Multiplicity Issues in Statistical Genetics

Yoav Benjamini

Tel Aviv University

www.math.tau.ac.il/~ybenja

Based on joint works

with

D. Yekutieli, A. Reiner,

and

N. Lee, N.Letwin, G. Elmer, N. Kafkafi

With the help of

R. Yakobov, R. Gutman, A. Sakov, D. Poder

Work partially funded by F.I.R.S.T foundation

Outline

Introduction

QTL analysis

- History
- Recent Advancements

Gene Expression

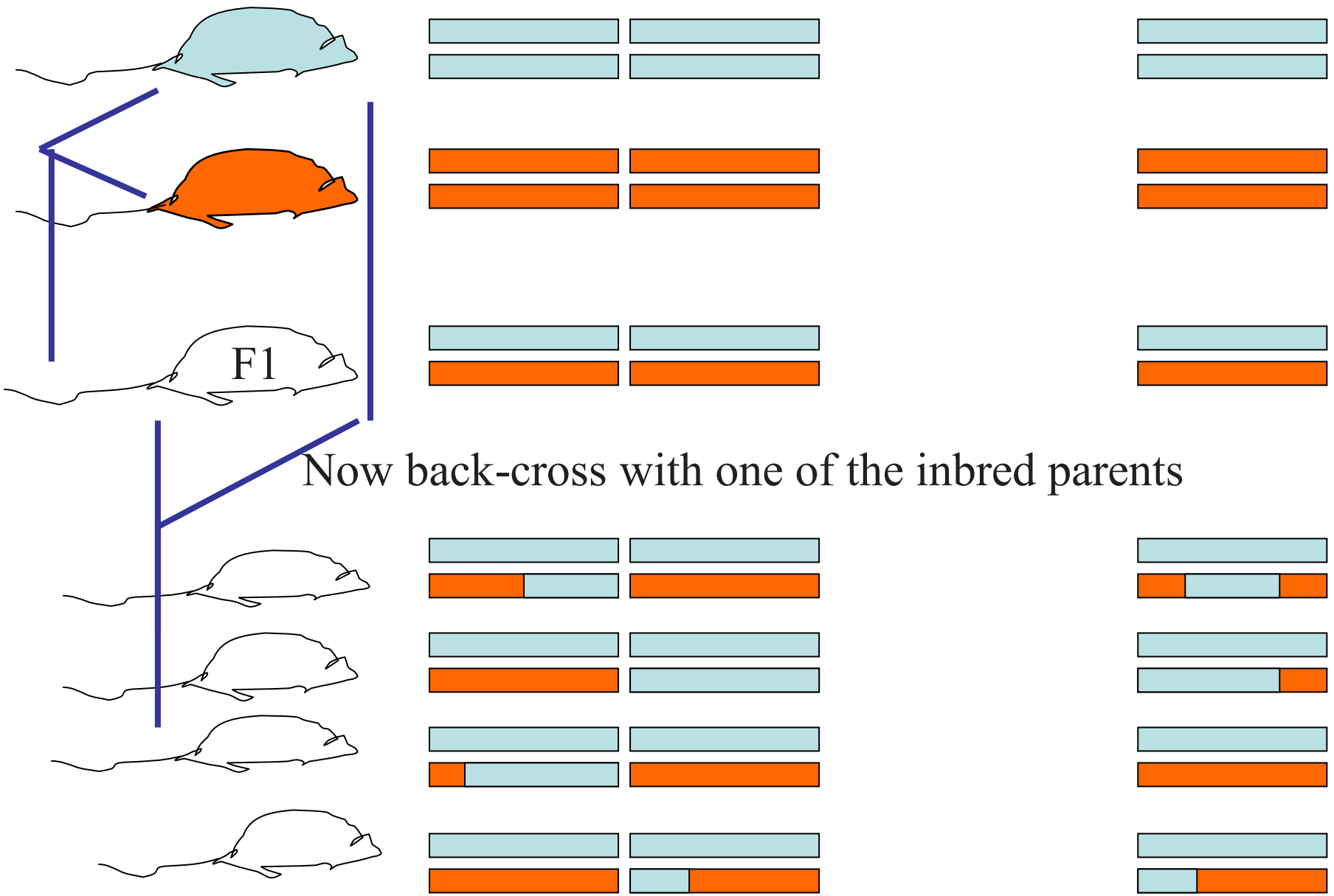
- History
- Recent Advancements

Moral of Stories

QTL analysis of complex traits in backross

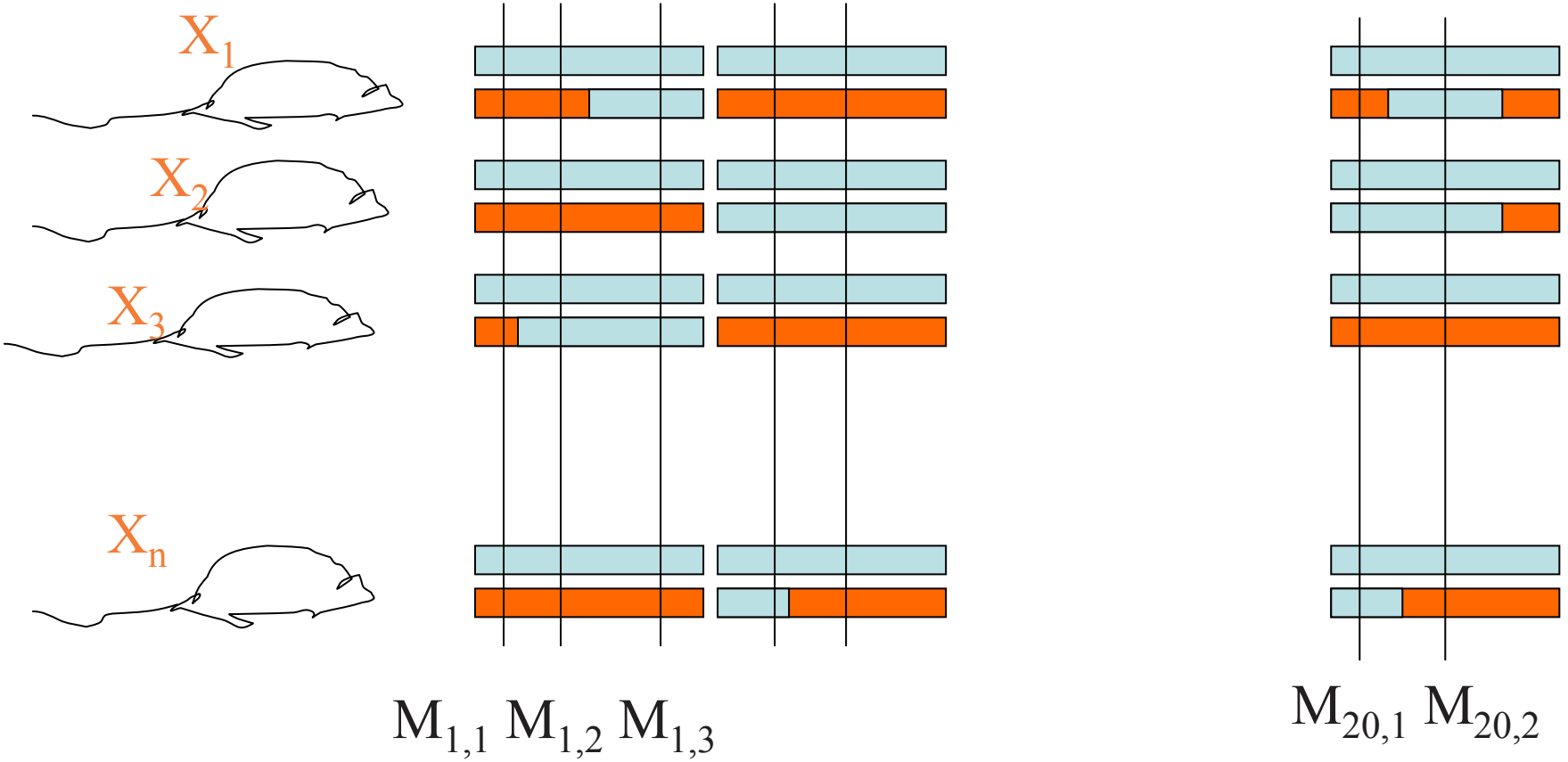
- **Q**uantitative **T**rait **L**oci (**QTL**) – finding genetic loci affecting quantitative traits.
- Generalized to the study of complex traits
- See slides following Science '94 illustration

Quantitative Trait Loci (QTL) Analysis



Quantitative Trait Loci (QTL) Analysis

On each mouse measure the trait



For each marker test linkage of trait to locus by (X-X)/s

Thresholding via multiple significance testing.

For each QTL (potential location on the genome)

construct a significance test for –

H_0: no linkage vs. H_1: Linkage

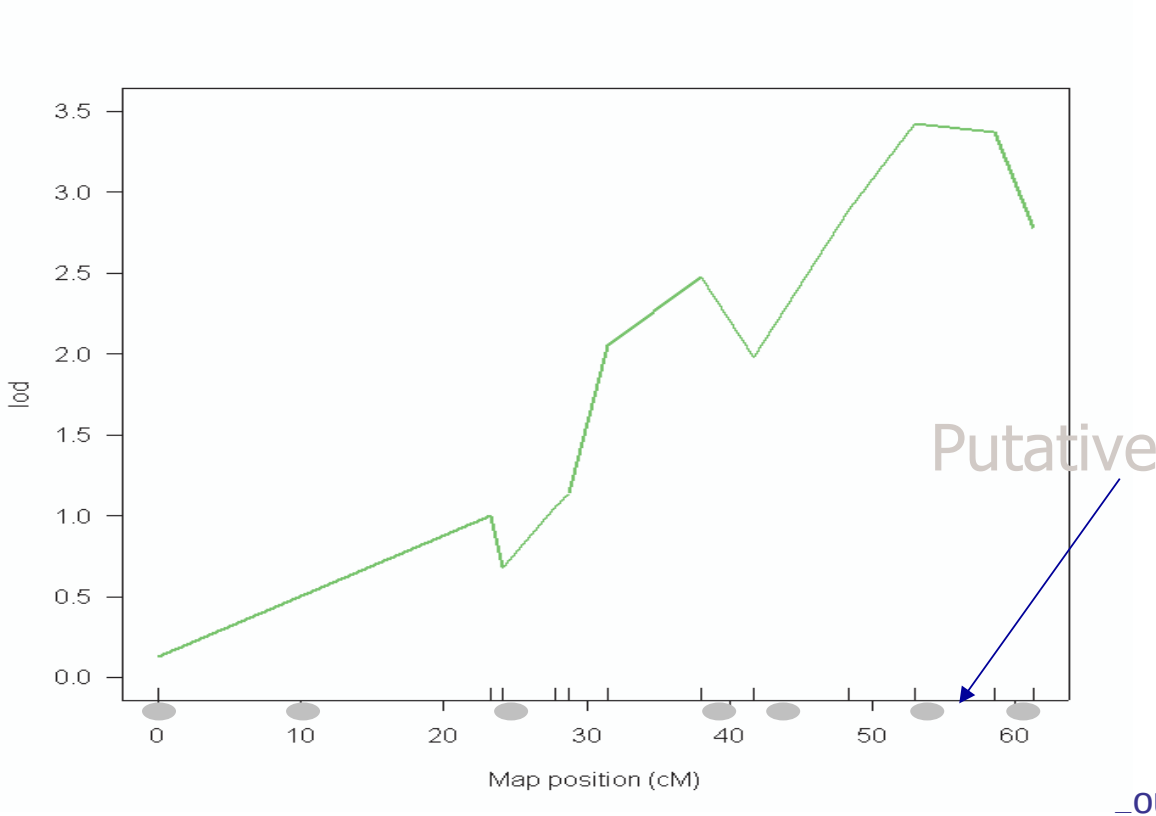
(false discovery = type I error)

For each trait and marker compute a **L**og **OD**ds score to test for linkage with trait:

.

1 trait

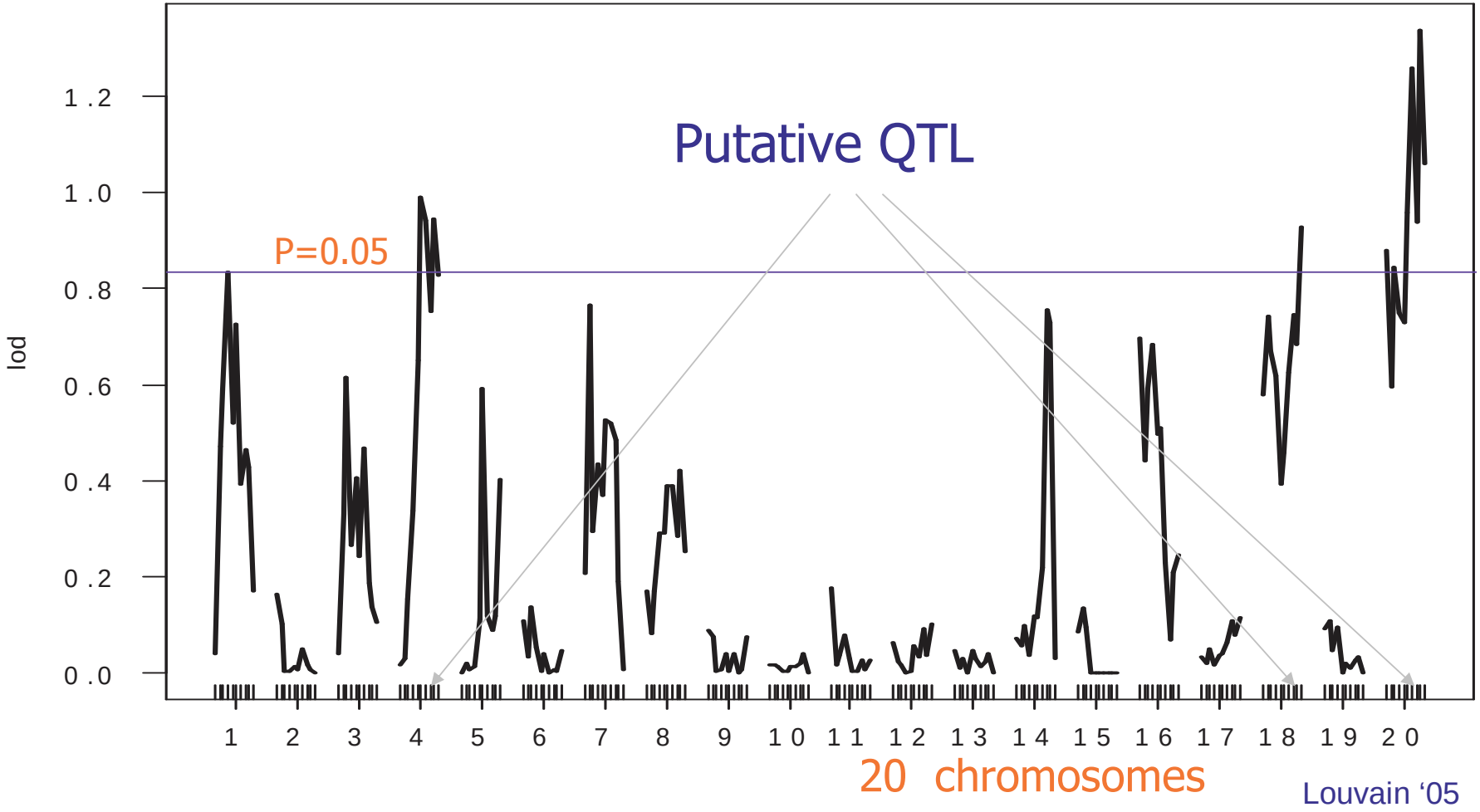
1 chromosome



_ouvain '05

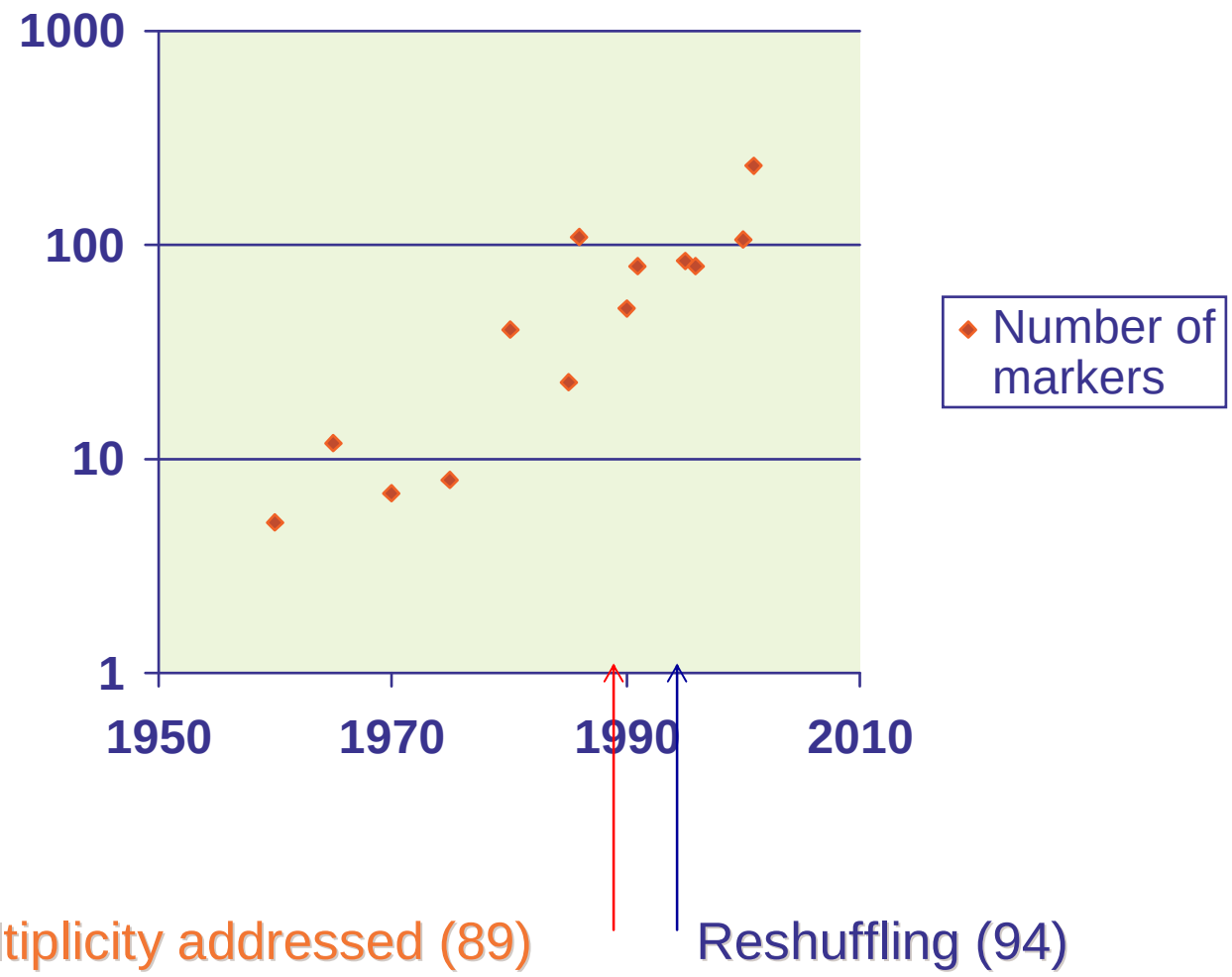
Testing each hypotheses at 0.05

Simulated genome scan- 1 trait, 20 chrom. No QTL exists



Louvain '05

QTL Analysis



Multiplicity as a threat on replicability

- Lander & Botstein (Genetics ‘89)

Proposed a threshold based on the theoretical limit distribution of the maximum of highly dense markers

- Churchill & Doerge (Genetics ‘94)

Proposed a threshold based on the re-randomization distribution of the maximum of the actual markers

- Lander & Kruglyak (Nat. Genetics ‘95)

“Genetic dissection of complex traits: guidelines for interpreting…”

- “Genetic dissection of complex traits: guidelines for interpreting…” Lander and Kruglyak
- “Adopting too lax a standard guarantees a burgeoning literature of false positive linkage claims, each with its own symbol… Scientific disciplines erode their credibility when substantial proportion of claims cannot be replicated…”

(… i.e. when the False Discovery Rate was too high!)

"Genetic dissection of complex traits: guidelines for interpreting…" Lander and Kruglyak '95

- *Significant QTL*: a QTL which is significant at the .05 FWE level in a genomwise scan.
- *Confirmed QTL*: A significant QTL which was confirmed in an independent study using nominal 0.05 level.

- *Suggestive QTL*: a QTL which is significant at a level, which will allow on the average one significance even if no QTL exists.

Suggestive QTL is equivalent to controlling at approximately 0.6 (~half!) FWE level.

Why report suggestive QTLs?

Why report suggestive QTLs?

“On the other hand, adopting too high a hurdle for reporting results runs the risk that nascent field will be stillborn.”

…i.e. to overcome loss of power

Our suggestion

Take seriously "…when substantial proportion of claims…"

Use the False Discovery Rate (FDR) as a criterion to set thresholds:

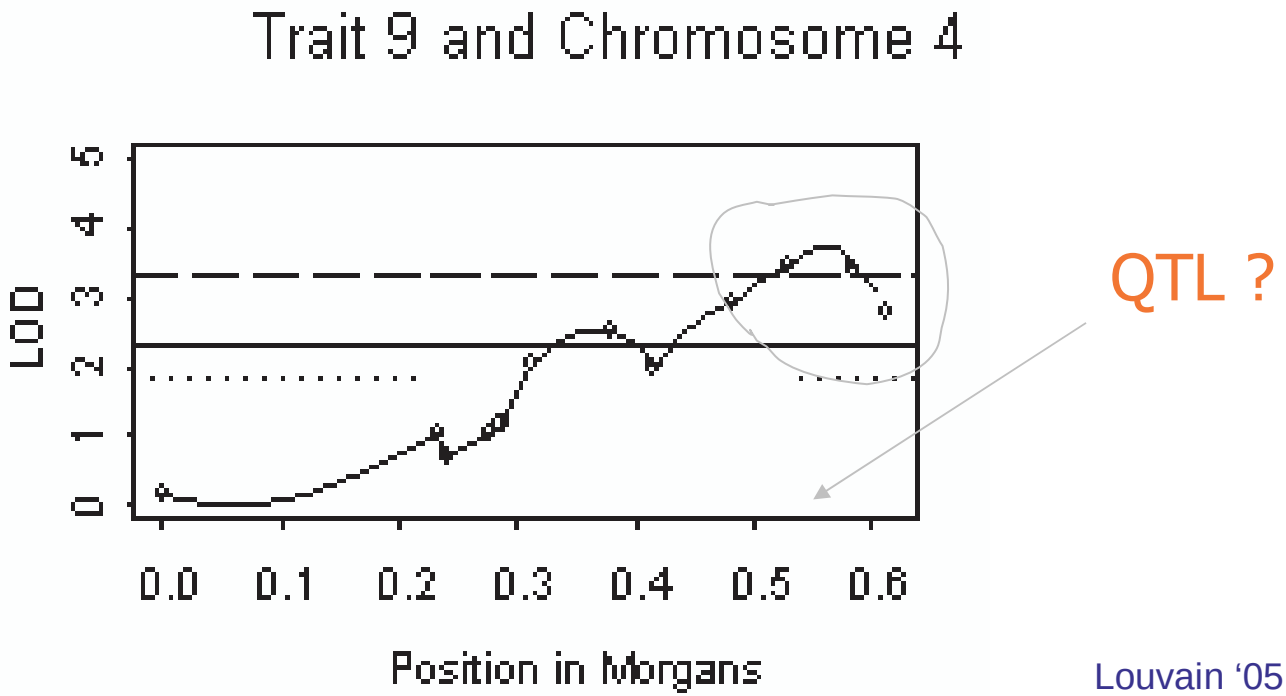
Use the linear stepup procedure (BH) to control the FDR

Analyzing QTL data: the gain in power

FWE = .05

Using the Linear Step-up procedure FDR =.05

Unadjusted = .05



The suggestion was hastily followed

- Weller et al (Genetics 1998) introduced the BH procedure to genome scanning. Some of their claims:
- "The BH procedure is always valid even under dependency"
- "An accurate prediction has been made of the proportion of hypotheses rejected in the first experiment that represent true effects"
- They discuss FDR as E(V)/R i.e. $mp_{(r)}/r$

The debate

- Zaykin, Young & Westfall (Genetics 2000) argued with Weller et al about the conditional interpretation

$$FDR = E(V/R) \leq E(V/R \mid R>0)$$

The latter is the positive FDR.

So far so good, but they also argued in favor of FWE:

- Weller responded that P(R>0) ~ 1
- Mosig et al (Genetics '01) offered $m_0 p_{(r)} / r$ *as a new*

"adjusted FDR criterion"

and claimed this should help to make Weller's point

Further claims

“ Because it is conditional on some true proportion of tests for which the nulls are false, we believe the FDR calculated that way is not subject to the critique of Zaykin et al”.

Note: this debate goes on in the Genetics literature, not in the Statistics literature

Suggested FDR guidelines for QTL analysis

- *FDR-Significant QTL*: a QTL which is significant at the .05 FDR level in a genomwise scan.
- *FDR-Confirmed QTL*: A FDR-significant QTL which was confirmed in an independent study using 0.05 FDR level. (overall level is 0.0025)
- *Suggestive QTL*: a QTL which is significant at a the 0.1 FDR level.

  Entirely drop " FWE Suggestive QTL"

Suggestive linkages can become FDR significant using a second study

Some practical issues:

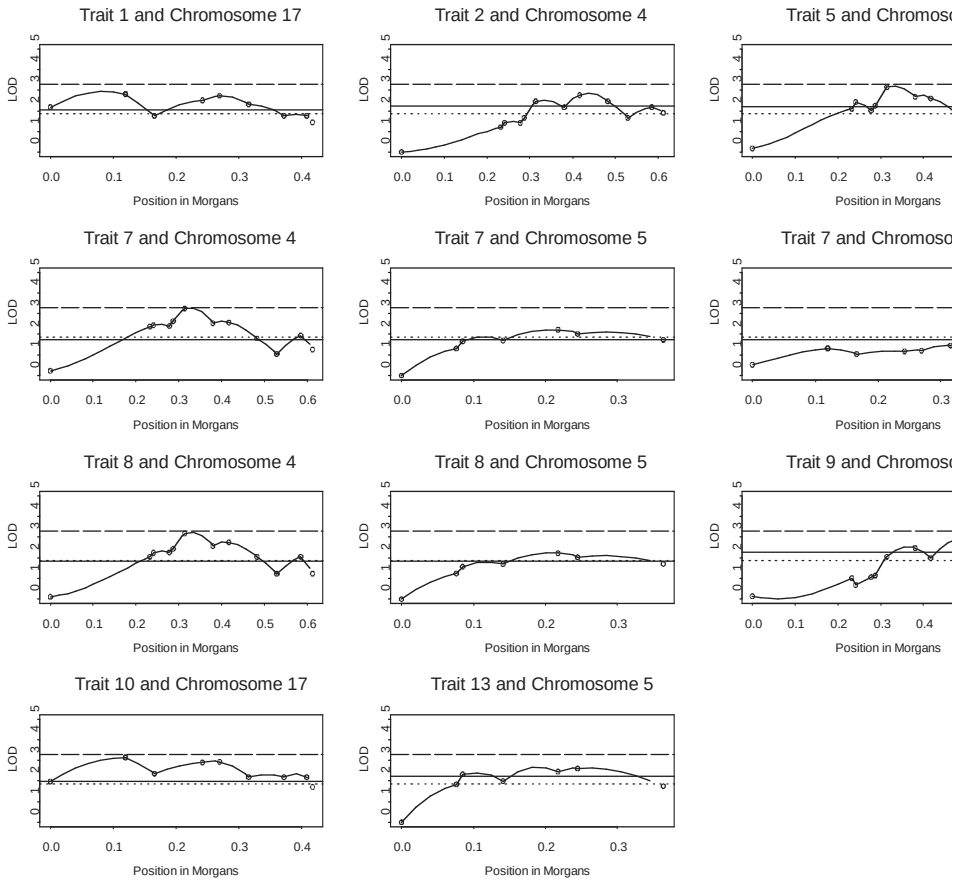
- Most mapping programs produce output that includes lod scores (MAPMAKER, GENEHUNTER).
- These can be easily inverted to p-values and BH can be applied

  (consult FDR homepage for Splus function)

- We have implemented FDR calculations into QTL cartographer, R/qtl of Carl Broman from JHU
- How should multiple traits be treated : jointly or separately?

Example: Initial results on hearing loss

Figure 1: Graphic results of QTL Interval Mapping.
Three thresholds are plotted Š FDR controlling threshold (solid line),
suggestive (short-dashed line) and significant (long-dash ed line)
linkage thresholds of Lander and Krug lyak (1995).



Multiple traits

	Resamp. FDR	BH procedure	Resamp. FWE
Single trait	3.06	3.08	3.81
2 traits	2.95	2.92	4.01
4 traits	2.96	2.92	4.15
8 traits	2.90	2.89	4.36

Comparison of FDR and FWE procedures - multiple traits

With FDR control

No built-in penalty for a more informative study

QTL analysis: Beyond testing

A recent complain (by Soller): FDR based QTL analysis is “too successful”, in that the regions over the chromosome identified as linked to the trait are too wide.

This is a problem with the formulation of QTL analysis as a pure testing problem.

No QTL on chromosome -

any discovery on the chromosome is false

A QTL on chromosome:

any discovery made on the chromosome is true

because of genetic linkage

.

QTL analysis: Beyond testing

This is a problem with the formulation of QTL analysis as a pure testing problem. We really want a confidence region on the chromosome for the location of the linked gene.

While such procedures exist, they are not adjusted with controlling against the effect of multiplicity

Solution I: selective confidence intervals (CIs) for the regions containing the genes being linked, which control the False Coverage-statement Rate (FCR)

The aims: (1) Many CIs (2) Short CIs (3) Low FCR

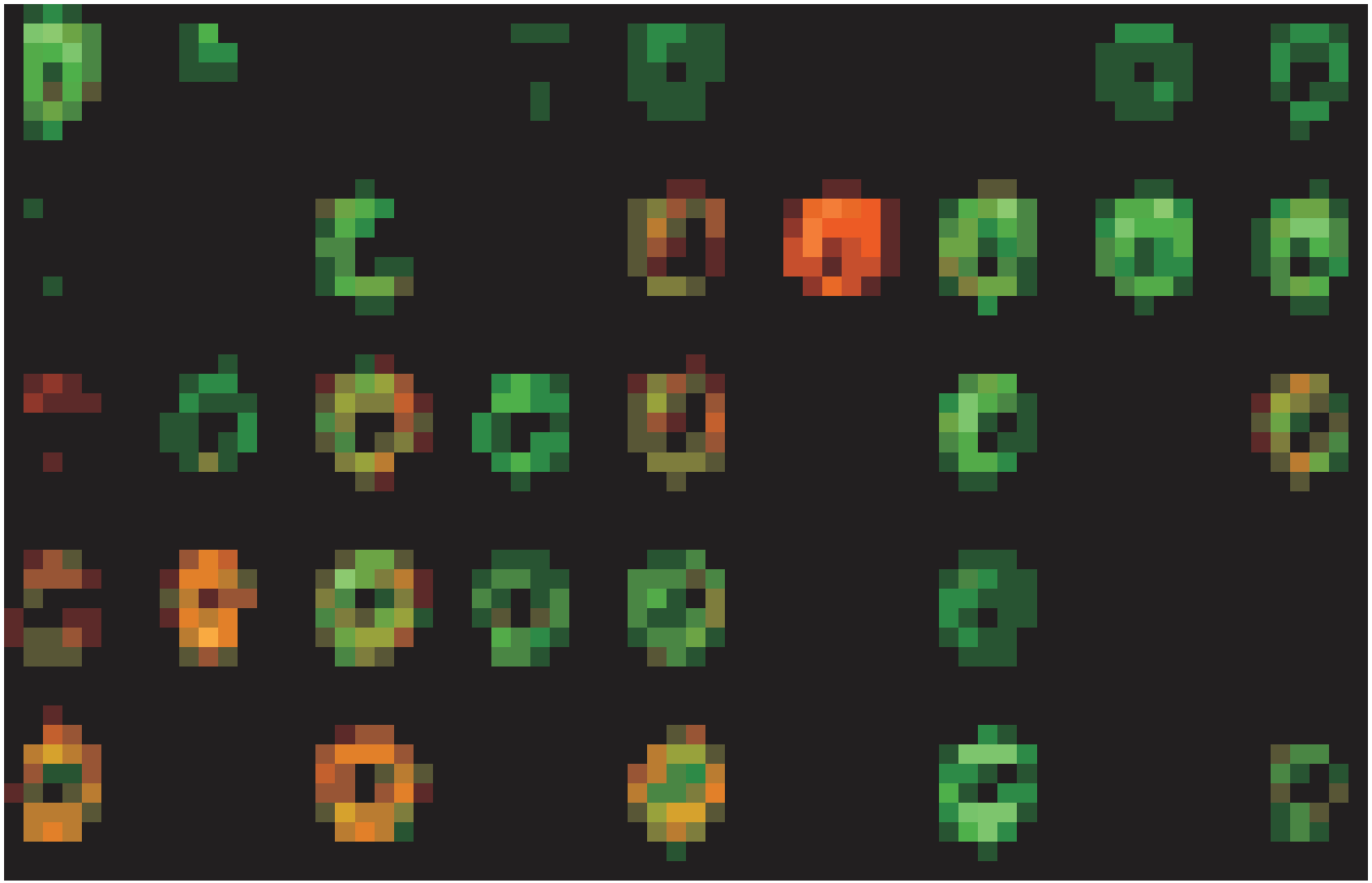
Solution II: Multi resolution genome scan

Solution II: Multi resolution genome scan

1. Methods exist for testing linkages to specific regions by conditioning on value of central marker
2. But: the smaller the region the less power to detect linkage.
3. Therefore: Work hierarchically to the maximum resolution yielding significances: Chromosome level, 1/2 chromosome level,1/4….
4. Appropriately calibrated FDR testing controls the overall FDR

…More later if you wish…

2. Gene-expression Data



Gene-expression micro-arrays

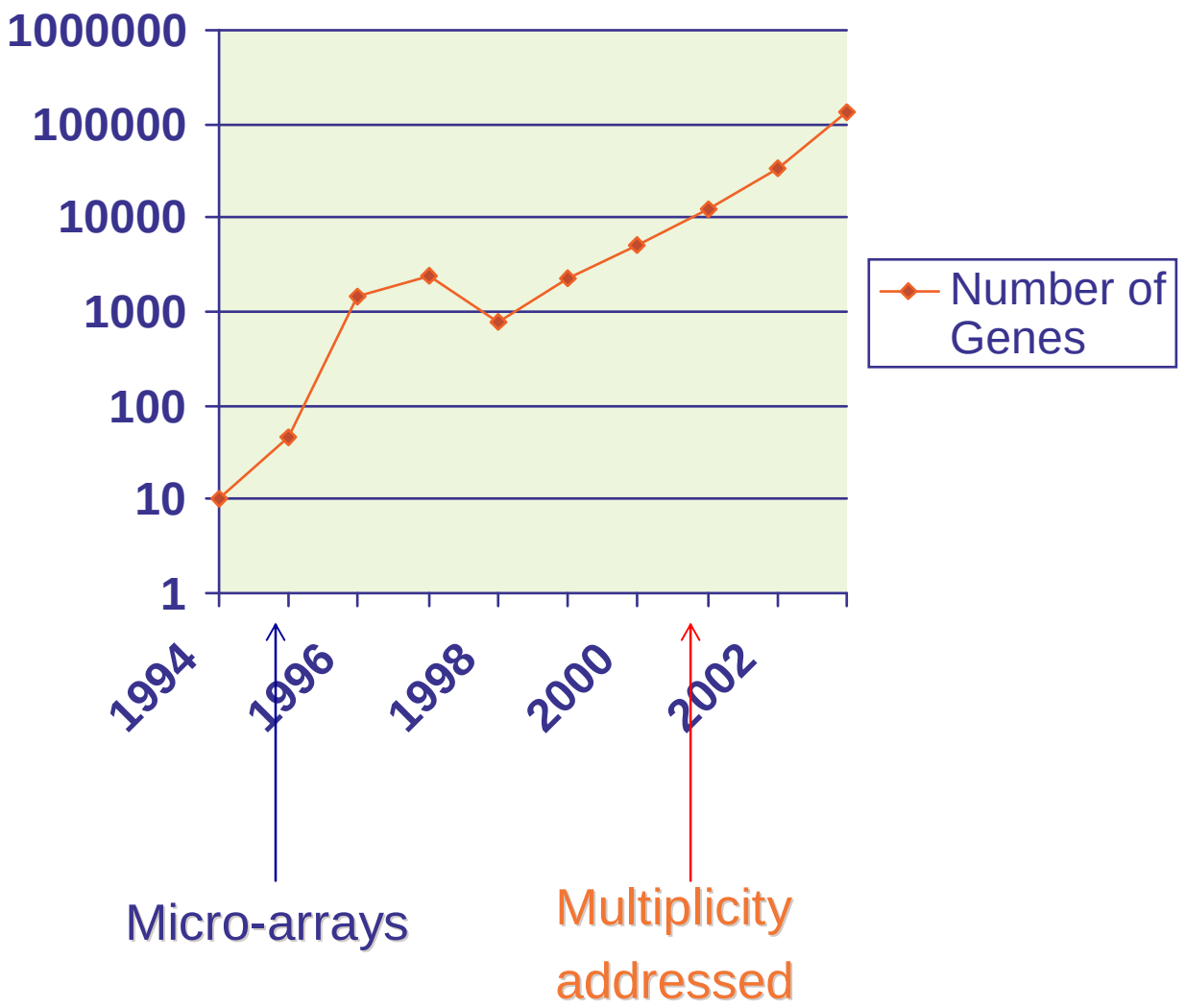
- Example: Dudoit, Yang, Callow, Speed (2001): Statistical analysis of a lipid metabolism study in mice.
- Treatment: 8 low HDL level knockout mice
- Control: 8 inbred mice
- Purpose: Identification of single differentially expressed genes in replicated cDNA microarray experiments.

For the technology see
http://www.bio.davidson.edu/courses/genomics/chip/chip.html

Microarrays and their Statistical Analysis

- The micro-array data consisted in this case of 6359 individual DNA sequences.
- Both treatment and control on a single chip
- The ratio of the fluorescence intensity measured at each spot is indicative of the relative abundance of the corresponding DNA sequence in the two samples.
- Data was suitably standardized using lowess smoother.
- A t-statistic is calculated for comparing gene expression mean between the control and treatment groups.

Gene-expression Data



2.1 The Speed Shop

- Used Westfall and Young's resampling procedure:

$$P_{adj} = Prob_{H0}(P^*_{(i)} \geq P_{(i)})$$

So instead of a single threshold there is sequence of decreasing thresholds as more genes are discovered

- Dudoit et al (2001) considered FDR but did not use because it "requires independence".

Recent work of Dudoit and van der Laan (2004) try to ease FWE control by controlling

Prob(V ≥ k) for some k>1 using resampling.

But how do you choose k?

2.2 The TAU Approach

We experiment with several FDR controlling procedures

- Use the p-values from t-tests in the BH procedure
- Use the (marginal) p-values from resampling in the BH procedure
- The FDR resampling “point estimate”-procedure
- The FDR resampling “upper-bound”- procedure

both estimate by resampling the dist’n of:

$$V^*(t)/(V^*(t)+s(t))$$

(both in Yekutieli and Benjamini 1999)

values–Adjusted P

A convenient way to present the results of a multiple testing procedure is by adjusted p-values

e.g., for Bonferroni, define $p^{BON}_{(i)} = m\, p_{(j)}$

and compare $p^{BON}_{(i)}$ to any desired α

For the Linear stepup procedure,

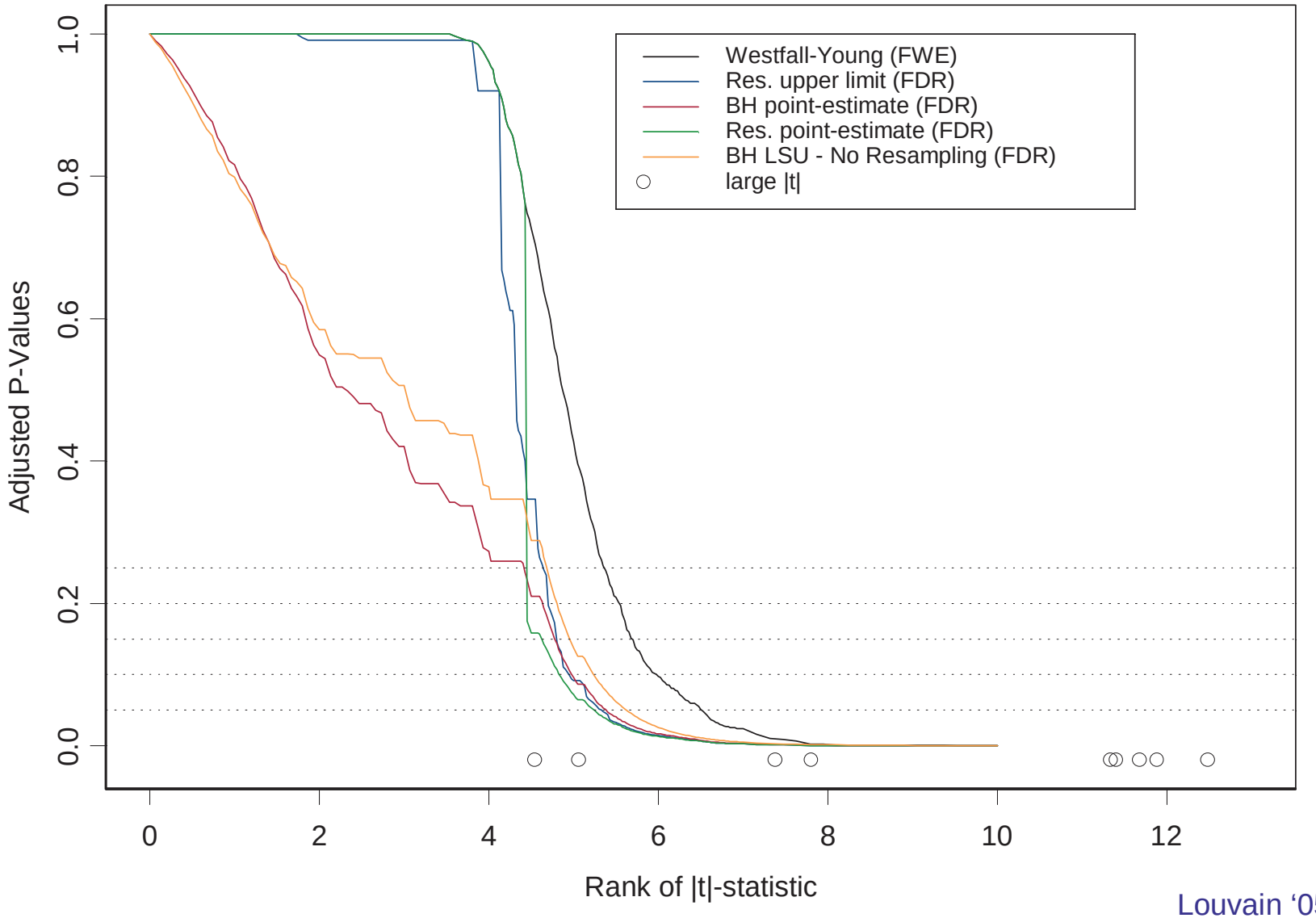
Define $p^{BH}_{(i)} = min\ \{ p_{(j)} m/j,\ j \geq i \}$

Obviously,

$p^{BH}_{(i)} \leq q$ <=> *for some* $j \geq i$, $p(j) \leq qj/m$ <=>

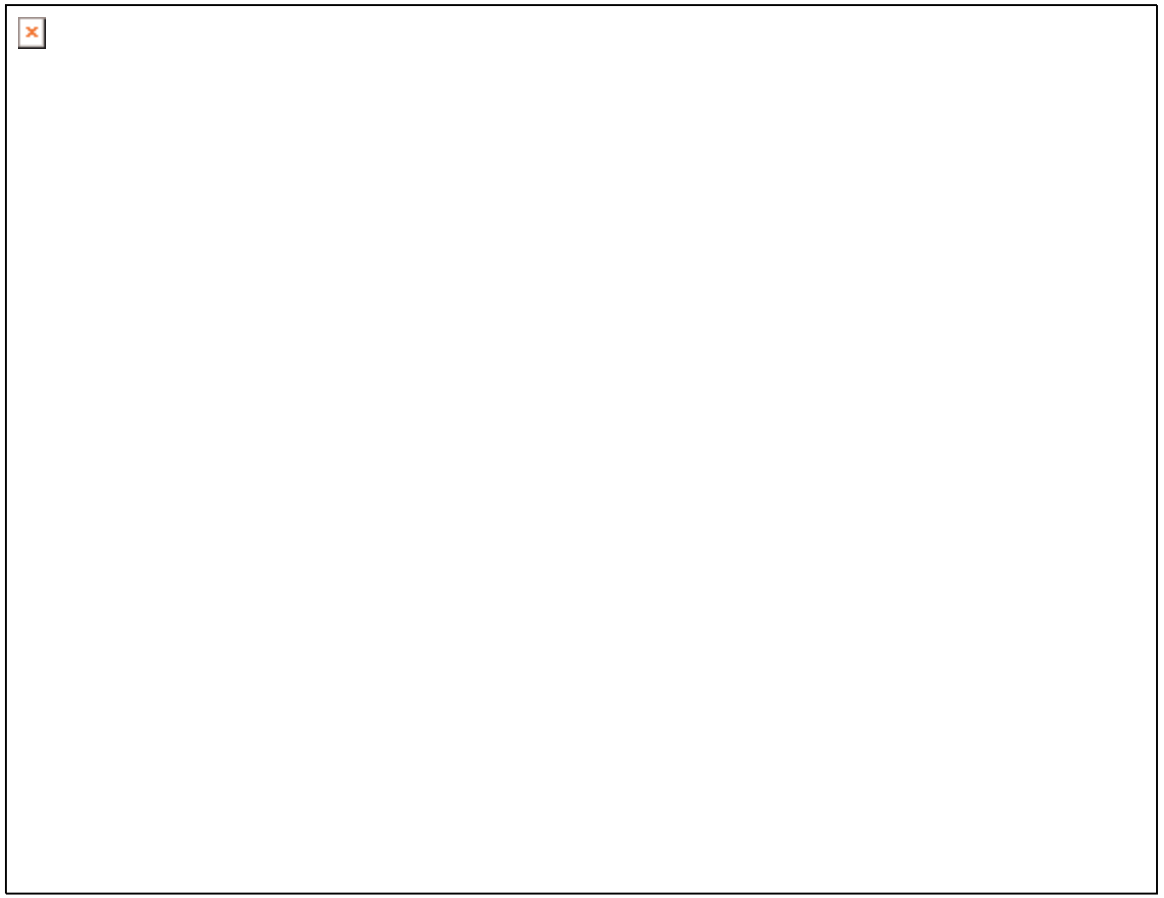
<=>$H_{(i)}$ *is rejected at FDR level q*

values Original Data– Adjusted P

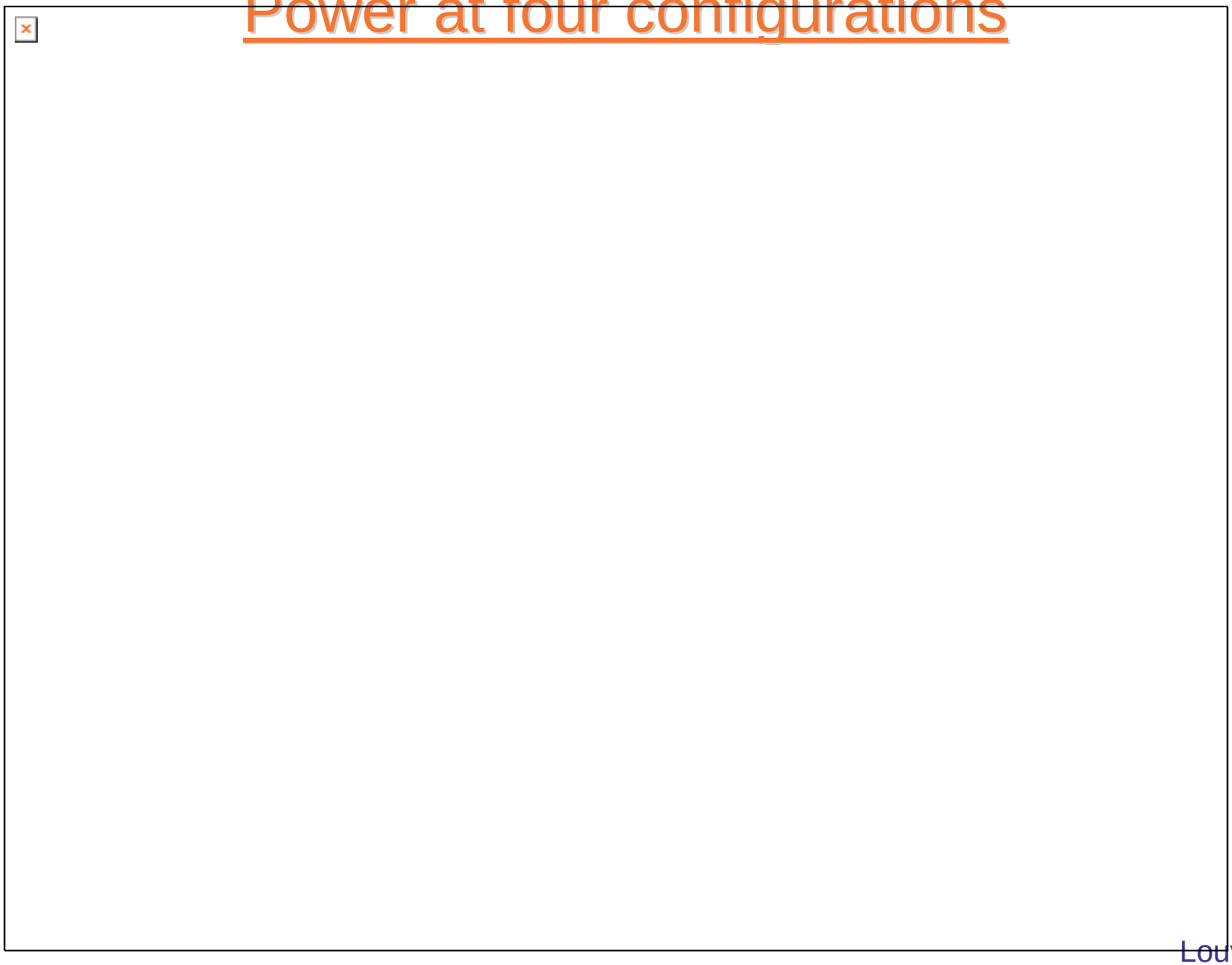


Louvain ‘05

Configurations of effects studied



Power at four configurations



Our practical conclusions at TAU

- Use FDR controlling procedure - even if on the conservative p-values from t-tests.

  Practically no computational effort involved

- Considerable power can be gained from using resampling to estimate marginal p-values

  Some computational effort involved

- Somewhat more power can be squeezed out of the full resampling scheme which makes use of the correlation structure

  Need our software (Splus) or write your own

  Details in Reiner, Yekutieli, YB (2003)

How about the adaptive procedures

In current microarray analyses m_0/m is too close to 1
for adaptive procedures that really control the FDR to be helpful i.e.

genes to be discovered / # genes tested < q

Thus, for the time being, I use $m_0/m = 1$

This may change as more focused microarrays are being offered (for neurological system, etc)

Or if dependency structure is better understood

(e.g. Storey Taylor & Siegmond ('04) works under independence and some kinds of dependence)

Implications of testimation to microarray analysis

Screen out genes which do not pass the FDR threshold
i.e.

For clustering (class forming),
use X^{FDR} rather than X

For class prediction
use complexity-penalty version of X^{FDR}

$$SSR(k) + \sigma^2 \sum_{i=1}^{k} z^2_{\frac{iq}{m}}$$

$$\approx SSR(k) + \sigma^2 k \cdot z^2_{\frac{kq}{m}} \approx SSR(k) + \sigma^2 k \cdot 2\log(m2/kq)$$

2.3 The Stanford Approach

Efron, et al (JASA, 2001) “Empirical Bayes Analysis of Microarrays Data”

- A common theoretical framework “The mixture model”:

  p_1= prob. gene is affected (that h_i =1),

  $f_1(Z)$ the density of its score

  p_0=1-p_1 = prob. gene is not affected (that H_i =0),

  $f_0(Z)$ the density for its score

  The mixture density of the scores is

$$f(Z)= p_0 f_0(Z) + p_1 f_1(Z)$$

Storey (‘01,’03) emphasized the

conditional FDR (positive FDR):

$$pFDR = E(V/R \mid R>0)$$

for fixed choice of thresholding p-value a,
showing under independence

$$pFDR(a)=Pr(H_i=0 \mid R_i(a)=1)$$

$$pFDR(a)=Pr(\ H_i=0 \mid R_i(a)=1)$$

$$=Pr(H_i=0 \ \& \ R_i(a)=1)/Pr(R_i(a)=1)$$

$$=Pr(H_i=0)Pr(\ R_i(a)=1 \mid H_i=0) \ / \ Pr(R_i(a)=1)$$

$Pr(H_i=0)=p_0$ Prior for null

$Pr(\ R_i(a)=1 \mid H_i=0)=a$ level of test

and $Pr(R_i(a)=1)$ is estimated by $R(a)/m$, so

estimated $pFDR(a)=p_0\, am \,/(\ R(a)\ (1-(1-\ a)^m)\ )$

Efron et al emphasized the local version:

$$fdr(Z)=p_0 f_0(Z)/ f(Z)$$

which they call the local false discovery rate,

the posterior probability that a gene with score Z is unaffected

In the Bayesian approach - specify f_1, p_0, then f

In the Empirical Bayes approach -

estimate p_0 $(=m_0/m)$ and f

In Efron ('04) f_0 is estimated as well.

The Stanford shop - SAM

The computational tool SAM - Significance assessment of microarrays (Tusher, Tibshirani, and Chu '01, Storey and Tibshirani '03)

- 1. Calculates test-statistics
- 2. Performs estimation of p-values using permutation sampling.
- 3. Estimates m_0 using (old) Storey's procedure
- 4. Uses the adaptive linear step-up procedure to get estimates of Q(a), by $q(a) = m_0 \, a \, / R(a)$
- In recent versions the largest Q(a) over all $a > p_{(i)}$ is calculated (and called estimated q-values)

This is simply the FDR adjusted p-values using the Linear Stepup and Storey's estimator for m_0

So

- SAM with $m_0/m=1$ is like non-adaptive linear step up (BH) with resampling based p-values (our option 2)
- With estimated m_0/m may run into (solvable) problems
- It is interesting to see how the statistical analysis of microarrays gave rise to so much research and developments in a seemingly theoretical statistical issues such as multiplicity control
- It is worthwhile to use procedures which are philosophically robust

The ultimate multiplicity challenge ?

- Rob Williams, U of Tennessee & Rebeca Doerge

  Discussed in JSM in NY - independently - projects in which gene expression levels are traits in QTL analysis!

- They were talking about 400,000 tests conducted simultaneously, with complicated dependency structure

"Every tool of our statistical witchcraft will be needed to handle such an analysis"

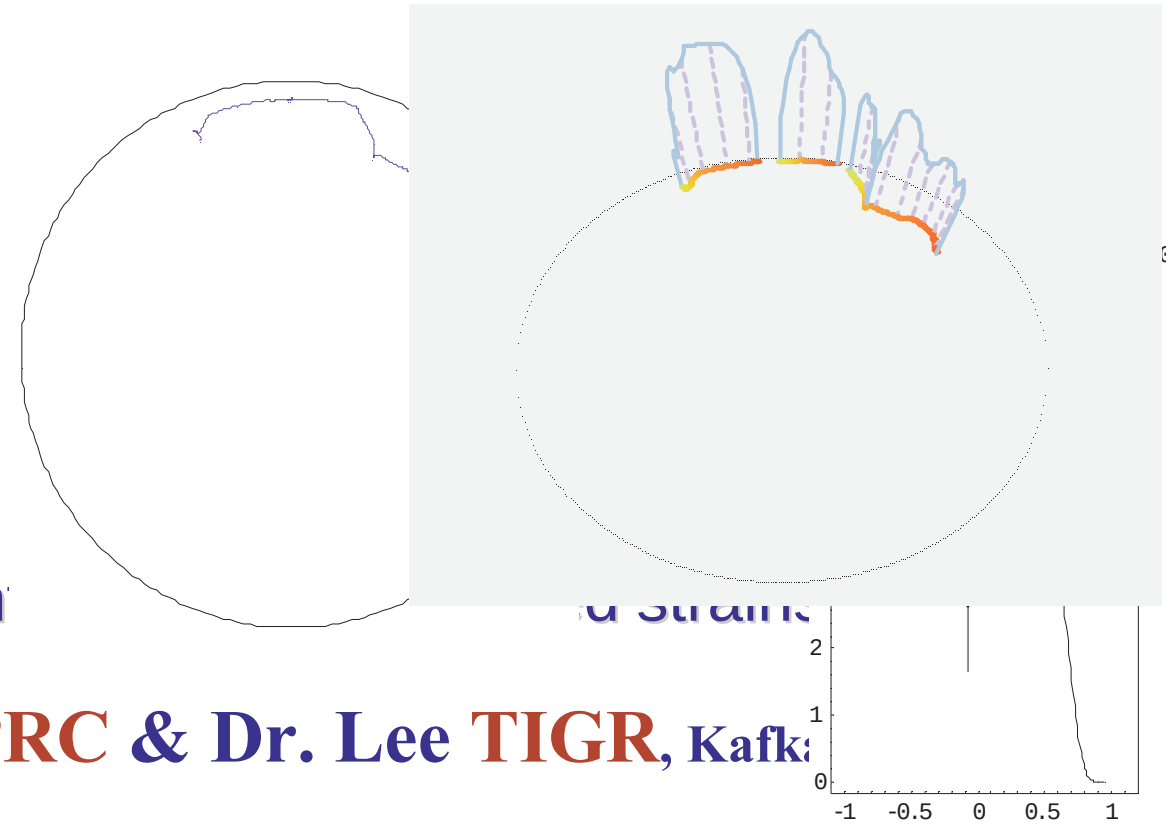
Never throw away the residues of your experiment

NIH: Phenotyping Mouse Behavior
High throughput screening of mutant mice

Behavior

Tracking

‚Stops, move segments and velocity profile



Comparing 17 endpoin

Golani TAU Elmer, MPRC & Dr. Lee TIGR, Kafk

The Brain and Behavior Example (BB)

Expression levels of ~27,000 genes, in 5 brain regions of the same mice from the 10 strains (two replicates per strain)

Research Question I: What genes exhibit strain differences in expression levels over entire brain?

Two-way ANOVA at each gene; p-values of strain effect

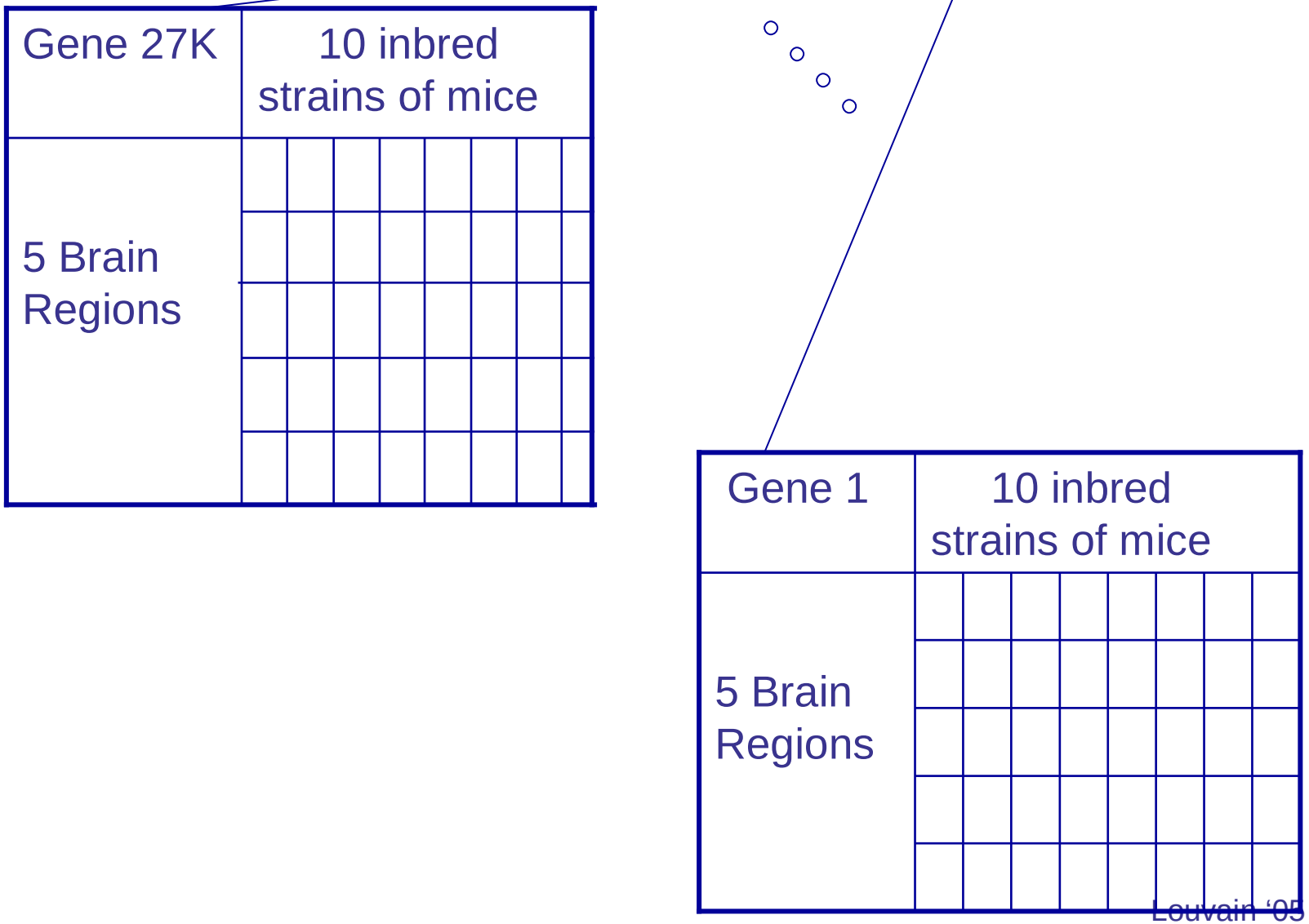
The Brain and Behavior Example (BB)

- 17 behavioral endpoints measured for 8 animals from each one of 10 strains of inbred mice (in 3 labs)
- Expression levels of ~27,000 genes, in 5 brain regions of the same mice from the 10 strains (two replicates per strain) by N. Lee TIGR

Research Question I: What genes exhibit strain differences in expression levels over entire brain?

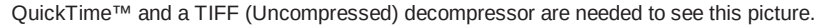
Two-way ANOVA at each gene; p-values of strain effect

Two-way ANOVA at each gene; p-values of strain effect



Sorted p-values vs their rank (BB)

Raw p-Value



q/m

Rank

FDR adjusted p-Values vs their rank

$$p^{BH}_{(i)} = min \{ p_{(j)} m/j, \ j \geq i \}$$

FDR-adjusted p-values

p-Value

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

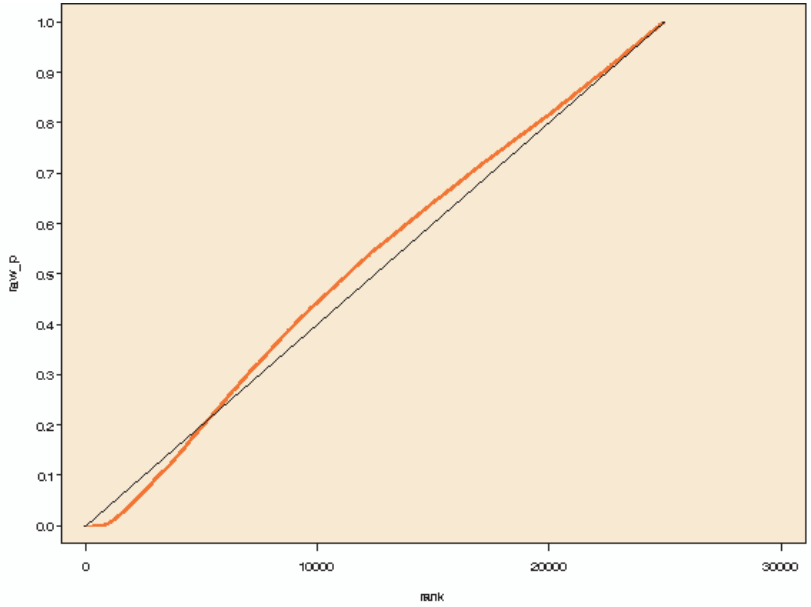
Raw p-values

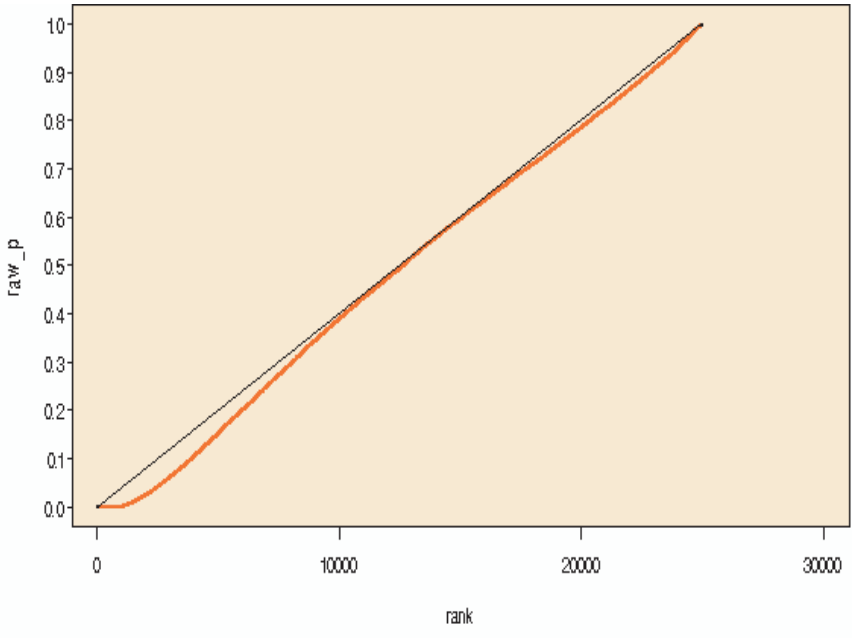
Rank

Louvain '05

F. P-P plots: Effect of Background Removal

4-way repeated medians related to physical layout





Before –

730 genes

(at FDR .05)

After -

~1000 genes
(at FDR .05)

More complex multiplicity challenges in BB: Research Questions II , III , and IV

- What genes exhibit strain differences in expression levels over specific brain regions?
    - Size of multiplicity problem 27K*5
- What strains are in fact different in their expression levels in specific genes?
    - Size of multiplicity problem 27K*45
- Are there any interactions of strain and brain region in specific genes?
    - Size of multiplicity problem 27K*50

Research Questions III

What pairs of strains are in fact different in their expression levels in specific genes?

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

A general approach for complex statistical analysis

- Analysis is divided into separate research questions
  - Many research questions
  - Research questions can be added at a later date
- A FDR tree fitted for each Complex research question
  - With new data more levels can be added to FDR tree
- FDR control for each research question

Our current (still somewhat vague) understanding is:

If # FDR trees < # discoveries then

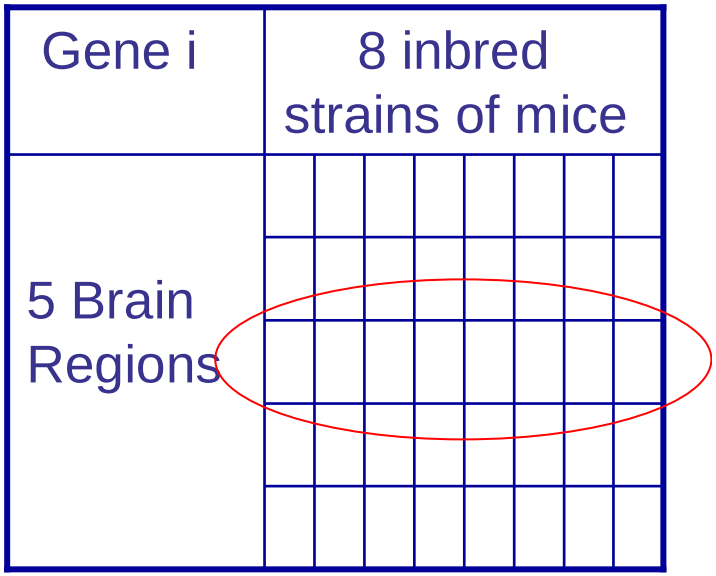
FDR is controlled for the entire study

Research Question V

- For what genes, and in what brain regions, are strain expression levels correlated with a specific behavioral trait?

Size of multiplicity problem 27K*5*17

~ 2.2 milion hypotheses



Two FDR controlling approaches:

1. Hierarchical testing
2. Subset selection

Method 1: Hierarchical testing

- First find genes-by-regions with significant strain differences using the Linear Step-up at level 0.025
- Then test each sub-family of hypotheses at each of the selected gene-by-region separately for correlation (with Spearman's test) using the Linear Step-up at level 0.025

  Number of traits-by-genes-by-region discoveries 123

Even though the latter pool is selected, and each sub family consists of 17 hypotheses only, the hierarchical tree procedure controls the FDR at least at

$$2*0.025*1.44 \sim 0.073$$ (BY & Yekutieli '02)

(Simulations show using .025 offers FDR <.05)

Method 2: A subset selection approach:

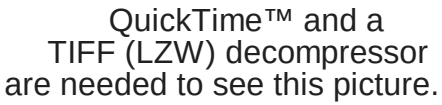
- First find genes&regions with significant strain differences

  using appropriate contrast in ANOVA per gene

- Then test for correlation (using Spearman's test) at all selected genes simultaneously, at some q.

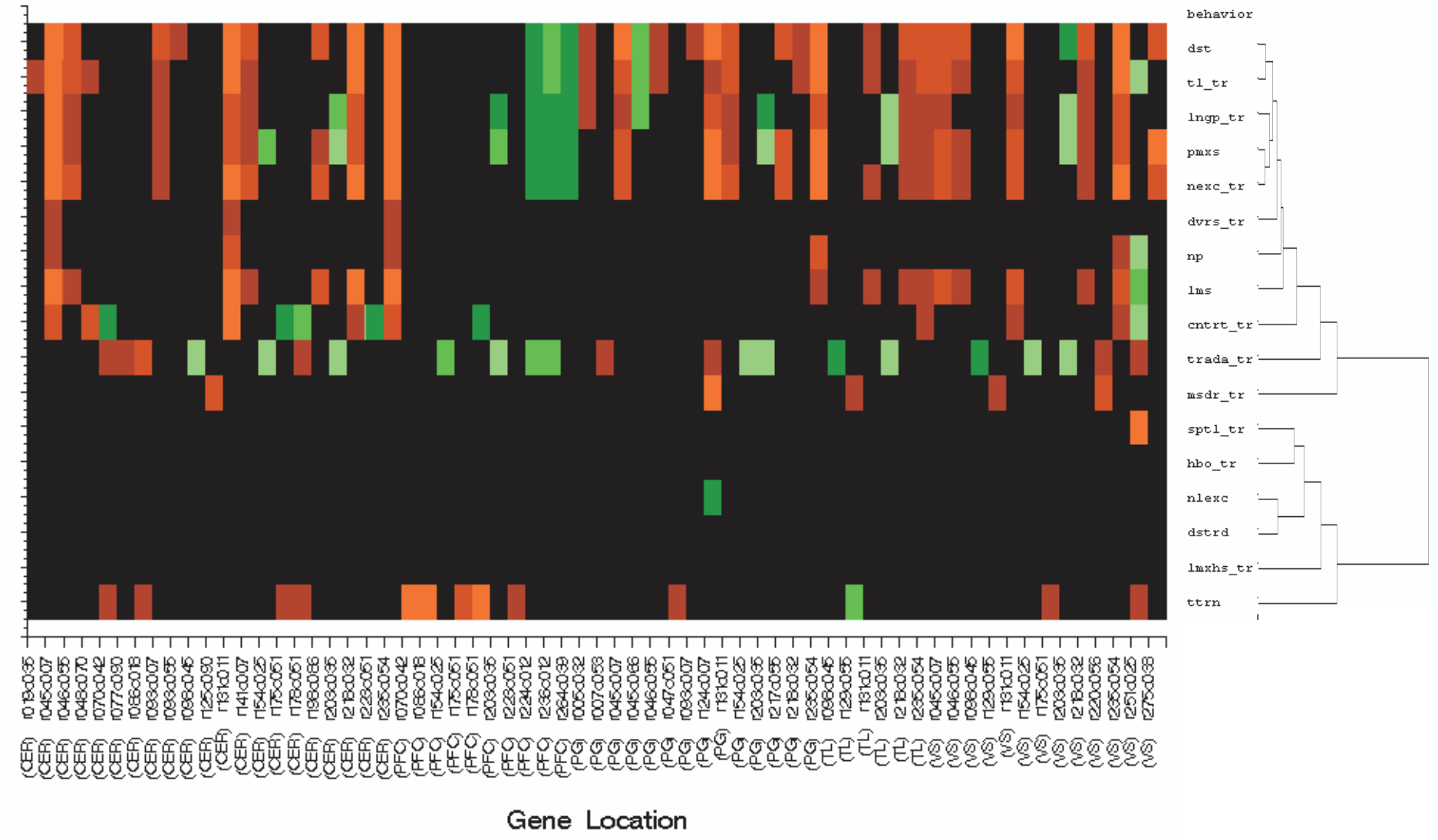
Even though the latter pool is selected, with only 465*17, the linear step-up procedure controls the FDR at q

The reason: approximate independence of tests


QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Gene Count for Cases Significant at FDR 0.025

	Brain Region					
	CER	PG	VS	TL	PFC	Total
Trait						
DST	8	6	6	2	1	23
TL_TR	5	4	4	1	1	15
LNGP_TR	5	3	2	1	-	11
PMXS	6	5	5	1	1	18
NEXC_TR	7	5	4	-	-	16
NP	1	1	1	-	-	3
LMS	5	-	4	-	-	9
CNTRT_TR	5	-	1	-	-	6
TRADA_TR	4	2	2	1	4	13
MSDR_TR	1	1	1	-	-	3
SPTL_TR	-	-	1	-	-	1
TTRN	-	-	-	1	4	5
Total	47	27	31	7	11	123


behavior
dst
tl_tr
lngp_tr
pmxs
nexc_tr
dvrs_tr
np
lms
cntrt_tr
trada_tr
msdr_tr
sptl_tr
hbo_tr
nlexc
dstrd
lmxhs_tr
ttrn
Gene Location
P Value
0.025 — 0.05 (neg. corr.)
0 — 0.01 (pos. corr.)
> 0.05
0.01 — 0.025 (neg. corr.)
0.01 — 0.025 (pos. corr.)
0 — 0.01 (neg. corr.)
0.025 — 0.05 (pos. corr.)

Louvain ‘05

Complex large-scale statistical analysis

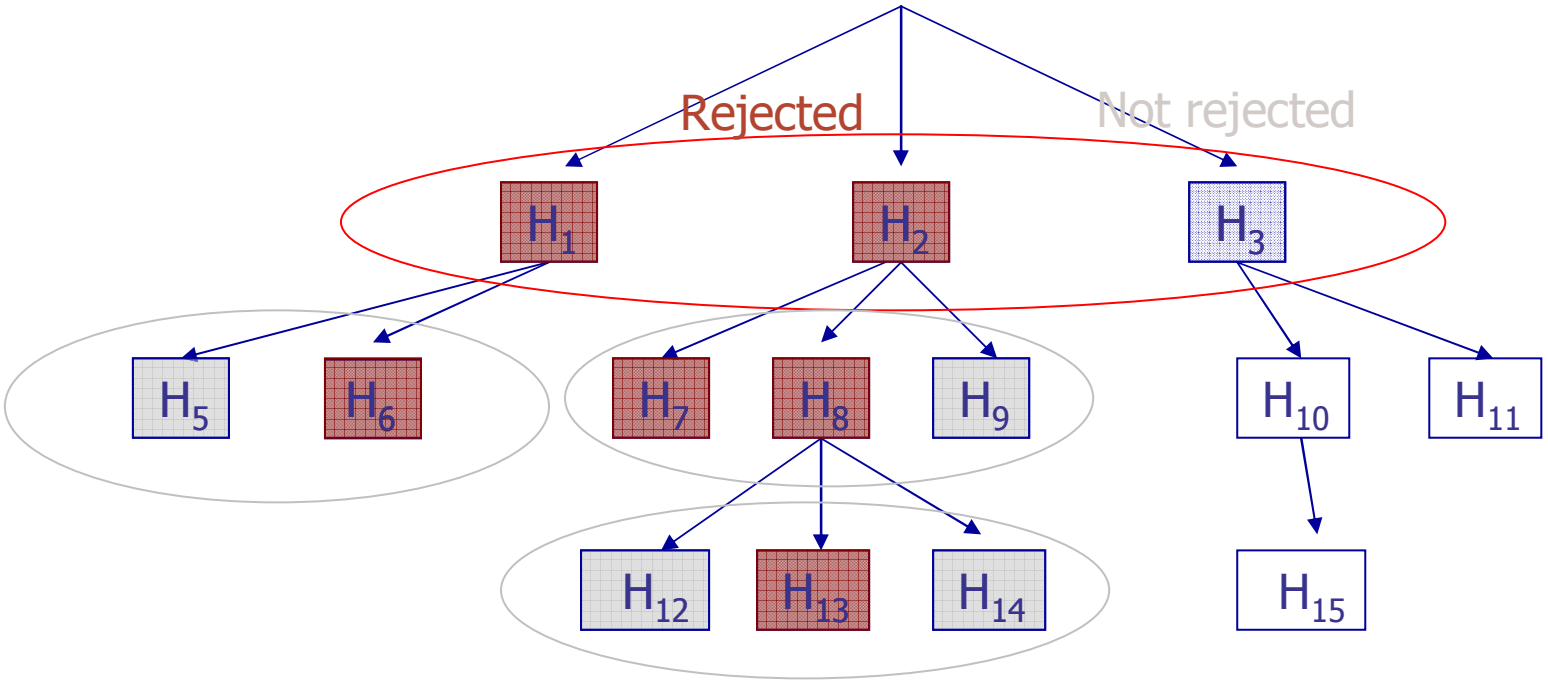
- Millions of tests may be conducted
- Many more tests considered
- Research questions posed after viewing the data
- Research questions are usually sequential in nature

"The appetite comes with the eating"

How to ensure reproducible results ?

G. FDR tree - hierarchical testing scheme

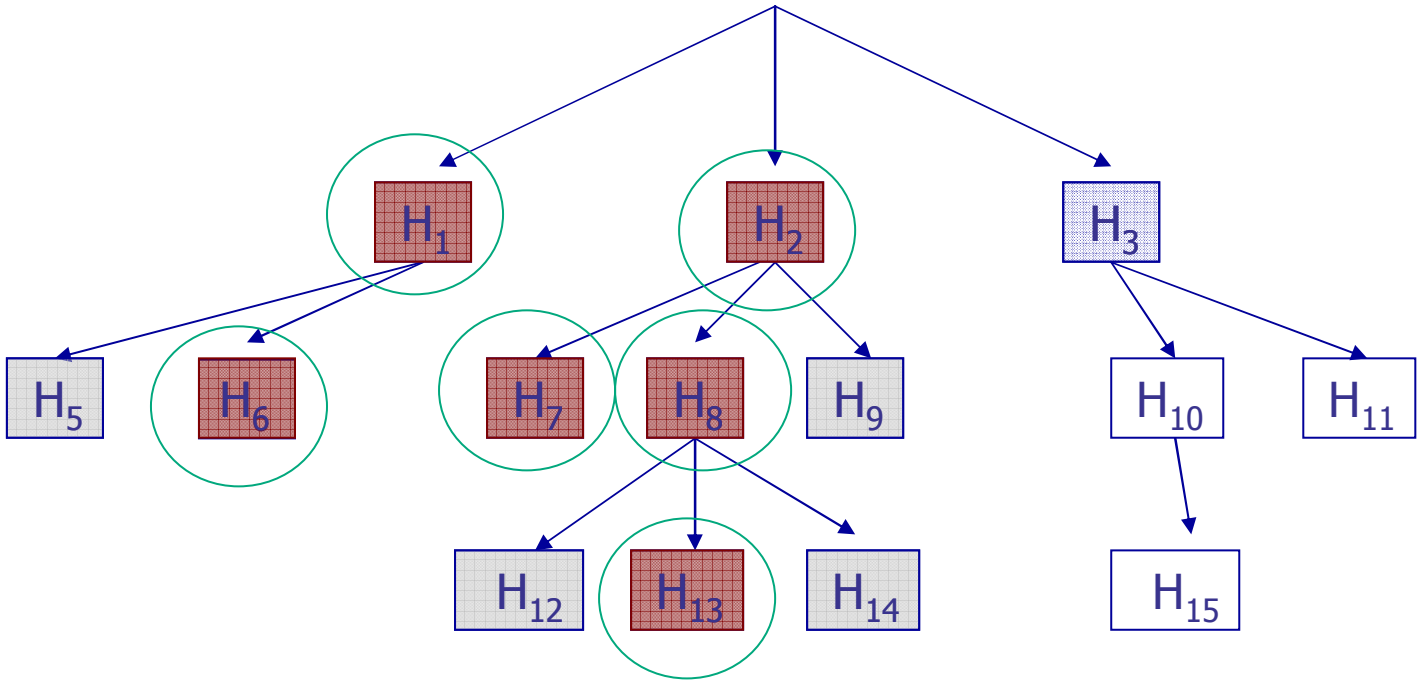
(BY & Yekutieli `02)



1. Arrange hypotheses in sub-families corresponding to a single parent hypotheses

2. Test sub-family of a rejected parent hypothesis by level q BH procedure

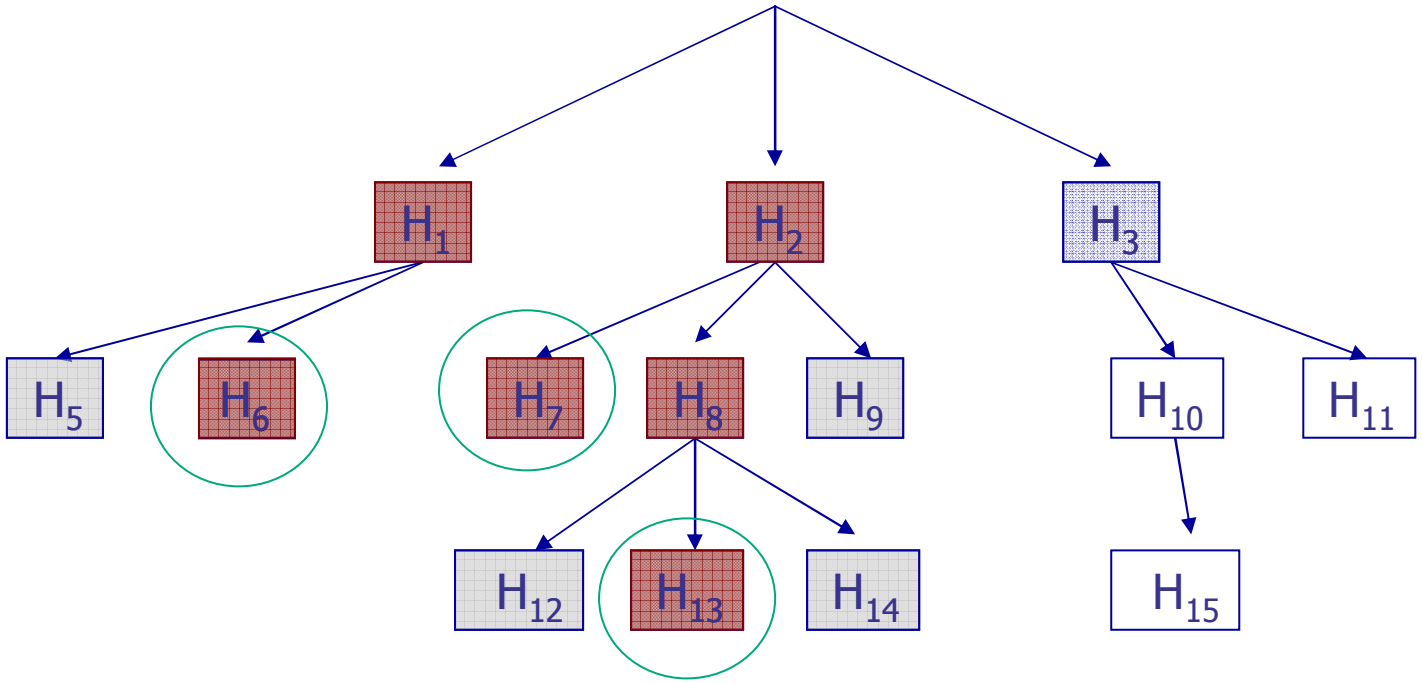
FDR tree: Full tree testing scheme



Theoretical results: independent test statistics FDR upper bound (any sized tree) FDR < $2 \times \delta^* \times q$, where $\delta^* < 1.44$

In more realistic settings and in simulations: FDR $\approx$ q

FDR tree – outer nodes testing scheme



Theoretical results: independent test statistics FDR upper bound:

$$FDR < 2 \times L \times \delta^* \times q, \quad L = \#\text{ of levels } (= 3)$$

FDR higher than in full tree scheme yet in "most" cases $\approx q$ Louvain '05

Inequalities are based on following lemma

$$FDR_t = \delta^* q \frac{m_{0t}}{m_t} E\left\{ I(D_t^{par}) \frac{R_t + 1}{R + 1} \right\}$$

Delicate and complicated proof due to Yekutieli

YB&Yekutieli (‘02+),Yekutieli (‘05+)

A general approach for complex statistical analysis

- Analysis is divided into separate research questions
  - Many research questions
  - Research questions can be added at a later date
- A FDR tree describes the structure of a complex research question
  - With new data more levels can be added to FDR tree
- FDR control for each research question

Our current (still somewhat vague) understanding is:

If # FDR trees < # discoveries then

FDR is controlled for the entire study

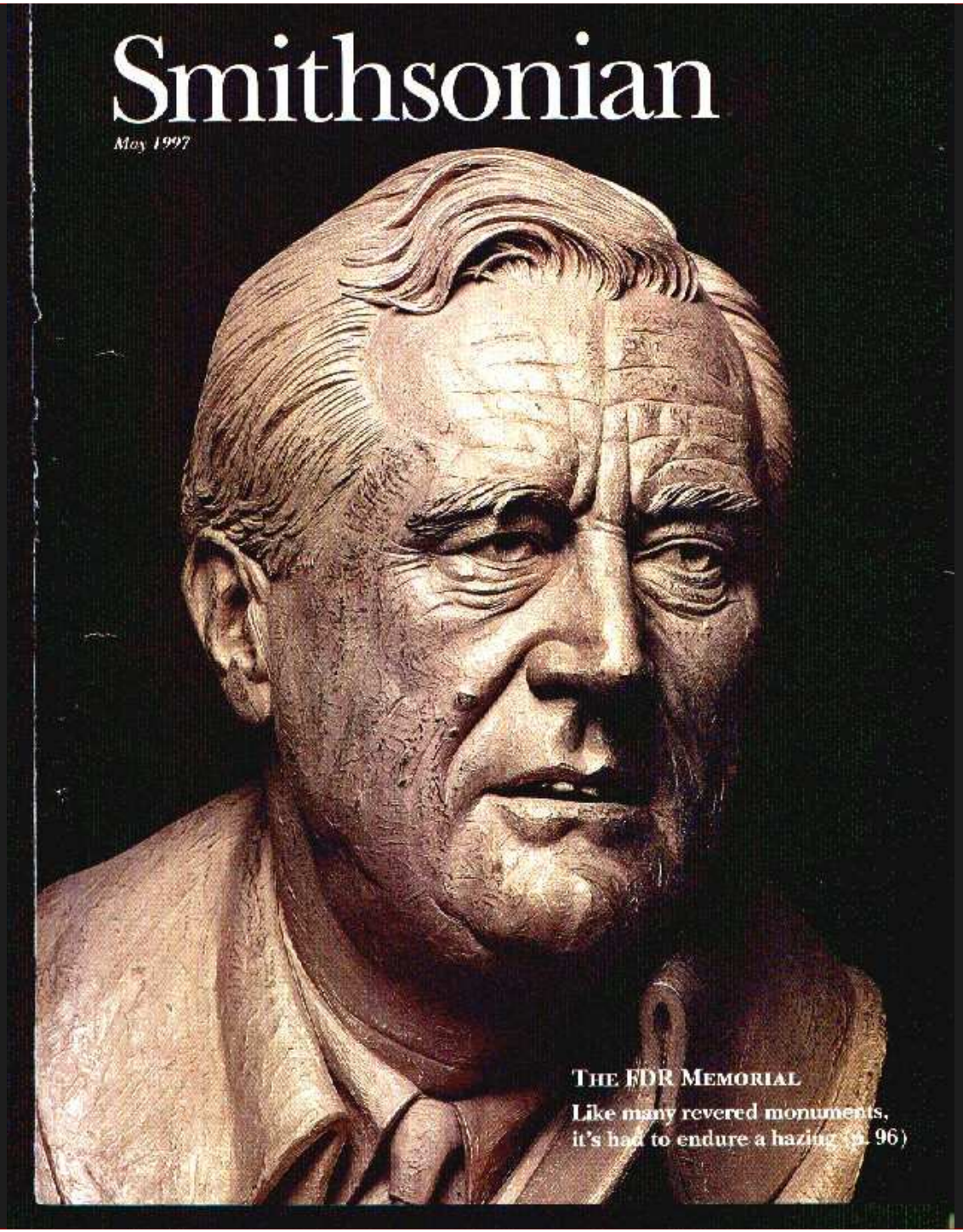
Moral of the stories

- By addressing multiplicity we are not merely “the guardians of science”
- Multiplicity control carries benefits for researchers
  - Reduces non-replicability
  - Reduces wasted follow-up efforts
- It takes for the problem to become very large before the advantages of addressing multiplicity becomes clear to researchers
- We (statisticians) should help in shortening this lag as such very large problems in Statistical Genetics become more common
- Many theoretical problems are waiting to be resolved

The FDR website www.math.tau.ac.il/~ybenja

Smithsonian

May 1997



THE FDR MEMORIAL

Like many revered monuments, it's had to endure a hazing (p. 96)

Solution II: Multi resolution genome scan

1. Methods exist for testing linkages to specific regions by conditioning on value of central marker
2. But: the smaller the region the less power to detect linkage.
3. Therefore: Work hierarchically to the maximum resolution yielding significances: Chromosome level, 1/2 chromosome level,1/4….
4. Appropriately calibrated FDR testing controls the overall FDR

Applying a 2 level binary FDR tree

Search for QTL modifying the effect of a deafness causing mutation on 15 behavioral traits

Level 1 –

Test the 300 chrom. level hypotheses

(15 traits $\times$ 20 chrom)

using the BH procedure with q=0.05.

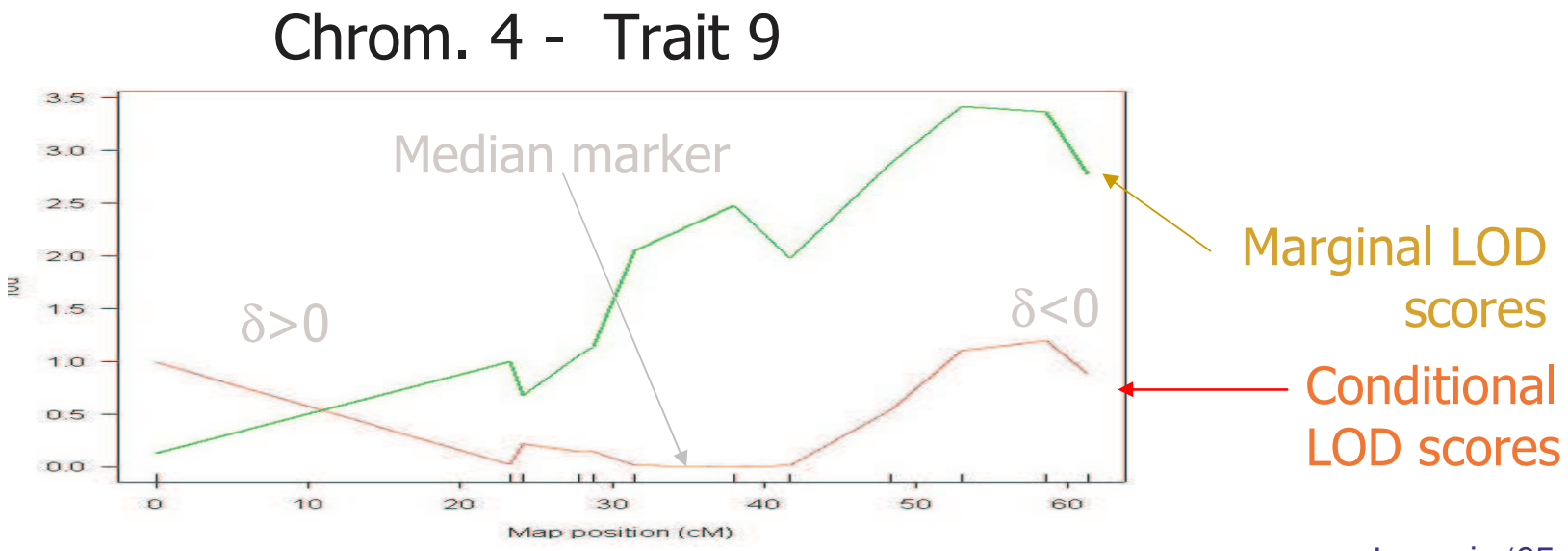
14 Chromosome level discoveries

	chr. 4	chr. 5	chr. 17
• Trait 1			0.0006
• Trait 2	0.0007		
• Trait 5	0.0001		
• Trait 7	0.0001	0.001	
• Trait 8	0.0001	0.001	
• Trait 9	0.002		
• Trait 10			0.0002
• Trait 11	0.002	0.005	
• Trait 13		0.005	
• Trait 15	0.004	0.001	

For each of the 14 chrom. level discoveries separately test each pair of half chrom. Level hypotheses BH procedure with q=0.05

Trait 9 - 2 QTLs in repulsion on both halves of chrom 4:

left half a positive effect ; right half negative effect.



Results of a a 2 level binary FDR tree

Total: 15 discoveries.

Relevant discoveries:

13 Chrom. level + 2 half chrom. level

We show in simulations that FDR of procedure < 0.05.

In theory there is a bound on any tree of inferences the increase in FDR level

2*(#of levels)*1.44*0.05

BH & Yekutieli (‘02)

Conclusion about adaptive procedures

In current microarray analyses m_0/m is too close to 1 for adaptive procedures that really control the FDR to be helpful i.e.

genes to be discovered / # genes tested < q

Thus, for the time being, use $m_0/m = 1$

This may change as more focused microarrays are being offered (for neurological system, etc)

Positive dependency

- Positive Regression Dependency on the subset of true null hypotheses:
- If the test statistics are $\mathbf{X}=(X_1,X_2,\ldots,X_m)$:
  - For any increasing set D, and H_{0i} true
  - Prob(**X** in D | X_i=s) is increasing in s
- Important Examples
  - Multivariate Normal with positive correlation
  - Absolute Studentized independent normal
  - (Studentized PRDS distribution, for q<.5)
- Covered but with no full theoretical proof
  - Pairwise comparisons
  - two sided correlated normal

Results on interval mapping

Table 2

Traits	Methods		
	Interval mapping (2 cM)	Interval mapping (1 cM)	Single marker regression
1	0.002173	0.002173	0.001725
2	0.001283	0.001167	0
3	0	0	0
4	0	0	0
5	0.001432	0.001258	0.001478
6	0	0	0
7	0.004513	0.004513	0.002512
8	0.003708	0.003136	0.002596
9	0.001966	0.001966	0.001371
10	0.002557	0.002557	0.001749
11	0	0	0
12	0	0	0
13	0.0013	0.0013	0
14	0	0	0
15	0	0	0

Gene-expression micro-arrays

- Example: Dudoit, Yang, Callow, Speed (2001): Statistical analysis of a lipid metabolism study in mice.
- Treatment: 8 low HDL level knockout mice
- Control: 8 inbred mice
- Purpose: Identification of single differentially expressed genes in replicated cDNA microarray experiments.

Microarrays and Multiplicity

- Neglecting multiplicity issues, i.e. working at the individual 0.05 level, would identify, on the average, 6359*0.05=318 differentially expressed genes, even if really no such gene exists.
- Addressing multiplicity with Bonferroni at 0.05 identifies 8 .

Table 1: First 12 Largest T-Statistics [1,2]

T-Statistic	P-Value (df=14)
-20.6	$7.0*10^{-12}$
-12.5	$5.6*10^{-9}$
-11.9	$1.1*10^{-8}$
-11.7	$1.3*10^{-8}$
-11.4	$1.8*10^{-8}$
-11.3	$1.9*10^{-8}$
-7.8	$1.8*10^{-6}$
-7.4	$3.6*10^{-6}$
5.0	$1.8*10^{-4}$
-4.5	$4.6*10^{-4}$
-4.5	$4.9*10^{-4}$
-4.4	$6.5*10^{-4}$

1. The t-statistics were ranked according to their absolu
2. Bonferroni adjusted p-value is $1.6*10^{-4}$.

values–Adjusted P .C

A convenient way to present the results of a multiple testing procedure is by adjusted p-values

e.g., for Bonferroni, define $p^{BON}_{(i)} = m\, p_{(j)}$

and compare $p^{BON}_{(i)}$ to any desired α

For the Linear stepup procedure,

Define $p^{BH}_{(i)} = min\ \{ p_{(j)} m/j,\ j \geq i \}$

Obviously,

$p^{BH}_{(i)} \leq q$ <=> *for some* $j \geq i$, $p(j) \leq qj/m$ <=>

<=>$H_{(i)}$ *is rejected at FDR level q*

They are the same as Storey's q-values

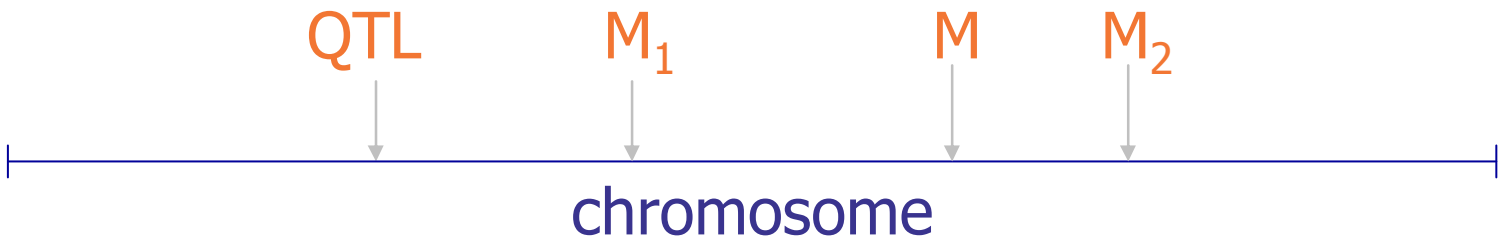
Solution II: Multi resolution genome scan

Zeng (1994) - Composite interval mapping
Improve specificity *within chromosome* by conditioning on flanking markers

If there are no QTL between markers M_1 and M_2 then conditioning on M_1 and M_2 , for any

$$M_1 < M < M_2$$

Marker M and trait are unlinked.



Solution II: Multi resolution genome scan

1. Chrom. level – no conditioning!

Num of hypotheses tested = num of chrom. × num of traits.

Discovery: a QTL somewhere on the chromosome.

2. Half chrom. Level – condition on M2

Discovery: limit QTL to within half a chromosome

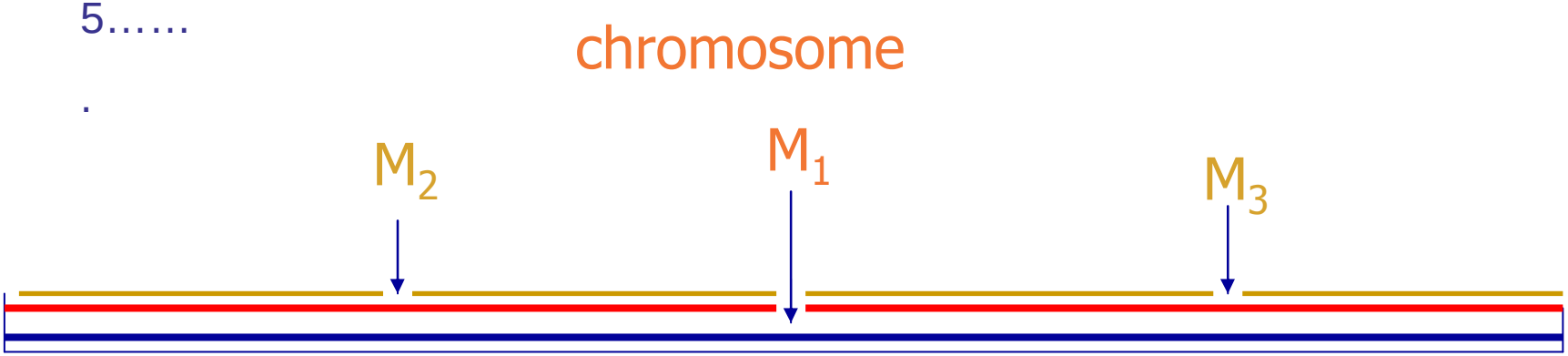
3. Quarter chrom. Level – condition on M1, M2, M3

Discovery: limit QTL to within quarter of a chromosome

4. ….

5……

.



Multi resolution genome scan

Problems:

1. Since there is very little power to make discoveries in high resolution levels - how to work at maximum resolution while not missing discoveries

2. How do we control the FDR ?

Solution 2: Hierarchical FDR procedures.

Behavioral Genetics

- Study the behavioral traits: hearing, sight,smell, alcoholism, exploratory behavior
- Compare inbred strains, knockouts, …
- From “Behavior Genetics in transition” (Mann, Science ‘94)

“…jumping too soon to discoveries..” (and press discoveries) raises the issue of *Replicability*

- Mann states statistical troubles as a major source of the replicability problem, yet did not mention lack of control of multiplicity as one.

When viewing as a simple multiple testing problem:

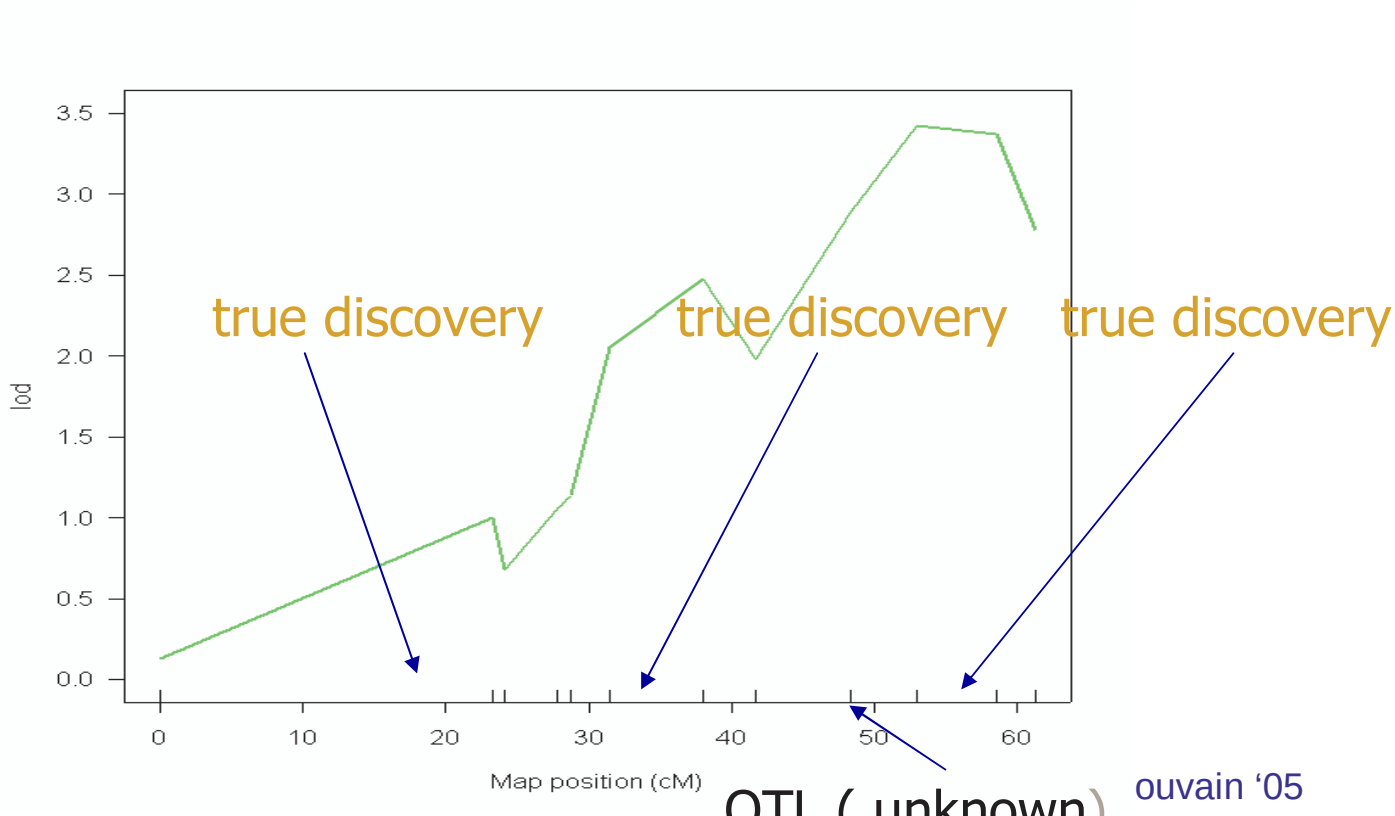
No QTL on chromosome -

any discovery made is false

If there is a QTL on chromosome:

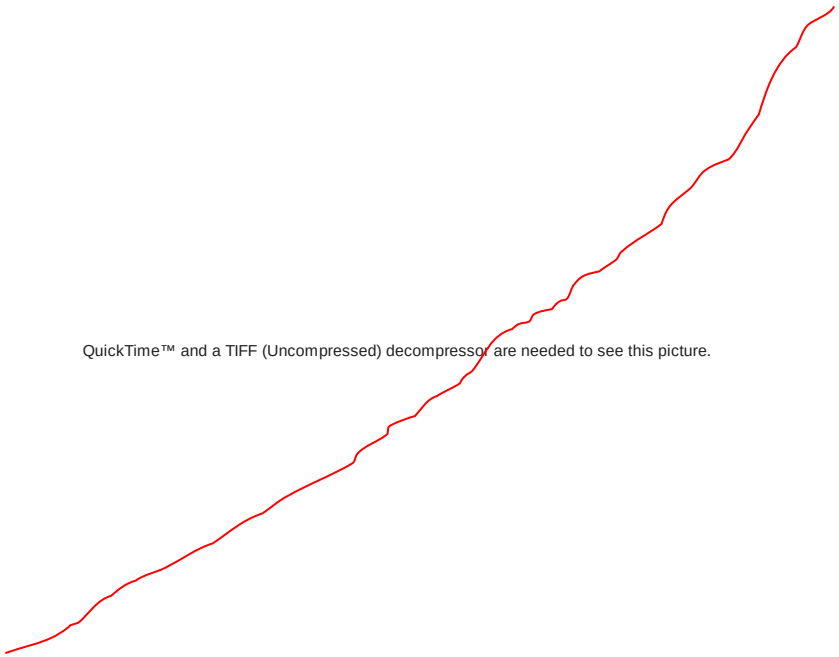
any discovery made on the chromosome is true

because of genetic linkage



ouvain '05

Questions IV: Are there any interactions of strain and brain region in specific genes?

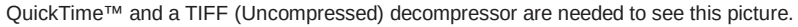


Beware!

- p-values too large! Problem?? Inference still valid!
- Check calculations and preprocessing. Improve preprocessing. Use resampling to increase power.

The distribution of the p-values of Spearman's tests

Raw p-Value



Rank

FDR-adjusted p-values (Spearman's)

Raw p-Value

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

Rank

Two stage linear step-up procedure

The approach described is somewhat too optimistic when connected to the linear stepup procedure

Especially under dependency

A Simple adaptive procedure with proven FDR control

under independence, and demonstrated FDR control

under positive dependence BY,Krieger&Yekutieli('01,?)

Stage I: Use the Linear Stepup with q/(1+q), rejecting r_1;

if r_1=0 stop

Stage II: Estimate m_0 = (m- r_1)(1+q),

Then use it again with q*= q m/ m_0

FDR in High Throughput Screening:

Makes sense in screening experiments which are followed by an independent study

- First study FDR is controlled at q_1.
- Second study can be conducted on the set of identified genes, controlling for FDR/FWE at level q_2.

still the overall FDR/FWE level is $q_1 * q_2$

(so 0.25*0.2 =.05)

Outline

Introduction

QTL analysis

- History
- Recent Advancements

Gene Expression

- History
- Recent Advancements

Behavioral Genetics

- History
- Recent Advancements

Moral of Stories

3. Behavioral Genetics

- Crabbe et al (Science '99) raise the issue of replicability across laboratories issue. They refuse to control for multiplicity, continuing the argument in their website.
- At a meeting in Koln (winter 2000) on standardization issues in mouse behavior measurements:

About 20 posters presented in the poster session.

# of comparison was 4-120, with median at about 28.

None (but one) controlled for the effect of multiplicity

- The mutagenesis project