted Pair-Group Method Average (UPGMA)

Applied to μarray data by Eisen et al. (1998)

Measure of distance = \( r_{ij} \) (correlation in expression between genes \( i \) and \( j \), or tissue \( i \) and \( j \))

Iterate on:

1) Maximal \( r \) ⇒ Next node.
2) New observation computed as the average expression levels of joined genes.
3) Recompute \( r \) for remaining pairs.

The UPGMA method was widely used in phylogeny ⇒ rooted tree.

**The nice appearance of the result (dendrogram) is one of the main reasons for its success**
Human breast tumours are diverse in their natural history and in their responsiveness to treatments. Variation in transcriptional programs accounts for much of the biological diversity of human cells and tumours. In each cell, signal transduction and regulatory systems transduce information from the cell's identity to its environmental status, thereby controlling the level of expression of every gene in the genome. Here we have characterized variation in gene expression patterns in a set of 65 surgical specimens of human breast tumours from 42 different individuals, using complementary DNA microarrays representing 8,102 human genes. These patterns provided a distinctive molecular portrait of each tumour. Twenty of the tumours were sampled twice, before and after a 16-week course of doxorubicin chemotherapy, and two tumours were paired with a lymph node metastasis from the same patient. Gene expression patterns in two tumour samples from the same individual were almost always more similar to each other than either was to any other sample. Sets of co-expressed genes were identified for which variation in messenger RNA levels could be related to specific features of physiological variation. The tumours could be classified into subtypes distinguished by pervasive differences in their gene expression patterns.
Figure 1 Variation in expression of 1,753 genes in 84 experimental samples. Data are presented in a matrix format: each row represents a single gene, and each column an experimental sample. In each sample, the ratio of the abundance of transcripts of each gene to the median abundance of the gene's transcript among all the cell lines (left panel), or to its median abundance across all tissue samples (right panel), is represented by the colour of the corresponding cell in the matrix.

a, Dendrogram representing similarities in the expression patterns between experimental samples. All ‘before and after’ chemotherapy pairs that were clustered on terminal branches are highlighted in red; the two primary tumour/lymph node metastasis pairs in light blue; the three clustered normal breast samples in light green. Branches representing the four breast luminal epithelial cell lines are shown in dark blue; breast basal epithelial cell lines in orange, the endothelial cell lines in dark yellow, the mesenchymal-like cell lines in dark green, and the lymphocyte-derived cell lines in brown.

b, Scaled-down representation of the 1,753-gene cluster diagram; coloured bars to the right identify the locations of the inserts displayed in c–j.

c, Endothelial cell gene expression cluster; d, stromal/fibroblast cluster; e, breast basal epithelial cluster; f, B-cell cluster; g, adipose-enriched/normal breast; h, macrophage; i, T-cell; j, breast luminal epithelial cell.

Hierarchical Clustering: A note of caution

Results depend very much on distance used.
Results may depend largely on some observations (bootstrap required to assess stability).
The method imposes a hierarchical structure on the data that may not reflect reality.
Day 2. cDNA microarray analysis

1. Basic techniques
   Clustering

2. Prediction of phenotype given cDNA pattern
   Partial Least Squares

3. Genetical genomics
   Heat shock proteins (rats)
   Whole genome (yeast)
   Combining expression and markers for gene detection

Learning ⇔ Phenotype prediction

The issue:

\[ X = \{\text{cDNA measurements}\} \]
\[ y = \{\text{probability of phenotype, say disease status}\} \]
\[ \quad \text{qualitative or quantitative}\] 
\[ y = f(X, \theta)? \]
Partial Least Squares (PLS)
Wold (1975)

Dimension reduction strategy in a situation where we want to relate a set of response variables \( Y \) to a set of predictors variables \( X \).

\[
t_h = X w_h^* \quad \text{(orthogonal } X\text{-components)} \\
u_h = Y c_h \quad \text{(orthogonal } Y\text{-components)}
\]

such that max. \( \text{Cov}(t_h, u_h) \).

There may be many more variables than observations.

In PLS-DA the \( Y \) are binary classificatory variables.

Widely used in chemometrics, some examples in \( \mu \)array analysis (Nguyen & Rocke, 2002; Datta 2002; Pérez-Enciso & Tenenhaus, 2003).

\[
y_k = \sum_{h=1,k} X w_h^* c_h + e = X W^* c + e
\]

\( w_h^* \) = p dimension vector with the weights given to each original variable in the k-th component.

\( c_h \) = the regression coefficient of \( y_k \) on \( h \)-th \( X \)-component variable.
84 tissues
(11 tumoral cell cultures, 65 breast cancer and 3 normal breast samples)
1753 cDNA clones

1. disease status (tumoral / normal)
2. before and after chemotherapy treatment
3. estrogen receptor (ER) status
4. tumor classification.

Disease status: principal components
81 cancer / 3 normal, all 1753 variables
disease status: PLS-DA

81 cancer / 3 normal, all variables

But ...

Models of very poor predictive abilities

Subset of variables (cDNA levels) preselected according to its variable importance in prediction (VIP), sort of weighed correlation.
81 cancer / 3 normal, 21 cDNA levels selected

Prediction of disease status:
PLS-DA

21 variables

Obs. 1 normal, 4-35 tumor predicted
Before / After chemotherapy
PLS-DA
48 observations
19 cDNA levels

ER positive / negative
PLS-DA
60 observations
30 cDNA levels
Tumor class
PLS-DA

54 observations
11 cDNA levels

class 2: 2 - 5 cm
class 3: > 5 cm
class 4: > 5 cm & infiltration

Genes involved in disease status

<table>
<thead>
<tr>
<th>VIP</th>
<th>Symbol</th>
<th>Name</th>
<th>Cluster/Gene Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td>AQP7</td>
<td>Aquaporin 7</td>
<td>Cluster g of Perou related to adipocytes in tumoral tissues</td>
</tr>
<tr>
<td>1.20</td>
<td>ITGA7</td>
<td>Integrin, alpha-7</td>
<td></td>
</tr>
<tr>
<td>1.20</td>
<td>CDK5R1</td>
<td>Cyclin dependent kinase 2</td>
<td></td>
</tr>
<tr>
<td>1.13</td>
<td>FOSB</td>
<td>FJB osteosarcoma oncogene homolog B</td>
<td>Altered in cancer</td>
</tr>
<tr>
<td>1.13</td>
<td>COL14A1</td>
<td>Undulin 4</td>
<td></td>
</tr>
<tr>
<td>1.11</td>
<td>PFKFB3</td>
<td>6 Phosphofructo-2-kinase</td>
<td></td>
</tr>
<tr>
<td>1.06</td>
<td>674</td>
<td>-</td>
<td>Oncogenes</td>
</tr>
<tr>
<td>1.06</td>
<td>GPD1</td>
<td>Glycerol 3 P dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>LPL</td>
<td>Lipoprotein lipase</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>767</td>
<td>-</td>
<td>Ovarian cancer</td>
</tr>
<tr>
<td>.97</td>
<td>FOS</td>
<td>FJB osteosarcoma oncogene homolog 2</td>
<td></td>
</tr>
<tr>
<td>.96</td>
<td>ADH2</td>
<td>Alcohol dehydrogenase 2</td>
<td></td>
</tr>
<tr>
<td>.95</td>
<td>GPD1</td>
<td>Glycerol 3 P dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>.94</td>
<td>GPX3</td>
<td>Glutathione peroxidase</td>
<td></td>
</tr>
<tr>
<td>.93</td>
<td>CNN1</td>
<td>Calponin 1</td>
<td></td>
</tr>
<tr>
<td>.90</td>
<td>FOS</td>
<td>FJB osteosarcoma oncogene homolog</td>
<td></td>
</tr>
<tr>
<td>.89</td>
<td>50</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>.88</td>
<td>CDKNC1C</td>
<td>Cyclin dependent kinase</td>
<td></td>
</tr>
<tr>
<td>.83</td>
<td>647</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>.83</td>
<td>760</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
### Genes involved in chemotherapy status

<table>
<thead>
<tr>
<th>VIP</th>
<th>Symbol</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.22</td>
<td>RCV1</td>
<td>Recoverin</td>
</tr>
<tr>
<td>1.21</td>
<td>FOS</td>
<td>FJB osteosarcoma oncogene homolog 1</td>
</tr>
<tr>
<td>1.15</td>
<td>HBA1</td>
<td>Hemoglobin alpha 1</td>
</tr>
<tr>
<td>1.08</td>
<td>CTGF</td>
<td>Connective tissue growth factor</td>
</tr>
<tr>
<td>1.06</td>
<td>TCEB3</td>
<td>Transcription elongation factor B</td>
</tr>
<tr>
<td>1.06</td>
<td>DCT</td>
<td>Dopachrome tautomerase</td>
</tr>
<tr>
<td>1.05</td>
<td>FOS</td>
<td>FJB osteosarcoma oncogene homolog 1</td>
</tr>
<tr>
<td>1.04</td>
<td>CTGF</td>
<td>Connective tissue growth factor</td>
</tr>
<tr>
<td>1.02</td>
<td>GEM</td>
<td>GTP-binding mitogen-induced t-cell protein</td>
</tr>
<tr>
<td>1.00</td>
<td>NR4A1</td>
<td>Nuclear receptor subfamily 4</td>
</tr>
<tr>
<td>0.98</td>
<td>CDK5R1</td>
<td>Cyclin dependent kinase 1</td>
</tr>
<tr>
<td>0.98</td>
<td>DPYSL3</td>
<td>Dihydropyrimidinase-like 3</td>
</tr>
<tr>
<td>0.94</td>
<td>FY</td>
<td>Blood group-duffy system</td>
</tr>
<tr>
<td>0.91</td>
<td>ATF3</td>
<td>Activating transcription factor 3</td>
</tr>
<tr>
<td>0.90</td>
<td>CDKN1A</td>
<td>Cyclin-dependent kinase</td>
</tr>
<tr>
<td>0.86</td>
<td>COPEB</td>
<td>Core promoter element-binding protein</td>
</tr>
<tr>
<td>0.85</td>
<td>EGR2</td>
<td>Early growth response 2</td>
</tr>
</tbody>
</table>

### Main genes involved in ER status

<table>
<thead>
<tr>
<th>Symbol</th>
<th>VIP</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.17</td>
<td>GATA3</td>
<td>GATA-binding protein 3</td>
</tr>
<tr>
<td>1.14</td>
<td>ESR1</td>
<td>Estrogen receptor 1</td>
</tr>
<tr>
<td>1.12</td>
<td>GATA3</td>
<td>GATA-binding protein 3</td>
</tr>
<tr>
<td>1.11</td>
<td>PES1</td>
<td>Pescadillo 1</td>
</tr>
<tr>
<td>1.08</td>
<td>ITP3</td>
<td>Inositol 1,4,5-triphosphate receptor, type 3</td>
</tr>
<tr>
<td>1.07</td>
<td>GATA3</td>
<td>GATA-binding protein 3</td>
</tr>
<tr>
<td>1.06</td>
<td>GATA3</td>
<td>GATA-binding protein 3</td>
</tr>
<tr>
<td>1.00</td>
<td>DSC2</td>
<td>Desmocollin 2</td>
</tr>
<tr>
<td>1.00</td>
<td>GRO1</td>
<td>Growth regulated protein precursor</td>
</tr>
<tr>
<td>1.00</td>
<td>CCNE1</td>
<td>Cyclin E1</td>
</tr>
<tr>
<td>1.00</td>
<td>TFF1</td>
<td>Trefoil factor 1</td>
</tr>
<tr>
<td>0.99</td>
<td>SLC7A8</td>
<td>Solute carrier family 7</td>
</tr>
<tr>
<td>0.98</td>
<td>ORM1</td>
<td>Orosomucoid 1</td>
</tr>
<tr>
<td>0.97</td>
<td>PFKP</td>
<td>Phosphofructokinase, platelet type 9</td>
</tr>
<tr>
<td>0.97</td>
<td>LRP8</td>
<td>Low density lipoprotein receptor-related protein 8</td>
</tr>
<tr>
<td>0.96</td>
<td>HNMT</td>
<td>Histamine n-methyltransferase</td>
</tr>
<tr>
<td>0.96</td>
<td>HNF3A</td>
<td>Hepatocyte nuclear factor 3-alpha</td>
</tr>
<tr>
<td>0.94</td>
<td>NAT1</td>
<td>N-acetyltransferase 1</td>
</tr>
<tr>
<td>0.94</td>
<td>HMG1</td>
<td>High mobility group protein 1</td>
</tr>
<tr>
<td>0.91</td>
<td>PTK7</td>
<td>Tyrosine-protein kinase-like 7 precursor 0.90</td>
</tr>
</tbody>
</table>

Ocular tumors chemotherapy changed

- Oncogene
- Growth factors, cyclines
- Transcription factors

Well known

Upregulated in cancer, induced by estrogens

Also in West

Also in Gruvbeger
Day 2. cDNA microarray analysis

1. Basic techniques
   Clustering

2. Prediction of phenotype given cDNA pattern
   Partial Least Squares

3. Genetical genomics
   Heat shock proteins (rats)
   Whole genome (yeast)
   Combining expression and markers for gene detection

<table>
<thead>
<tr>
<th>VIP</th>
<th>Symbol</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.35</td>
<td>COL14A1</td>
<td>Undulin 1 altered in cancer</td>
</tr>
<tr>
<td>1.21</td>
<td>1244</td>
<td>-</td>
</tr>
<tr>
<td>1.08</td>
<td>LOX</td>
<td>Protein-lysine 6-oxidase</td>
</tr>
<tr>
<td>1.00</td>
<td>CRIP2</td>
<td>Cysteine-rich intestinal protein 2</td>
</tr>
<tr>
<td>.96</td>
<td>767</td>
<td>-</td>
</tr>
<tr>
<td>.93</td>
<td>459</td>
<td>-</td>
</tr>
<tr>
<td>.93</td>
<td>TFAP2B</td>
<td>Transcription factor AP2-beta</td>
</tr>
<tr>
<td>.90</td>
<td>1542</td>
<td>-</td>
</tr>
<tr>
<td>.86</td>
<td>ARHB</td>
<td>RAS homolog gene family, member B</td>
</tr>
<tr>
<td>.84</td>
<td>1017</td>
<td>-</td>
</tr>
<tr>
<td>.79</td>
<td>MRSPSZ7</td>
<td>KIAA protein tumor progression</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transcription, growth factors</td>
</tr>
</tbody>
</table>
Aim

Studying expression levels as any other quantitative trait

1. Which is the transcriptome's genetic architecture?
2. Can mRNA levels be used to refine QTL position estimates?

QTL for mRNA levels

Dumas et al. 2002
Brem et al. 2002
Pérez-Enciso 2004
Dumas et al. (2000)

Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains.

Biological Background
Heat shock proteins (hsp) are highly conserved, they are induced by several stressors, protect other proteins from denaturalization. HSPs are mediated by heat shock transcription factors (hsf) 1 and 2.
Stress susceptibility is correlated with future high blood pressure.

Methods

- 20 recombinant inbred lines BN.Lx with SHR.
- cDNA probes for 5 hsp.
- 3 Tissues: kidney, heart, and adrenal tissue.
- 4 rats / line.
- 475 polymorphic markers, ~ 20 markers / chr.
- Analysis with MapManager, no statistical details provided (single marker analysis?).
Dumas et al. 2000

Table 1: Total genome scan of hsp expression in the adrenal, heart and kidney of RIS

<table>
<thead>
<tr>
<th>Marker</th>
<th>Adrenal</th>
<th>Heart</th>
<th>Kidney</th>
<th>D7 marker, Pooled器官</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsp27</td>
<td>D7 0.47</td>
<td>0.04</td>
<td>0.85</td>
<td>0.032 0.40 0.06</td>
</tr>
<tr>
<td></td>
<td>Myh3</td>
<td></td>
<td></td>
<td>0.50 0.02 -</td>
</tr>
<tr>
<td>hsp70</td>
<td>D7 0.63</td>
<td>0.03</td>
<td>0.01</td>
<td>0.02 0.48 0.05</td>
</tr>
<tr>
<td></td>
<td>YC1v</td>
<td></td>
<td>0.35</td>
<td>0.13 0.80 0.805</td>
</tr>
<tr>
<td>hsp44</td>
<td>D7 0.42</td>
<td>0.07</td>
<td>-</td>
<td>0.35 0.14 -</td>
</tr>
<tr>
<td></td>
<td>D9 0.57</td>
<td>0.59</td>
<td>-</td>
<td>-    -    -</td>
</tr>
<tr>
<td>hsp61</td>
<td>D7 0.58</td>
<td>0.007</td>
<td>0.44</td>
<td>0.06 0.53 0.02</td>
</tr>
<tr>
<td></td>
<td>YC1v</td>
<td></td>
<td>0.30</td>
<td>0.33 0.83 0.803</td>
</tr>
<tr>
<td>hsp105</td>
<td>D7 0.53</td>
<td>0.02</td>
<td>0.42</td>
<td>0.06 0.4 0.08</td>
</tr>
<tr>
<td></td>
<td>YC1v</td>
<td></td>
<td>0.30</td>
<td>0.30 0.83 0.830</td>
</tr>
<tr>
<td></td>
<td>Myh3</td>
<td></td>
<td>0.39</td>
<td>0.39 0.83 0.830</td>
</tr>
<tr>
<td></td>
<td>D12</td>
<td></td>
<td>0.55</td>
<td>0.31 0.52 0.02</td>
</tr>
</tbody>
</table>

Correlations equal or higher than 0.3 are displayed. Suggestive (P < 0.001) or significant linkages (P < 0.0002 appear in bold type. Abbreviations: D7, D7Cebpa.18f20.057.0.0.32 marker; DA, DlaCol16.G22.012 marker; Myh3, Myh3 gene. Heart (ambiophonic) marker (chromosome 10); YC1v, Y chromosome.

*To increase statistical power, correlations on the three organs were combined for analysis of individual hsp at the D7 marker.
Main results

- Wide variability in expression levels despite uniformity in founder strains.
- No QTL (except evidence of 1) mapped to the gene itself.
- High correlation in expression levels for the same gene between tissues.
- The largest effect QTL region contained the hsft1 gene (chr. 7).
- And also the same QTL affected the expression of all hsps.
Brem et al. (2002)

- Comparison of two S. cerevisiae strains, lab and wild types
- Large differences in gene expression: 1528 / 6215 (P < 0.005)
- Genotyping with microarrays in tetrads, 3312 SNPs, > 99% genome
- Test for linkage between every marker and every cDNA level: Wilcoxon-Mann-Whitney test and P level assigned by permutation.

Main results

308 / 1528 (20%) cDNA levels showed linkage with at least one marker (P<10^-5)

262 mRNA levels not different between strains but linkage to some marker (as in Dumas et al’s results).

1220 (80%) mRNA levels were different but no significant linkage: evidence of multiple loci affecting message level, probably > 5 loci according to simulation.

Is the linked marker located close (< 10 kb) of the gene encoding the mRNA? 185 / 570 = 32% yes action in cis

For the remaining (trans-acting) markers, small number of marker affects many mRNA levels, or many markers each affecting a few mRNAs?: 10 bins contained more than 5 levels (impossible by random), ranging from 7 to 87 levels.
Expression levels of parents and segregants for two genes that show linkage. In each panel, the first column shows expression levels for all 40 segregants, and the second and third columns show expression levels for six replicates of each parent. The fourth and fifth columns show expression levels for segregants that inherited the linked marker from BY and RM, respectively. (A) The gene is YLL007C, and the marker lies in YLL009C. (B) The gene is XBP1 (YIL101C), and the marker lies in YIL060W. Note that, in this example, the effect of the locus is in the opposite direction from the difference between the parents, illustrating transgressive segregation.

The number of linkages plotted against genome location. The genome is divided into 611 bins of 20 kb each, shown in chromosomal order from the start of chromosome I to the end of chromosome XVI. The dashed line is drawn at 5 linkages; no bin is expected to contain 5 linkages by chance. The regions with an unusually large number of linkages are marked 1 through 8 and correspond to the groups in Table 1.
Table 1. Groups of messages linking to loci with widespread transcriptional effects. The location of the center of the linked bin is shown as chromosome:base pair. Lists of genes in each group are available as supplementary information (32).

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of messages</th>
<th>Common function</th>
<th>Linkage bin</th>
<th>Putative regulator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>Budding, daughter cell separation</td>
<td>II:</td>
<td>CST13</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>Leucine biosynthesis</td>
<td>III:</td>
<td>LEU2</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>Mating</td>
<td>III:</td>
<td>MAT</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>Uracil biosynthesis</td>
<td>V:</td>
<td>URA3</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>Heme, fatty acid metabolism</td>
<td>XII:</td>
<td>HAP1</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>Subtelomerically encoded helicases</td>
<td>XII:</td>
<td>SIR3</td>
</tr>
<tr>
<td>7</td>
<td>94</td>
<td>Mitochondrial</td>
<td>XIV:</td>
<td>Unknown</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>Msn2/4-dependent induction</td>
<td>XV:</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Conclusions

- Most levels affected by several loci
- Many regions in cis
- Small number of alleles trans-acting and affecting many mRNA levels simultaneously
Pérez-Enciso
(Genetics, 2004)

1. QTL 'hotspots' reliability.
2. Estimates' stability.

Traditional simulations

1. Model specification (M) and its associated parameters (θ)
2. Data simulation (y) given the model (M) and θ

\[ p(y | θ, M) \]

- phenotypes
- heritability
- Additivity, # loci
Simulación no tradicional

1. Given real expression data ($y_1$)
2. Genotype simulation ($y_2$)
3. Assignment of genotypes randomly to phenotypes
4. Data analysis

\[ p(y_2, \theta_2 | y_1) \]

Real data

• Rosenwald et al. 2002 (NEJM)
  • 240 individuals with lymphoma
  • 7399 probes (lymphochip)
  • ~ 10% missing data

• Whitney et al. 2003 (PNAS)
  • n= 76 (blood)
  • 3441 probes
  • ~ 4% missing data
Simulation: haplotypes

1. Coalescence (programa ms de Hudson): 3000 chrs., 100.000 SNPs, \( \rho = 4Ne r = 1000 \)
2. Gene dropping 1000 gens., \( Ne = 1500, 1 \) Morgan
3. \( \sim 25,000 \) SNPs in \( t = 1000 \)
4. Only SNPs frec > 0.10 analyzed (\( \sim 20,000 \))

Simulation (contd.)

Random assignment of two chrs. to each individual
For each mRNA (j), QTL (k) position is estimated by maximum likelihood (ML)

\[
L_{jk} = \prod_{i=1}^{N} \phi (y_{ij} - \mu_{ijk}, \sigma_{jk}^2)
\]

- i-th indiv., j-th mRNA level
- mean genotype ijk
- Residual variance jk (constant ∀ genotype)
- Significance if \( P < 10^{-6} \)
Simulation plausibility

Patil (2001)


Disequilibrium pattern
Results (1): hotspots

A bin is made of 50 consecutive SNPs.
Results (1): *hotspots*

**Rosenwald**

**Whitney**

Schadt et al. 2003 (mouse)
Results (2): estimates’ reliability

Noise added to the system by:

1. Randomly elimination of 16% of individuals
2. Elimination of 9 out of 10 consecutive SNPs (remaining ~ 2000)
What happens with less individuals?

What happens when we run out of money for genotyping?
### Adding noise (Rosenwald)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>R-Ind</th>
<th>R-Ind-SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>% P&lt;10^{-6}</td>
<td>67</td>
<td>40</td>
</tr>
<tr>
<td>δ=0(%)</td>
<td>70</td>
<td>23</td>
</tr>
<tr>
<td>δ&lt;10(%)</td>
<td>73</td>
<td>50</td>
</tr>
<tr>
<td>δ&lt;100(%)</td>
<td>77</td>
<td>60</td>
</tr>
<tr>
<td>δ&gt;10^4(%)</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

### Adding noise (Brem, yeast)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>B-Ind</th>
<th>B-Ind-SNP</th>
<th>BR-Ind</th>
</tr>
</thead>
<tbody>
<tr>
<td>% P&lt;10^{-6}</td>
<td>72</td>
<td>42</td>
<td>73</td>
</tr>
<tr>
<td>δ=0(%)</td>
<td>76</td>
<td>17</td>
<td>58</td>
</tr>
<tr>
<td>δ&lt;10(%)</td>
<td>94</td>
<td>70</td>
<td>59</td>
</tr>
<tr>
<td>δ&lt;100(%)</td>
<td>100</td>
<td>99</td>
<td>64</td>
</tr>
<tr>
<td>δ&gt;10^4(%)</td>
<td>0</td>
<td>0.1</td>
<td>6</td>
</tr>
</tbody>
</table>
How an association profile looks like?
Results ML vs. ANOVA

Conclusions

- *QTL hotspots* should be interpreted with caution
- LD/association profiles in outbred populations can be extremely complex
- Unstability in ~ 40% QTL
Refining gene positions

- Wayne & McIntyre 2002
- Mootha et al. 2003
- Pérez-Enciso et al. 2003

Wayne & McIntyre (2002)
Combining mapping and arraying:
An approach to candidate gene identification

*Drosophila* ovariole number: related to fecundity and varies with latitude.

QTL analysis in RIL of Oregon-R and 2b strains (⇒ 5286 candidate genes).

Deletion mapping (⇒ 548 candidate genes).

Differences in mRNA levels between strains (⇒ 1 to 25 candidates). Pools of 25 individuals were assayed, 3 replicates per line. Analysis via ANOVA.
The black arrow highlights the recombinational map position of the candidate genes CG17327, yellow-f, and Su(fu). Red curves indicate the value of the test statistic for the presence of QTL. Blue triangles indicate cytological markers used in the QTL experiment. Horizontal bars are the deficiencies that were tested; gold bars showed a significant interaction across parents and genotypes, whereas green bars did not.

Mootha et al. (2003): Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics

Leigh syndrome (French-Canadian type) is relatively common in a Quebec region (1/23 incidence, 1/2000 newborn are affected).

Shown previously to be associated to a region in chr. 2p16-21.

A single founder haplotype was evidenced.
Fig. 2. Microsatellite markers and genetic distances are shown to the left of the chromosome map. Genes with varying levels of annotation support are shown with different colors (RefSeq gene, blue; Ensembl gene, green; human mRNA, orange). Genes represented in mRNA expression sets are indicated with a check to the right of the gene names.

Microarray analysis

Mitochondria neighborhood index ($N_R$): number of mitochondrial genes among the $R$ most similar genes in expression pattern.

Distance between expression levels measured by the Euclidean distance.

Public data were used.
Validation of $N_R$

**EXAMPLE:** $N_{10} = 5$ because there are five mitochondrial genes within the query's 10 nearest-neighboring genes.

Distribution of $N_{100}$ values. The blue histogram shows the distribution of $N_{100}$ for all genes, and the red histogram plots $N_{100}$ for known mitochondrial genes. *, the histogram bin containing LRPPRC.

Combining data

Among the candidate genes, LRPPRC had a remarkably high $N_R$.

Different peptides from the LRPPRC gene were identified in the mitochondrial fraction; no other candidate gene could explain the observed protein pattern.
Representative tandem mass spectrum showing y-ion and b-ion series along with the deduced peptide sequence. (C) The predicted LRPPRC amino acid sequence with high-scoring peptides, identified by organelle proteomics, marked in red.

Identifying the mutation

The gene was initially sequenced in two patients, a parent and an unrelated control.

A single mutation was identified in all patients and in no control, resulting in a missense mutation.

A deletion was found in an additional single patient. This patient was doubly heterozygous for both mutations.
Fig. 5. Mutations identified in LRPPRC. LRPPRC has 38 exons (blue) predicted to encode a 1,394-aa protein. The amino acid sequence corresponding to exons 9 and 35 are shown as well as the aligned sequences from mouse, rat, and Fugu. The exon 9 missense mutation, A354V, and the exon 35 truncation, C1277STOP, are shown in red. Conserved residues are shaded in gray. *, a stop codon.

Can microarray data be used to refine gene positions?

Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study
Pérez-Enciso et al. (2003) Genetics, accepted

Expression data could be used to improve QTL mapping if the following two conditions were met:

1. Some of the gene expression levels must be under (at least partial) genetic control of the QTL
2. Some of these heritable gene expression levels must be related to the trait.

Otherwise, accommodating expression data in a statistical model would reduce power of tests.
Underlying genetic model

\[ P(y_i = 1 \mid h_i) = \frac{\exp(h_i)}{1 + \exp(h_i)} \]

where

\[ h_i = \omega' x_i \]

The QTL shifts the expected value of \( h \) (affects simultaneously several expression levels)

How can we simulate realistic data?
Unusual simulation procedure

1. Specify a subset of parameters ($\theta_1$)
2. Simulate disease phenotypes ($y_2$) and rest of parameters ($\theta_2$) given expression data ($y_1$) and $\theta_1$

$$p(y_2, \theta_2 | y_1, \theta_1)$$

- incidence, allelic frequencies
- affected / non affected
- $\omega$ vector
- microarray data

The procedure

1. Characterize $\omega$
2. Simulate disease status (Binom.)
3. Determine QTL parameters
4. Sample QTL genotype
5. Sample surrounding haplotype

Real microarray data

$y=0$          $y=1$

Haplotype simulation
1. Choosing weights to expression levels

Most of elements in $\omega$ will be zero

$n_g$ mRNAs were chosen among those with no missing values

- **'Diffuse'** scenario: mRNAs with $\omega \neq 0$ chosen independently at random
- **'Clustered'** scenario: first mRNA at random, successive chosen with a probability that was proportional to the correlation with the first mRNA
- **'Uniform'** scenario: weights $\omega$ chosen from a uniform (-1, 1).
- **'Exponential'** scenario: weights $\omega$ chosen from an exponential $\mu=1$.

Weights were found by trial and error, setting the restriction $E(y)=0.50\pm0.05$, to mimic a case/control study.

2. Generating disease status

For each indiv.,

$$P(y_i = 1 | h_i) = \frac{\exp(h_i)}{1 + \exp(h_i)}$$

Binomial sampling
3/4. Generating QTL parameters and genotypes

Diallelic QTL

\[ f(h | g) = N(\mu_g, \sigma^2) \]

Given \( a = (\mu_{g=AA} - \mu_{g=BB})/2\sigma \) and \( \sigma \):

\[ P(g_k | h_i) = \frac{P(g_k) f(h_i | g_k)}{\sum_j P(g_j) f(h_i | g_j)} \]

The within genotype variance was obtained solving iteratively from:

\[ \text{Var}(h) = E_g [\text{Var}(h|g)] + \text{Var}_g [E(h|g)] \]

5. Generating the haplotype

10 Nearby SNPs were generated assuming that a founder haplotype carrying the mutant QTL allele appeared 500 generations ago using an exponential growth model.

Minor SNP allele = 0.3.
Data used

Sorlie et al. (2001) PNAS 98:10869-10874
http://genome-www5.stanford.edu/MicroArray/SMD/
85 breast cancer samples
456 mRNA clones (their 'intrinscis set')
Log2 ratios between the sample and a control are reported.
71 mRNAs did not have any missing record, and were thus eligible to be in h.

Parameters used

\[ n_g = 1, 5, 10, 20 \]
\[ a = 0.5, 1, \text{and} 1.5 \text{ SD} \]
QTL genotype frequencies:
\[ 0.5/0.5 \text{ & } 0.25/0.50/0.25 \]
Scenarios: D/U, D/E, C/U, C/E
500 simulations per case
Analysis strategy

• No \( \mu \)array data: ANOVA on phenotypes and markers as classifying variable.

• \( \mu \)array data used: ANOVA on estimated liability and markers as classifying variable. Liability estimated using Partial Least Squares (PLS) logistic regression.

Logistic regression with PLS (Esposito-Vinci & Tenenhaus, 2001)

For each variable \( j = 1, 2, \ldots, q \) compute its significance in a logistic regression, each variable in turn using the model \( P(y_i = 1) = \frac{\exp(b_0 + \beta_j x_{ij})}{1 + \exp(b_0 + \beta_j x_{ij})} \).

Select those variables that are significant; The first ‘supergene’ is defined, for each \( i \)-th individual, as \( t_{i1} = w_1' x_i \), with \( w_{ij} = \beta_{ij} / C_1 \sum_{\ell \in \mathcal{R}_1} \beta_{ij}^2 \).

The regression coefficient \( b_1 \) is obtained from fitting \( P(y_i = 1) = \frac{\exp(b_0 + b_1 t_{i1})}{1 + \exp(b_0 + b_1 t_{i1})} \).

The next PLS component is obtained by testing again each of the original \( q \) variables plus the previous ‘supergene’ \( P(y_i = 1) = \frac{\exp(b_0 + b_1 t_{i1} + \beta_{ij} x_{ij})}{1 + \exp(b_0 + b_1 t_{i1} + \beta_{ij} x_{ij})} \), \( j = 1, 2, \ldots, q \). Once it is determined the new set of significant variables, the second ‘supergene’ is obtained from \( t_{i2} = w_2' x_i \), with \( w_{ij} = \beta_{ij} / C_2 \sum_{\ell \in \mathcal{R}_2} \beta_{ij}^2 \).
# of significant mRNAs

% of significant mRNAs that are causal

![Graphs showing the relationship between clone number and significant mRNAs under different conditions.](image_url)
LD profile

Variability in LD profiles

\( \hat{h} \)

individual mRNA components
Main conclusions

1) The usefulness of microarray data for gene mapping increases when both the number of mRNA levels in the underlying liability and the QTL effect decrease, and when genes are coexpressed.

2) The correlation between estimated and true liability is large.

3) It is unlikely that mRNA clones identified as significant with PLS are the true responsible mRNAs, especially as the number of clones in the liability increases.

4) The number of significant mRNA levels increases critically if mRNAs are co-expressed in a cluster; however, the proportion of true causal mRNAs within the significant ones is similar to that in a no co-expression scenario.

5) Data reduction is needed to smooth out the variability encountered in expression levels when these are analyzed individually.

Literature

Nature Genetics
december 2002 & january 1999
special issues


