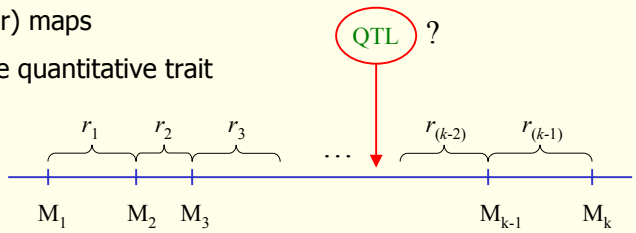
GENETICAL GENOMICS

- Quantitative Trait Locus (QTL), Major genes
- Overview of linkage analysis and QTL mapping
  - Single marker analysis
  - Genetic map construction
  - Interval mapping and Composite interval mapping
  - Different designs
- Combining Marker information and Gene expression data
- Examples

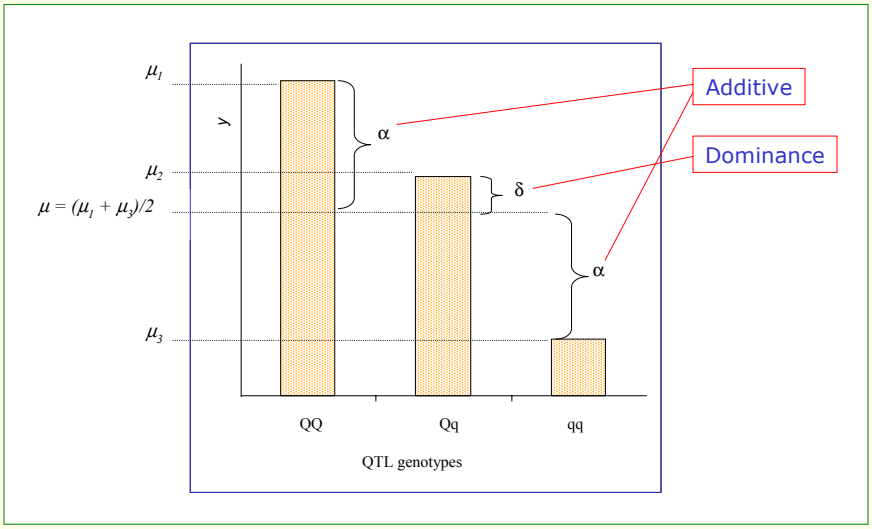
QTL MAPPING

- Methods based on linkage disequilibrium between markers and QTL (line crossing or segregating population)
- Requirements:
  - ① Linkage (marker) maps
  - ② Variation for the quantitative trait
  - ③ Polymorphism



- Methods for mapping QTL
  - ① Single marker analysis
  - ② Interval mapping
  - ③ Composite interval mapping
  - ④ Bayesian methods

SINGLE MARKER ANALYSIS



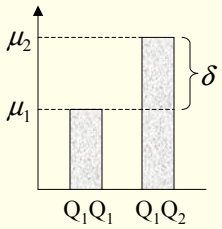
SINGLE MARKER ANALYSIS

➡ Simple example with candidate gene and backcross population

$Q_1Q_1 \times Q_2Q_2$

$Q_1Q_2 \times Q_1Q_1$

$Q_1Q_1 \quad Q_1Q_2$



Genotype	Obs.	Mean	STD
Q_1Q_1	n_1	m_1	s_1
Q_1Q_2	n_2	m_2	s_2

⇨ $H_0: \delta = 0$ vs $H_1: \delta \neq 0$

$$t = \frac{m_1 - m_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$CI[\delta;(1-\alpha)]: \ (m_2 - m_1) \pm t_{(n_1+n_2-2;\alpha/2)}\sqrt{\frac{s^2}{n_1 + n_2 - 2}}$$

SINGLE MARKER ANALYSIS

- Simple QTL and marker (M); recombination frequency = *r*

Genotype	Freq.	E[y]	Marker group	Freq.	E[y]
$M_1M_1Q_1Q_1$	(1-*r*)/2	μ_1	M_1M_1	½	$(1-r)\mu_1 + r\mu_2$
$M_1M_1Q_1Q_2$	*r*/2	μ_2			
$M_1M_2Q_1Q_1$	*r*/2	μ_1	M_1M_2	½	$r\mu_1 + (1-r)\mu_2$
$M_1M_2Q_1Q_2$	(1-*r*)/2	μ_2			

$$\begin{matrix} M_1|M_1 \\ Q_1|Q_1 \end{matrix} \quad \begin{matrix} M_1|M_2 \\ Q_1|Q_2 \end{matrix}$$

$$\begin{matrix} M_1|M_1 \\ Q_1|Q_1 \end{matrix} \quad \begin{matrix} M_1|M_1 \\ Q_1|Q_2 \end{matrix} \quad \begin{matrix} M_1|M_1 \\ Q_1|Q_1 \end{matrix} \quad \begin{matrix} M_1|M_2 \\ Q_1|Q_2 \end{matrix} \quad \begin{matrix} M_1|M_2 \\ Q_1|Q_2 \end{matrix}$$

Difference between marker group expected values

$$r\mu_1 + (1-r)\mu_2 - (1-r)\mu_1 - r\mu_2$$

$$= (1-2r)(\mu_2 - \mu_1) = (1-2r)\delta$$

SINGLE MARKER ANALYSIS

(EXAMPLE)

- *Brassica napus*; Flowering time
- 10 Markers

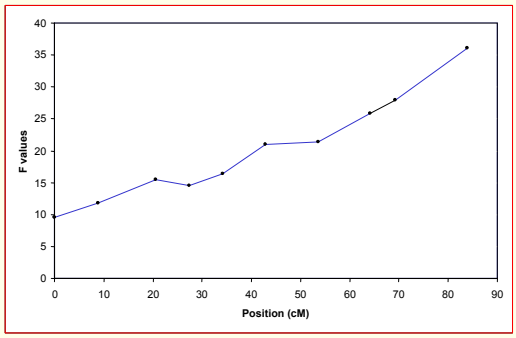
  (positions: 0, 8.8, 20.6, 27.4, 34.2, 42.9, 53.6, 64.1, 69.2, 83.9 cM)

- 104 individuals; Double haploid

```
3.0204    -1    -1    -1    -1    -1    -1    -1    -1   -99    -1
2.9704    -1    -1    -1    -1   -99    -1    -1    -1    -1     1
2.7408    -1    -1     1     1     1     1     1     1     1     1
  ⋮        ⋮     ⋮     ⋮     ⋮     ⋮     ⋮     ⋮     ⋮     ⋮     ⋮
3.3673     1     1     1     1    -1    -1    -1    -1    -1     1
3.0681     1     1     1     1   -99     1     1     1    -1    -1
3.2771    -1   -99    -1    -1    -1    -1    -1    -1    -1    -1
```

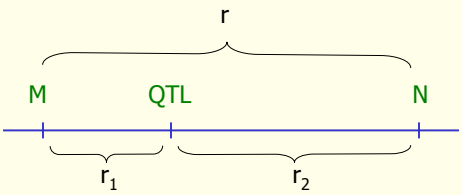
(Satagopan et al. *Genetics* 144: 805-816, 1996)

Chrom.	Marker	μ	τ	LRT	F	p-value
1	1	3.184	-0.202	9.379	9.624	0.002 **
1	2	3.204	-0.230	11.378	11.789	0.001 ***
1	3	3.232	-0.266	14.706	15.485	0.000 ***
1	4	3.229	-0.259	13.885	14.562	0.000 ***
1	5	3.240	-0.276	15.554	16.446	0.000 ****
1	6	3.259	-0.307	19.518	21.041	0.000 ****
1	7	3.252	-0.302	19.747	21.312	0.000 ****
1	8	3.257	-0.318	23.450	25.775	0.000 ****
1	9	3.258	-0.330	25.156	27.884	0.000 ****
1	10	3.252	-0.362	31.518	36.059	0.000 ****

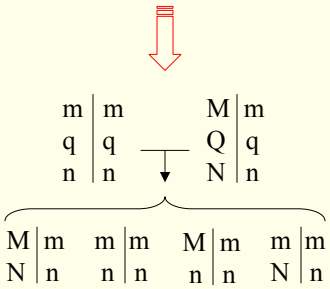


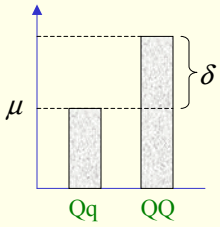
INTERVAL MAPPING

(Lander & Botstein, 1989)



Backcross





$$y_i = \mu + q_i\delta + \varepsilon_i$$

phenotype — QTL genotype — residual

$$q_i = \begin{cases} 0 \text{ , if qq} \\ 1 \text{ , if Qq} \end{cases}$$

INTERVAL MAPPING

If $\varepsilon_i \sim N(0,\sigma^2) \rightarrow y_i \mid q_i \sim N(\mu + q_i\delta, \sigma^2)$

$$p(y_i \mid q_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu - q_i\delta)^2\right\}$$

$$L(\mu,\delta,\sigma^2,\lambda,\boldsymbol{q} \mid \boldsymbol{y}) \propto \prod_{i=1}^{N}\left[f(y_i \mid q_i = 0)\Pr(q_i = 0) + f(y_i \mid q_i = 1)\Pr(q_i = 1)\right]$$

$$L(\mu,\delta,\sigma^2,\lambda,\boldsymbol{q} \mid \boldsymbol{y}) \propto \prod_{i=1}^{N}\left[\frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}\Pr(q_i = 0 \mid \lambda)\right.$$

QTL position

$$\left. + \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu - \delta)^2\right\}\Pr(q_i = 1 \mid \lambda)\right]$$

INTERVAL MAPPING

$\Pr(q_i \mid \lambda)$ is modeled in terms of recombinations between flanking markers and QTL:

Marker Genotypes	Pr(q_i = QQ)	Pr(q_i = Qq)
M,N	$(1 - r_1)(1 - r_2)/(1 - r)$	$r_1 r_2/(1 - r)$
M,n	$(1 - r_1)\, r_2 / r$	$r_1 (1 - r_2)/ r$
m,N	$r_1 (1 - r_2)/ r$	$(1 - r_1)\, r_2 / r$
m,n	$r_1 r_2/(1 - r)$	$(1 - r_1)(1 - r_2)/(1 - r)$

Approximation:

(no double recombination)

Markers	Pr(q_i = QQ)	Pr(q_i = Qq)
M,N	1	0
M,n	(1 - p)	p
m,N	p	(1 - p)
m,n	0	1

$$p = \frac{r_1}{r}$$

INTERVAL MAPPING

- Likelihood estimation: EM algorithm to estimate parameters, including λ (position of QTL).
- Alternatively: Fix λ (grid search) and evaluate LOD.

$$\mathrm{LOD}_\lambda = \log_{10}\left[\frac{L_\lambda(\hat{\mu},\hat{\delta},\hat{\sigma}^2,\hat{\mathbf{q}}\,|\,\mathbf{y})}{L_\lambda(\hat{\mu},\hat{\sigma}^2,\hat{\mathbf{q}}\,|\,\mathbf{y},\delta=0)}\right]$$

- A QTL is detected whenever the LOD score gets larger than a threshold; estimated position of the QTL maximizes LOD.

INTERVAL MAPPING
Regression Approach
(Haley & Knott, 1992)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ \vdots & \vdots \\ p_{N1} & p_{N2} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

alternatively

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & p_{12} \\ 1 & p_{22} \\ \vdots & \vdots \\ 1 & p_{N2} \end{bmatrix} \begin{bmatrix} \mu \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

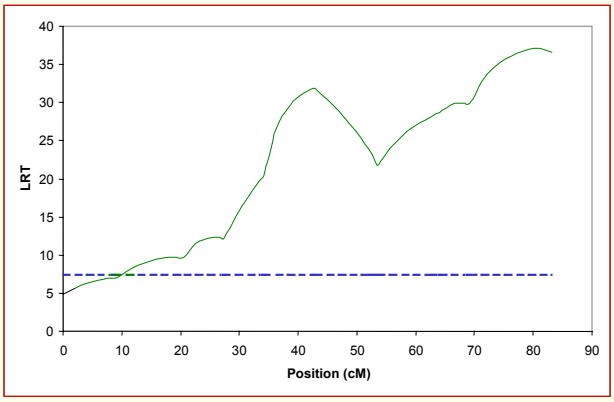
Residual Sum of Squares:

$$\mathrm{RSS} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

Estimated position of the QTL minimizes RSS.

INTERVAL MAPPING
(Example)

- *Brassica napus*; Flowering time (Satagopan et al., 1996)



INTERVAL MAPPING

- COMMENTS:

① Backcross to both parental lines, or use F2 design, to estimate additive and dominance effects.

② Threshold; multiple testing; false positives

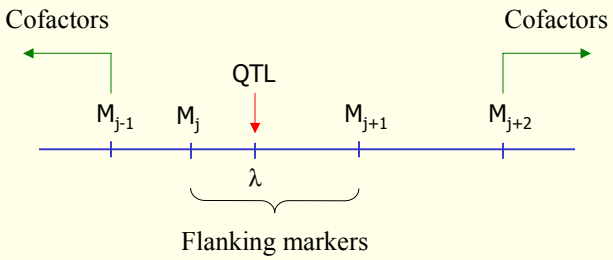
③ Confidence intervals

④ Multiple QTL, ghost QTL

COMPOSITE INTERVAL MAPPING

(Zeng, 1993, 1994)

- Interval analysis adding marker cofactors (to account for the effects of unlinked QTLs); combination of single interval mapping and multiple linear regression.



COMPOSITE INTERVAL MAPPING

(Zeng, 1993, 1994)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Dummy variables

$$y_i = \beta_0 + \beta^* x_{ij} + \sum_{k \neq j, j+1} \beta_k w_{ik} + \varepsilon_i$$

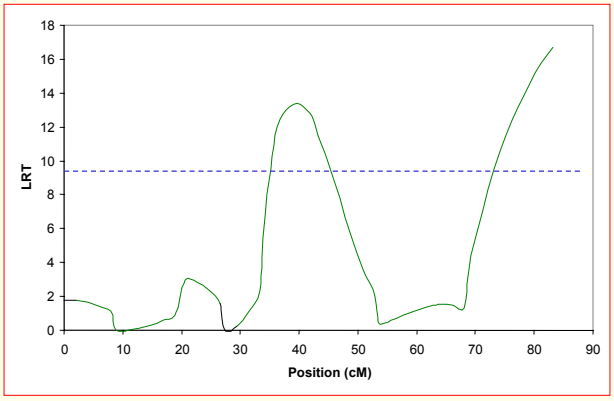
Intercept

Genetic effect of the putative QTL (between markers j and j+1)

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1j} & w_{11} & \cdots & w_{1p} \\ 1 & x_{2j} & w_{21} & \cdots & w_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{Nj} & w_{N1} & \cdots & w_{Np} \end{bmatrix}$$

COMPOSITE INTERVAL MAPPING
(Example)

➡ *Brassica napus*; Flowering time (Satagopan et al., 1996)



SOFTWARE

⇨ QTL Cartographer

(http://statgen.ncsu.edu/qtlcart/cartographer.html)

⇨ MapMaker/QTL

(http://www-genome.wi.mit.edu/genome_software/)

⇨ QTL Express

(http://qtl.cap.ed.ac.uk/)

⇨ PlabQTL

(http://www.uni-hohenheim.de/~ipspwww/soft.html)

➲ Others:

http://www.stat.wisc.edu/biosci/linkage.html#linkage)

GENETICAL GENOMICS

➡ Integration of genetic marker data and gene expression profiling

- Linkage analysis
- Linkage disequilibrium
- QTL mapping

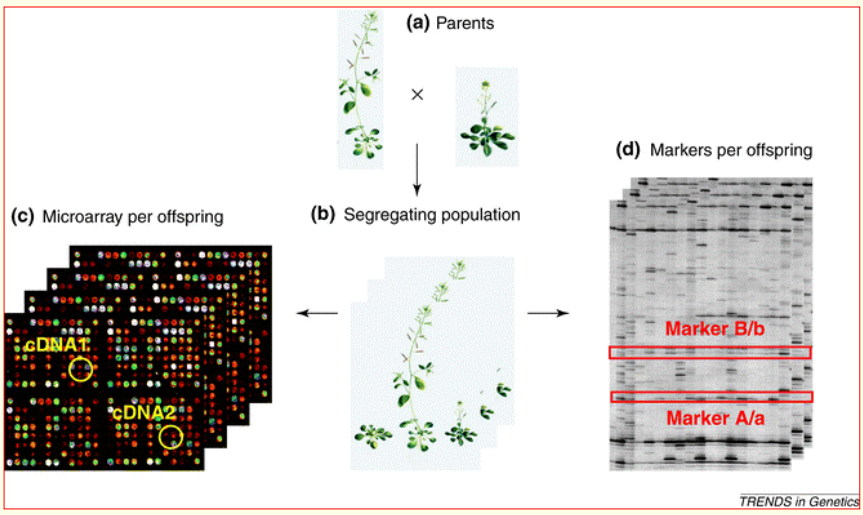
Correlation study between DNA sequence differences and phenotypic variation

- Hypothesis testing
- Classification
- Clustering
- Correlation analysis
- Metabolic pathways

Studying variation in the genes being switched on/off

Genome-wide genetic analyses of gene expression data

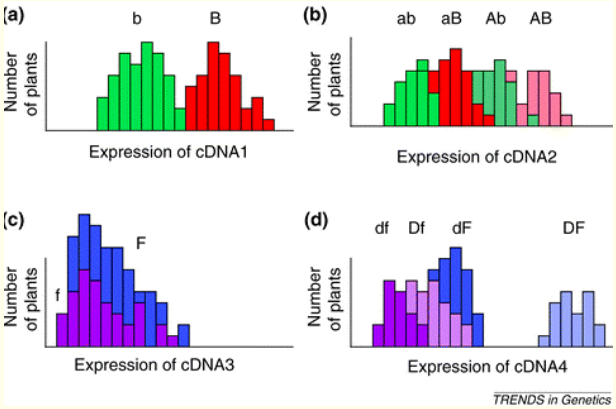
- Genetic variation between individuals in a segregating population
- Analytical tools available for molecular markers and for genome-wide expression profile data.



Expression profiling in combination with molecular marker analysis of a segregating population makes it possible to use QTL analysis for identification of influential genes. (a) A line cross experiment may be used to generate a segregating population (b), and each individual is used for (c) analysis by microarray profiling and (d) molecular marker analysis. (Source: Jansen and Nap, 2001)

➡ Idea: Expression profile of each gene in the population as a quantitative trait.

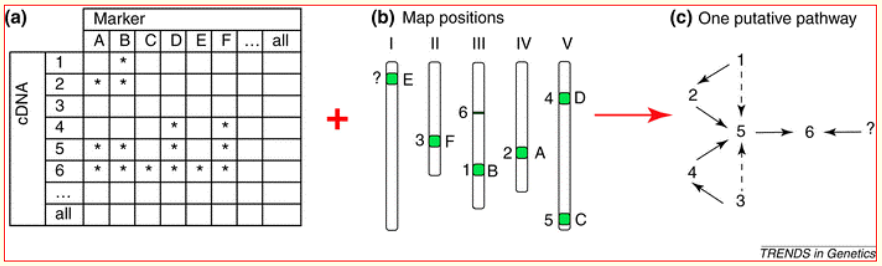
➡ Result: Substantial additional insight into the function and interrelation of gene products and gene action (Jansen and Nap, 2001).



Example of four hypothetical cDNAs; The microarray data is plotted in a histogram and analysed in combination with the molecular marker data. (Source: Jansen and Nap, 2001)

Note: (a) Qualitative expression (aside of noise): may be used as marker
(b-d) Multigenic variation (quantitative)

Suppose the segregating population has been screened for a phenotypic trait of interest and shows QTLs for that trait. Genetical genomics approach can help identifying the gene(s) responsible for such QTLs, and unraveling genes and gene products that are involved in metabolic and regulatory pathways (Jansen and Nap, 2001)



Pathway reconstruction and pathway 'memory'. The table (a) gives the hypothetical results of analyzing the gene expression profile data. (b) Genetic map of markers and position of each cDNA. In combination, these data can be used to analyze which genes influence the expression of other genes and to deduce in what order this influence is exerted (c).

• For each gene (cDNA) or gene product analysed (e.g. using proteomics and metabolomics) in the segregating population, QTL analysis will pinpoint the regions of the genome influential for its expression (eQTL)

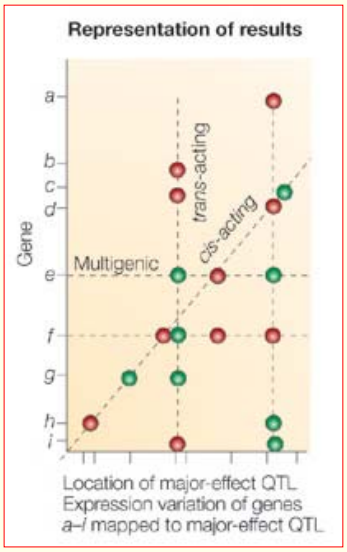
• If available, the sequence and annotation of that genomic region would be helpful for the identification of the genes involved.

• Identification of candidate genes by combining QTL information from all genes and gene products that are analysed.

cis-acting factors: variation in gene expression that maps to the gene themselves

trans-acting factors: variation in gene expression that maps to other genomic locations.

eQTL hotspots: chromosomal regions containing multiple eQTLs.



• Candidate genes: overlaps between differentially expressed genes, QTLs, and eQTLs found in common chromosomal regions (Darvasi, 2003)

• Candidate genes (genomic regions) can involve also genes not present in the microarray, genes with very low expression levels, or genes with influential expression at a time (long) before sampling of RNA, protein or metabolite (Jansen and Nap, 2001)

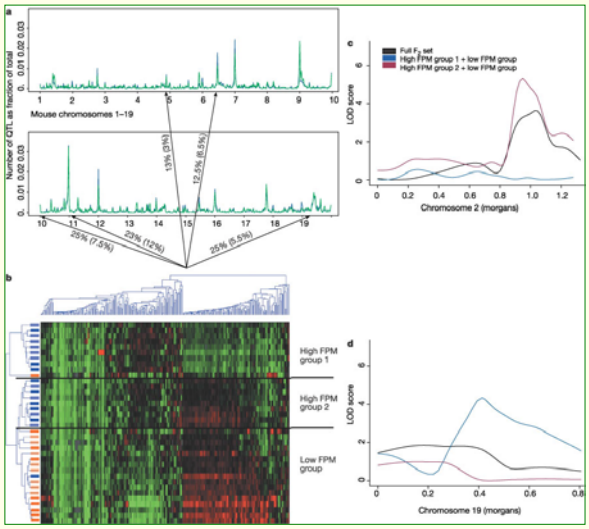
EXAMPLE

(Schadt et al., 2003)

➡ Data set:

- Liver tissue; mice under high-fat diets
- n = 111 F_2 (two inbred strains, C57B2/6 and DBA/2)
- Oligonucleotide array: 23,574 genes
  (7,861 significantly differentially expressed ($p<0.05$) between parental strains)
- QTLs found for 2,123 genes (LOD > 80.0; $p < 0.00005$)
  (25% of F_2 variation)

Gene expression quantitative trait loci (eQTL) distributions and the molecular basis for fat pad mass (FPM) in the F_2 cross.

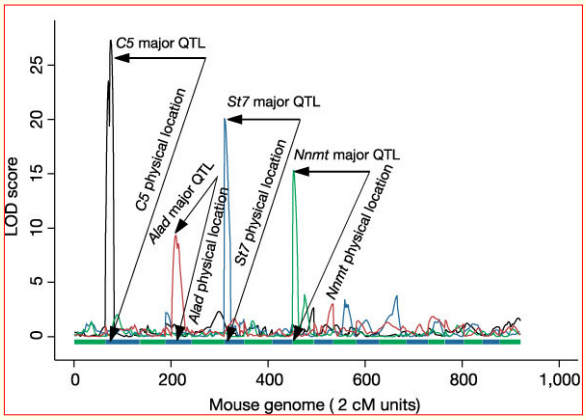


(a) Percentage of eQTL on autosomal chromosomes at two LOD score threshold.

(b) Heat map for hierarchically clusters genes (x) and mice (y).

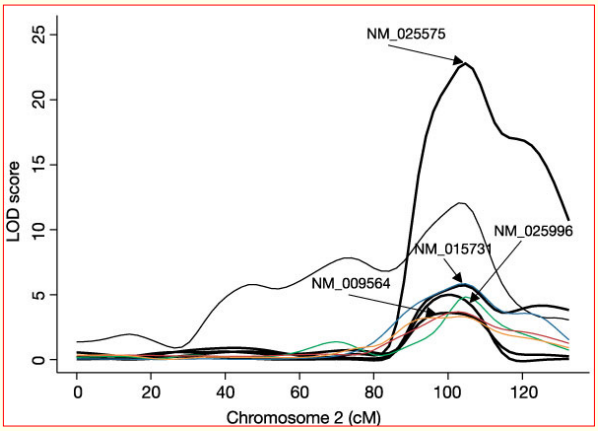
(c,d) Chromosomes 2 and 19 LOD score profiles for the FPM trait.

- Clustering of eQTL to specific loci, and the relationship between these genes with respect to expression, lead to highly significant and interesting patterns that can be associated with phenotypes related to common diseases.
- In general, cis-acting eQTL present higher LOD scores than trans-acting ones.



eQTL LOD scores of four genes: (1) black curve: C5; (2) red curve: Alad; (3) blue curve: St7; and (4) green curve: Nnmt. The alternating green and blue strip at the base of the curves represents chromosomes (1 to 19) boundaries. Source: Schadt et al. (2003).

- eQTL hotspots and association with clinical (quantitative) traits.



LOD score curves for four obesity-related traits (blue curve: subcutaneous FPM; green curve: perimetrial FPM; red curve: omental FPM; orange curve: adiposity; thin black curve: joint multivariate analysis for the four traits) and four candidate genes (thick black curves)

FINAL REMARKS

- Combining gene expression profiling, molecular markers data and phenotypic scores in a segregating population can help identifying the drivers of the pathways or the causal factors underlying phenotypic variation.
- Genetical genomics is a powerful tool for fine mapping and candidate gene discovering
- Advance in technology will allow the (faster and cheaper) screening of large sample sizes.
- Multidisciplinary and interdisciplinary area of research.

REFERENCES

Doerge, R. W. (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nature Reviews Genetics 3 (1): 43-52.

Darvasi, A. (2003) Gene expression meets genetics. Nature. 422, 269-270.

Jansen, R. C. and J. P. Nap (2001) Genetical genomics: the added value from segregation. Trend Genet. 17, 388-391.

Jansen, R. C. (2003) Studying complex biological systems using multifactorial perturbation. Nature Reviews 4, 145-151.

Schadt, E. E. et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302.