Expression QTLs and Mapping of Complex Trait Loci

Paul Schliekelman
Statistics Department
University of Georgia

Definitions: Genes, Loci and Alleles

- A gene codes for a protein. Proteins due everything.
- A gene position on a chromosome is called a locus.
- The alternate forms of the gene at a locus are called alleles.

Definitions, continued:

- A causative locus for a trait is a locus with DNA sequence variation between individuals that contributes to variation in the trait between them.
- A quantitative trait locus (QTL) is a causative locus for a quantitative trait (e.g. height, IQ).

Definitions: Genetic Linkage

- Genes on the same chromosome tend to be inherited together. This is called **linkage.**
- The closer together two genes lie, the more often they are inherited together.

Genetic Markers

- A **genetic marker** is a known and detectable variation in a gene or other stretch of DNA.
- The presence of a marker can be detected in an individual organism's genome and its inheritance can be followed.
- A marker that is linked to a locus affecting a trait will appear more often in individuals with that trait.

Gene Mapping

- Goal of gene mapping is to locate genetic loci that are responsible for variation in traits of interest (e.g. occurrence of disease, intelligence).
- Based on searching for correlations between markers and trait of interest.

Gene Mapping Successes

- Mendelian traits have inheritance patterns consistent with a single causative locus.
- Hundreds of genes affecting known Mendelian traits (e.g. cystic fibrosis, breast cancer) in humans have been mapped.
- Some cases have led to new treatments and/or screening methods.

Gene Mapping Failures

- Complex traits do not follow simple one-locus Mendelian expectations.
- Assumed to be caused by mutations at multiple loci.
- Examples include many common diseases: asthma, bipolar disorder, diabetes, prostate cancer, etc.

Mapping of complex disease genes – little success

Altmuller *et al,* 2001: Reviewed 101 full genome scans of 31 complex diseases.

- 67% did not show significant linkage to any marker.
- of the significant linkages, very few have been reproduced.

Low power for mapping complex trait loci

- Power to detect genes decreases as the number of loci affecting the trait increases.
- By definition, complex diseases should be harder to map.
- We don’t know how many genes underlie complex diseases, so don’t know whether to be surprised by lack of success.

The central problem of human disease genetics is to solving this dilemma.

New Buzzword:

Genetical Genomics: the combination of molecular marker data and genome-wide expression data to elucidate the genetics of complex traits

Expression QTLs (eQTLs)

- Gene expression levels can be treated as quantitative traits and their quantitative trait loci (QTLs) mapped.
- Transcript: gene whose expression level is being measured.
- An eQTL for that transcript refers to a genetic locus with DNA sequence variation causing variation in the expression level.

Expression QTLs (eQTLs)

- The eQTL could be at the transcript locus itself, or elsewhere in the genome.
- By combining microarrays and marker data, eQTLs can be simultaneously generated for thousands of gene expression levels

Recent eQTL surveys

- Numerous recent eQTL surveys have found that expression levels frequently exhibit intermediate to high heritabilities.

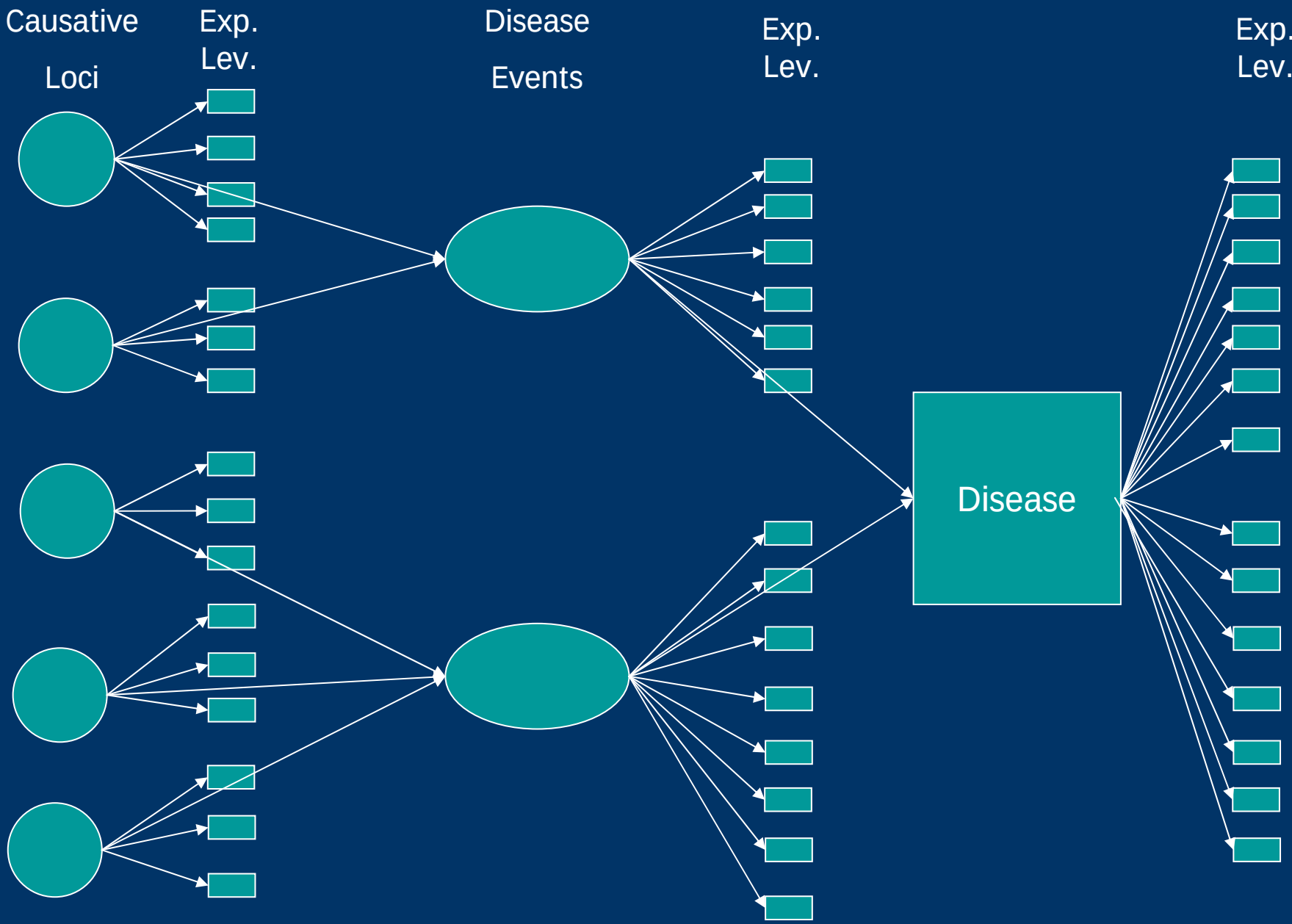
- E.g. Schadt *et al* 2003 found 4339 eQTLS over 3701 genes with log-of-odd scores>4.3 and 11,021 genes with eQTL LOD scores>3.0

Recent eQTL surveys

- Thus, it is relatively easy to map QTLs for many expression levels.

Will eQTLs be useful for dissecting complex physiological traits (e.g. disease)?

- Presumably, disease causative loci are also eQTLs for some gene expression levels.
- Two task must be accomplished:
  1. Show that transcript is related to disease.
  2. Map eQTL for transcript. Then this eQTL should be causative locus for disease.


Causative
Loci
Exp.
Lev.
Disease
Events
Exp.
Lev.
Exp.
Lev.
Disease

Model for Gene Expression

Assume

- L disease causative loci.
- M expression levels ($X_1 \ldots X_M$) tested.
- Each expression level X_i depends on some subset of the L causative loci (often empty).
- T = indicator variable for disease status.

Model for Gene Expression

$$P(X \mid T=1) = \frac{P(T=1 \mid X)P(X)}{P(T=1)}$$

$$= \frac{\sum_{G} P(T=1 \mid G)P(X \mid G)P(G)}{K}$$

$K = Disease \;\; prevalence$

Expression Levels

- Assume expression levels are normally distributed on some scale (e.g. log scale).
- Mean and variance determined by genotype at one or more of the L causative loci.
- Expression level distribution is then a mixture of normal distributions with weights determined by genotype distribution conditioned on disease status.

Multiplicative Model

$$P(T=1|G)=P(T_1=1|G_1)P(T_2=1|G_2)....P(T_L=1|G_L)$$

Can show:

$$P(X|T=1)=\frac{\sum_{g_1,g_2,...g_c}P(X|g_1,g_2,...g_c)u_1(g_1)u_2(g_2)....u_c(g_c)P(g_1)P(g_2)...P(g_c)}{K_1K_2...K_c}$$

$$P(X|T=0)=$$

$$\frac{\sum_{g_1,g_2...g_c}P(X|g_1,g_2,...g_c)\left(1-u_1(G_1)u_2(G_2)...u_c(G_c)K_{c+1}...K_L\right)P(g_1)...P(g_c)}{1-K}$$

$$Where\ K_i=\sum_{G_i}P(T_i=1|G_i)P(G_i)\ and\ u_i(G_i)=P(T_i=1|G_i)$$

If *c*=1 (expression controlled by one locus) and Hardy-Weinberg Equilibrium (alleles independent):

$$P(X \mid T=1) = (1-p_1)^2 u_1(dd)\varphi(X \mid \mu_{dd}, \sigma_{dd}) + 2p_1(1-p_1)u_1(Dd)\varphi(X \mid \mu_{Dd}, \sigma_{Dd}) + p_1^2 u_1(DD)\varphi(X \mid \mu_{DD}, \sigma_{DD})$$

Where p_1 is the allele frequency

D=disease allele and *d*=“normal” allele

$\phi(X \mid \mu_{DD}, \sigma_{DD})$ is normal pdf

Additive Model

$$P(T=1|G)=P(T_1=1|G_1)+P(T_2=1|G_2)+....+P(T_L=1|G_L)$$

Can show:

$$P(X|T=1)=\frac{\sum_{G_1}P(X|G_1)(u_1(G_1)+K_2+...K_L)P(G_1)}{K}$$

$$P(X|T=0)=\frac{\sum_{G_1}P(X|G_1)(1-u_1(G_1)-K-_2K_3...-K_L)P(G_1)}{1-K}$$

Expression Level Mean

- Expression level mean assumed to have either multiplicative or additive dependence on genotype:

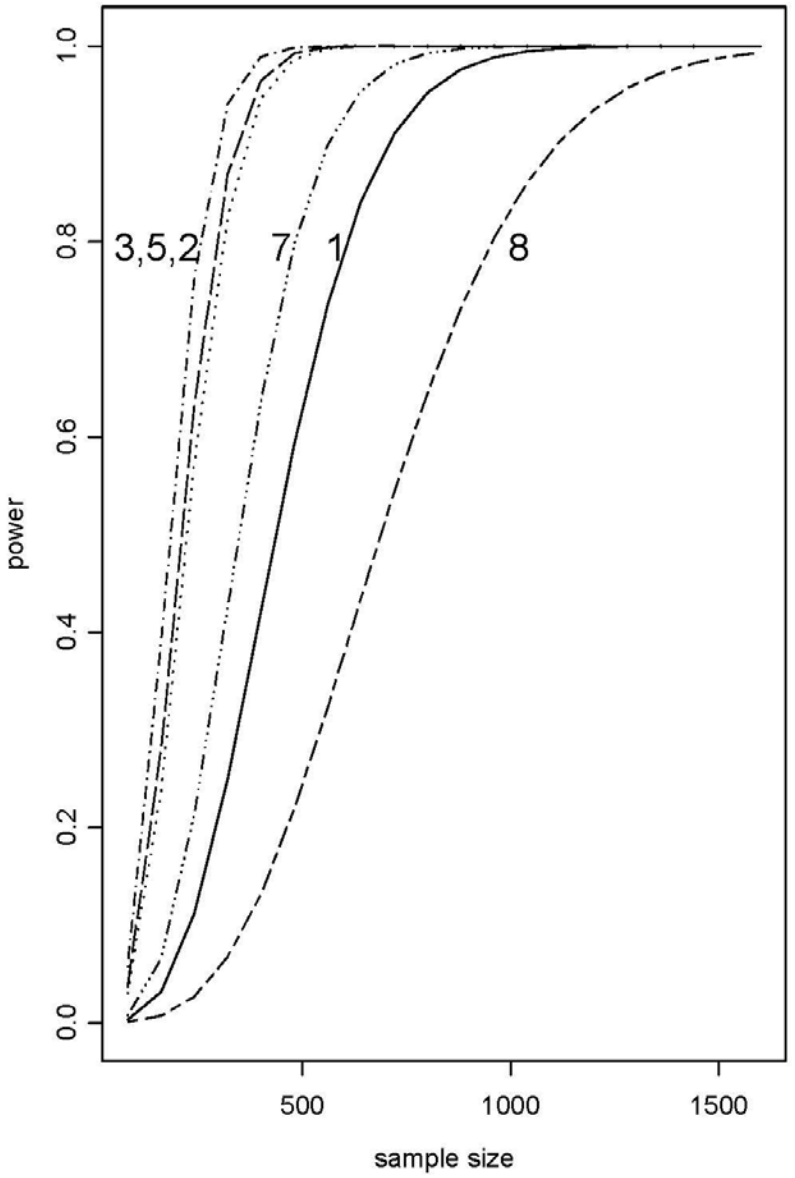
$$\mu(g_1,\ldots,g_c) = \mu_1(g_1)\mu_2(g_2)\ldots\mu(g_c)$$

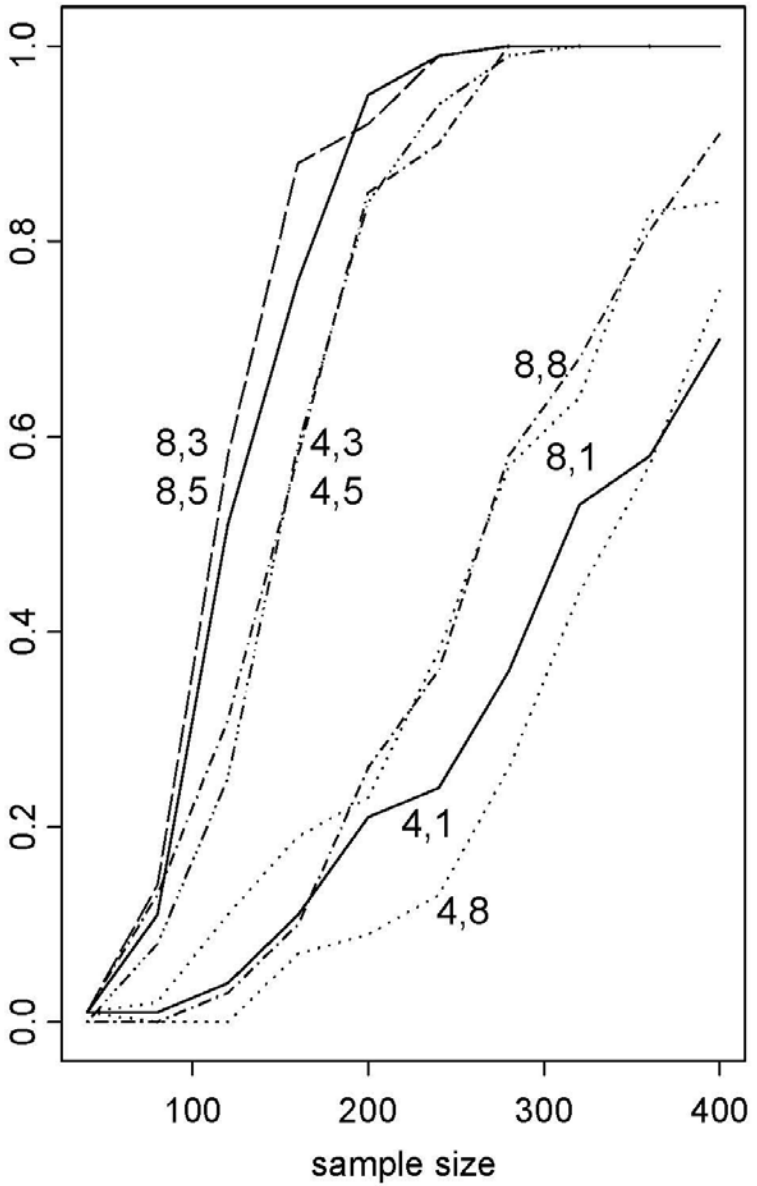
or

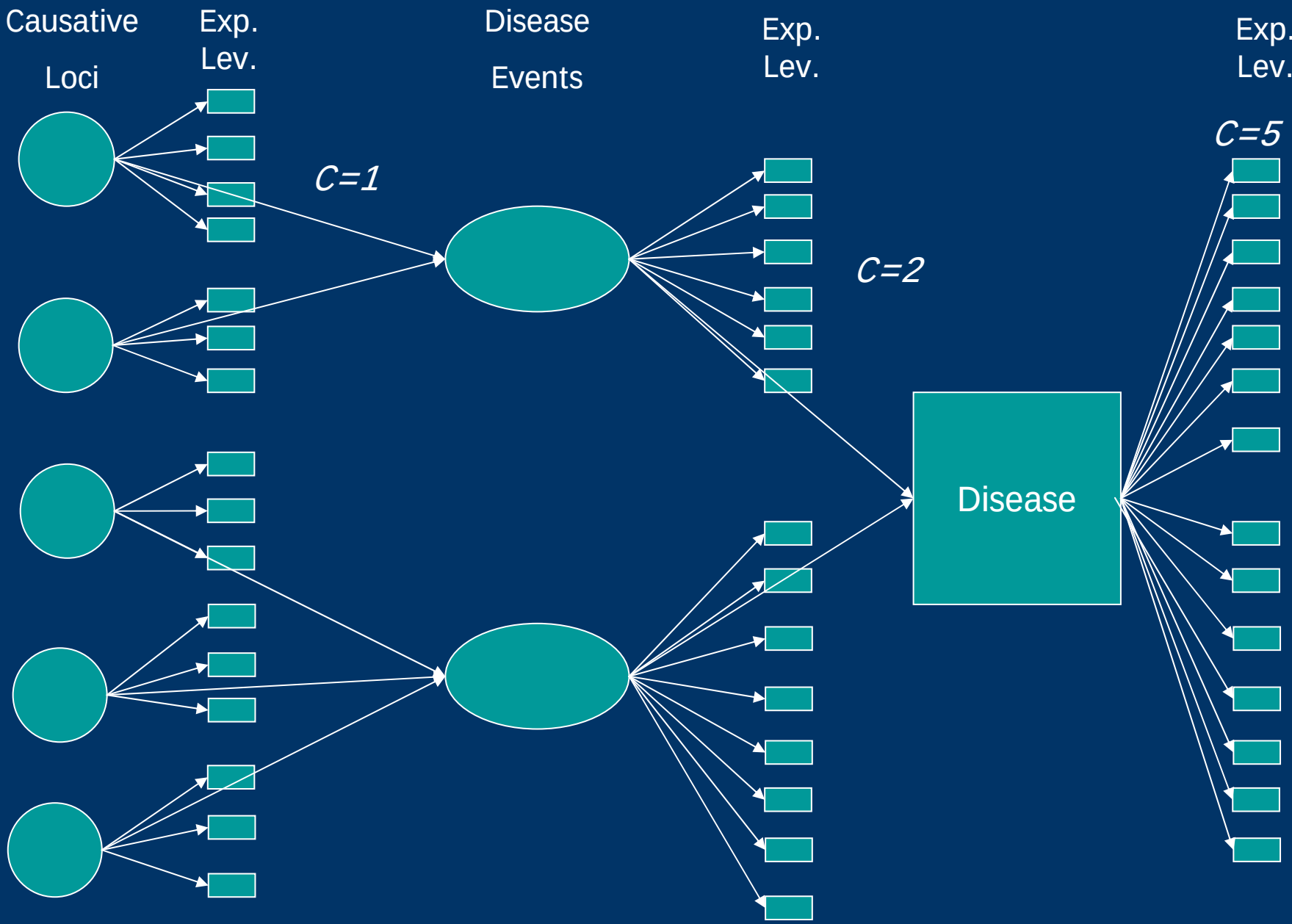
$$\mu(g_1,\ldots,g_c) = \mu_1(g_1)+\mu_2(g_2)+\ldots+\mu(g_c)$$

Power Calculations

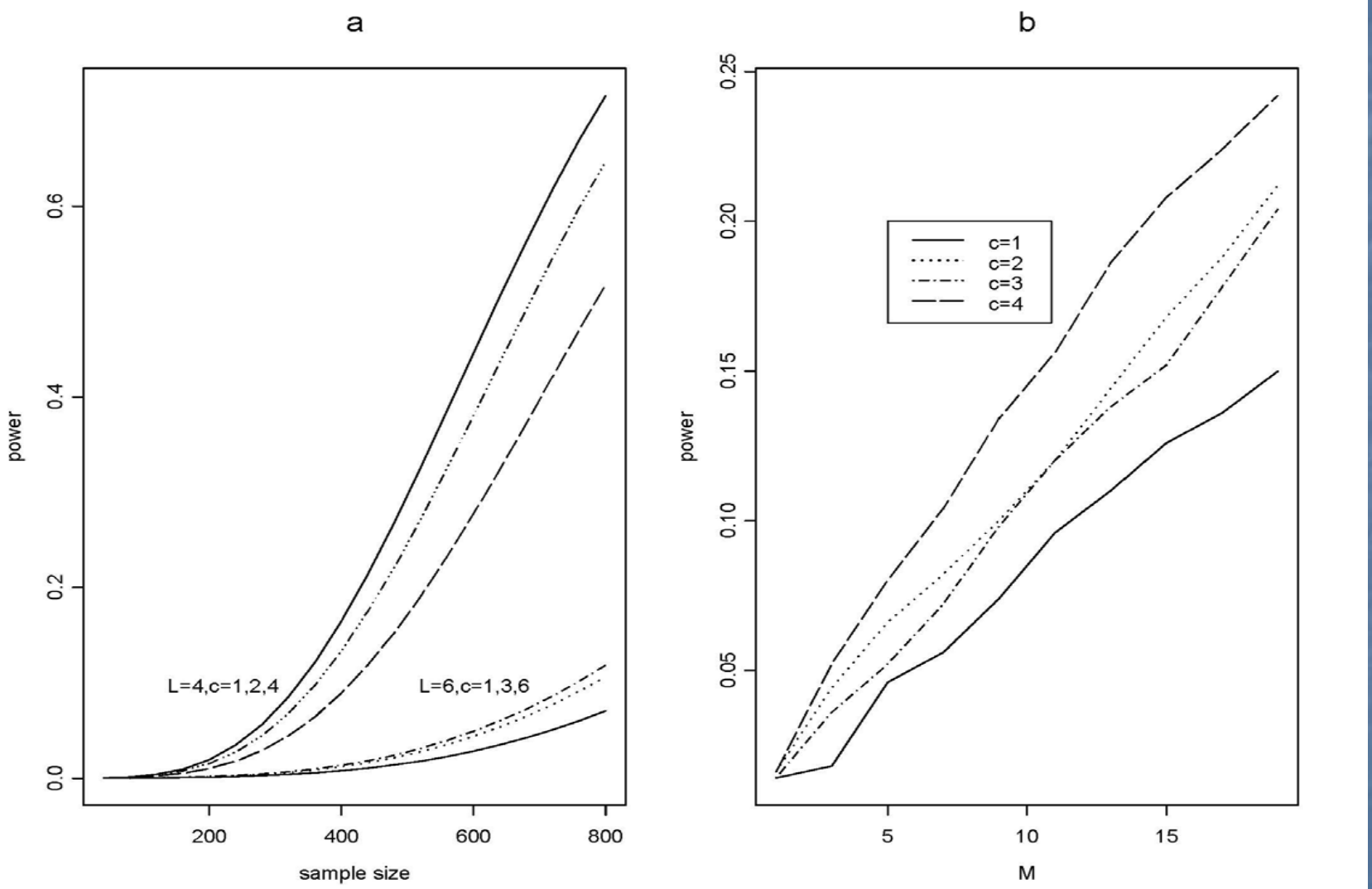
- Two groups: disease affected and unaffected.
- t-test for group difference conducted for each gene.
- Calculate power to detect differential expression.
- Interested in relationship between power and genetic model.


power
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
sample size
3,5,2
7
1
8


0.0
0.2
0.4
0.6
0.8
1.0
100
200
300
400
sample size
8,3
8,5
4,3
4,5
8,8
8,1
4,1
4,8


Causative
Loci
Exp.
Lev.
C=1
Disease
Events
Exp.
Lev.
C=2
Disease
Exp.
Lev.
C=5

Additive Model



Detecting expression level differences: Conclusions

- Power to detect expression level differences is poor for a multiplicative model with c=1, but reasonable for c=2-5.
- Power to detect expression level differences is very poor for an additive model.

Power for Mapping eQTLs

Power for mapping QTLs in natural populations deteriorates quickly as the number of QTLs increases. Power poor even for $c=2$.

Conclusions – Natural Populations

- Power to detect expression level differences extremely poor if disease probability is additive.
- For multiplicative model there is no value of *c* (number of controlling loci) where power is good for detecting both linkage and expression level differences.

Many Unknowns

- Distribution of relationships between expression level and causative loci.
- Dominance relationships.
- Form of gene interactions.
- Allele frequencies.

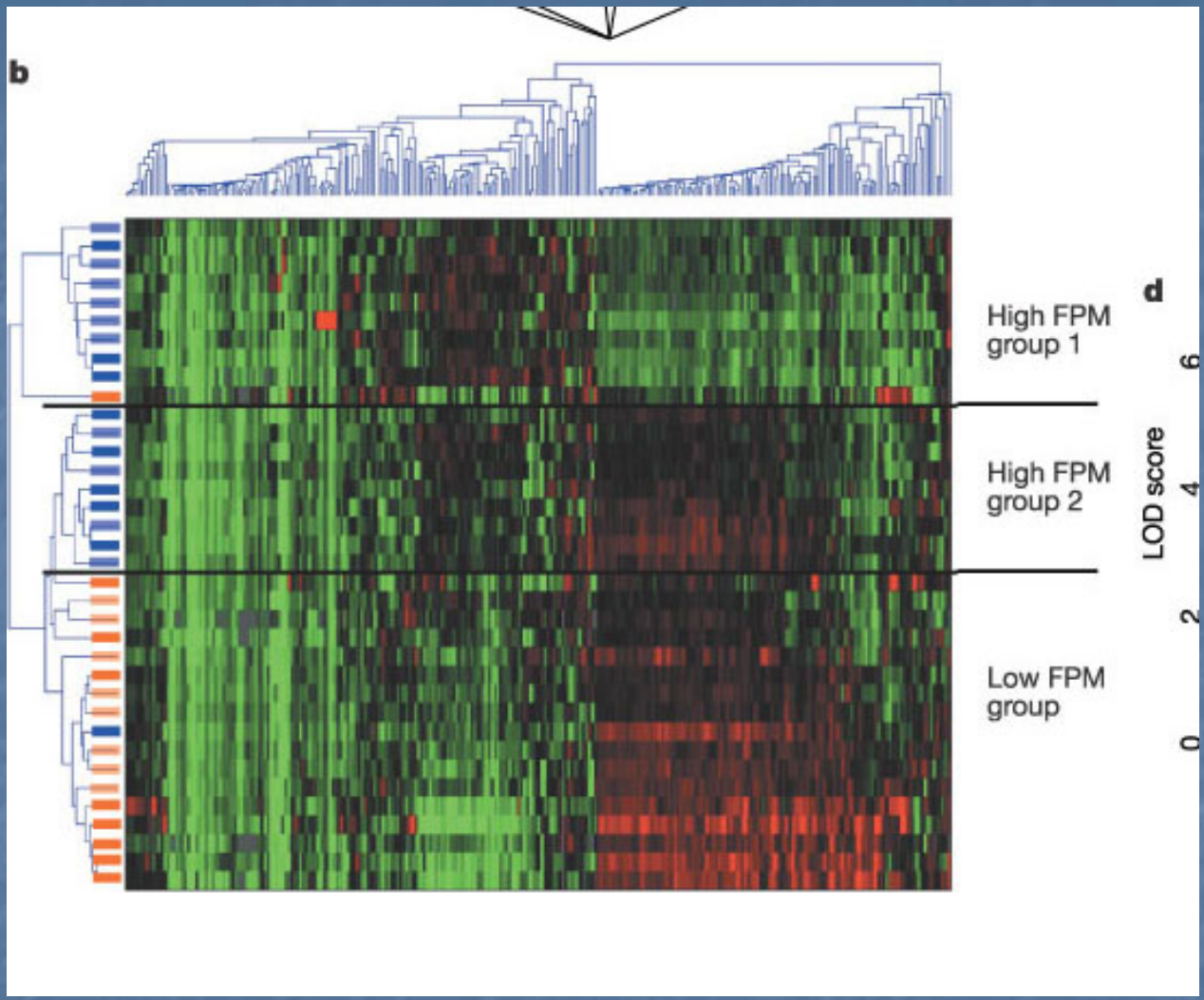
Much Potential For Improved Power

- There is much information in the joint behavior of expression levels.
- Better experimental designs are possible (e.g. utilize family structure).

How can we use joint expression information?

Schadt et al 2003:

- Crossed two mouse strains: one susceptible to diabetes and weight gain on high fat diet and the other not.
- 111 offspring from cross were put on high fat diet for 4 months.
- Liver tissue from each was profiled using microarrays.
- Genotypes for several hundred markers were obtained for each mouse.



From Schadt *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302.

Results:

- Microarray data splits high FPM data into two groups.
- Gene mapping comparing separate high-FPM groups to low-FPM produced significant linkage for a obesity locus.
- Gene mapping comparing combined high-FPM groups to low-FPM groups do not find significant linkage.

Genetic Heterogeneity

	Affected Individuals													
Locus	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1		x				x								x
2				x				x	x			x		
3					x		x							
4			x								x			
5				x	x			x						
6										x				x

Major variation in which disease mutations are carried by affected individuals

	Affected Individuals													
Locus	**2**	**6**	**14**	**4**	**8**	**9**	**12**	**5**	**7**	**3**	**11**	**13**	**10**	**1**
1	x	x	x											
2				x	x	x	x							
3								x	x					
4										x	x			
5				x	x			x				x		
6			x										x	x

Goal: A method for identifying the genetic heterogeneity using genomewide expression data

Gene mapping power could be dramatically improved if this heterogeneity could be accounted for

Proposed EM Algorithm Based Method:

Data:

- N_A affected and N_U unaffected individuals.
- Microarray for each individual
- Disease status for each individual

		Transcript					
Sample Indiv.	Disease Status	1	2	3	4	5	6
1	Y_1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
2	Y_2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
3	Y_3	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}
4	Y_4	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	X_{46}
5	$Y\text{-}_5$	X_{51}	X_{52}	X_{53}	X_{54}	X_{55}	X_{56}

Unobserved Variables

Z_i=controlling locus for transcript i, i=1 to M

V_{jk}=genotype on controlling locus j in individual k.

Controlling loci 1 to L are the disease loci.

Controlling loci L+1 to $L+c$ are "null loci" that do not affect disease.

More specific goal:

Cluster sample individuals by their values of V_{jk} (that is, their genotype on the disease loci.).

Likelihood

$$L\left(\vec{\theta} \mid \vec{X}, \vec{Y}, \vec{Z}, \vec{V}\right) =$$

$$P\left(\vec{X} \mid \vec{Z}, \vec{V}, \vec{\theta}\right) P\left(\vec{Y} \mid \vec{V}, \vec{\theta}\right) P\left(\vec{Z}, \vec{V} \mid \vec{\theta}\right)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{M} P\left(X_{ij} \mid Z_j, V_{iZ_j}, \vec{\theta}\right) \prod_{k=1}^{N} P\left(Y_k \mid \vec{V}_k, \vec{\theta}\right) P\left(\vec{Z}, \vec{V} \mid \vec{\theta}\right)$$

$$P\left(X_{ij} \mid Z_j, V_{iZ_j}, \vec{\theta}\right) = \textit{Expression level distn. for gene } j \textit{ in indiv. } i$$

$$P\left(Y_k \mid \vec{V}_k, \vec{\theta}\right) = \textit{Disease probability for indiv. } k$$

$$P\left(\vec{Z}, \vec{V} \mid \vec{\theta}\right) = \textit{Controlling locus and genotype probs.}$$

Parameter Estimates:

Expression level mean for transcript *a* with controlling locus genotype *b*:

$$\hat{\mu}_{ab} = \frac{\sum_{t=1}^{L}\sum_{i=1}^{N} x_{ia} P\left(Z_a = t, V_{it} = b \mid \vec{X}, \vec{Y}, \vec{\theta}_0\right)}{\sum_{t=1}^{L}\sum_{i=1}^{N} P\left(Z_a = t, V_{it} = b \mid \vec{X}, \vec{Y}, \vec{\theta}_0\right)}$$

Expression level variance for transcript *a* with controlling locus genotype *b*:

$$\hat{\sigma}^2_{ab} = \frac{\sum_{t=1}^{L}\sum_{i=1}^{N}\left(x_{ia}-\mu_{ab}\right)^2 P\left(Z_a = t, V_{it} = b \mid \vec{X}, \vec{Y}, \vec{\theta}_0\right)}{\sum_{t=1}^{L}\sum_{i=1}^{N} P\left(Z_a = t, V_{it} = b \mid \vec{X}, \vec{Y}, \vec{\theta}_0\right)}$$

Probability that the expression of gene *r* has controlling locus *q*)

$$\hat{\alpha}_{rq} = P\left(Z_r = q \mid \vec{X}, \vec{Y}, \vec{\theta}_0\right)$$

Probability that controlling locus *s* has genotype *t*

$$\hat{\lambda}_{st} = \frac{\sum_{j=1}^{N} P\left(V_{js} = t \mid \vec{X}, \vec{Y}, \vec{\theta}_0\right)}{N}$$

Disease penetrance parameters

The genotype specific risk parameters are found by numerical maximization.

Assigning Genotypes:

$$P\left(v_{ij}=k \mid Y_i, \vec{W}_{ij}\right)=\frac{P\left(Y_i \mid v_{ij}=k\right)P\left(\vec{W}_{ij} \mid v_{ij}=k\right)}{P\left(Y_i, \vec{W}_{ij}\right)}$$

$\vec{W}_{ij}$ = vector containing the values of expression levels in individual i that are controlled by locus j.

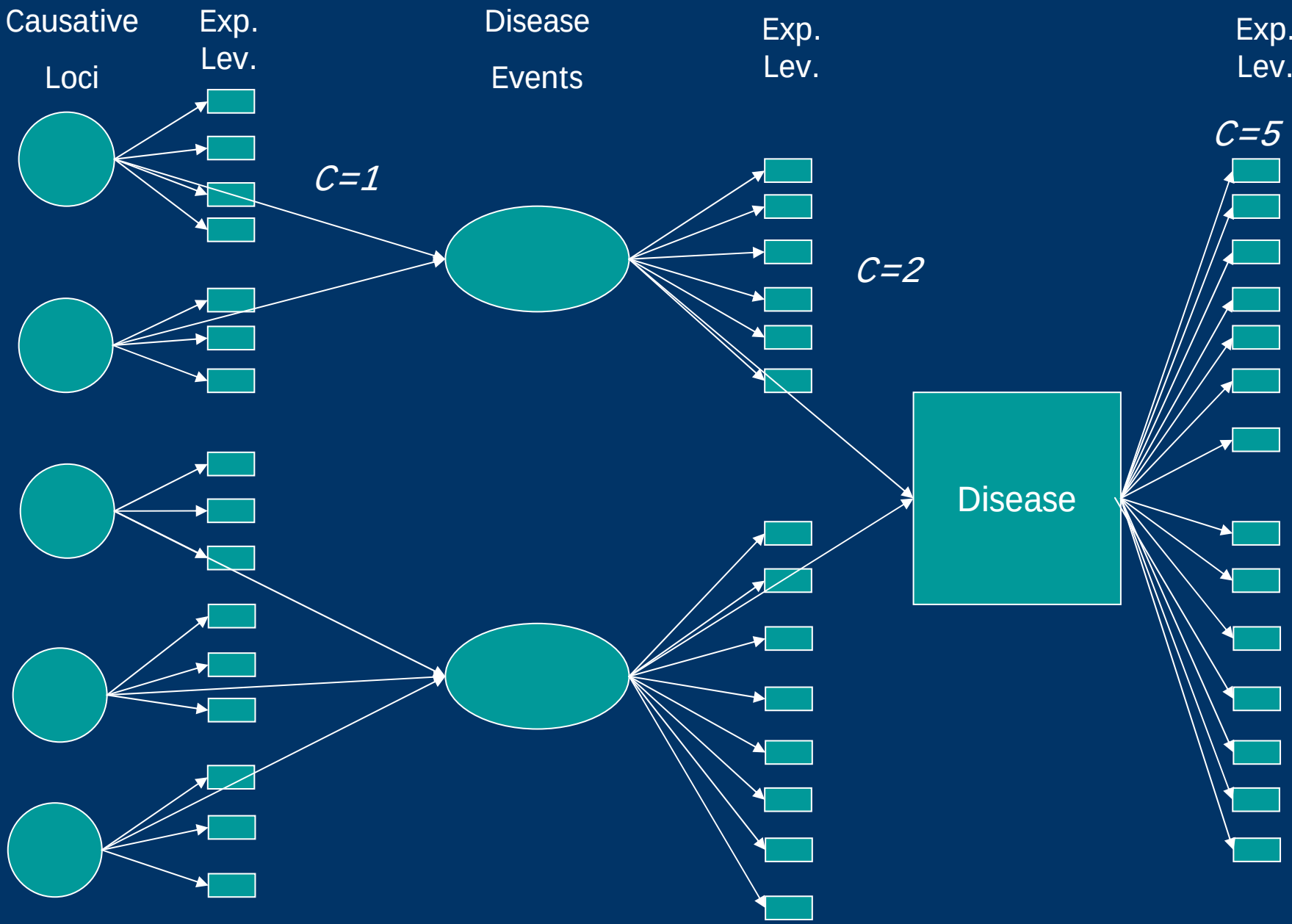
- Transcripts assigned to same controlling locus if their expression is correlated in a way that is consistent with genetic model.
- Sample individuals assigned to same genotype at controlling locus if controlled expression levels are similar.

How well will it work?

- Potentially much information about genotype available in expression levels.

On the other hand:

- Highly assumption laden.
- Many parameters to estimate.
- High-level transcripts may overwhelm analysis.


Causative
Loci
Exp.
Lev.
C=1
Disease
Events
Exp.
Lev.
C=2
Disease
Exp.
Lev.
C=5

Mapping Power - multiplicative