Permutation tests

Karl Broman

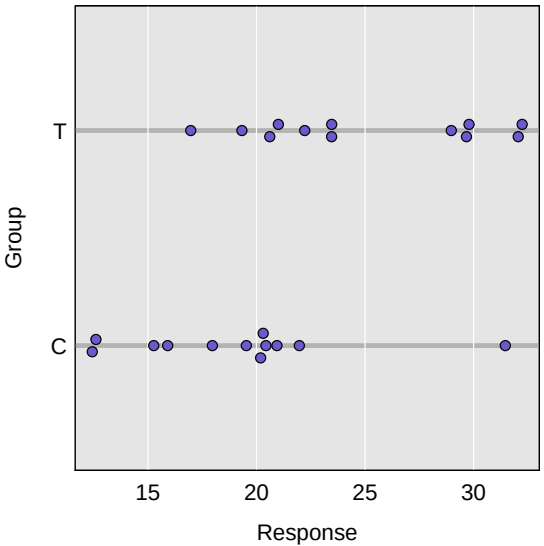
Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

Randomized experiment

Treatment groups				Responses			
C	T	T	T	12.6	32.1	21.0	29.8
T	T	C	C	23.5	17.0	19.5	15.3
C	C	T	T	31.5	22.0	29.7	19.3
T	T	C	C	22.2	20.6	20.9	12.4
C	C	T	C	20.4	20.3	32.2	18.0
T	C	C	T	23.5	20.2	15.9	29.0

Experimental results



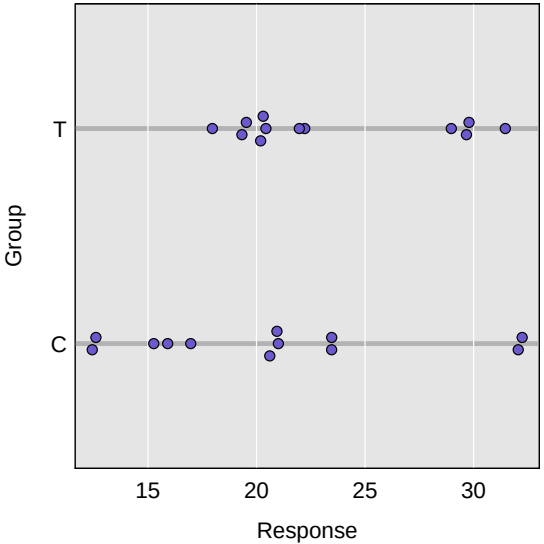
$\overline{Y}_T - \overline{Y}_C = 5.9$

$\hat{SE} = 2.1$

$t = 2.79$ $P = 0.01$

95% CI = (1.5, 10.3)

Permuted results



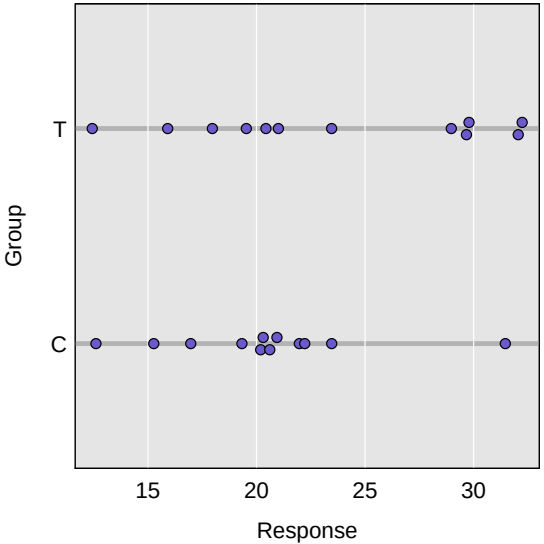
$\overline{Y}_T - \overline{Y}_C = 2.9$

$\hat{SE} = 2.4$

$t = 1.22$

95% CI $= (-2.0, 7.9)$

Permuted results



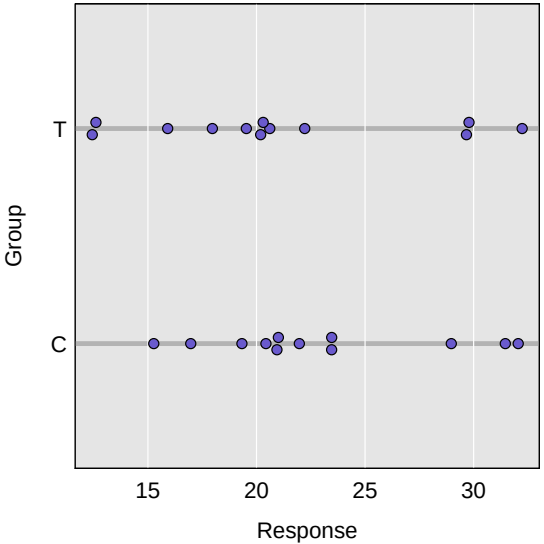
$\overline{Y}_T - \overline{Y}_C = 3.2$

$\hat{SE} = 2.4$

$t = 1.34$

95% CI $= (-1.8, 8.1)$

Permuted results



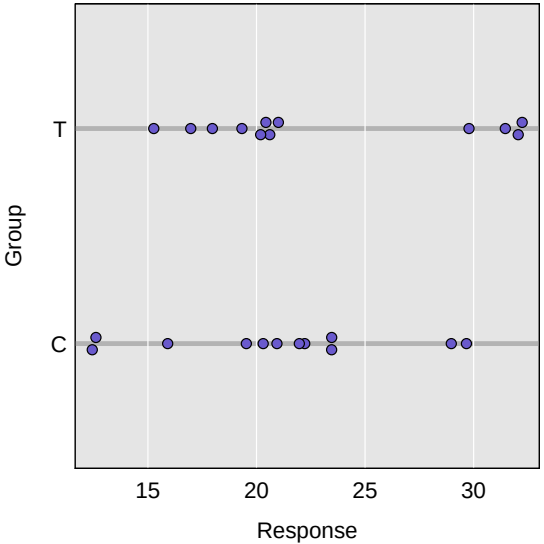
$\overline{Y}_T - \overline{Y}_C = -1.8$

$\hat{SE} = 2.4$

$t = -0.75$

95% CI $= (-6.9, 3.2)$

Permuted results



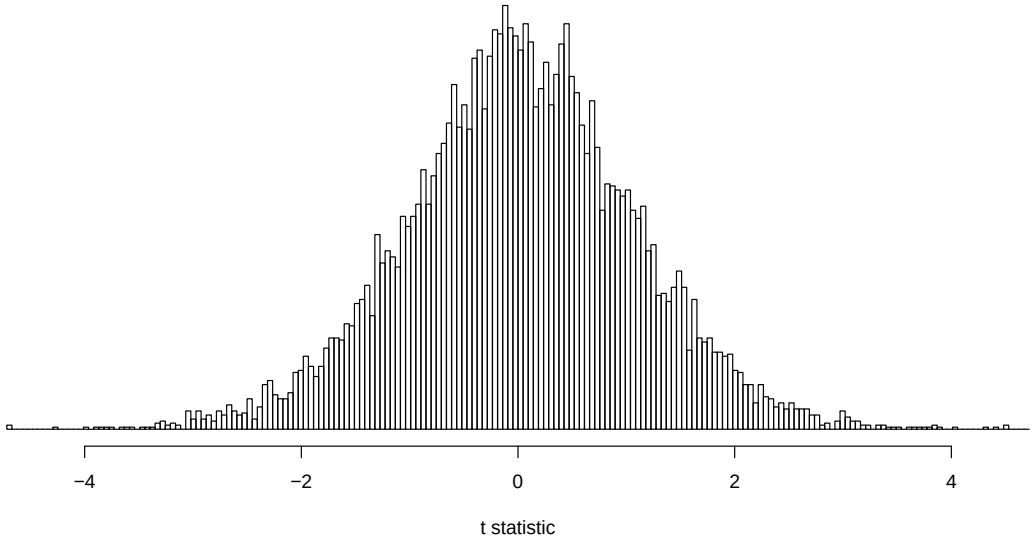
$\overline{Y}_T - \overline{Y}_C = 2.2$

$\hat{SE} = 2.4$

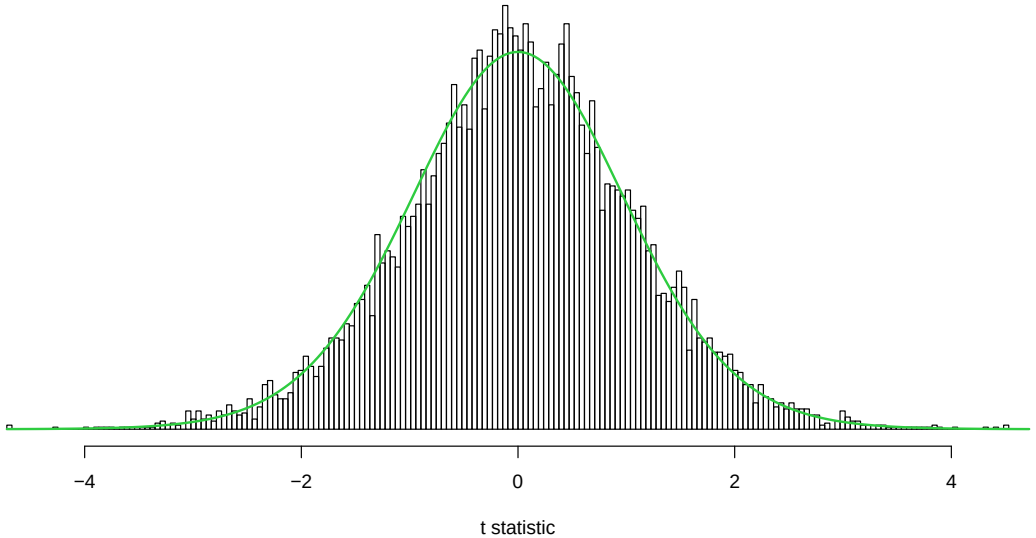
$t = 0.89$

95% CI $= (-2.9, 7.2)$

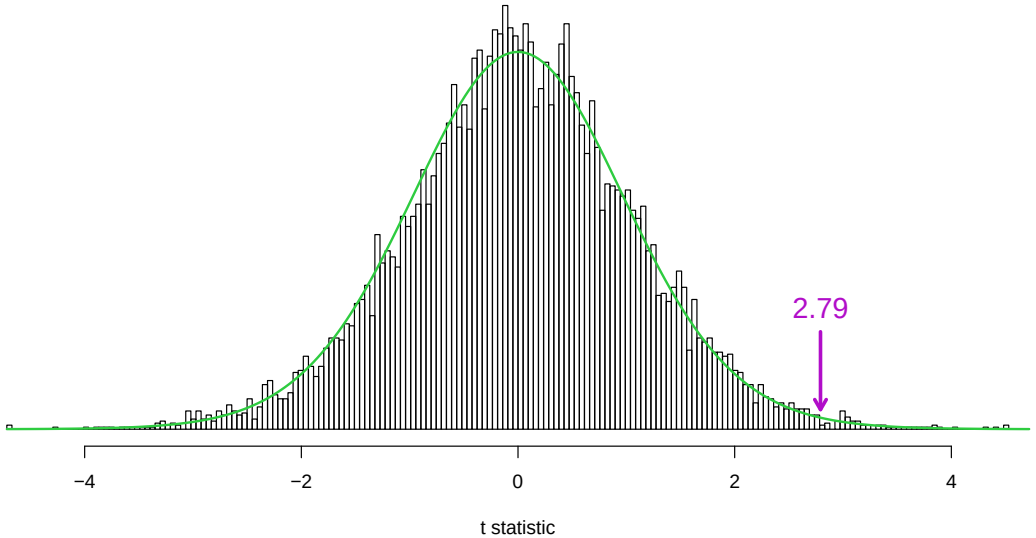
10,000 permutations



10,000 permutations



10,000 permutations



Assumptions for the permutation test

The observations are exchangeable
under the null hypothesis.

What test statistic?

- Anything will be valid
- Focus on power
- Robustness can still be important

  For example, resistance to outliers

How many permutations?

- Typically n = 1,000 or 10,000
- Focus on getting a good estimate of the p-value
- X = number of permutations $\geq$ observed value

  $\sim$ binomial(n, p) where p = true p-value
- With small datasets, may be able to do an exhaustive enumeration.

Empirical Threshold Values for Quantitative Trait Mapping

G. A. Churchill and R. W. Doerge

Biometrics Unit, Cornell University, Ithaca, New York 14853

Manuscript received April 22, 1994
Accepted for publication July 25, 1994

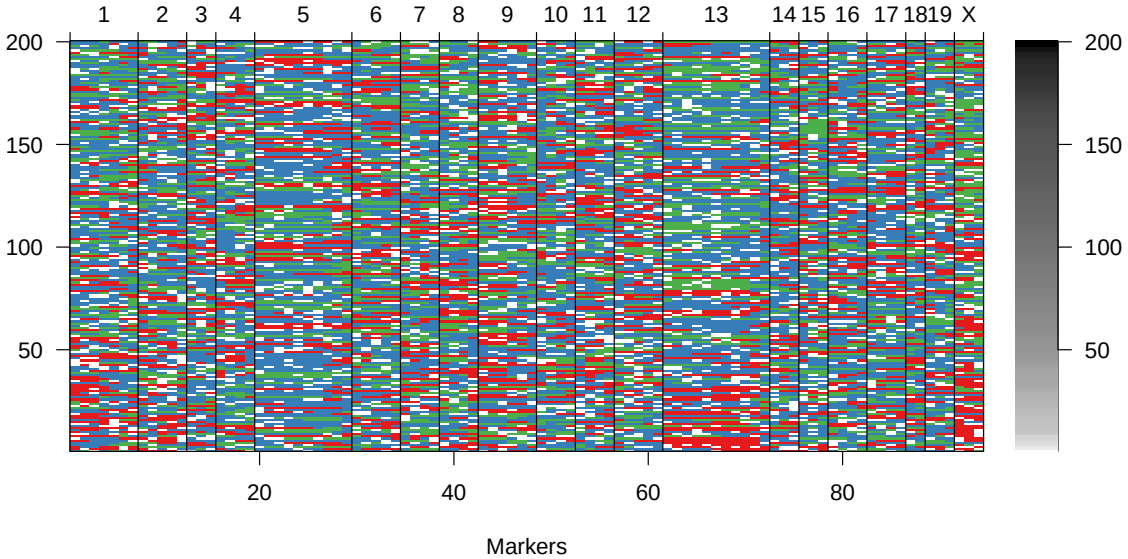
ABSTRACT

The detection of genes that control quantitative characters is a problem of great interest to the genetic mapping community. Methods for locating these quantitative trait loci (QTL) relative to maps of genetic markers are now widely used. This paper addresses an issue common to all QTL mapping methods, that of determining an appropriate threshold value for declaring significant QTL effects. An empirical method is described, based on the concept of a permutation test, for estimating threshold values that are tailored to the experimental data at hand. The method is demonstrated using two real data sets derived from F_2 and recombinant inbred plant populations. An example using simulated data from a backcross design illustrates the effect of marker density on threshold values.

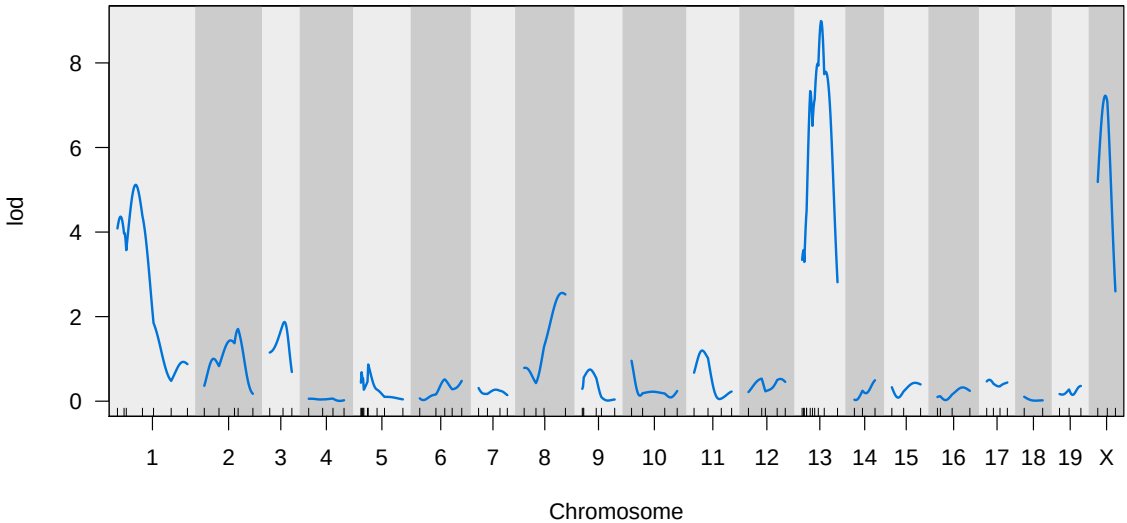
METHODOLOGICAL research on the problems of detecting and locating quantitative trait loci (QTL) has received considerable attention over the past several years. A variety of methods have been developed

The problem of determining appropriate threshold values is made even more difficult because there are many factors that can vary from experiment to experiment and can influence the distribution of the test sta-

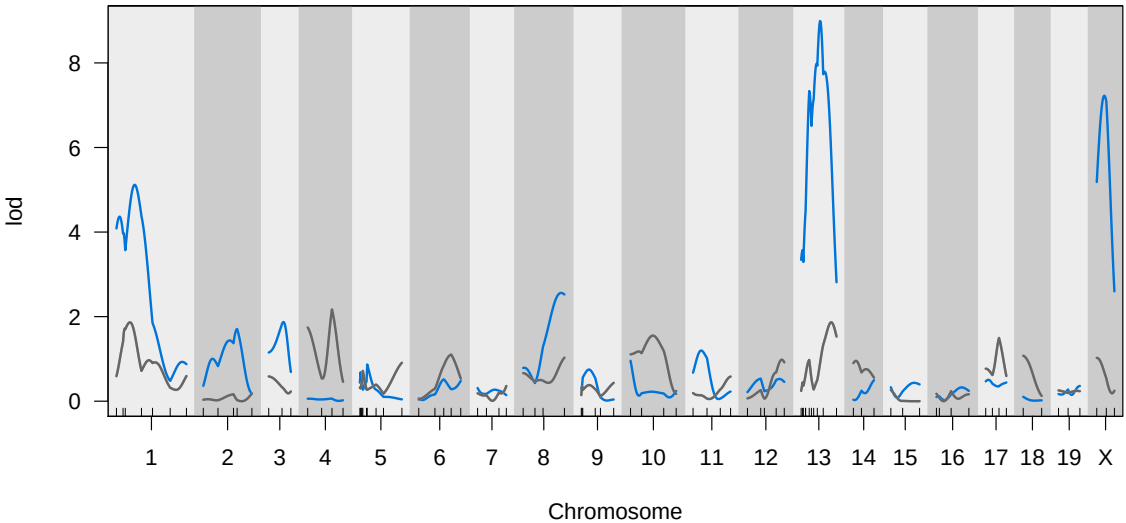
QTL data


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
200
150
100
50
20
40
60
80
Markers
200
150
100
50

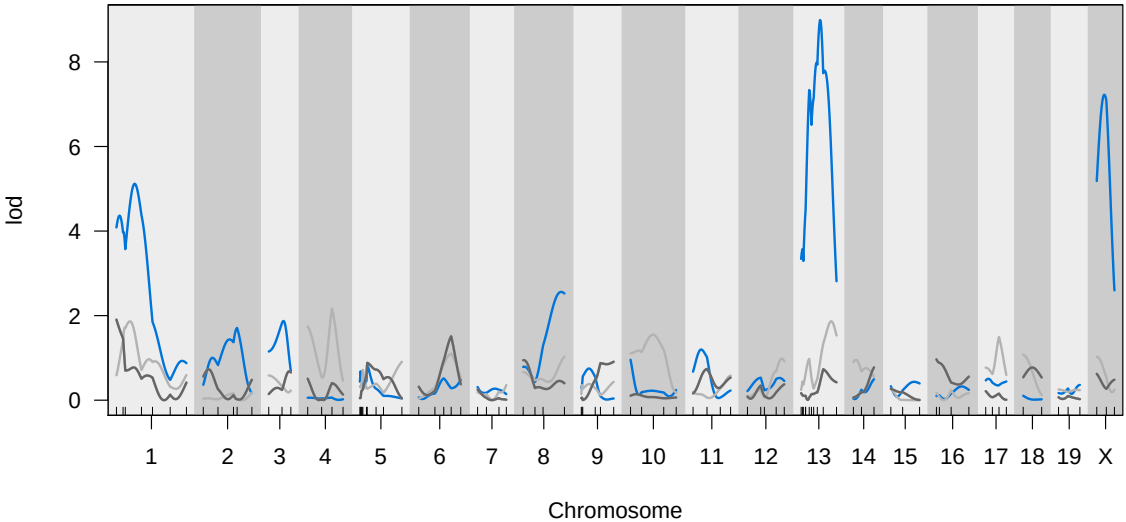
QTL genome scan



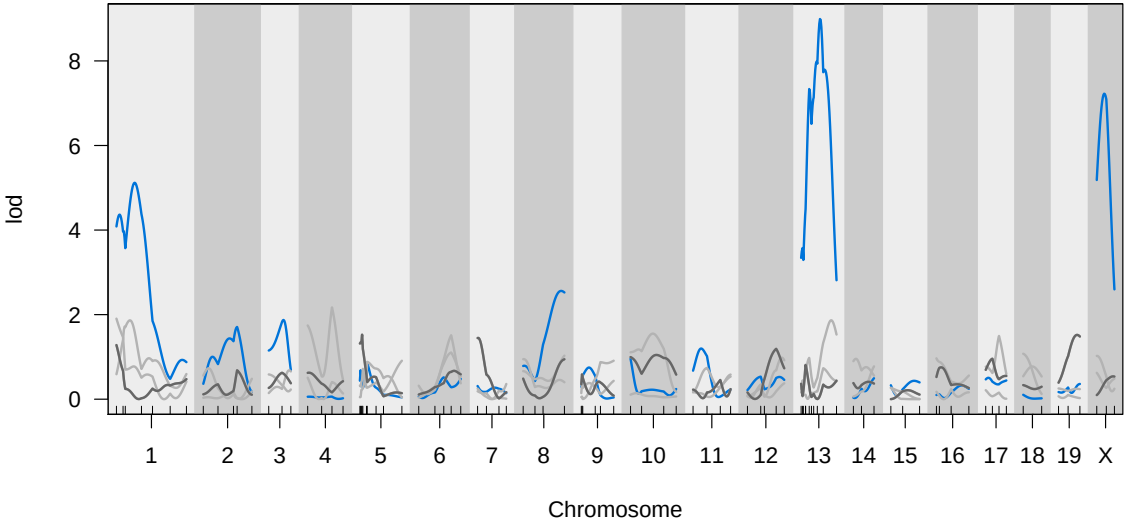
QTL genome scan



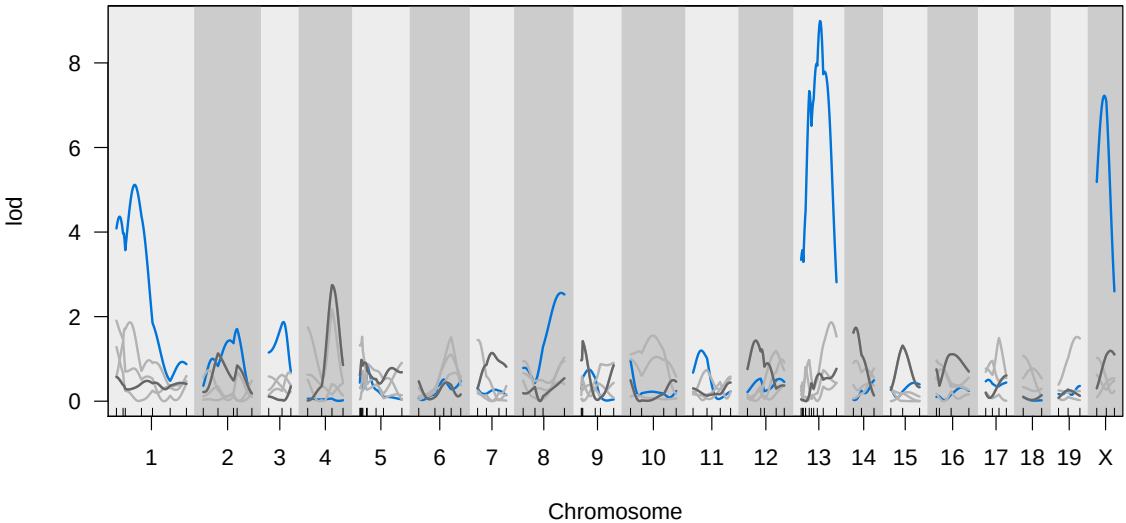
QTL genome scan



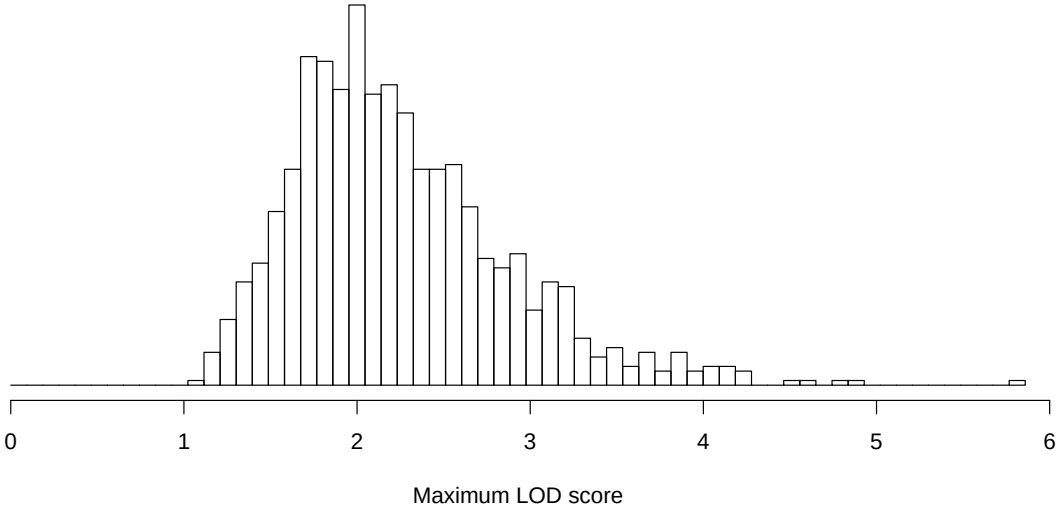
QTL genome scan



QTL genome scan



Permutation results



Multiple testing

- Many examples
  - gene expression or proteomic studies
  - genome-wide association studies
  - 1000s of predictors in an epi study
- Most stringent approach: control family-wise error rate (FWER)
- A Bonferroni adjustment can be too conservative
- Take max statistic in each permutation replicate

If test statistic varies

- taking $\max(X_j)$ assumes that the X_j have a common null distribution
- if not, you'd want to normalize so they do
- One approach: use the permutation results to do so
  - for each column of permutation results, turn values into ranks
  - then find the maximum rank in each row
  - find where the observed statistics rank within each column
  - This gives adjusted p-values that account for the search

Abuse of p-values

- Focusing on strict, arbitrary thresholds like 0.05
- Not looking at the confidence interval for the effect
- Ignoring multiple comparisons
- Turning science into true/false questions

Abuse of p-values

- Focusing on strict, arbitrary thresholds like 0.05
- Not looking at the confidence interval for the effect
- Ignoring multiple comparisons
- Turning science into true/false questions

But I still like p-values.

Abuse of p-values

- Focusing on strict, arbitrary thresholds like 0.05
- Not looking at the confidence interval for the effect
- Ignoring multiple comparisons
- Turning science into true/false questions

But I still like p-values.

It's useful to ask, "Could this just be noise?"

Randomized block design

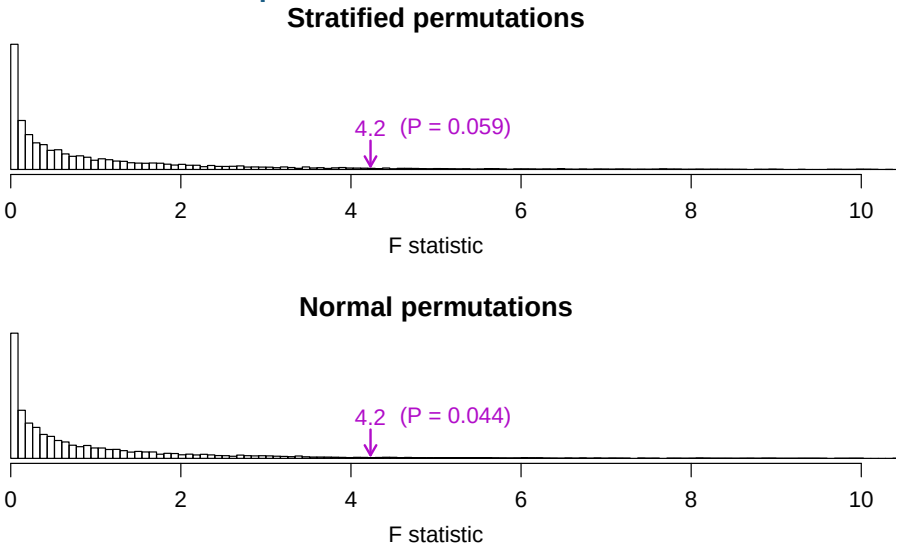
Treatment groups

T	T	T	C
C	C	T	C
T	C	T	T
C	T	C	C
T	T	C	C
C	C	T	T

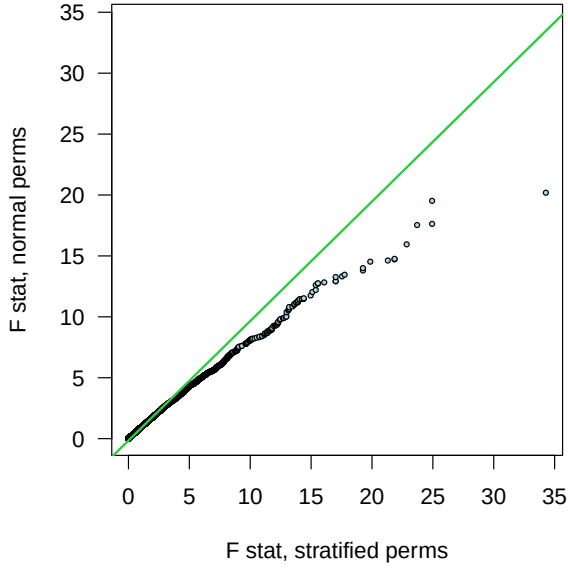
Responses

12.1	20.0	9.0	19.5
7.7	18.0	14.9	21.5
16.7	16.3	18.4	23.2
19.4	17.6	7.5	11.5
18.4	17.3	9.8	13.9
8.3	12.2	24.9	28.7

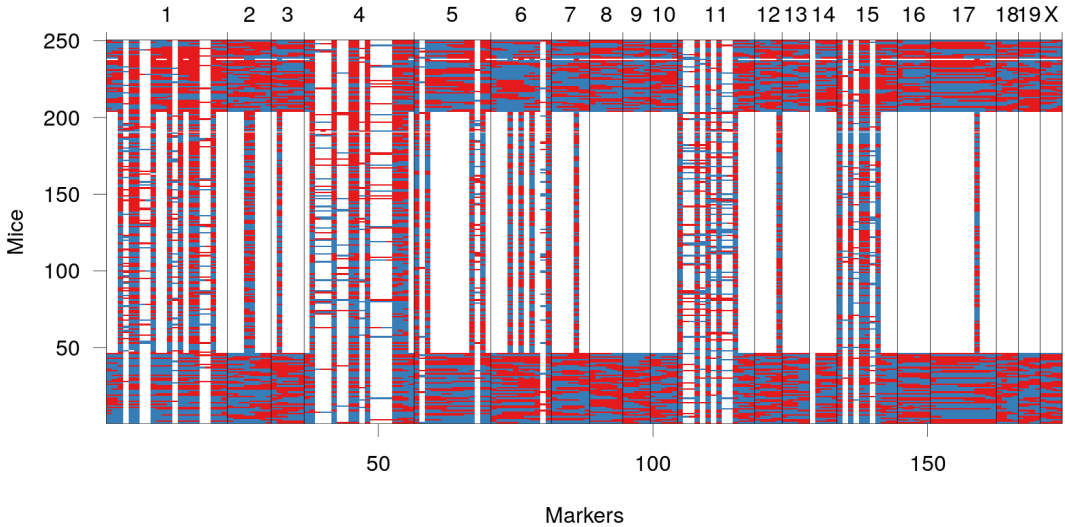
Stratified vs normal permutations



Stratified vs normal permutations



Selective genotyping



DOI: 10.1534/genetics.107.080093

Note

Significance Thresholds for Quantitative Trait Locus Mapping Under Selective Genotyping

Ani Manichaikul,* Abraham A. Palmer,† Śaunak Sen‡ and Karl W. Broman*,[1]

**Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, †Departments of Human Genetics and Psychiatry, University of Chicago, Chicago, Illinois 60637 and ‡Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94107*

Manuscript received August 6, 2007
Accepted for publication August 21, 2007

ABSTRACT

In the case of selective genotyping, the usual permutation test to establish statistical significance for quantitative trait locus (QTL) mapping can give inappropriate significance thresholds, especially when the phenotype distribution is skewed. A stratified permutation test should be used, with phenotypes shuffled separately within the genotyped and ungenotyped individuals.

Summary

- Permutation tests, when appropriate, are the most natural of significance test.
- Permutation tests can make it easy to control for multiple testing.
- Stratified permutation tests accommodate a common non-exchangeable situtation.
- Many are quite negative about p-values, but I still like them.