Mapping the Genome: Mathematical and computational cha llenges

A. Korol
korol@research.haifa.ac.il
Institute of Evolution, University of Haifa, Israel

Berlin, Weierstrass Institute, October 2006

Outline

Biological Background

- **Organization of genetic material (DNA)**
- **Recombination**

Building multilocus maps

- **Reduction to *TSP***
- **Consensus mapping as "synchronous" *TSP***

Mapping quantitative traits

- **Univariate and multivariate formulations**
- **Using novel chip technologies**
- **The challenge of high dimensionality**

Some background

Life, Cell, and the Genome

Life primary principles:
- Metabolism (catalyzed)
- Reproduction (inheritance)
- Evolution

Genome as evolving “program”

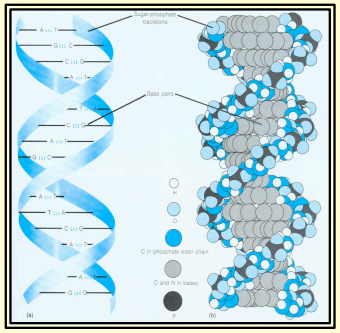
Genetic material: Molecular organization

Double helix DNA (Watson & Crick, 1953):

(a) location - mainly in chromosomes (nucleus)

(b) structure - a long double helix molecule

(c) coding elements - cytosine (C) & thymine (T)
adenine (A) & guanine (G)

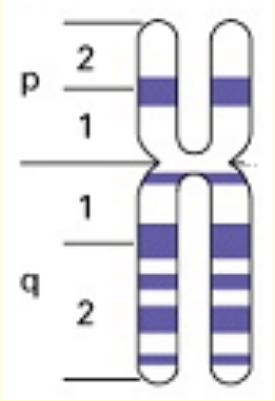


pairs: A ↔ T and G ↔ C

***Complementary* pairing**

Genes encoding for proteins and other molecules are using this 4-letter alphabet across life

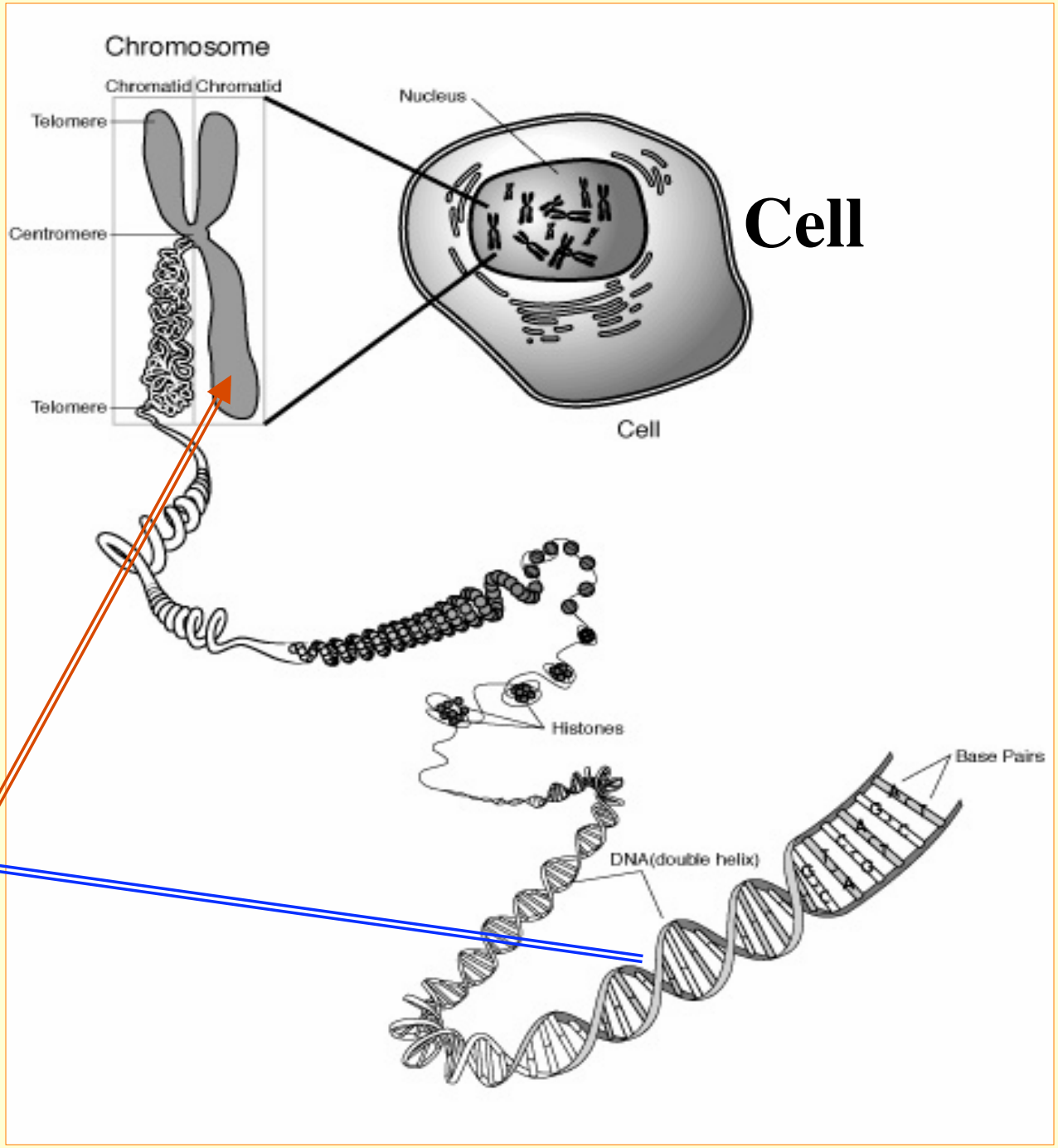
Organization of genetic material in eukaryotes



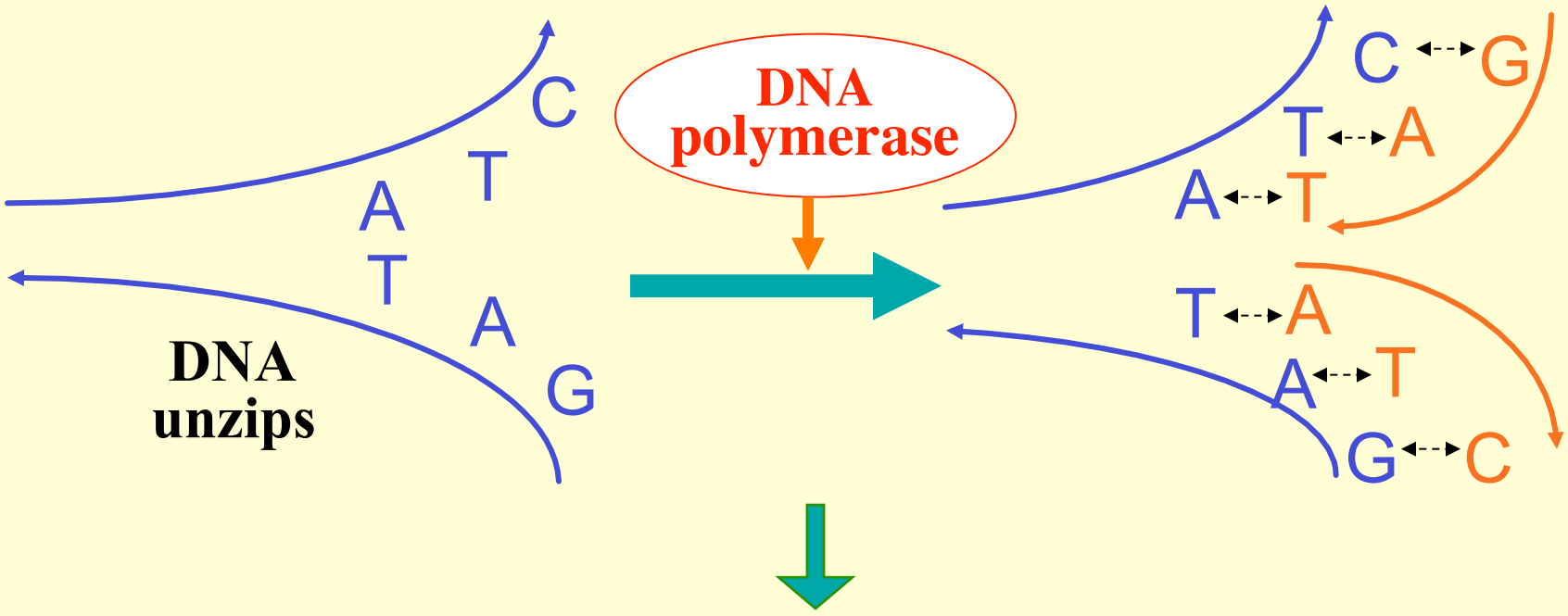
Several levels of DNA folding

from ***cm***

to ***micron***

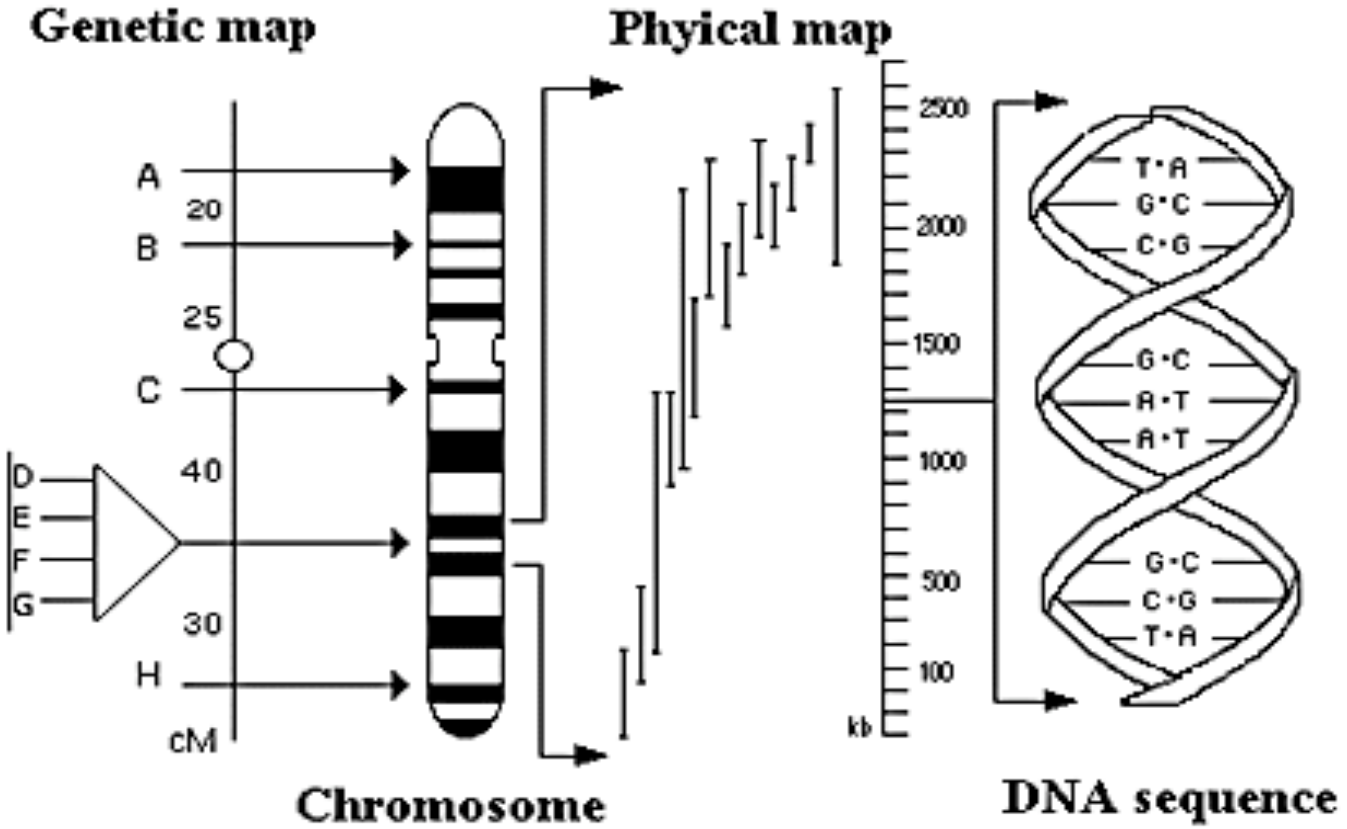


DNA replication: Forming DNA for new cells



Semi-conservative replication: 2 double-stranded DNA molecules for 2 new cells

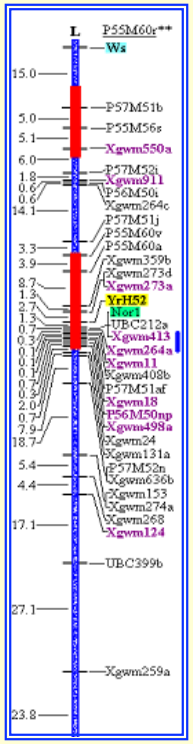
Structural genomics includes: genetic mapping, physical mapping and sequencing of entire genomes



Results of consecutive steps of structural genomics

Genome mapping (genetic and physical mapping)

Genome mapping is a major part of genome projects and precondition for most of the genomic applications

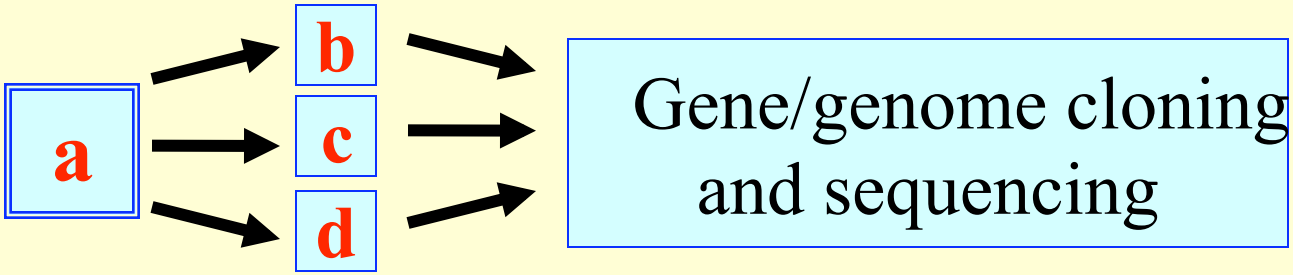


a. Positioning of DNA markers → genetic maps

b. Positioning DNA pieces → physical maps

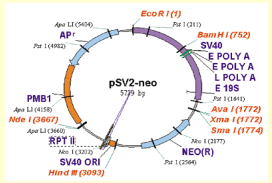
c. Locating *Mendelian* genes relative to markers

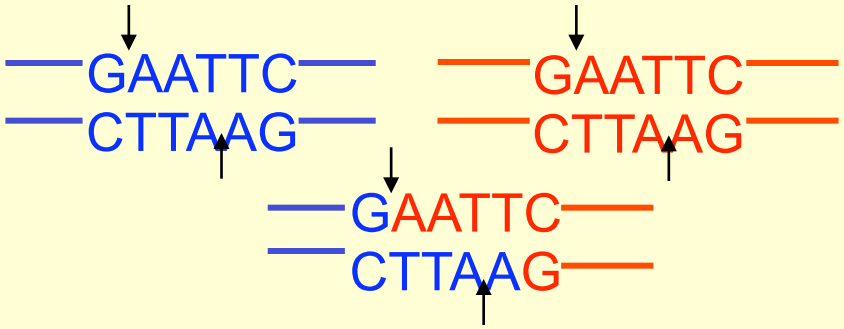
d. Mapping quantitative trait loci (QTL maps)



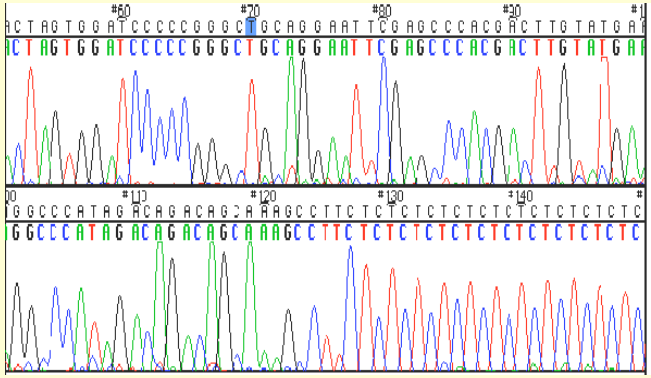
Major technological breakthroughs

Recombinant DNA

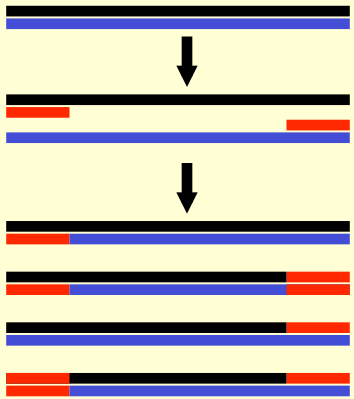




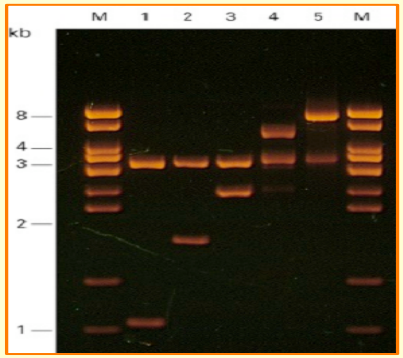
DNA sequencing



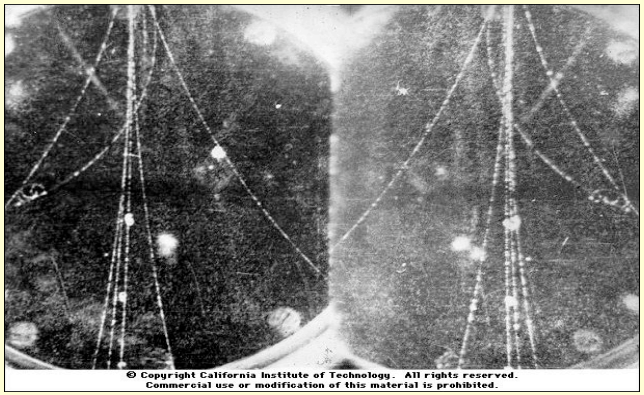
PCR



Separation of molecules



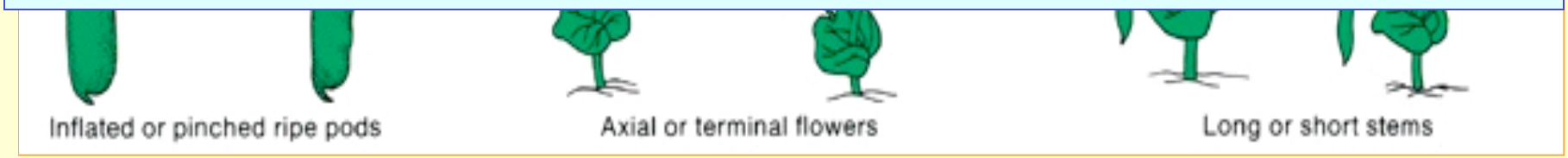
Like in particle physics



Mendel laws of genetics were discovered based on pairs of contrasting inherited pea *phenotypic* traits

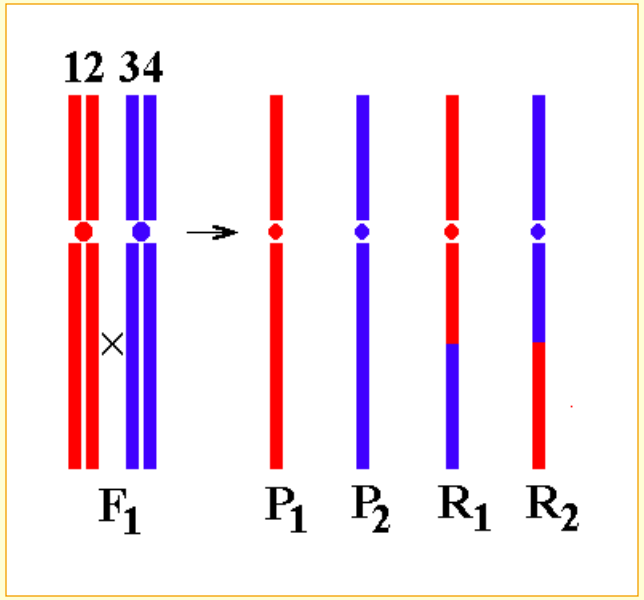
In the progeny of hybrids between carriers of these traits Mendel found **new combinations**, in proportions fitting independent segregation model (Mendel 3rd law). Unlike such situations with ***unlinked genes*** that belong to different chromosomes, transmission of ***linked genes*** is not independent.

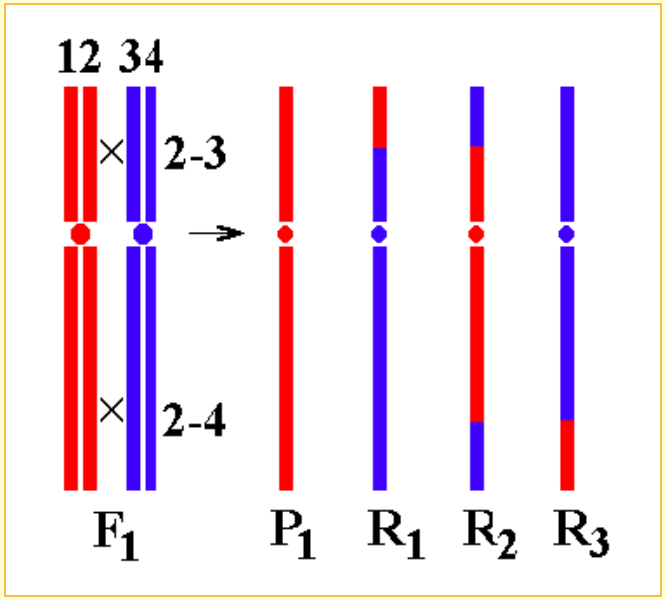
Studies of linked genes in fruit fly lead Morgan to discovery of ***genetic recombination***.



Recombination (crossing-over) is the central event of sex

occurs at *meiosis*, during formation of sexual cells





single-exchange double-exchange

meiotic configurations

Recombination: the basis of genetic mapping

Genetic mapping: a procedure of revealing the order of genes in chromosomes. It uses a notion of *genetic distance*. But in fact, mapping is based on *recombination rates*.

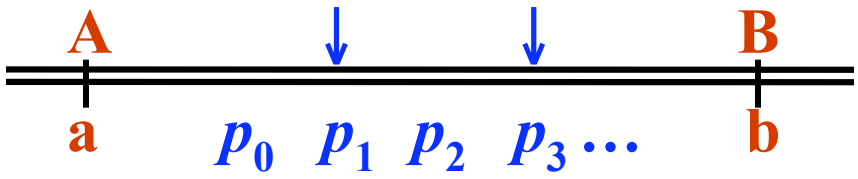
$$\frac{a\ b}{a\ b} \times \frac{A\ B}{A\ B} = \frac{A\ B}{a\ b} \longrightarrow \frac{A\ B}{a\ b} \times \frac{a\ b}{a\ b}$$

homo *homo* *hetero* "two-point back-cross"

	Parental types	Recombinants
AB/ab meiosis → sperm	$(1-r_m)\{AB + ab\}$	$r_m\{Ab + aB\}$
AB/ab meiosis → eggs	$(1- r_f)\{AB + ab\}$	$r_f\{Ab + aB\}$

Recombination rate and genetic map distance

***Genetic Distance*: $x = d$ (a,b) -** average number of recombination events in the segment over many meiotic cells



where p_k – prob. of k ($k = 0, 1, \ldots$) exchanges in the interval.

Thus $x = \sum_{k=0}^{\infty} k p_k$, but $r = \sum_{k=0}^{\infty} p_{2k+1}$

recombination rate *r* is the proportion of recombinant gametes

Problem: ***observed*** vs. ***occurred***: Only **uneven exchanges** result in **recombinants** that can be registered.

Constructing dense and reliable genetic maps (ordering the markers)

- The 3rd generation of human map includes ~ $2 \cdot 10^4$ loci
- A maize mapping project (Iowa, 2005) → ~ 10^4 loci
- 12% (!) of markers on cattle maps proved erroneously positioned

A B C D E F G H …

a b c d e f g h …

Different approaches of multilocus ordering

Genetic distance

$x_{ij} \geq 0$

$x_{ij} = x_{ji}$

$x_{ik} + x_{kj} = x_{ij}$

Recombination rate

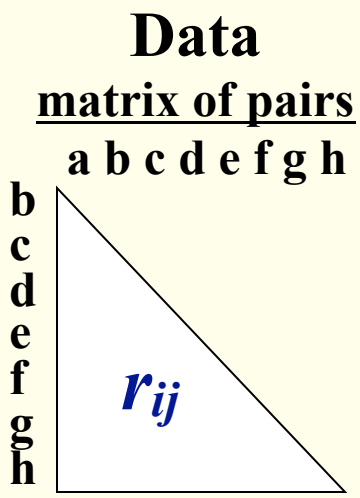
$r_{ij} \geq 0$

$r_{ij} = r_{ji}$

$r_{ik} + r_{kj} \geq r_{ij}$

A B C D E F G H

a b c d e f g h



A Multilocus likelihood analysis: calculates probabilities of orders

B Stepwise mapping by adding a marker at each step ("empiric")

C Treatment of the full matrix of pair-wise distances (our approach)

Constructing dense genetic maps

(reliable multilocus ordering)

- Objectives
  - ♦ Building multilocus maps (with ~10^3 markers/chr)
  - ♦ Verification of the orders (and removing "bad guys")
  - ♦ Building consensus maps
- Method and technology
  - ♦ Reduction to the *Traveler Salesman Problem* (TSP)
  - ♦ *Evolutionary strategy optimization* algorithms

ES algorithm as a simulation analogue of evolutionary adaptation models

Natural elements	Simulation elements
Chromosome **Individual,** a set of chromosomes **Mutation,** a small change of the chromosome **Population,** set of individuals **Fitness,** quantitative characteristic of organism's "fitness" **Selection,** choosing the fittest individual(s) for the next generation	Variable value x_i Solution vector $\boldsymbol{x} = (x_1,\ldots,x_n)$ Operator **M:** $\boldsymbol{x}^k \rightarrow \boldsymbol{x}^{k+1}$ Set $\boldsymbol{P}$ of solution vectors $\{\boldsymbol{x}^k\}$ Criterion value $f(\boldsymbol{x}^k)$ Operator **S**: $f(\boldsymbol{x}^k) \rightarrow \min$

ES algorithm for ordering multilocus maps

Order 1:	a b c d e f g h k l m n	l_1
Order 2:	b a c d e f g h k l m n	l_2
	.....	
Order *i*:	f c m h e a g n k l b d	l_i

Let order O_i be considered a '***genotype***', and its '***fitness***' b

e defined as: $w_i = l\,(\mathbf{O}_i) = 1/l_i$ (or $-l_i$)

'Progeny' is produced via ***mutations*** (changed orders).
A 'child' replaces its parent if its fitness is higher.
To build the map we need only the (ML) estimates of
pair-wise recombination rates for all pairs of markers

Building multilocus maps: Sources of complexity

- **½ *n*!** orders possible. As a solution we need to find not j ust any order with a small total map length. Rather, the goal is to reveal **the real** order (i.e. *unique solution*)
- Sampling variation of r_{ij}, missing data, data errors
- Small sample size relative to the number of markers
- Genetic interference (inter-dependence of cross-overs along the chromosome)

Re-sampling for quality control

The best way to **check / verify** the map is to show that the obtained solution does not depend on:
(a) sampling data variation, and **(b)** starting points

By taking sub-samples, one can build **repeated** maps and test whether/where marker ordering remains the same.

↓

Detecting trouble-making markers

Major unsolved problem: We believe that we can reach the unique solution for a given data set.

However, we have no regular procedure that leads to the best subsets of markers allowing for:

- stable ordering
- combined with highest "map coverage"
- combined with minimal gaps along the map

Removing one marker out of n → n ways

two markers → ½ $n(n\text{-}1)$ ways

…

Initial ordering:
Unstable neighborhoods

Stable neighborhoods: after removing problematic markers

Assembling multilocus consensus maps

Objective: Building multilocus maps based on data from different labs and different mapping populations

Requirements:

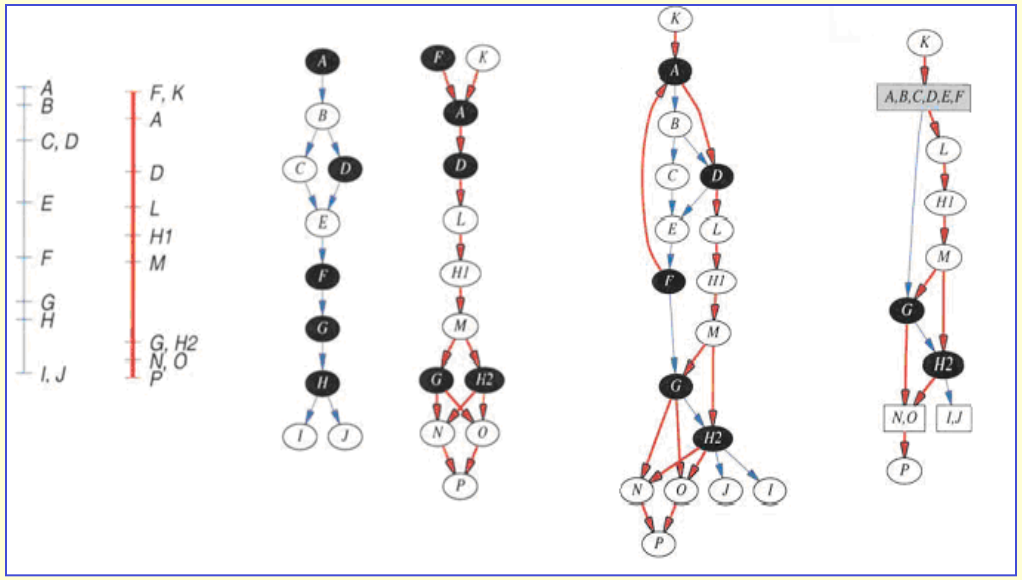
- Shared markers must be in an identical order
- The resulting consensus ordering must be verified via re-sampling

Proposed strategy:

- **Re-building maps** under the constraint of *identical order for shared markers*, instead of **looking for shared orders in pictures** of previously build maps

Graph-theoretical approach for reconciling two orders, received from different sources

"Giving credit" to individual multilocus maps: Yap et al., *Genetics*, 2003



Our strategy: re-building the maps

Re-analysis of raw data by reduction to ***synchronous TSP***

→ Parallel discrete optimization for multiple data sets with the foregoing constraint

Consensus mapping (with 100% shared markers)

simulated example, 6 families each with n=100

Pop	Mthd	Marker position																		
1	NSO	20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	SO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	NSO	1	2	3	4	5	6	7	9	8	10	11	12	13	14	15	16	17	18	19	20
	SO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	NSO	15	14	13	12	11	10	9	8	7	6	5	4	3	1	2	16	17	18	19	20
	SO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	NSO	1	2	3	4	5	6	7	8	9	11	10	12	13	14	15	16	17	18	19	20
	SO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	NSO	1	2	3	4	5	6	7	8	9	10	11	12	13	18	17	16	15	14	19	20
	SO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	NSO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	18	20
	SO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

NSO - non-synchronized optimization

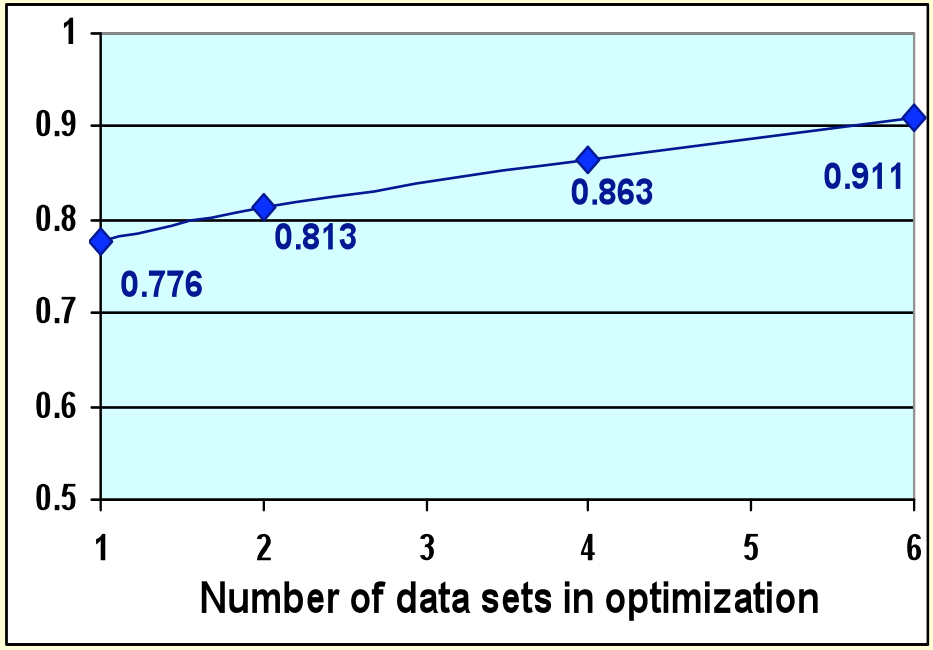
SO - synchronized optimization

False order

True order

Quality of multilocus ordering as a function of the proportion of utilized data sets

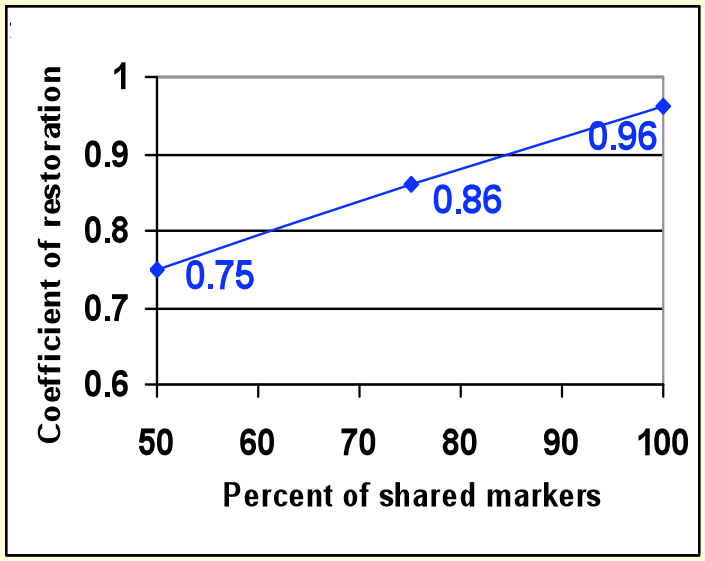
(results of tests with six data sets)



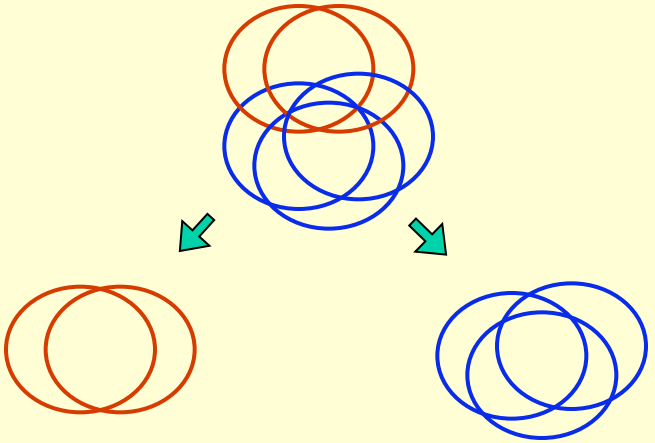
Thus, the information from additional data sets allows reaching better map quality

However, we have a problem →

Dependence of the quality of consensus map on the proportion of shared markers



The need in re-structuring the synchronous mapping problem



Unsolved problems:

- How to subdivide the entire set into subsets
- How to build the jackknifing procedure

Genetic dissection of quantitative traits

genotype → (development / environment) → phenotype + markers

↓

observations (data)

↓

QTL mapping

↑

Physiology & Medicine
Molecular Biol. & Genetics
Appl. Math., Statistics,
Computer Science

↓

"explaining" the phenotype

↓

Applications

↓

QTL of economic or medical importance; fitness-related QTL; gene expression as ***molecular phenotype*** (eQTL analysis)

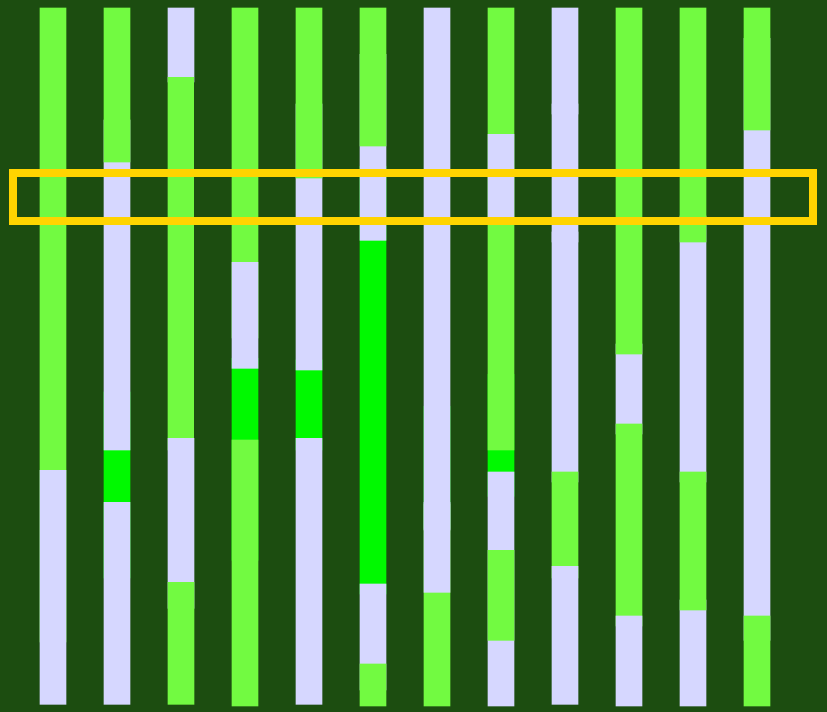
Analysis of the genetic composition of segregating recombinant genotypes

Individual recombinant chromosomes

Segments from Parent A

:

Segments from Parent B

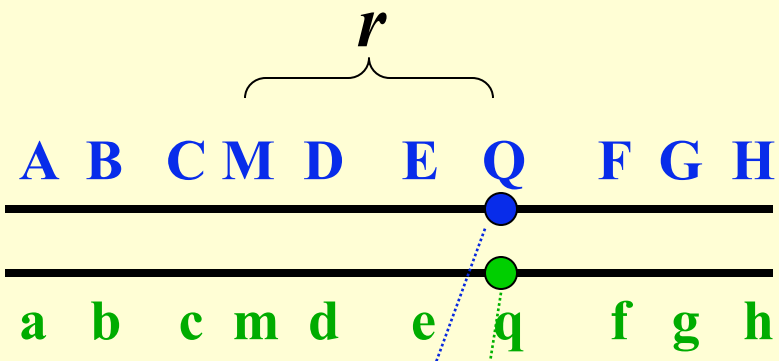


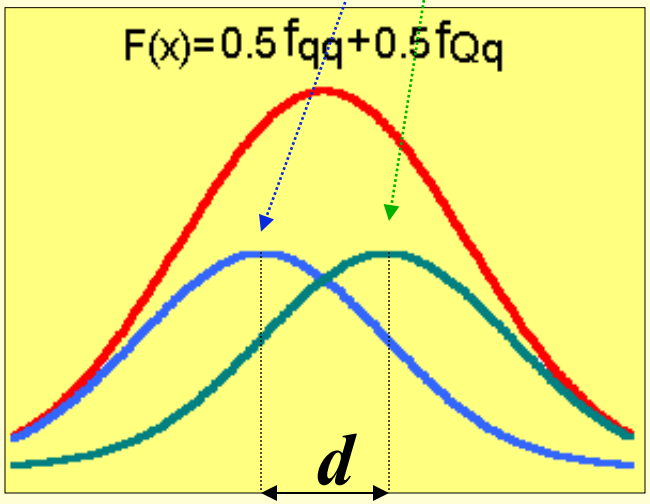
Looking for loci affecting quantitative traits by using DNA markers

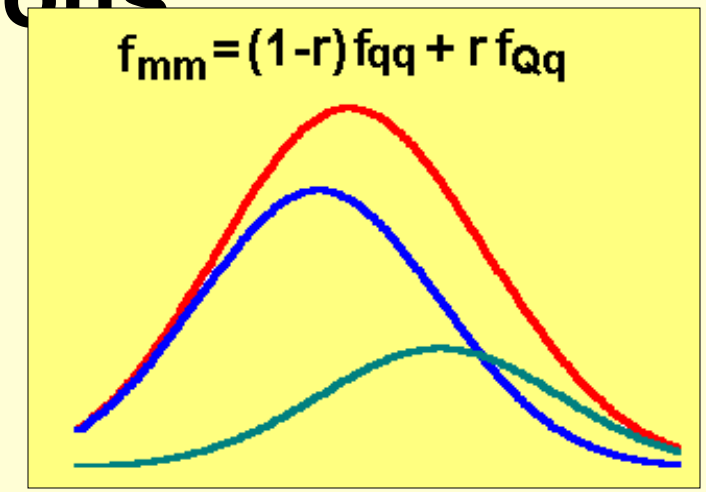
QTL mapping

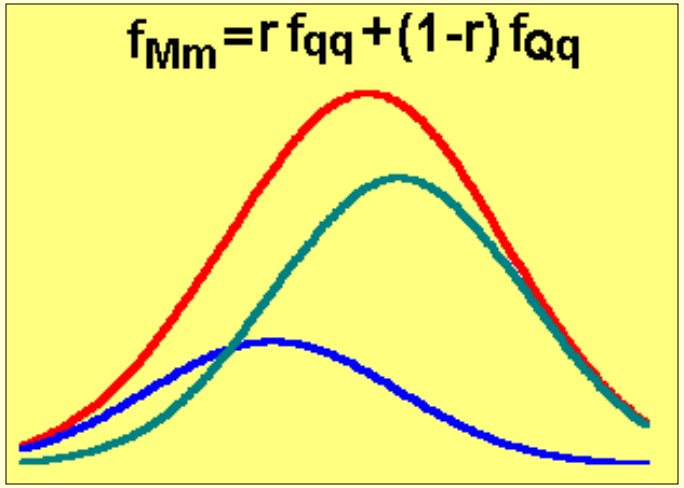
Beat Keller, Institut für Pflanzenbiologie, Universität Zürich

QTL analysis – dealing with distributions



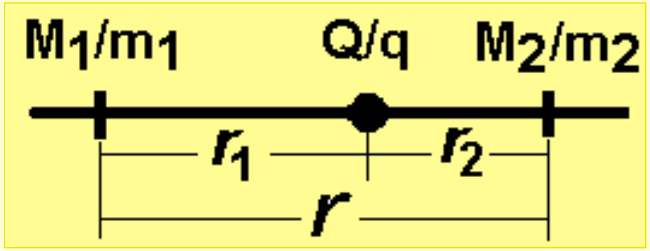






QTL Interval Mapping

Expected distributions of the trait in the flanking marker groups are mixtures of non-recombinants and recombinants



$$f_{M_1M_2} = [(1-r_1)(1-r_2)f_Q + r_1 r_2 f_q]/(1-r)$$

$$f_{M_1m_2} = [(1-r_1) r_2 f_Q + r_1(1-r_2) f_q]/r$$

$$f_{m_1M_2} = [r_1(1-r_2) f_Q + r_2(1-r_1) f_q]/r$$

$$f_{m_1m_2} = [r_1 r_2 f_Q + (1-r_1)(1-r_2) f_q]/(1-r)$$

The model of QTL effect

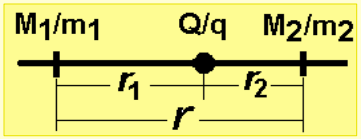
For additive QTL effect: → $x = m + dg_q + \xi$

where g_q = **-1** for **qq**, and **+1** for **QQ**; $E\xi=0$, $\sigma\xi = \sigma$,

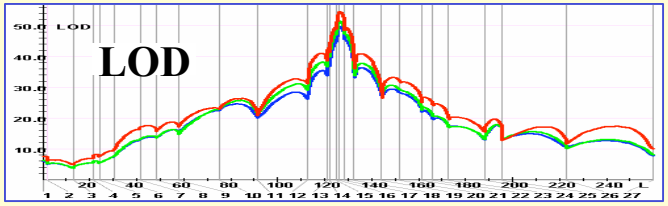
and $d=(\mu_{QQ} - \mu_{qq})/2$, $\mu_{qq} = m-d$, $\mu_{QQ} = m+d$

ML-estimation in QTL interval analysis

$$L(r, m, d, \sigma) = \prod_{i=1}^{4} \prod_{j=1}^{N_i} f_i(r, m, d, \sigma \mid x_{ij}) \rightarrow \max$$



ML-estimates: r^*, m^*, d^*, σ^*



Log-likelihood ratio test (for H_1 vs H_0)

What do one expect from analytical tools ?

To extract maximum mapping information from the experimental data

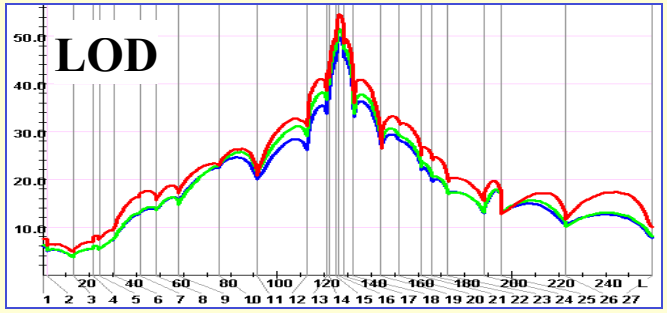
The main questions in QTL analysis:

- QTL detection power (detect the effect when it exists)
- Minimum "false positives" (high significance)
- Accuracy of parameter estimates

For single-trait analysis:

$ELOD = -\frac{1}{2} N \log (1 - H^2)$,

$H^2 = d^2/(d^2+\sigma^2)$ is "*heritability*"

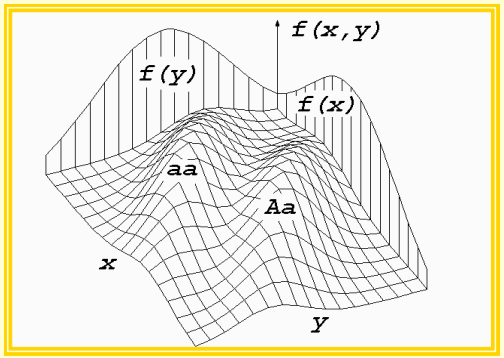


What could be the benefit from a transition to multiple-trait analysis ?

For single-trait analysis:

$$ELOD_x = - \tfrac{1}{2} N \log(1- H^2_x)$$

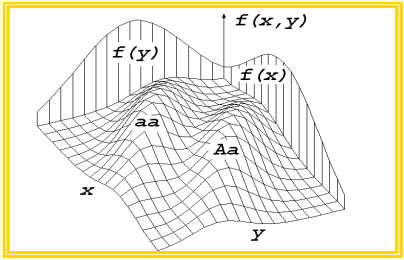
The same holds in two-trait analysis, upon $H^2_x \Longrightarrow H^2_{xy}$

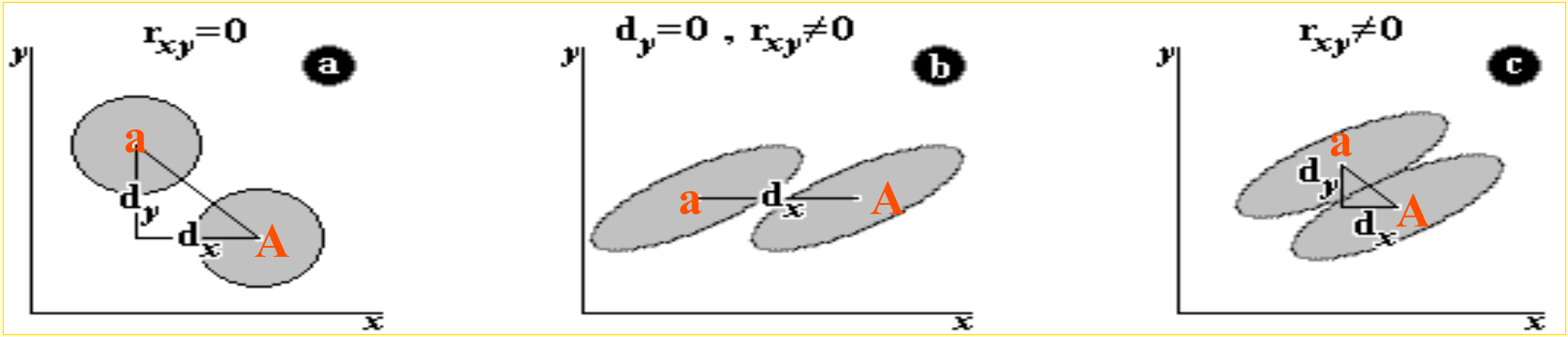


$$H^2_{xy} = 1 - \frac{\sigma^2_x \sigma^2_y (1- R^2_{xy})}{(\sigma^2_x + d^2_x/4)(\sigma^2_y + d^2_y/4) - \sigma^2_x \sigma^2_y [R_{xy} + d_x d_y/(4\sigma_x \sigma_y)]^2}$$

It appears that $H^2_{xy} \geq H^2_x \rightarrow ELOD_{xy} \geq ELOD_x$

The main sources of statistical superiority of two-trait analysis



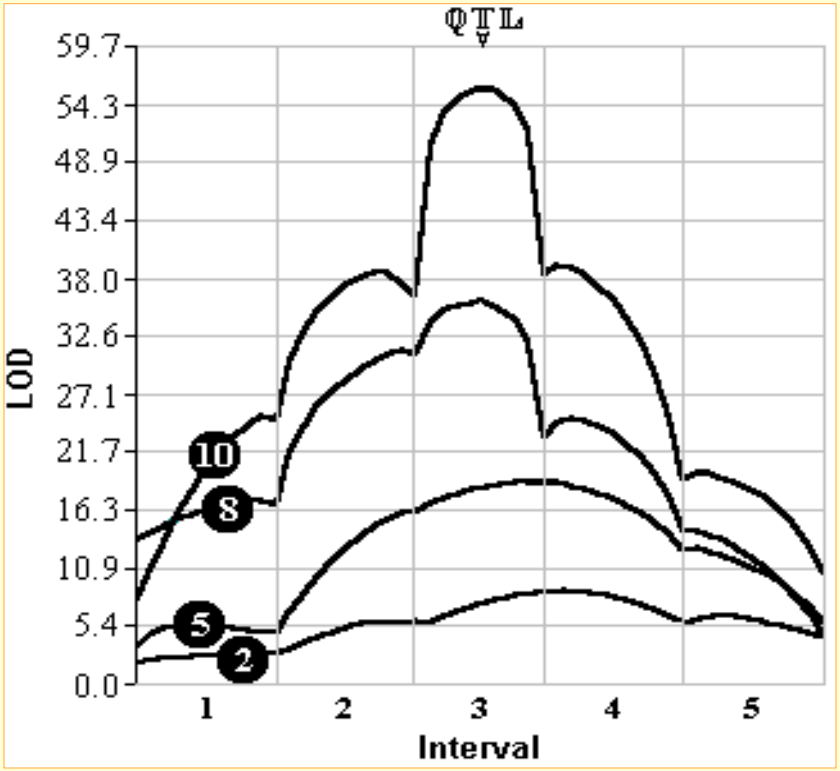


any d_x & d_y

any d_x & r_{xy}

$r_{xy}\, d_x d_y < 0$

Effect of the number of traits on the efficiency of QTL mapping

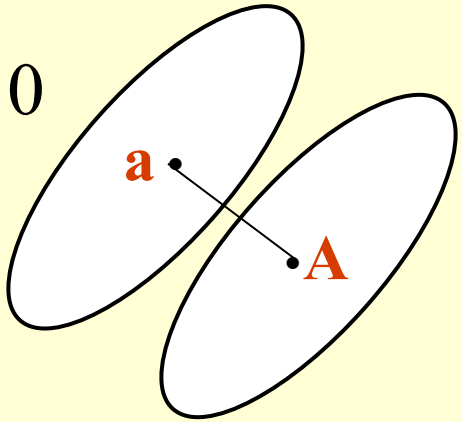


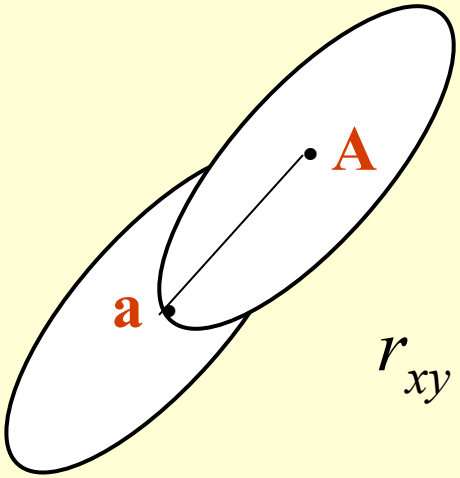
Based on *interval-specific* multivariate analysis

Multiple-trait analysis does not necessarily improve the quality of QTL analysis

$r_{xy}\, d_x d_y < 0$

(1)



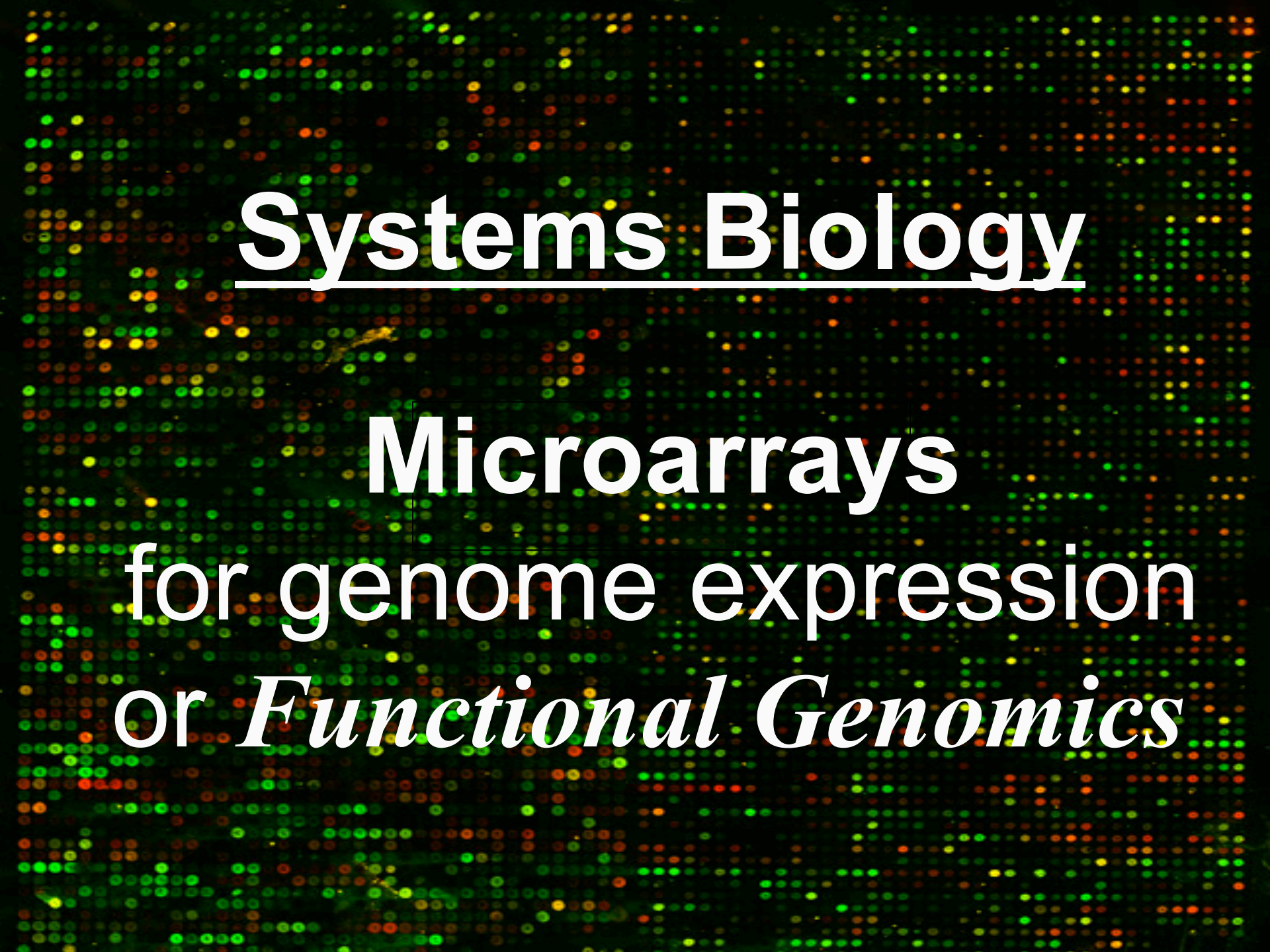


(2)

$r_{xy}\, d_x d_y > 0$

With the same overlapping of marginal distributions, the bivariate distributions of QTL groups **a** and **A** overlap less in **(1)** than in **(2)**

Required: Extension of the above criterion for arbitrary numbers of traits. To allow selecting of sub-sets with improved resolution… (for $n \sim 10^2$ or even 10^4)

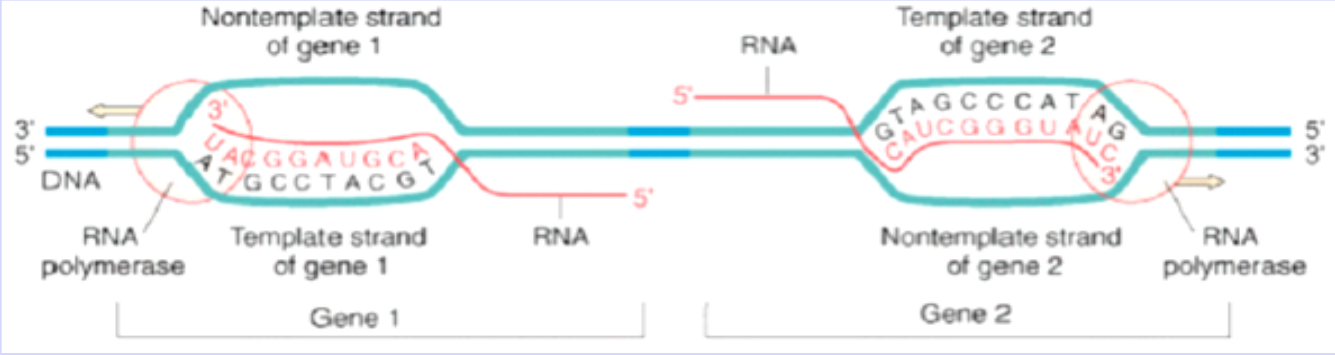

Systems Biology
Microarrays
for genome expression
or Functional Genomics

How genes are expressed in the cell ?

Gene sequences encoding for proteins are non-overlapping texts that begin from *start signal* and end by *stop signal*.

DNA transcription mRNA translation protein

A gene can be *transcribed* many times. The resulting mRNA can be *translated* many times → many copies of the enzyme. Each synthesized enzyme molecule can *catalyze* the target reaction thousands times → strong "signal amplification"



All mRNAs of the genome

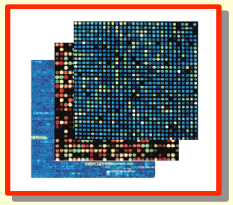
Expression of each gene can be scored as a quantitative trait in a mapping population ($n \sim 10^2$-10^3) and tested for association with DNA markers across the genome ($k \sim 10^2$-10^5) → eQTL mapping

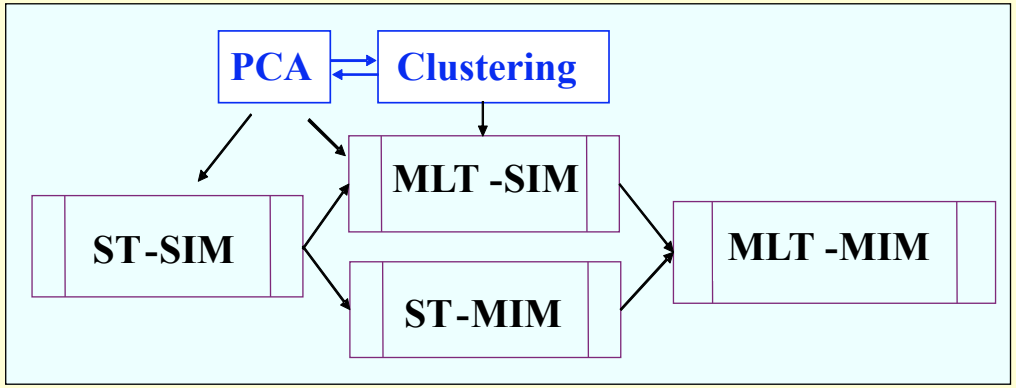
<u>The challenge of the problem size</u>: With $N \sim 10^4$ genes, the number of data points reaches $\sim 10^8$-10^{12}

Multiple-trait QTL analysis of the $N \sim 10^4$ expression traits ?
An urgent need in "dimensionality reduction" methods

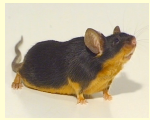
pathological states, aging, evolution, etc.?

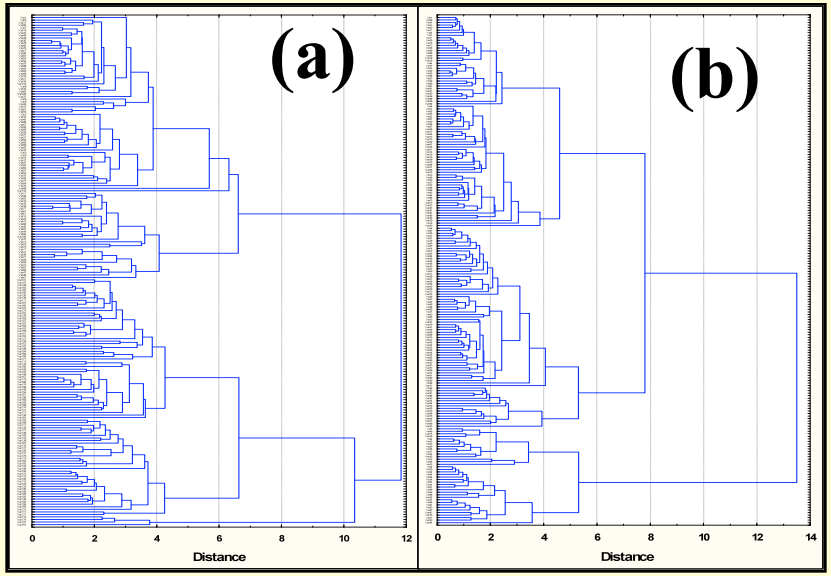
Expression scores as a vector of quantitative traits: Dealing with high dimensionality in multiple-trait QTL mapping♪



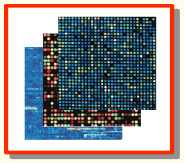


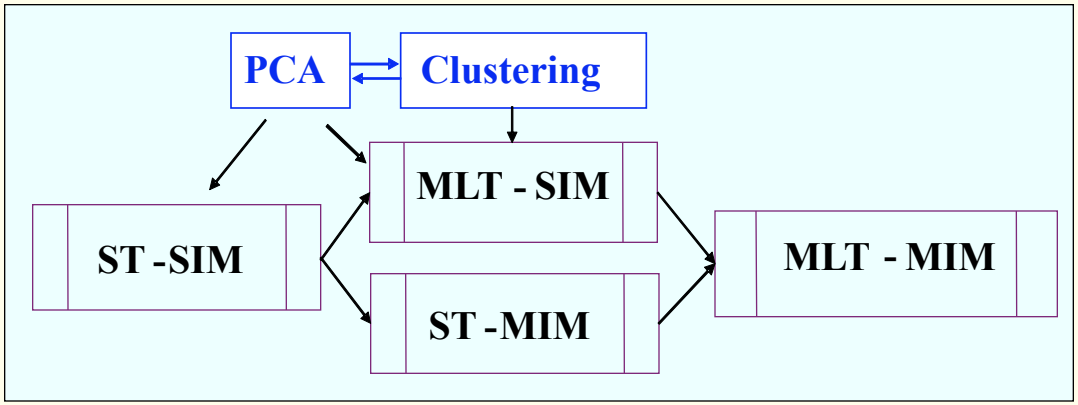
Clustering of the chosen 400 genes: **(a)** 100+100 up- and down-regulated, **(b)** 100+100 plus- and minus-correlated to obesity genes

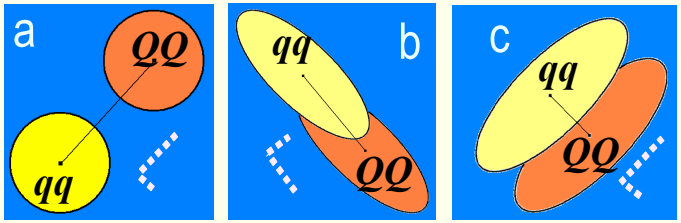




Expression scores as a vector of quantitative traits: Dealing with high dimensionality♪







The first PCs may (a & b) or may not (c) correspond to the direction of multivariate QTL effects

Multiple-trait vs. single-trait eQTL mapping

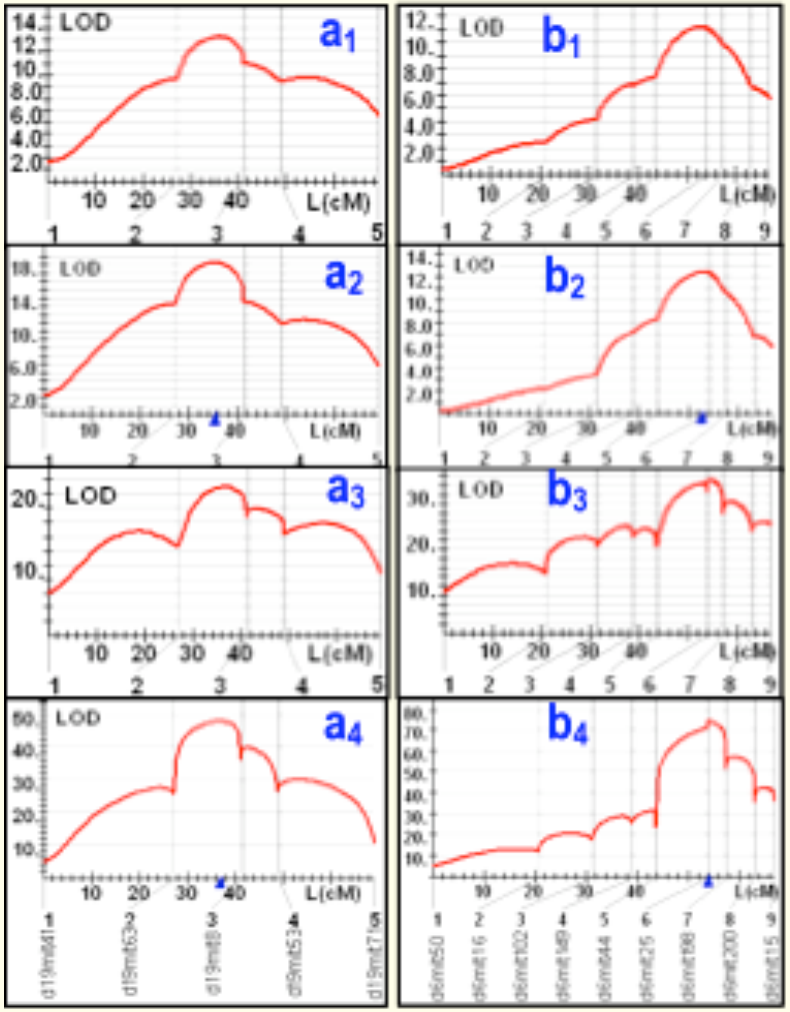
dealing with clusters (on an example of mouse obesity

Mapping for 2 sub-clusters: **(a) up-** or **down-regulated, (b) positively** or **negatively correlated** with obesity.

For these groups, the estimated QTL location L and SD(L) were, for chr. 19 and 6, respectively:

	(a)	**(b)**
SIM-ST	36.9±6.8	51.7±2.8 cM
MIM-ST	35.0±3.3	52.4±1.9 cM
SIM-MLT	39.6±8.4	52.4±5.9 cM
MIM-MLT	36.7±2.6	53.9±0.7 cM

Using data from Ghazalpour et al., 2005



Summary (what we have been talking about)

- ***Genome mapping*** - reduction to TSP
- ***Consensus mapping*** - synchronous TSP
- ***QTL mapping***
- ***Multiple-trait QTL analysis*** - looking for best sub-sets
- ***Microarray analysis*** - expression QTL (eQTL)
- ***Multi-trait eQTL mapping*** - dimensionality reduction

Acknowledgments

Y. Ronin, D. Minkov, D. Mester,
M. Korostishevesky, A. Itzkovich,
J. Peng, O. Orion, Z. Frenkel, M. Soller,
J. Weller, J. Hillel, J. Beckmann