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 successsss !

Impressive, extremely powerful technology

Potentially very useful in Human genetics

Many data publicly available !

Some webpages

Yandell's references
http://www.cs.wisc.edu/~yuanj/gene/array.html

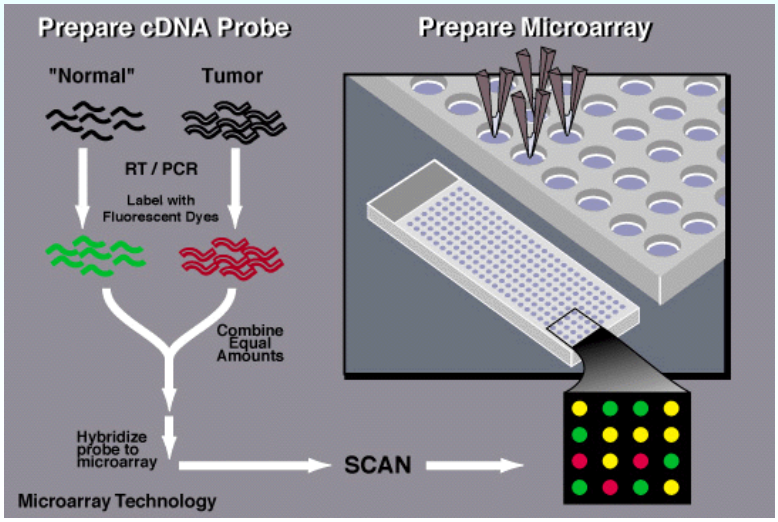
Stanford Microarray Database
http://genome.www5.stanford.edu/MicroArray/SMD/

Lymphoma/Leukemia Molecular Profiling Project
http://llmpp.nih.gov/lymphoma/index.shtml

Gene Expression Omnibus
NIH / NCBI

EISEN's lab
Treeview

cDNA microarray principle



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

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\ \sqrt{\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

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 = $r_{i,j}$ (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

Example

Molecular portraits of human breast tumours

CHARLES M. PEROU, THERESE SORLIE, MICHAEL B. EISEN, MATT VAN DE RIJN, STEFANIE S. JEFFREY, CHRISTIAN A. REES, JONATHAN R. POLLACK, DOUGLAS T. ROSS, HILDE JOHNSEN, LARS A. AKSLEN, OYSTEIN FLUGE, ALEXANDER PERGAMENSCHIKOV, CHERYL WILLIAMS, SHIRLEY X. ZHU, PER E. LONNING, ANNE-LISE BORRESEN-DALE, PATRICK O. BROWN & DAVID BOTSTEIN

* Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA
‡ Department of Genetics, The Norwegian Radium Hospital, N-0310 Montebello Oslo, Norway
§ Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA
Department of Surgery, Stanford University School of Medicine, Stanford, California 94305 , USA
¶ Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, USA
# Department of Pathology, The Gade Institute, Haukeland University Hospital, N-5021 Bergen, Norway
Department of Molecular Biology, University of Bergen, N-5020 Bergen, Norway
** Department of Oncology, Haukeland University Hospital, N-5021 Bergen, Norway
†† Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA
† These authors contributed equally to this work

Nature **406,** 747-752 (17 August 2000)

Perou et al. 2000

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.**

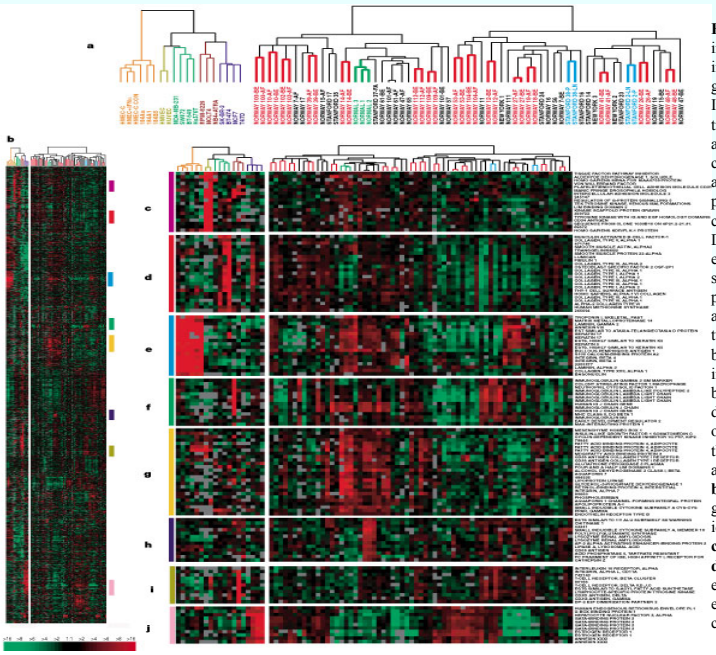


Figure 1 Variation in expression of 1,753 genes in 84 experimental samples. Data are presented in a matrix format: each row represents a single gene, and each column an experimental sample. In each sample, the ratio of the abundance of transcripts of each gene to the median abundance of the gene's transcript 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 mesynchemal-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:

$\mathbf{X} \equiv$ {cDNA measurements}

$\mathbf{y} \equiv$ {probability of phenotype, say disease status qualitative or quantitative}

$\mathbf{y} = f(\mathbf{X}, \theta)$?

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**.

$\mathbf{t}_h = \mathbf{X}\,\mathbf{w}_h^*$ (orthogonal **X**-components)

$\mathbf{u}_h = \mathbf{Y}\,\mathbf{c}_h$ (orthogonal **Y**-components)

such that max. Cov($\mathbf{t}_h$, $\mathbf{u}_h$).

There may be many more variables than observations

In PLS-DA the **Y** are binary clasificatory variables

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

$$\mathbf{y_k} = \Sigma_{h=1,k}\, \mathbf{X}\, \mathbf{w}^*_h\, c_h + \mathbf{e} = \mathbf{X}\, \mathbf{W}^*\, \mathbf{c} + \mathbf{e}$$

$\mathbf{w_h}^*$ = 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

Pérez-Enciso & Tenenhaus 2003

Perou et al. data reanalyzed

84 tissues

(11 tumoral cell cultures, 65 breast cancer and 3 normal breast samples)

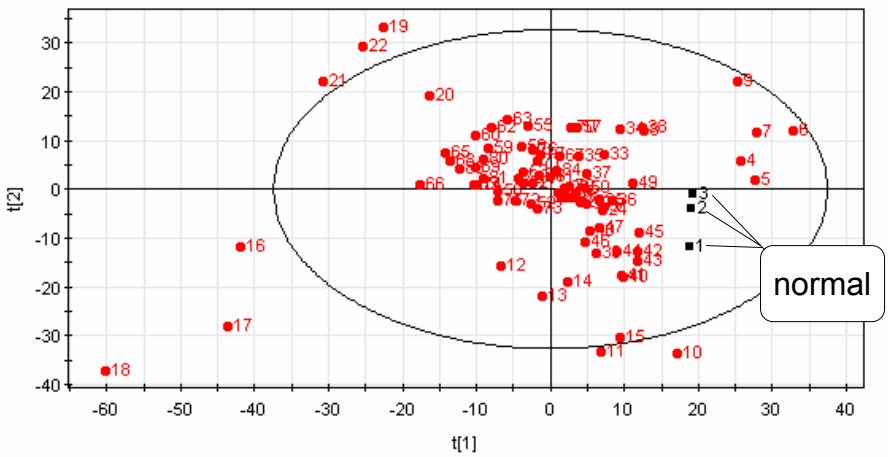
1753 cDNA clones

Discriminant analyses

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

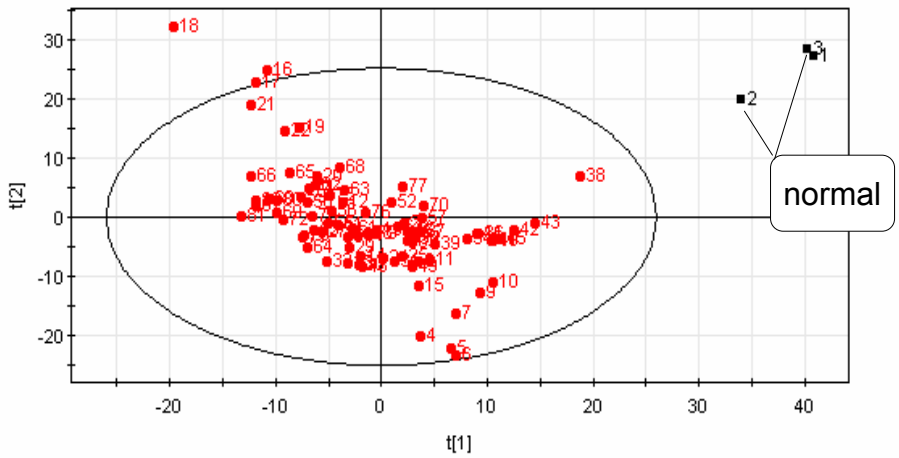
disease status: principal components

81 cancer / 3 normal, all 1753 variables



disease status:
PLS-DA

81 cancer / 3 normal,
all variables



But ...

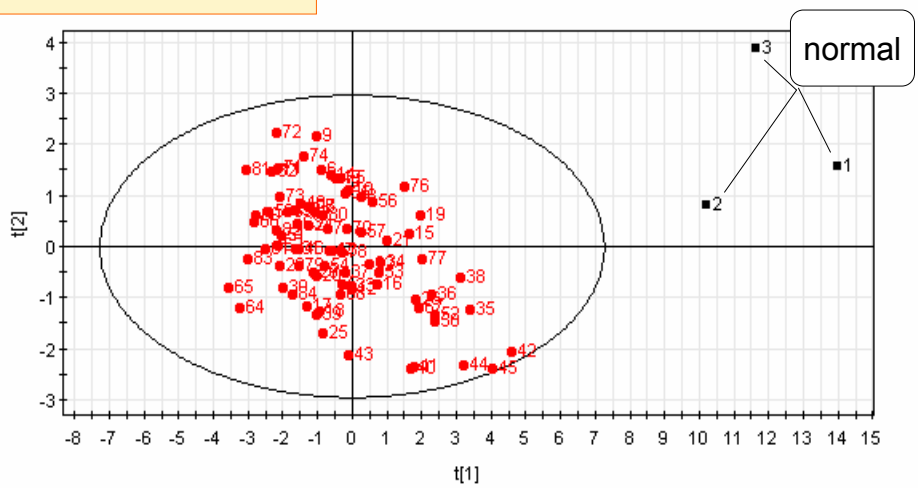
Models of very poor predictive abilities

Subset of variables (cDNA levels) preselected according to its variable importance in prediction (VIP), sort of weighed correlation.

disease status: PLS-DA

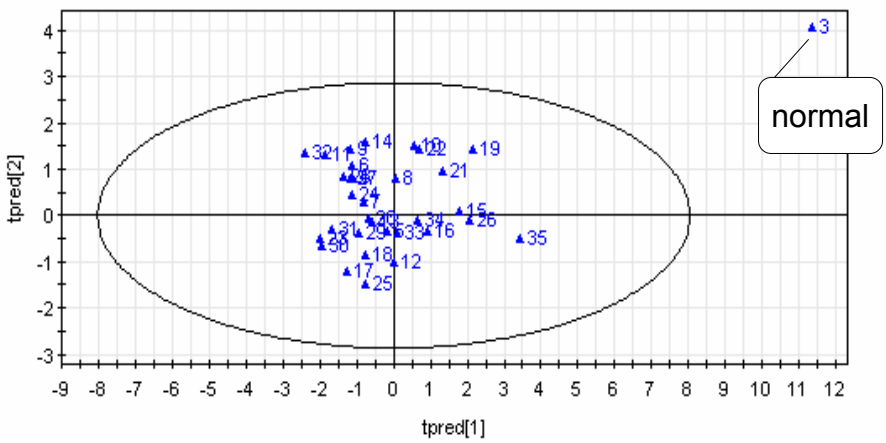
81 cancer / 3 normal,

21 cDNA levels selected



Prediction of disease status: PLS-DA

21 variables

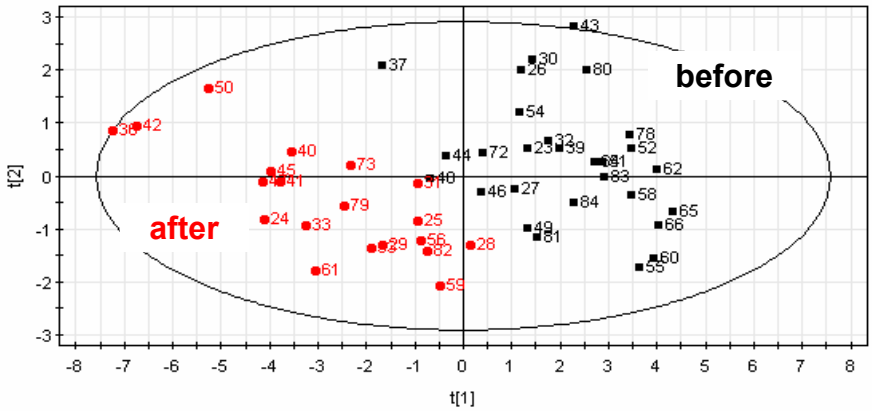


Obs. 1 normal, 4-35 tumor predicted

Before / After chemotherapy PLS-DA

48 observations

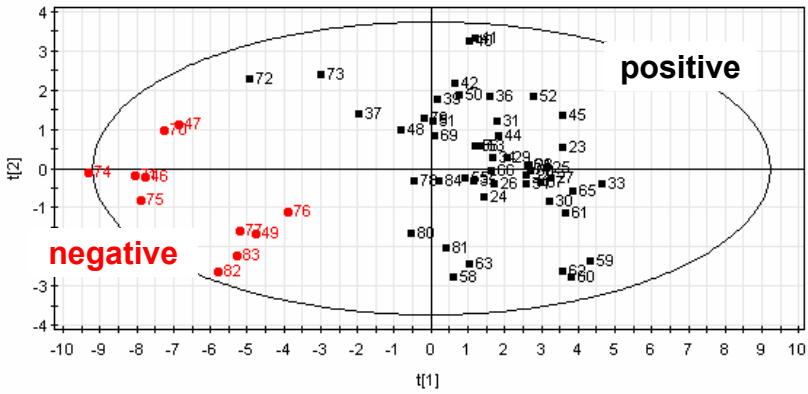
19 cDNA levels



ER positive / negative PLS-DA

60 observations

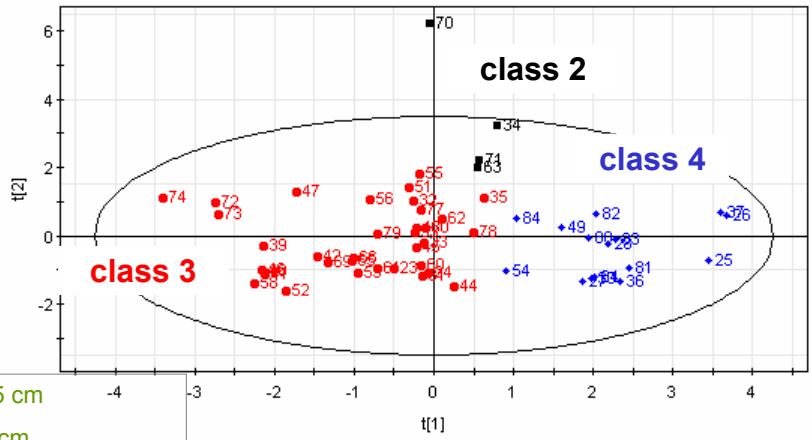
30 cDNA levels



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

VIP	Symbol	Name
1.20	AQP7	Aquaporin 7
1.20	ITGA7	Integrin, alpha-7
1.20	CDK5R1	Cyclin dependent kinase 2
1.13	FOSB	FJB osteosarcoma oncogene homolog B
1.13	COL14A1	Undulin 4
1.11	PFKFB3	6 Phosphofructo-2-kinase
1.06	674	-
1.06	GPD1	Glycerol 3 P dehydorgenase
1.00	LPL	Lipoprotein lipase
1.00	767	-
.97	FOS	FJB osteosarcoma oncogene homolog 2
.96	ADH2	Alcohol dehydrogenase 2
.95	GPD1	Glycerol 3 P dehydrogenase
.94	GPX3	Glutathione peroxidase
.93	CNN1	Calponin 1
.90	FOS	FJB osteosarcoma oncogene homolog
.89	50	-
.88	CDKNC1C	Cyclin dependent kinase
.83	647	-
.83	760	-

cluster g of Perou related to adipocytes in tumoral tissues

altered in cancer

oncogenes

ovarian cancer

Genes involved in chemotherapy status

VIP	Symbol	Name
1.22	RCV1	Recoverin
1.21	FOS	FJB osteosarcoma oncogene homolog [1]
1.15	HBA1	Hemoglobin alpha1
1.08	CTGF	Connective tissue growth factor
1.06	TCEB3	Transcription elongation factor B
1.06	DCT	Dopachrome tautomerase
1.05	FOS	FJB osteosarcoma oncogene homolog [1]
1.04	CTGF	Connective tissue growth factor
1.02	GEM	GTP-binding mitogen-induced t-cell protein
1.00	NR4A1	Nuclear receptor subfamily 4
0.98	CDK5R1	Cyclin dependent kinase [1]
0.98	DPYSL3	Dihydropyrimidinase-like 3
0.94	FY	Blood group-duffy system
0.91	ATF3	Activating transcription factor 3
0.90	CDKN1A	cyclin-dependent kinase
0.86	COPEB	Core promoter element-binding protein
0.85	EGR2	Early growth response 2

ocular tumors chemotherapy changed

oncogene

growth factors, cyclines

transcription factors

Main genes involved in ER status

Symbol	VIP	Name
1.17	GATA3	GATA-binding protein 3 [wg]
1.14	ESR1	Estrogen receptor 1 [wg]
1.12	GATA3	GATA-binding protein 3 [wg]
1.11	PES1	Pescadillo 1
1.08	ITPR3	Inositol 1,4,5-triphosphate receptor, type 3
1.07	GATA3	GATA-binding protein 3 [wg]
1.06	GATA3	GATA-binding protein 3 [wg]
1.00	DSC2	Desmocollin 2
1.00	GRO1	Growth regulated protein precursor
1.00	CCNE1	Cyclin E1
1.00	TFF1	Trefoil factor 1 [w]
0.99	SLC7A8	Solute carrier family 7 [g]
0.98	ORM1	Orosomucoid 1
0.97	PFKP	Phosphofructokinase, platelet type [g]
0.97	LRP8	Low density lipoprotein receptor-related protein8
0.96	HNMT	Histamine n-methyltransferase
0.96	HNF3A	Hepatocyte nuclear factor 3-alpha
0.94	NAT1	N-acetyltransferase 1
0.94	HMG1	High mobility group protein 1 [g]
0.91	PTK7	Tyrosine-protein kinase-like 7 precursor 0.90
	TRIP13	Thyroid hormone receptor interactor 13

well known

Upregulated in cancer, induced by estrogens

also in West

also in Gruvberger

Genes involved in tumor classification

VIP	Symbol	Name
1.35	COL14A1	Undulin [1]
1.21	1244	-
1.08	LOX	Protein-lysine 6-oxidase
1.00	CRIP2	Cysteine-rich intestinal protein 2
.96	767	-
.93	459	-
.93	TFAP2B	Transcription factor AP2-beta
.90	1542	-
.86	ARHB	RAS homolog gene family, member B
.84	1017	-
.79	MRSPSZ7	KIAA protein

altered in cancer

tumor progression

transcription, growth factors

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

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

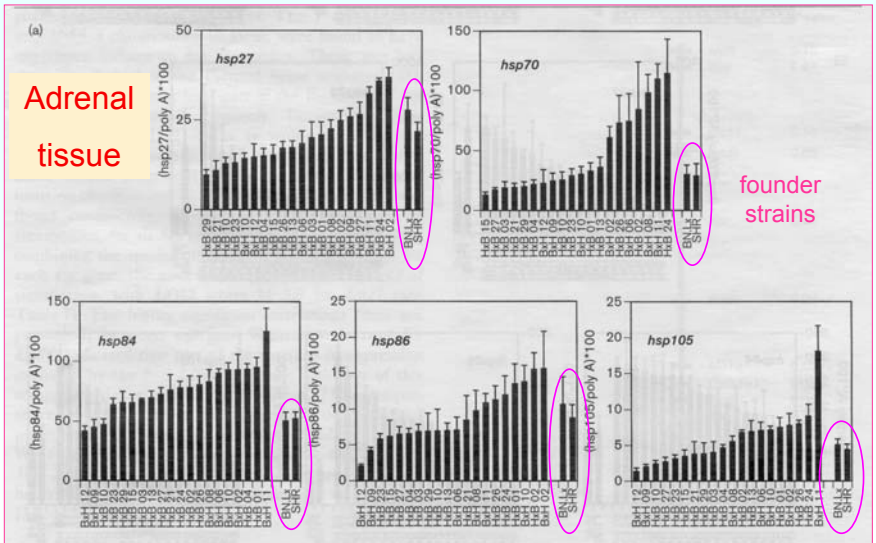
Heat shock proteins (hsp) are highly conserved, they are induced by several stressors, protect other proteins from denaturalization.

HSPs are mediated by heat shock transcription factors (hstf) 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 hsps.
- 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?).



Strain distribution pattern of *hsp* expression in the RIS adrenal (a), heart (b), and kidney (c). H × B, strain originating from female SHR × male BN.*Lx* progenitors. B × H, strain originating from female BN.*Lx* × male SHR progenitors. Each bar represents a strain. The results are the mean ± SEM of four rats per strain.

Dumas et al. 2000

Dumas et al. 2000

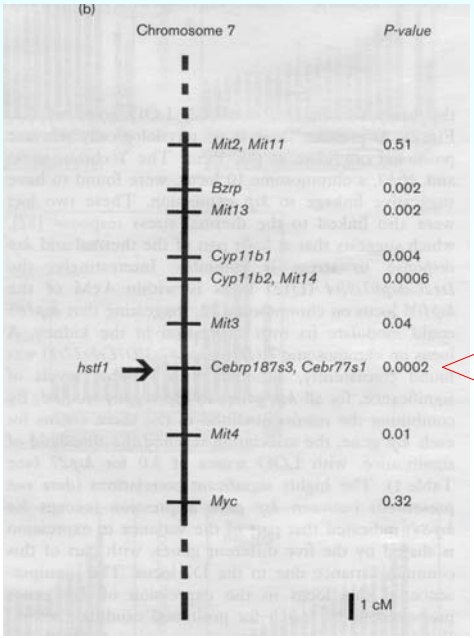
Table 1 **Total genome scan of *hsp* expression in the adrenal, heart and kidney of RIS**

	Marker	Adrenal *r*	Adrenal *P*	Heart *r*	Heart *P*	Kidney *r*	Kidney *P*	*D7* marker, Pooled organs* *r*	*D7* marker, Pooled organs* *P*
hsp27	*D7*	**0.47**	0.04	**0.65**	0.002	0.40	0.08	**0.46**	**0.0002**
	Myh3	–	–	0.50	0.02	–	–	(LOD score 3.0)	
hsp70	*D7*	0.63	**0.003**	0.51	0.02	0.48	0.03	**0.42**	**0.0008**
	Y Chr	–	–	0.35	0.13	**0.60**	**0.005**	(LOD score 2.4)	
	Myh3	–	–	–	–	0.44	0.05		
hsp84	*D7*	0.42	0.07	–	–	0.35	0.14	0.38	0.01
	D9	**0.57**	**0.009**	–	–	–	–	(Adrenal + kidney)	
	D4	–	–	**0.71**	**0.0002**	–	–		
				(LOD score 3.1)					
hsp86	*D7*	**0.58**	**0.007**	0.44	0.06	0.53	0.02	**0.35**	**0.007**
	Y Chr	–	–	–	–	**0.63**	**0.003**		
	Myh3	–	–	0.30	0.20	0.55	0.02		
hsp105	*D7*	0.53	0.02	0.42	0.06	0.4	0.08	0.36	0.005
	Y Chr	–	–	–	–	0.49	0.03		
	Myh3	–	–	0.55	0.01	–	–		
	D12	–	–	–	–	0.52	0.02		

Correlations equal to or higher than 0.3 are displayed. Suggestive ($P < 0.001$) or significant linkages ($P < 0.0003$) appear in bold type. Abbreviations: *D7*, *D7Cebrp187s3/D7Cebr77s1* marker; *D9*, *D9Cebr16C27s2* marker; D4, *D4Mit19* marker; *D12*, *D12Cebrp97s9/s4* marker; *Myh3*, myosin heavy chain (embryonic) gene marker (chromosome 10); *Y Chr*, Y chromosome.

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

Dumas et al. 2000



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. cerevisae 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.

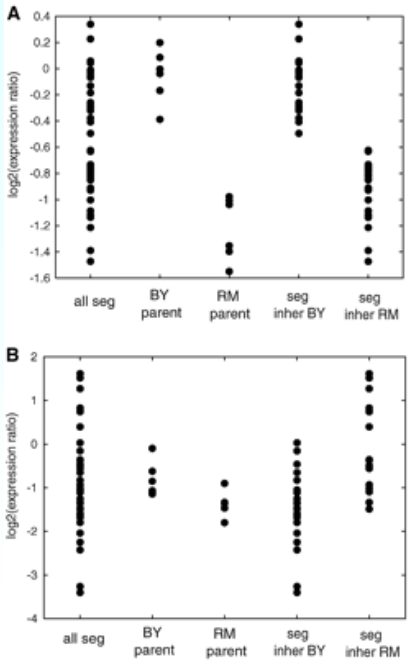
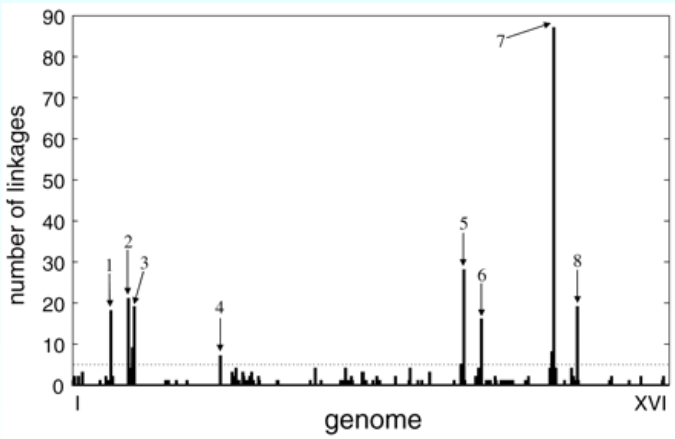


Figure 2

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**.

Figure 3



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).

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

Conclusions

- Most levels affected by several loci
- Many regions in cis
- Small number of alleles trans-acting and affecting many mRNA levels simultaneously

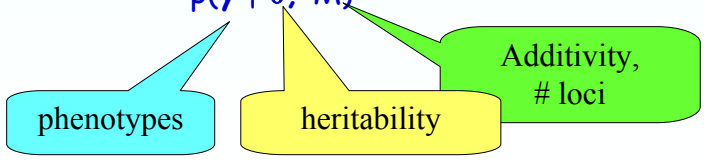
Pérez-Enciso (Genetics, 2004)

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

$p(\mathbf{y} \mid \theta, M)$



Simulación no tradicional

1. Given real expression data ($\mathbf{y}_1$)
2. Genotype simulation ($\mathbf{y}_2$)
3. Assignment of genotypes randomly to phenotypes
4. Data analysis

$$p(\mathbf{y}_2, \theta_2 \mid \mathbf{y}_1)$$

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

Real 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, ρ = 4Ne r = 1000
2. *Gene dropping* 1000 gens., Ne = 1500, 1 Morgan
3. ~ 25,000 SNPs in t = 1000
4. Only SNPs frec > 0.10 analyzed (~ 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)

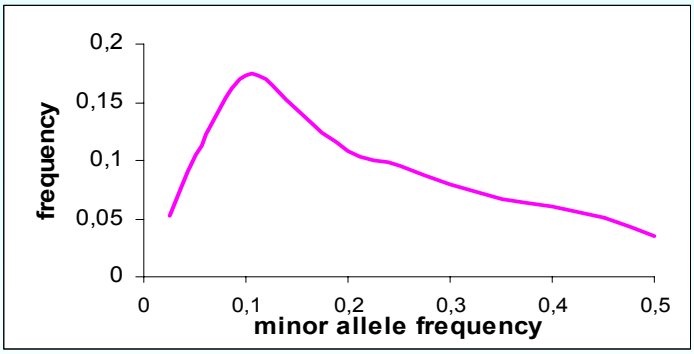
$$L_{jk} = \prod_{i=1}^{N} \phi(y_{ij} - \mu_{ijk}, \sigma^2_{jk})$$

i-th indiv., j-th mRNA level

mean genotype ijk

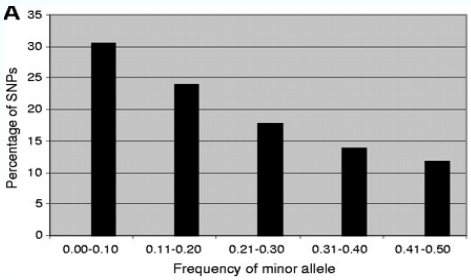
Residual variance jk (constant $\forall$ genotype)

Significance if $P < 10^{-6}$

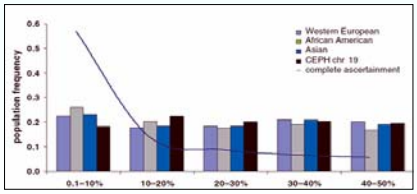


Simulation plausibility

here

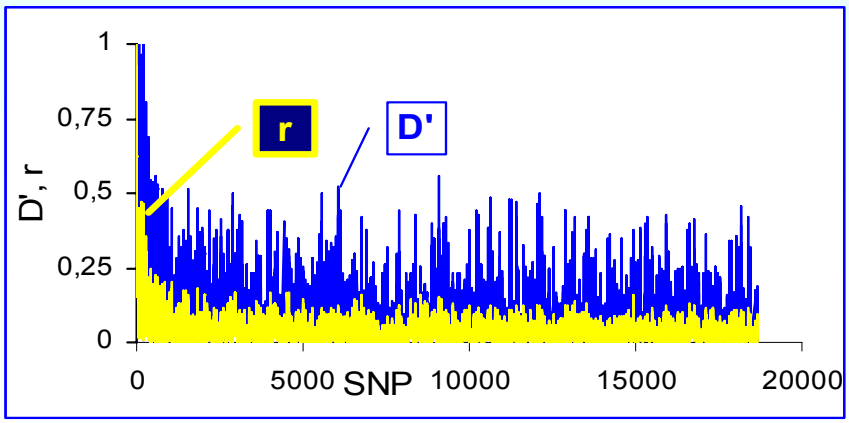


Patil (2001)

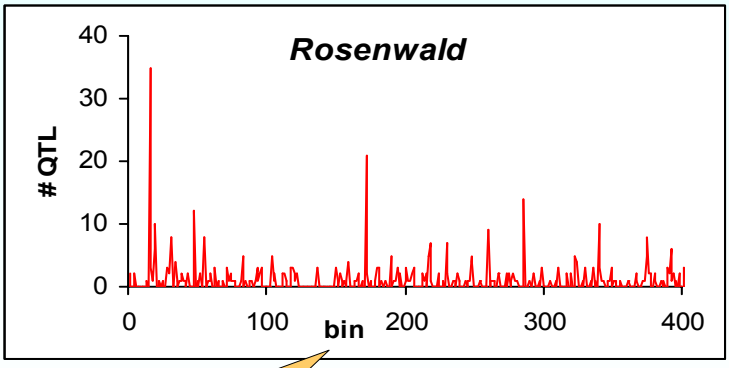


Phillips (2003)

Disequilibrium pattern

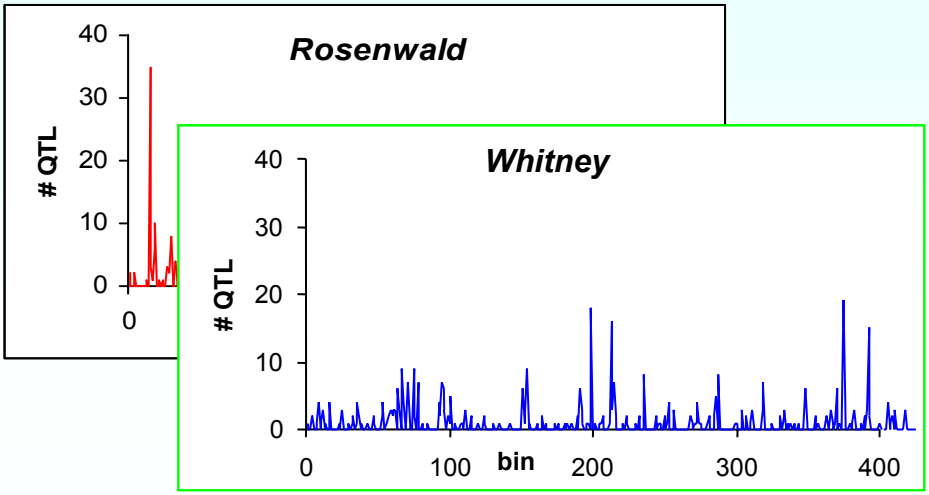


Results (1): *hotspots*

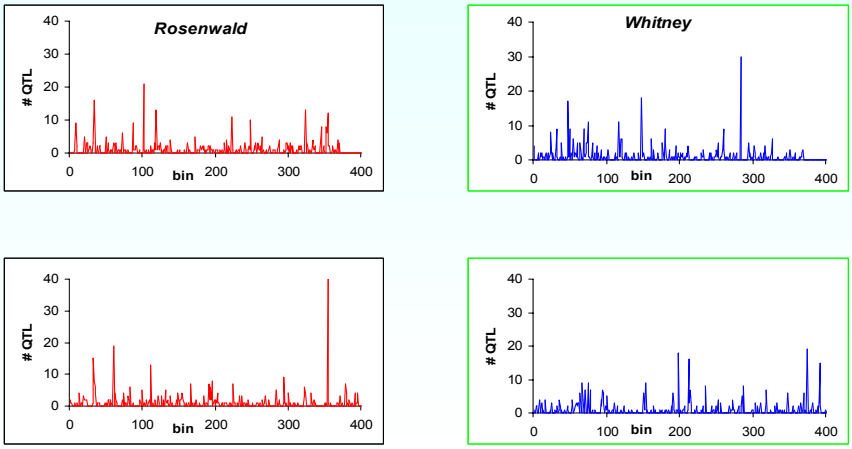


a bin is made of 50 consecutive SNPs

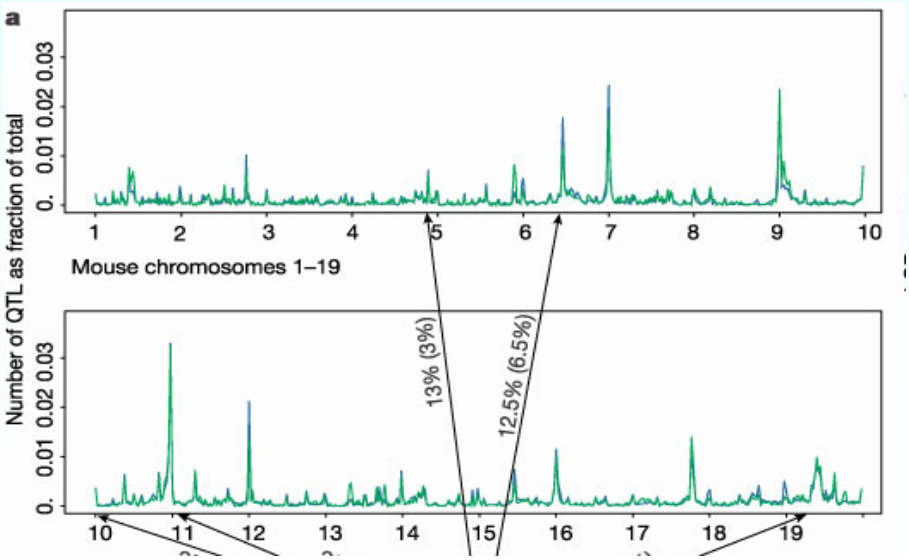
Results (1): *hotspots*



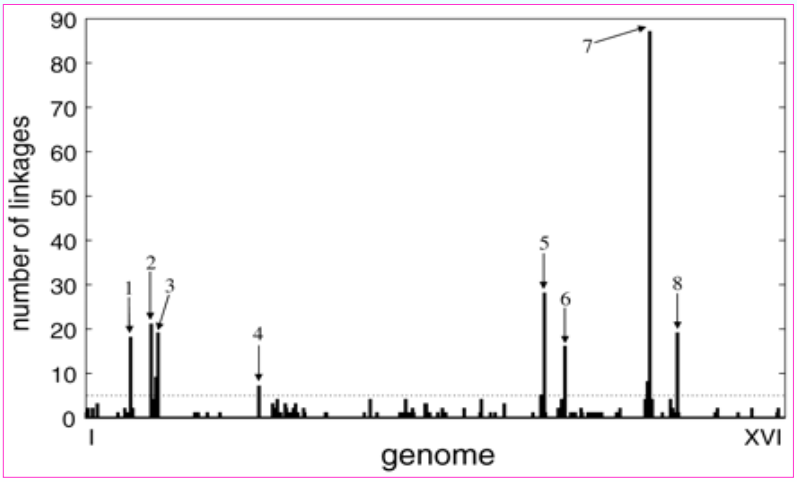
Results (1): *hotspots*



Schadt et al. 2003 (mouse)



Brem et al. 2002 (yeast)

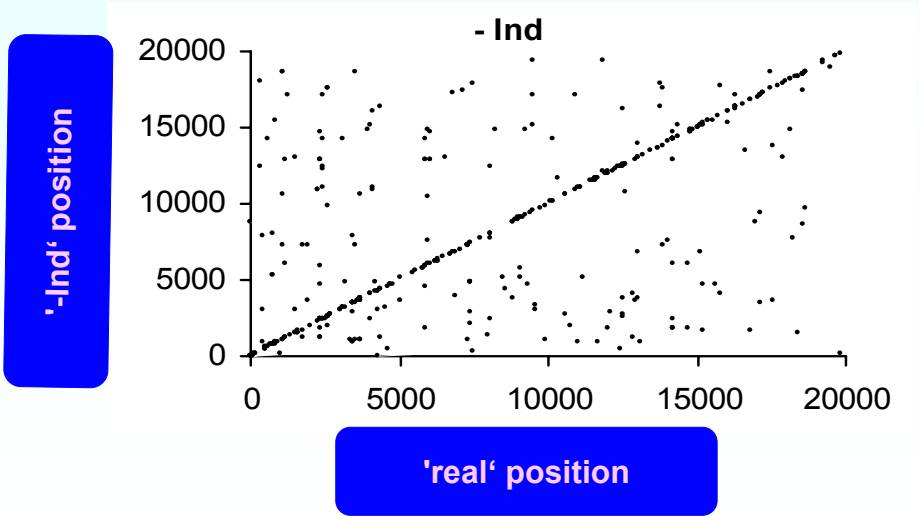


Results (2): estimates' reliability

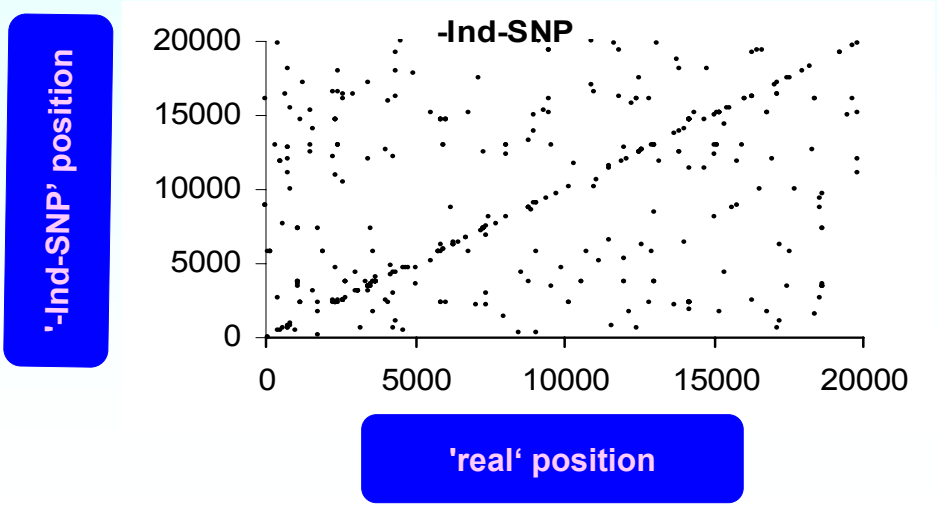
Noise added to the system by:

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)

Dataset	R-Ind	R-Ind-SNP
% $P<10^{-6}$	67	40
$\delta=0$(%)	70	23
$\delta<10$(%)	73	50
$\delta<100$(%)	77	60
$\delta>10^4$(%)	5	6

Power

Distance between 'true' and 'estimated' estimates (in SNPs)

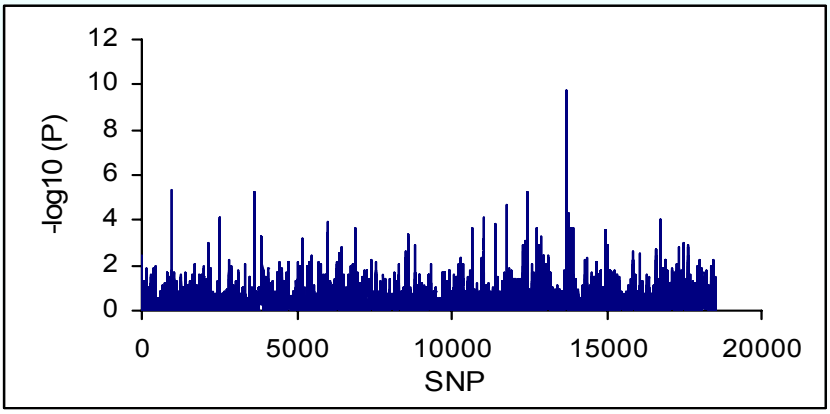
Adding noise (Brem, yeast)

Dataset	B-Ind	B-Ind-SNP	BR-Ind
% $P<10^{-6}$	72	42	73
$\delta=0$(%)	76	17	58
$\delta<10$(%)	94	70	59
$\delta<100$(%)	100	99	64
$\delta>10^4$(%)	0	0.1	6

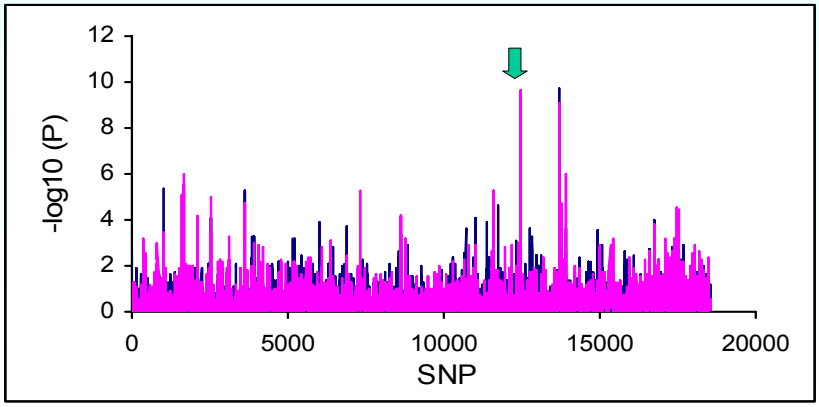
Power

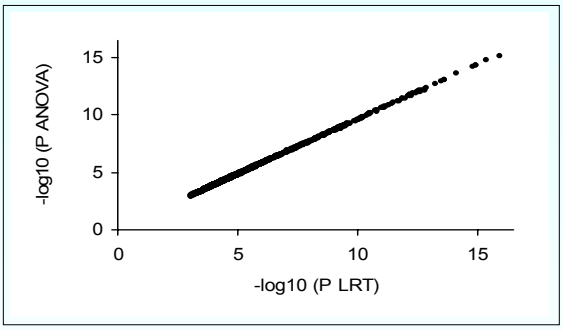
Distance between 'true' and 'estimated' estimates (in SNPs)

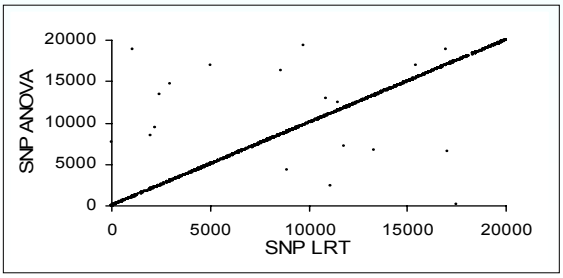
How an association profile looks like?



How an association profile looks like?







Results ML vs. ANOVA

Conclusions

- *QTL hotspots* should be interpreted with caution
- LD/associatio 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

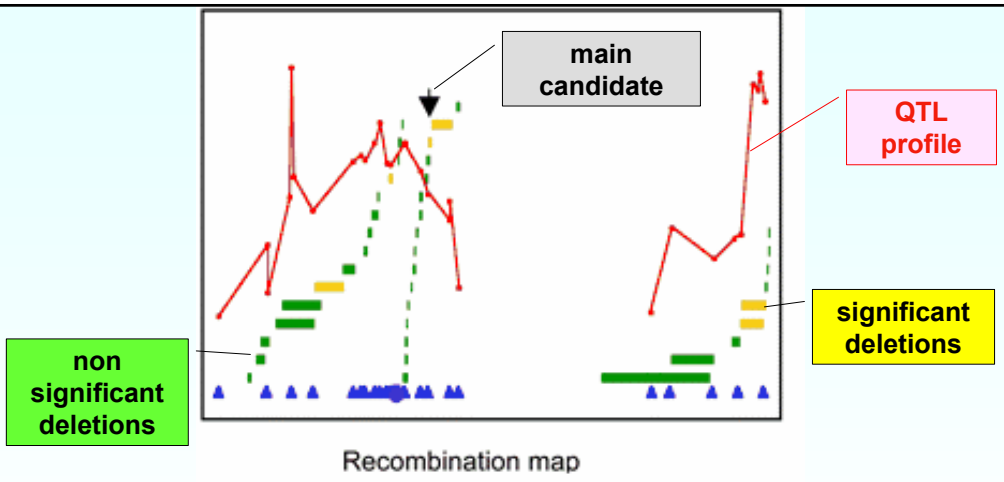
Wayne & McIntyre (2002)
Combining mapping and arraying: An approach to candidate gene identification

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

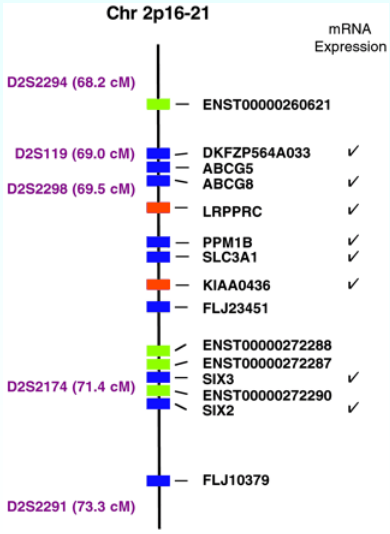
Mootha et al. (2003):

Identification of a gene causing human cytochrome *c* oxidase deficiency by integrative genomics

Leigh syndrome (French-Canadian type) is relatively comon 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.



Chr 2p16-21 region

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.

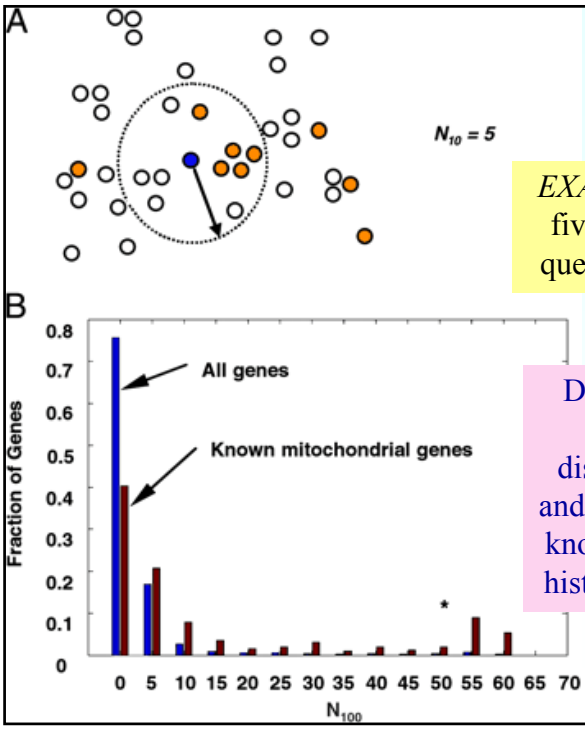
Microarray analysis

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



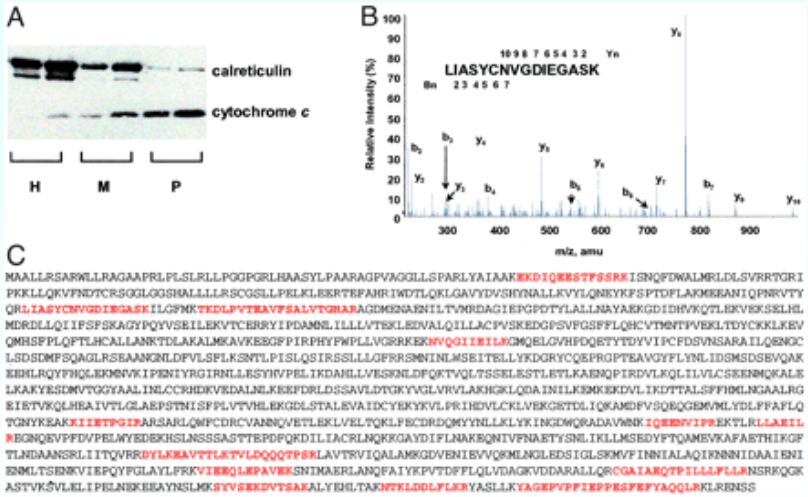
EXAMPLE: $N_{10} = 5$ because there are five mitochondrial genes within the query's 10 nearest-neighboring genes

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 mithocondrian fraction; no other candidate gene could explain the observed protein pattern.



(***B***) 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.

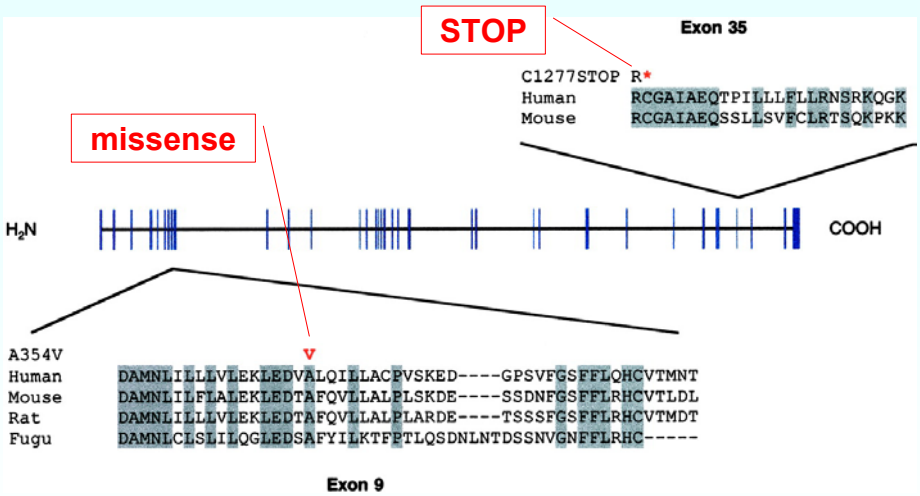


Fig. 5. Mutations identified in *LRPPRC*. *LRPPRC* has 38 exons (blue) predicted to encode a 1,394-aa protein. The amino acid sequence corresponding to exons 9 and 35 are shown as well as the aligned sequences from mouse, rat, and *Fugu*. The exon 9 missense mutation, A354V, and the exon 35 truncation, 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

logistic

$$P(y_i = 1 \mid h_i) = \exp(h_i) / [1 + \exp(h_i)]$$

underlying liability

$$h_i = \omega' \mathbf{x}_i$$

expression data indiv. i

unknown weights

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 ($\boldsymbol{\theta}_1$)
2. Simulate disease phenotypes ($\mathbf{y}_2$) and rest of parameters ($\boldsymbol{\theta}_2$) given expression data ($\mathbf{y}_1$) and $\boldsymbol{\theta}_1$

$$p(\mathbf{y}_2, \boldsymbol{\theta}_2 \mid \mathbf{y}_1, \boldsymbol{\theta}_1)$$

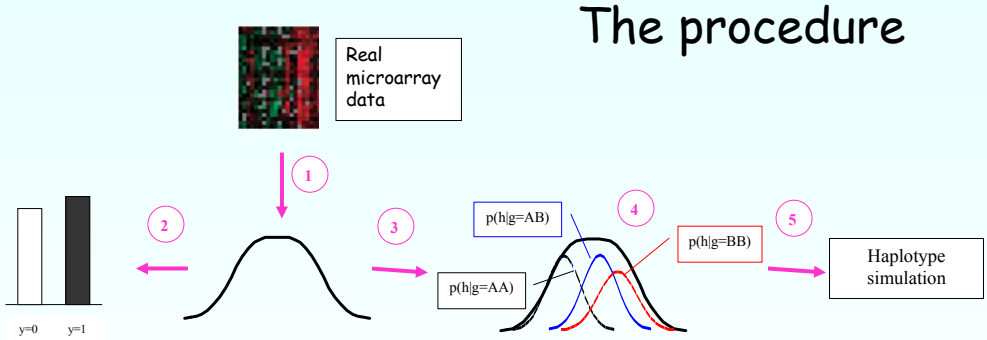
affected / non affected

ω vector

μarray data

incidence, allelic frequencies

The procedure



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

1. Choosing weights to expression levels

Most of elements in ω will be zero

$\mathbf{n_g}$ 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 ω chosen from a uniform (-1, 1).

'Exponential' scenario: weights ω chosen from an exponential μ=1.

Weights were found by trial and error, setting the restriction E(y)=0.50±0.05, to mimic a case/control study.

2. Generating disease status

For each indiv.,

$$P(y_i = 1 \mid h_i) = \exp(h_i) / [1 + \exp(h_i)]$$

Binomial sampling

3/4. Generating QTL parameters and genotypes

Diallelic QTL

$$f(h \mid g) = N(\mu_g, \sigma^2)$$

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

$$P(g_k \mid h_i) = P(g_k)\, f(h_i \mid g_k) \,/\, \Sigma_j\, [P(g_j)\, f(h_i \mid g_j)]$$

used for sampling individual genotypes

The within genotype variance was obtained solving iteratively from:

$$Var(h) = E_g\, [Var(h|g)] + Var_g\, [E(h|g)]$$

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 'intrincsic 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, and 1.5 SD

QTL genotype frequencies:

0.5/0/0.5 & 0.25/0.50/0.25

Scenarios: D/U, D/E, C/U, C/E

500 simulations per case

Analysis strategy

- No μarray data: ANOVA on phenotypes and markers as classifying variable.

- μarray 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 $j = 1, 2,..., q$ compute its significance in a logistic regression, each variable in turn using the model $P(y_i = 1) = \exp(b_0 + \beta_{1j} x_{ij}) / [1 + \exp(b_0 + \beta_{1j} x_{ij})]$,

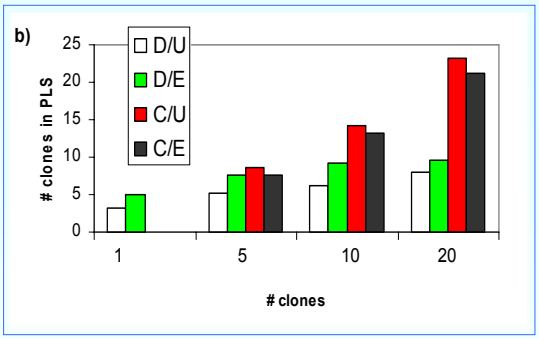
Select those variables that are significant; The first 'supergene' is defined, for each i-th individual, as $t_{1i} = \mathbf{w_1}' \mathbf{x_i}$, with $w_{1j} = \beta_{1j} / C_1$

$$\sqrt{\sum_{j \in \Re^1} \beta_{1j}^2}$$

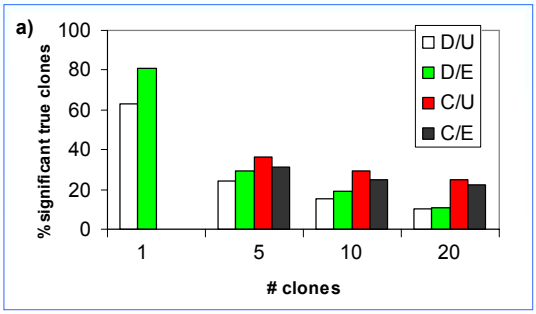
The regression coefficient b_1 is obtained from fitting $P(y_i = 1) = \exp(b_0 + b_1 t_{1i}) / [1 + \exp(b_0 + b_1 t_{1i})]$.

The next PLS component is obtained by testing again each of the original q variables plus the previous 'surpergene' $P(y = 1) = \exp(b_0 + b_1 t_1 + \beta_{2j} x_j) / [1 + \exp(b_0 + b_1 t_1 + \beta_{2j} x_j)]$, $j = 1, 2,..., q$. Once it is determined the new set of significant variables, the second 'supergene' is obtained from $t_{2i} = \mathbf{w_2}' \mathbf{x_i}$, with $w_{2j} = \beta_{2j} / C_2$

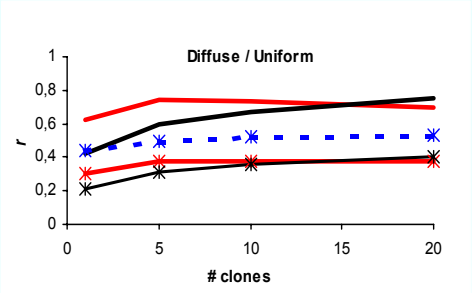
$$\sqrt{\sum_{j \in \Re^2} \beta_{2j}^2}$$

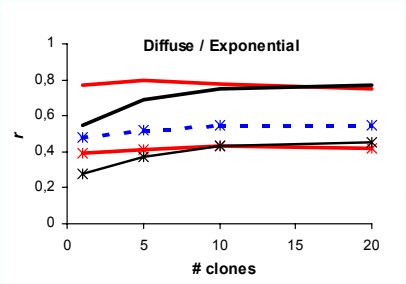


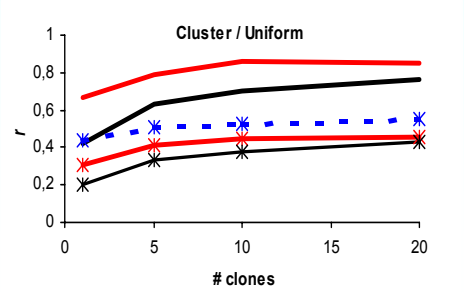
# of significant mRNAs

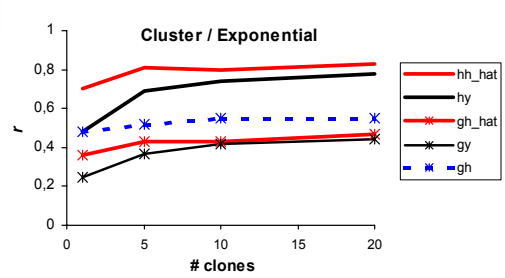


% of significant mRNAs that are causal

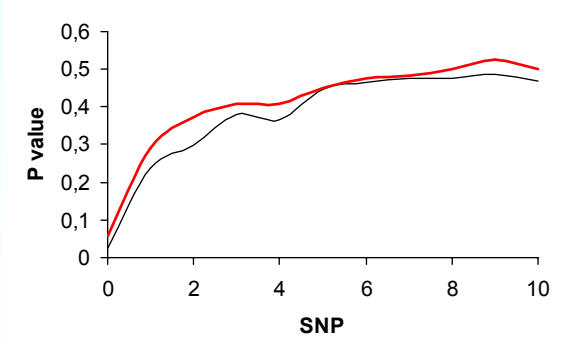




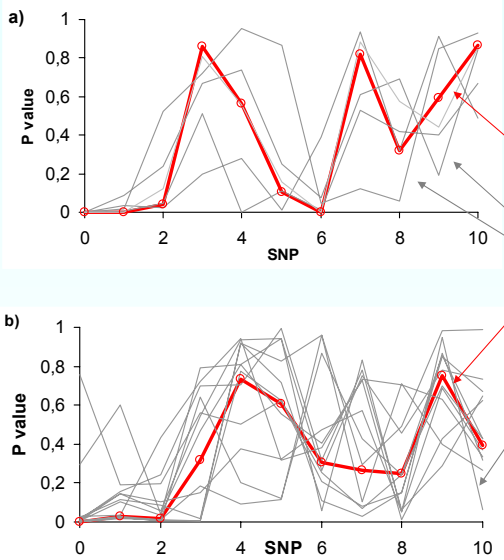




LD profile



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

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296,** 752-755.

Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21,** 33-37.

Dumas, P., Sun, Y., Corbeil, G., Tremblay, S., Pausova, Z., Kren, V., Krenova, D., Pravenec, M., Hamet, P., & Tremblay, J. (2000). Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains. *J Hypertens* **18,** 545-551.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95,** 14863-14868.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning, Springer Verlag, New York.

Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, Lander ES (2003) Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci U S A 100: 605-610

Nguyen DV, Rocke DM (2002) Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18:39-50

Pérez-Enciso M, Tenenhaus M (2003) Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Hum Genet 112: 581-92

Pérez-Enciso M., Toro MA, Tenenhaus M, Gianola D (2003). Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. Genetics 164:1597-1606

Pérez-Enciso, M. 2004. In silico assessment of genetic variation for the transcriptome in outbred populations. Genetics, in press.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747-752.

Rosenwald, A. *et al.* The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *N Engl J Med* 346, 1937-1947 (2002).

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98:10869-74

Tenenhaus, M. (1998). La régression PLS, Editions Technip, Paris.

Wayne ML, McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. Proc Natl Acad Sci U S A 99: 14903-6

Whitney,A.R. *et al.* Individuality and variation in gene expression patterns in human blood. *PNAS* 100, 1896-1901 (2003).