Why you should know about experimental crosses

Why you should know about experimental crosses

To save you from embarrassment

Why you should know about experimental crosses

To save you from embarrassment

To help you understand and analyse human genetic data

Why you should know about experimental crosses

To save you from embarrassment

To help you understand and analyse human genetic data

It’s interesting

Experimental crosses

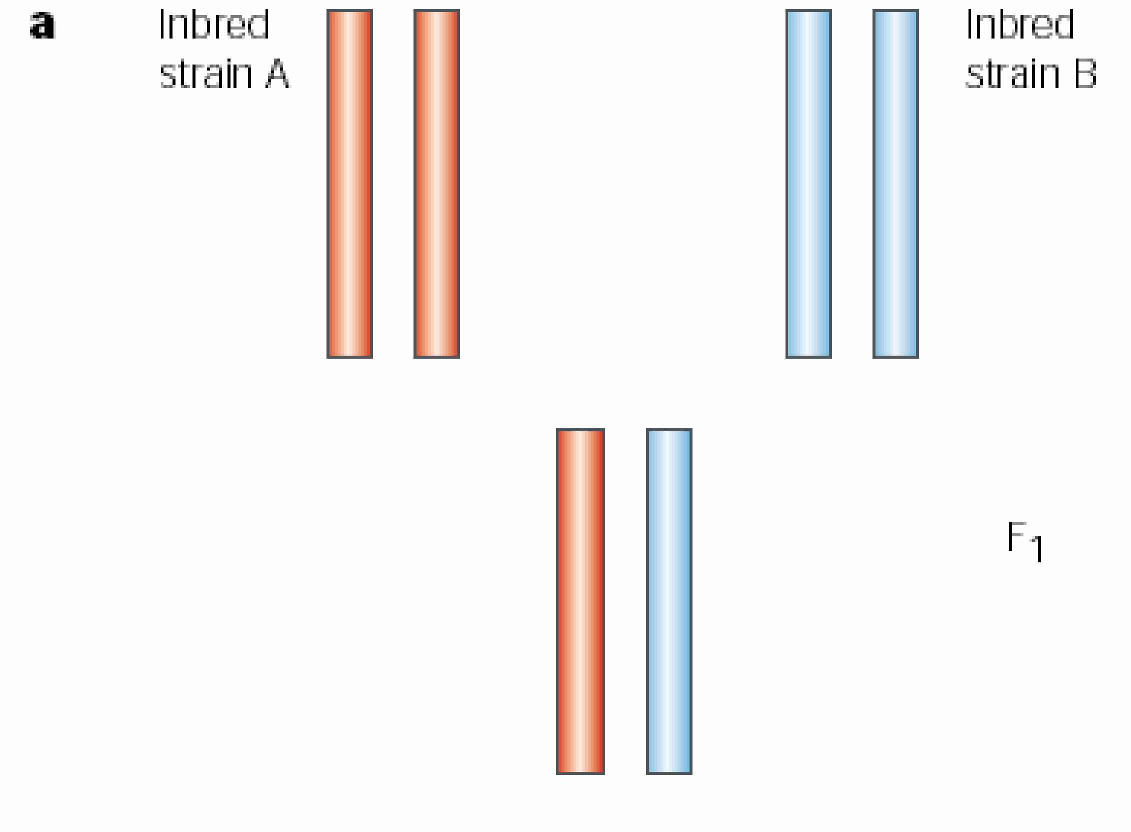
Experimental crosses

Inbred strain crosses

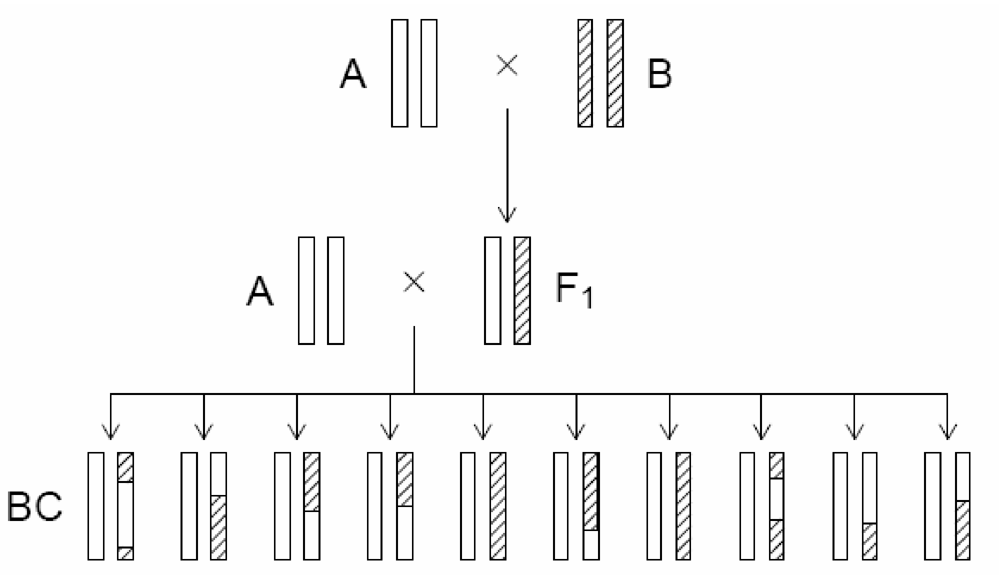
Recombinant inbreds

Alternatives

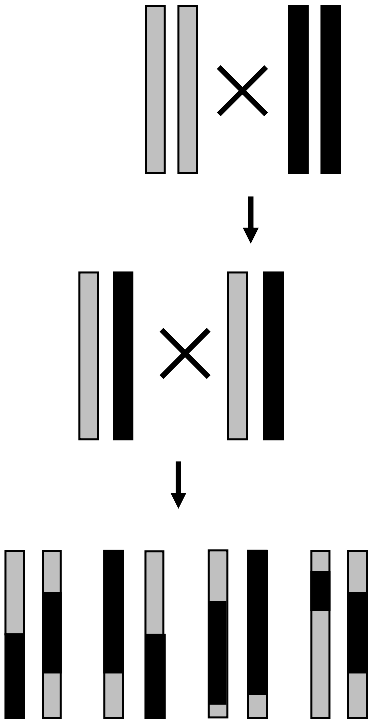
Inbred Strain Cross



Backcross



F2 cross



Generation

F0

F1

F2

Data conventions

AA = A

BB = B

AB = H

Missing data = -

Data conventions

Genotype file

ID	M1	M2	M3
A01231	A	A	A
A07612	B	-	H
A01812	H	H	A

Data conventions

Genotype file

ID	**M1**	**M2**	**M3**
A01231	A	A	A
A07612	B	-	H
A01812	H	H	A

Phenotype file

ID	**Phenotype**	**Covariate**
A01231	10	F
A07612	-	F
A01812	8	M

Data conventions

Genotype file

ID	**M1**	**M2**	**M3**
A01231	A	A	A
A07612	B	-	H
A01812	H	H	A

Phenotype file

ID	**Phenotype**	**Covariate**
A01231	10	F
A07612	-	F
A01812	8	M

Map file

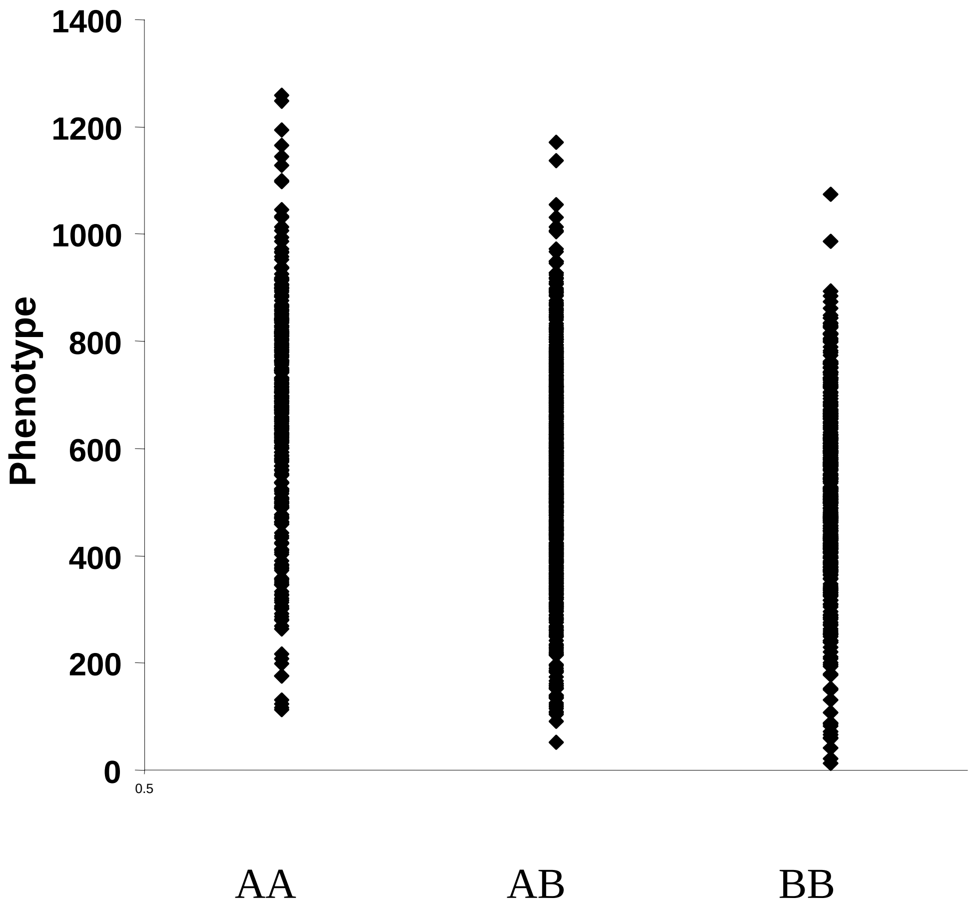
Map file

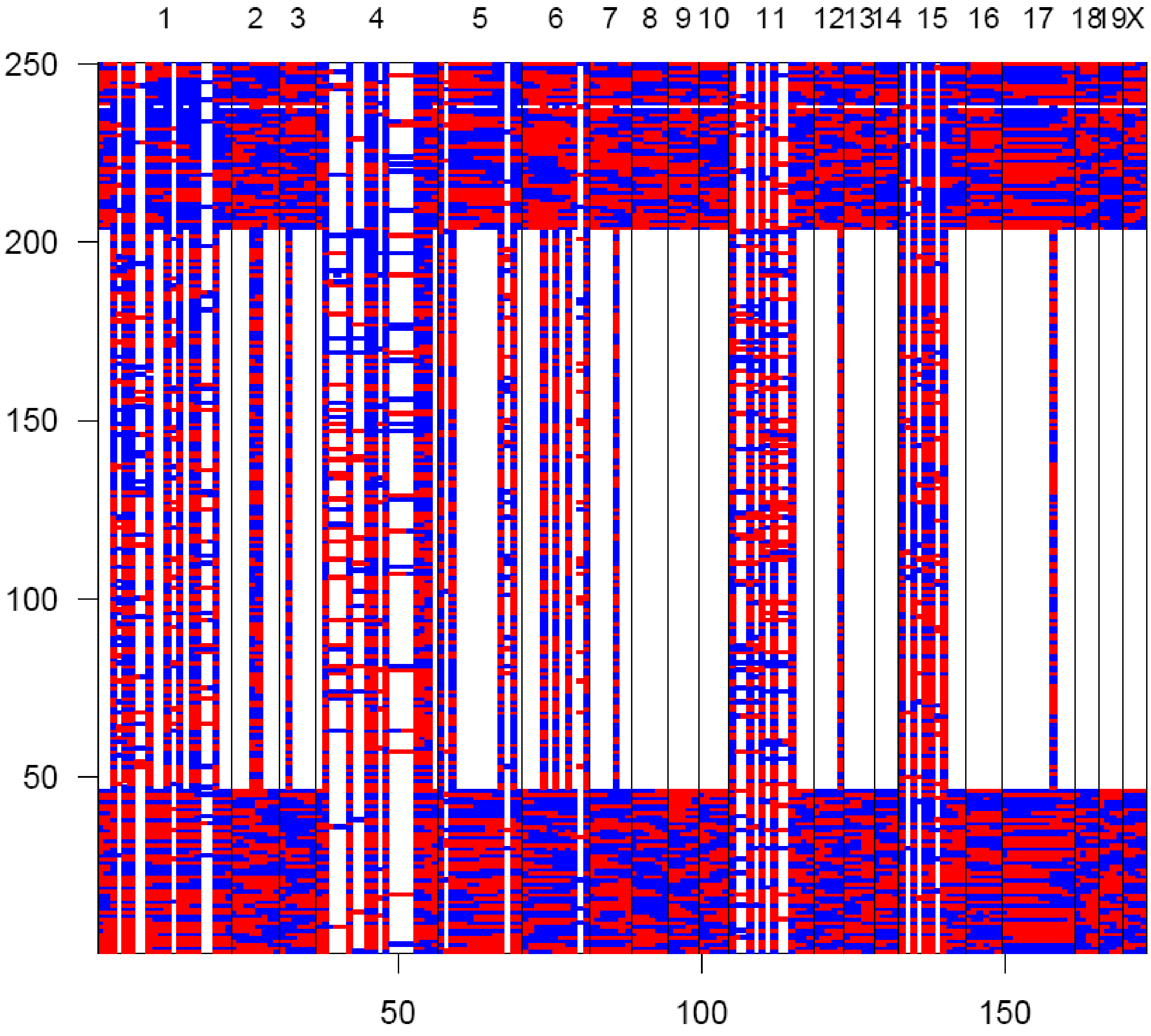
Use the latest mouse build and convert physical to genetic distance: 1 Mb = 1.6 cM

Use our genetic map: http://gscan.well.ox.ac.uk/

Analysis

If you can’t see the effect it probably isn’t there


1400
1200
1000
800
600
400
200
0
Phenotype
0.5
AA
AB
BB



Red = Hom

Blue = Het

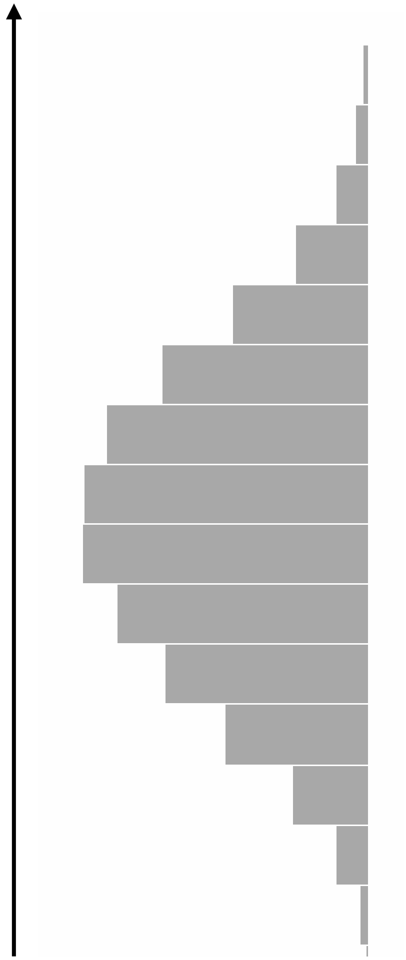
Backcross genotypes

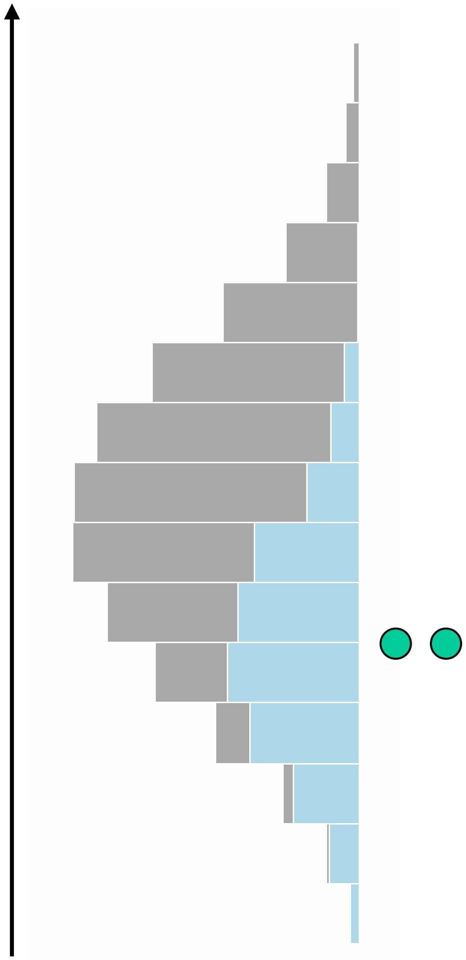
Statistical analysis

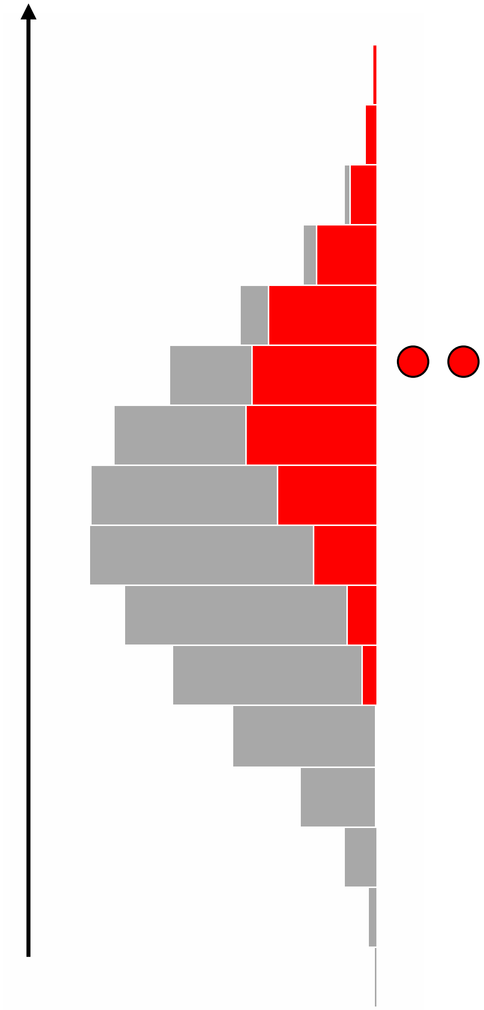
Linear models

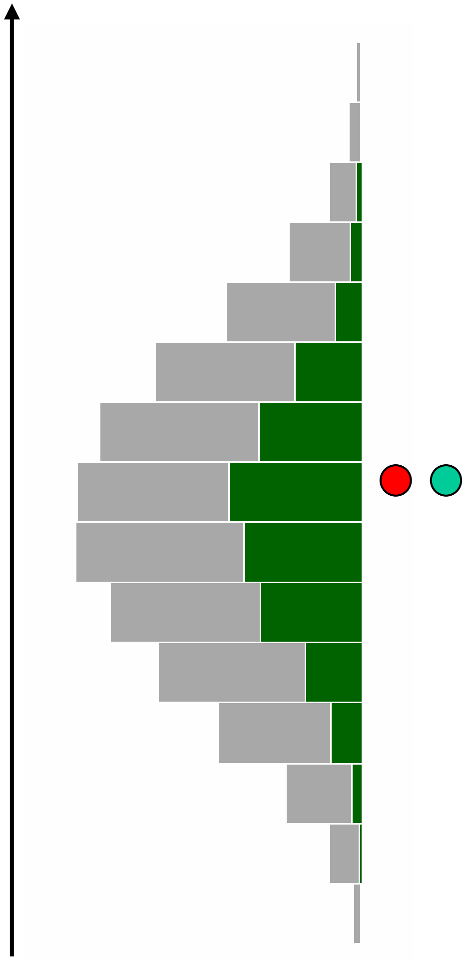
Also known as

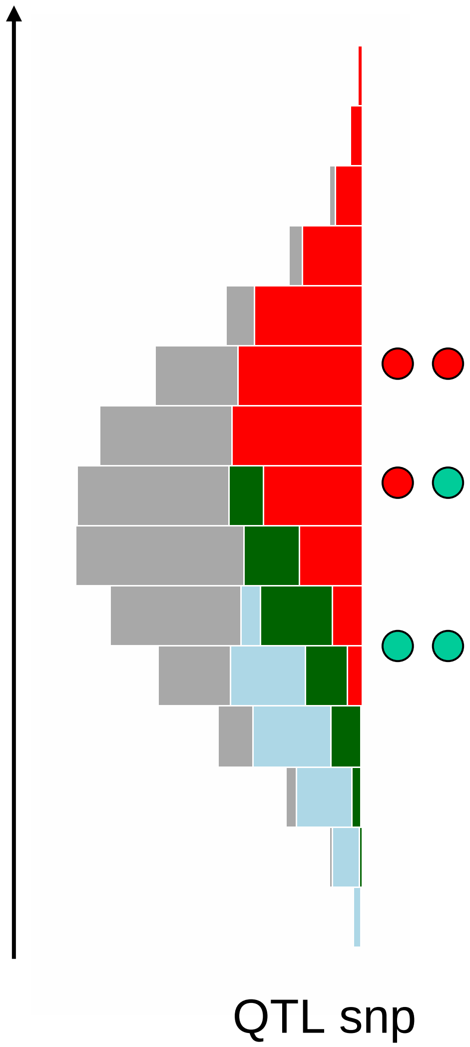
ANOVA
ANCOVA
regression
multiple regression
linear regression

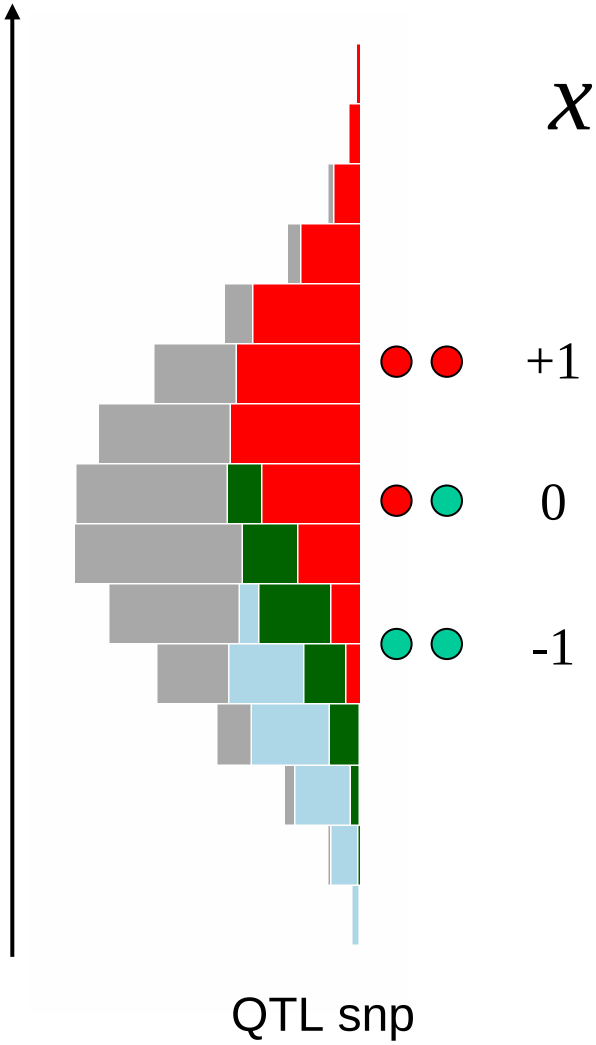


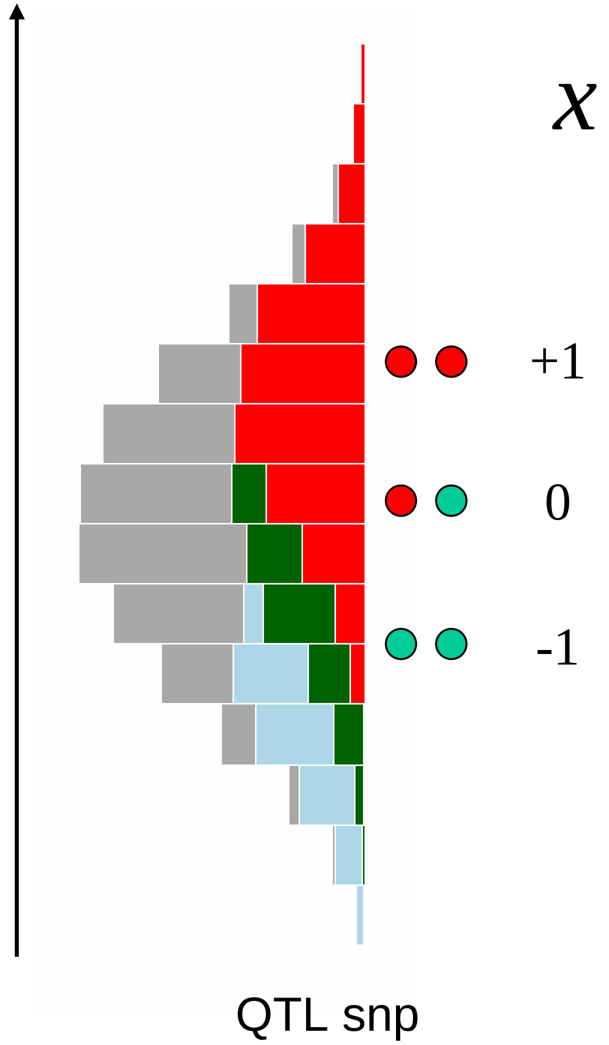




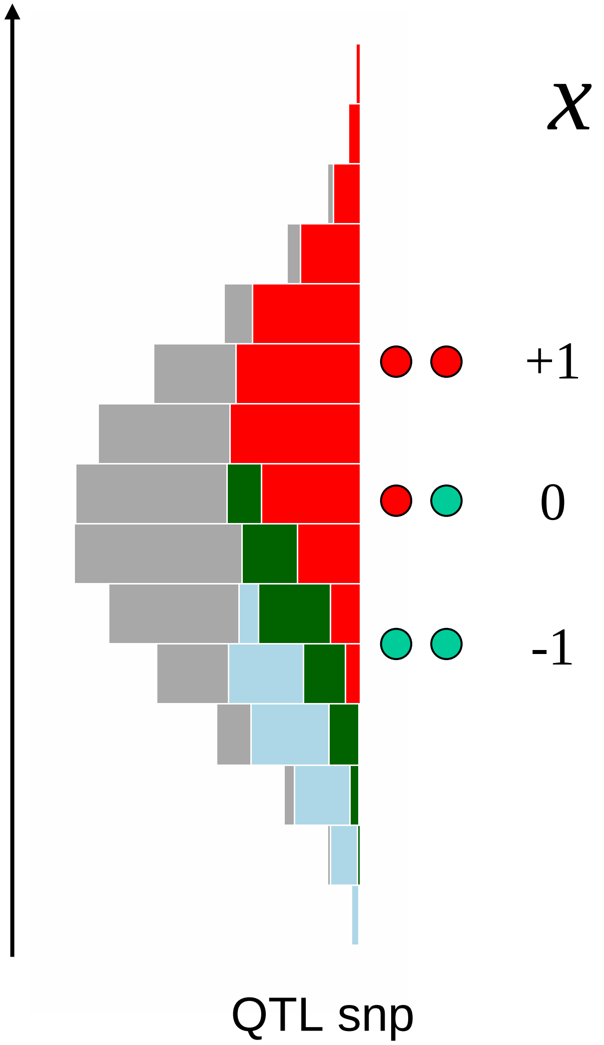



QTL snp

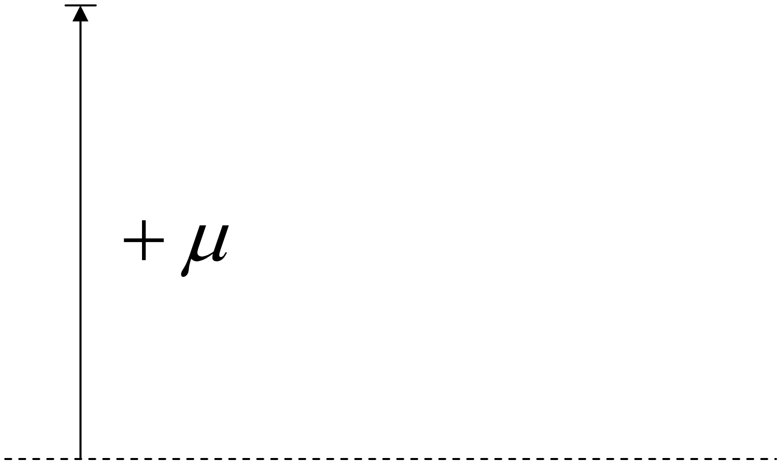

X
+1
0
-1
QTL snp

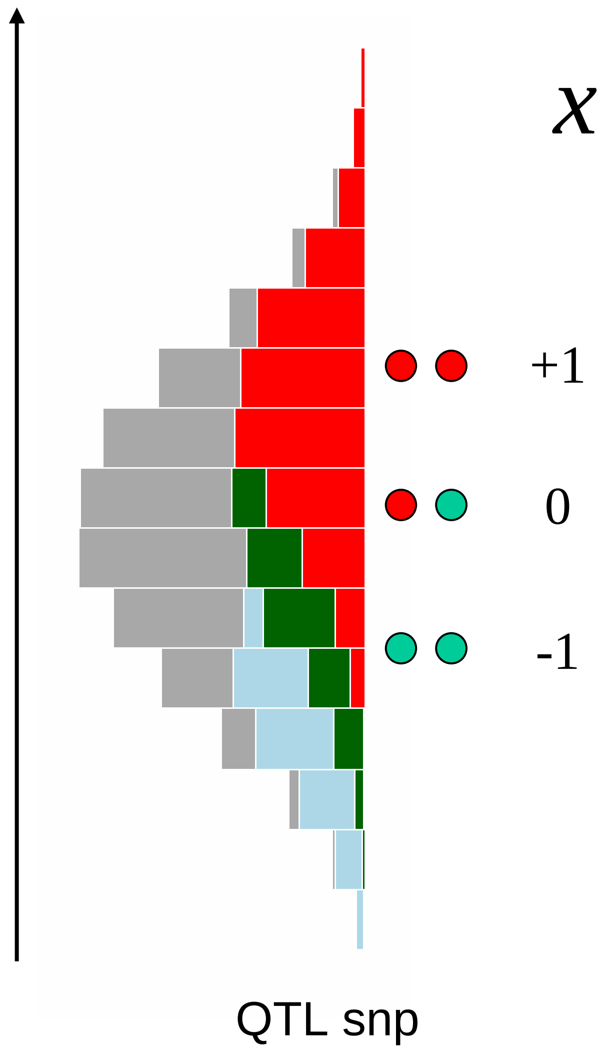


$$y = \mu + ax + \varepsilon$$

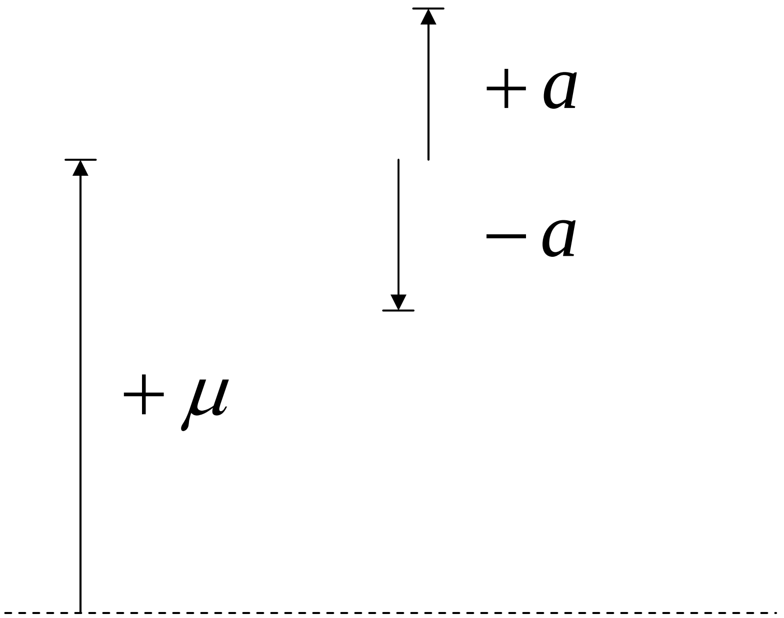


$$y = \mu + ax + \varepsilon$$

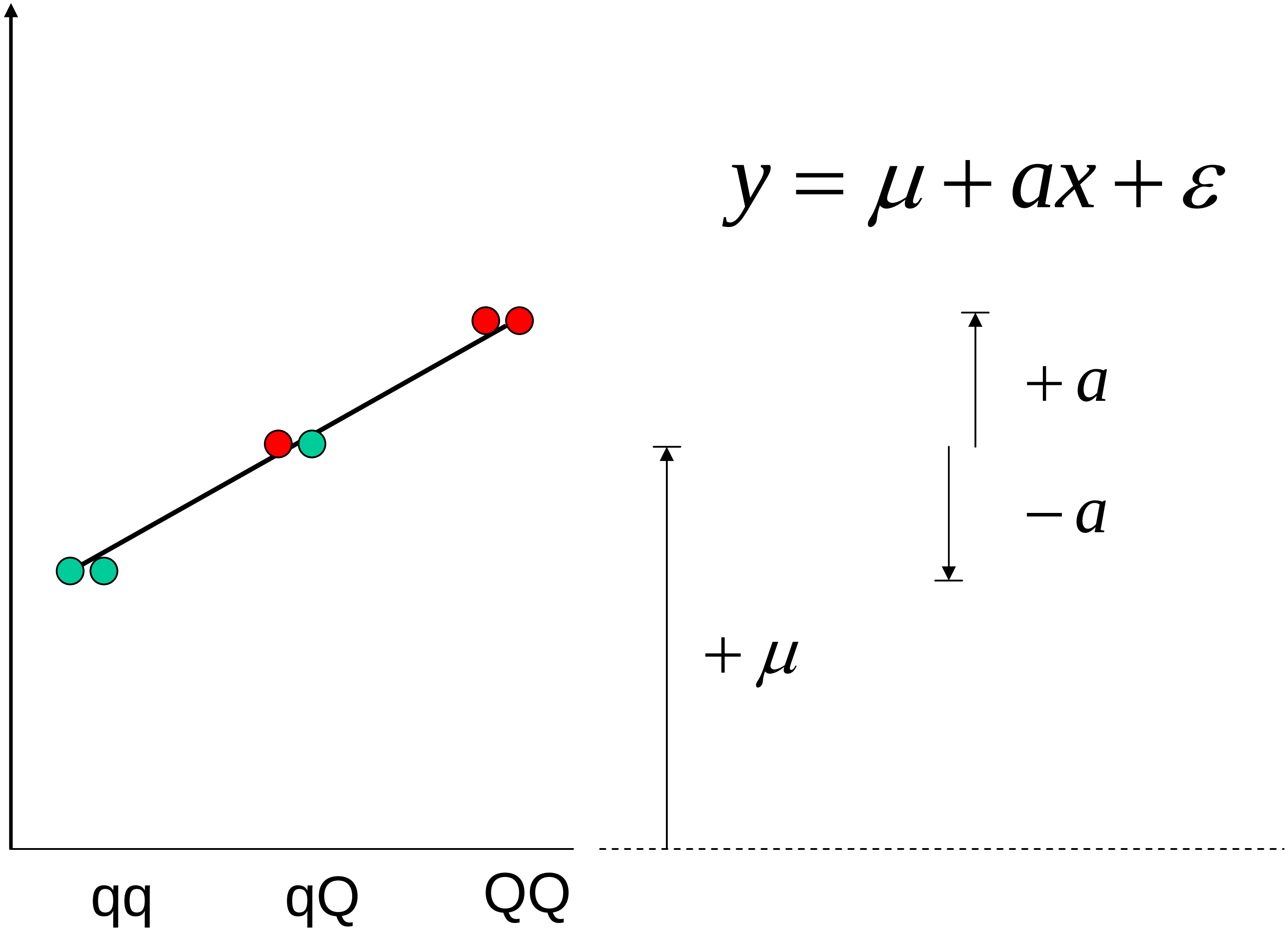


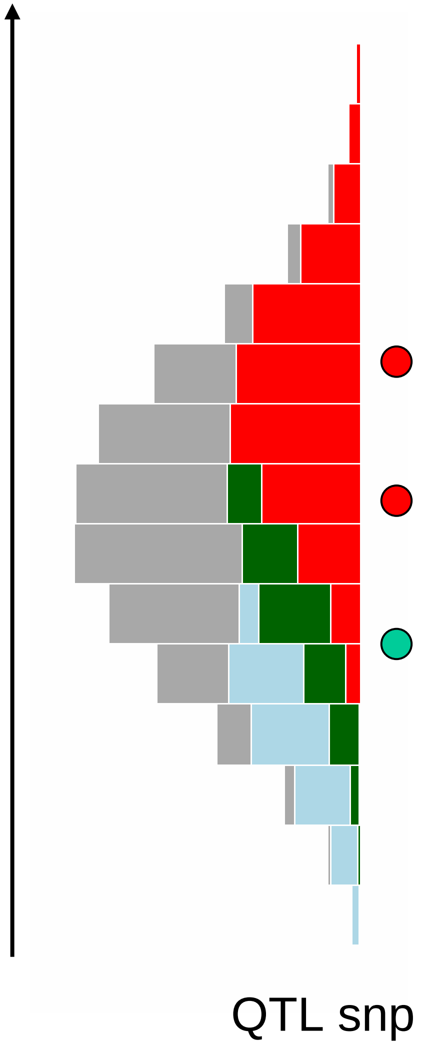


$$y = \mu + ax + \varepsilon$$



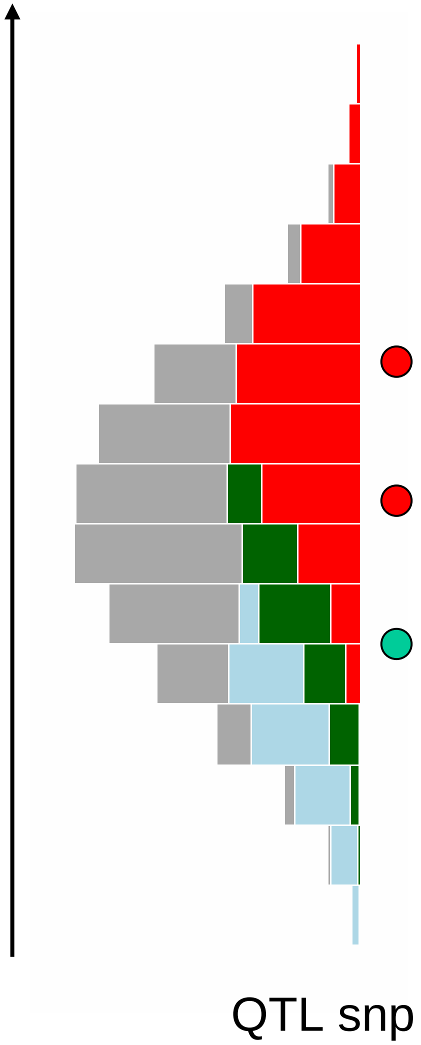
$$y = \mu + ax + \varepsilon$$



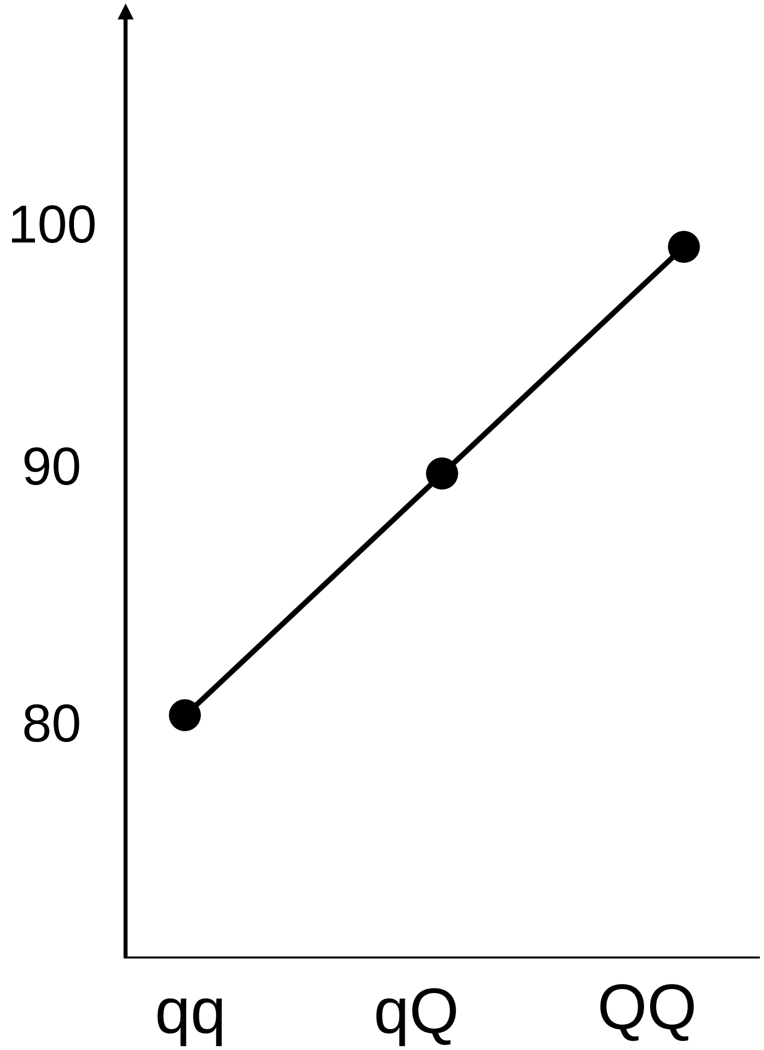


	x_A	x_D
🔴 🔴	+1	0
🔴 🟢	0	1
🟢 🟢	-1	0

$$y = \mu + a x_A + d x_D + \varepsilon$$

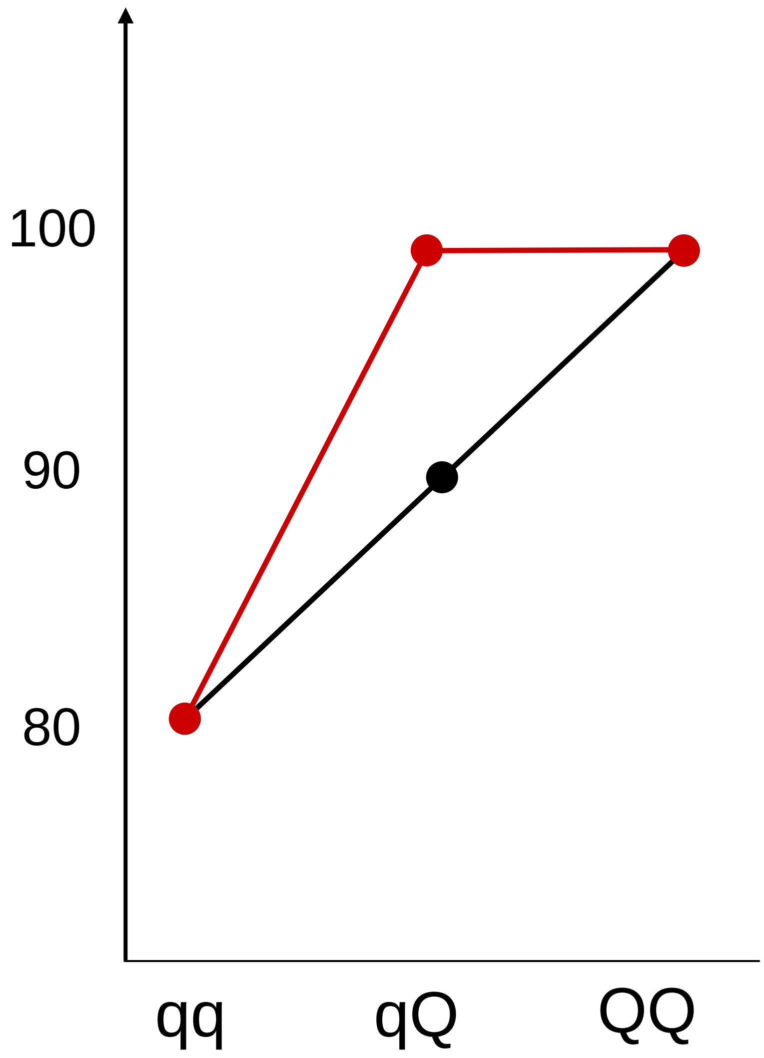


	x_A	x_D
🔴 🔴	+1	0
🔴 🟢	0	1
🟢 🟢	-1	0



$$y = \mu + a x_A + d x_D + \varepsilon$$

90 10 0



$$y = \mu + a x_A + d x_D + \varepsilon$$

90 10 0

90 10 10

Hypothesis testing

H_0: $$y = \mu + \varepsilon$$

H_1: $$y = \mu + ax + \varepsilon$$

Hypothesis testing

H_0: y ~1

$$y = \mu + \varepsilon$$

H_1: y ~ 1 + **x**

$$y = \mu + ax + \varepsilon$$

Hypothesis testing

H_0: y ~ 1

$$y = \mu + \varepsilon$$

H_1: y ~ 1 + **x**

$$y = \mu + ax + \varepsilon$$

H_1 vs H_0 : Does **x** explain a significant amount of the variation?

Hypothesis testing

H_0: y ~ 1

$$y = \mu + \varepsilon$$

H_1: y ~ 1 + **x**

$$y = \mu + ax + \varepsilon$$

H_1 vs H_0 : Does **x** explain a significant amount of the variation?

likelihood ratio

LOD score

Hypothesis testing

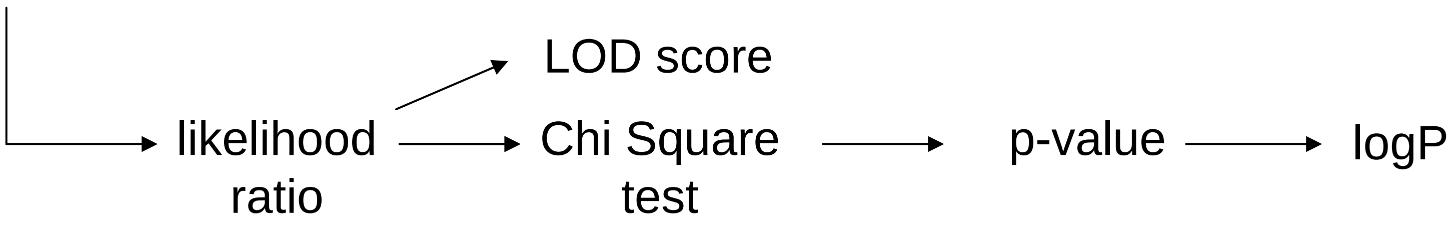
H_0: y ~ 1

$$y = \mu + \varepsilon$$

H_1: y ~ 1 + **x**

$$y = \mu + ax + \varepsilon$$

H_1 vs H_0 : Does **x** explain a significant amount of the variation?



LOD score

likelihood ratio → Chi Square test → p-value → logP

Hypothesis testing

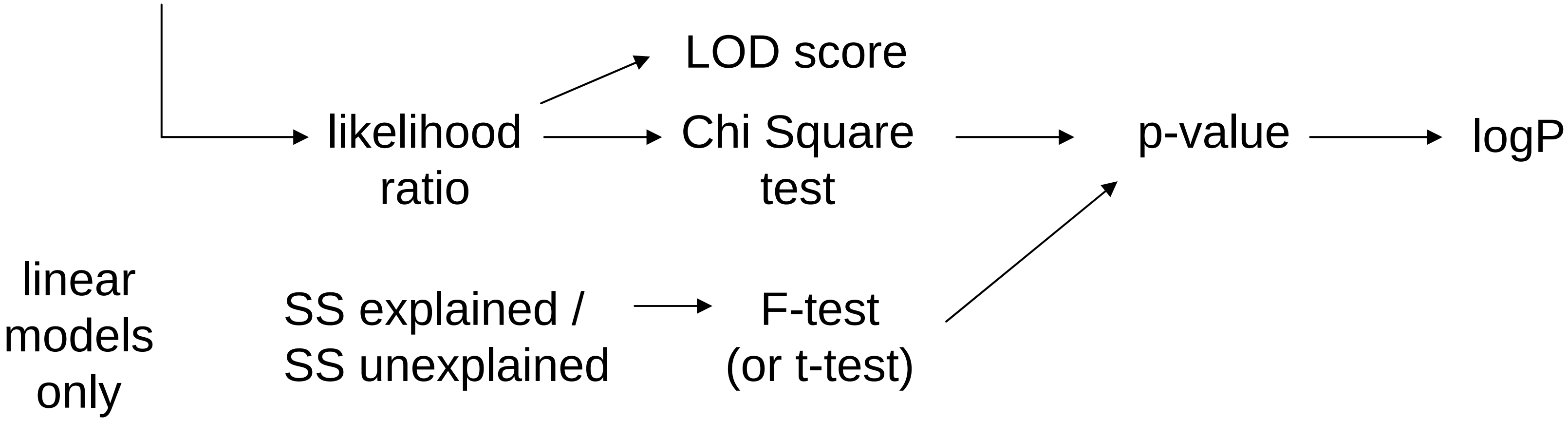
H_0: $y \sim 1$

$$y = \mu + \varepsilon$$

H_1: $y \sim 1 + x$

$$y = \mu + ax + \varepsilon$$

H_1 vs H_0 : Does **x** explain a significant amount of the variation?



likelihood ratio → Chi Square test → p-value → logP

likelihood ratio → LOD score

linear models only

SS explained / SS unexplained → F-test (or t-test) → p-value

Hypothesis testing

H_0: y ~ 1 + x

$$y = \mu + ax + \varepsilon$$

H_1: y ~ 1 + x + **x2**

$$y = \mu + ax_A + dx_D + \varepsilon$$

H_1 vs H_0 : Does **x2** explain a significant extra amount of the variation?

PRACTICAL: hypothesis test for identifying QTLs

To start:

1. Copy the folder faculty\valdar\AnimalModelsPractical to your own directory.
2. Start R
3. File -> Change Dir… and change directory to your AnimalModelsPractical directory
4. Open Firefox, then File -> Open File, and open “f2cross_and_thresholds.R” in the AnimalModelsPractical directory

H_0: phenotype ~ 1

H_1: phenotype ~ **a**

H_2: phenotype ~ a + **d**

Test:

H_1 vs H_0

H_2 vs H_1

H_2 vs H_0

PRACTICAL: Chromosome scan of F2 cross

Two problems in QTL analysis

Missing genotype problem

Model selection problem

Missing genotype problem

M1	**M2**	**M3**	**Q**	**M4**	**M5**
H	A	A	-	A	A
H	H	H	-	A	A
B	B	-	-	H	H

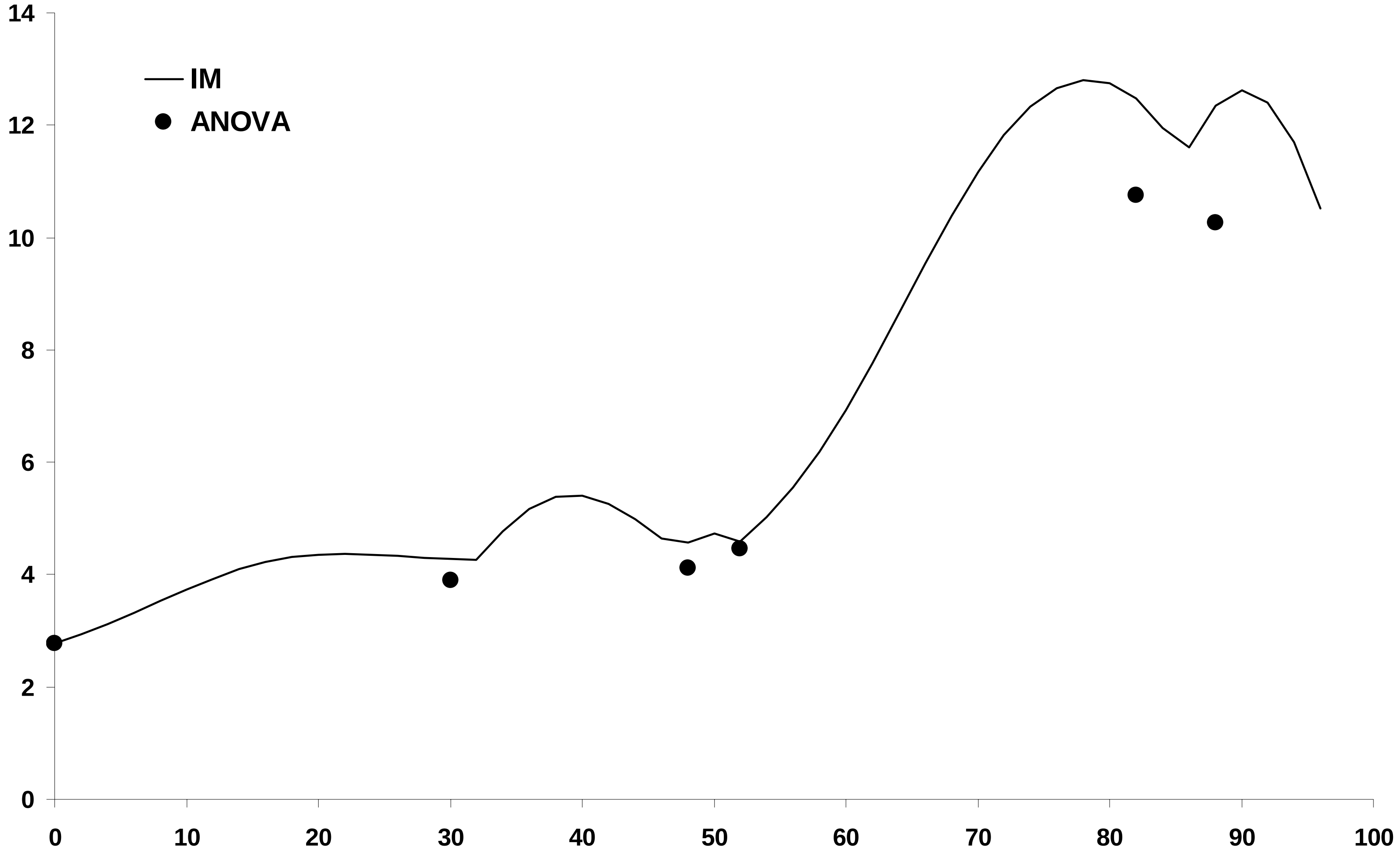
Solutions to the missing genotype problem

Maximum likelihood interval mapping

Haley-Knott regression

Multiple imputation

Interval mapping



Interval mapping

qq genotype $\longrightarrow$ 10

qQ genotype $\longrightarrow$ 20

Interval mapping

qq genotype $\longrightarrow$ 10 $\Pr(qq) = {}^{1}\!/_{2}$

qQ genotype $\longrightarrow$ 20 $\Pr(qQ) = {}^{1}\!/_{2}$

Interval mapping

qq genotype → 10 $\Pr(qq) = {}^{1}\!/_{2}$

qQ genotype → 20 $\Pr(qQ) = {}^{1}\!/_{2}$

Which is the true situation?

qq $y = 10 + \varepsilon$

qQ $y = 20 + \varepsilon$

Interval mapping

qq genotype $\longrightarrow$ 10 $\Pr(qq) = \frac{1}{2}$

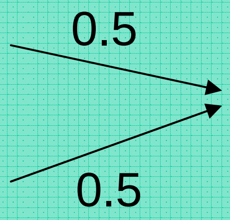
qQ genotype $\longrightarrow$ 20 $\Pr(qQ) = \frac{1}{2}$

Which is the true situation?

qq $$y = 10 + \varepsilon$$

qQ $$y = 20 + \varepsilon$$

Fit both situations and then weight them



"mixture" model

ML interval mapping

Interval mapping

qq genotype → 10 $\Pr(qq) = \frac{1}{2}$

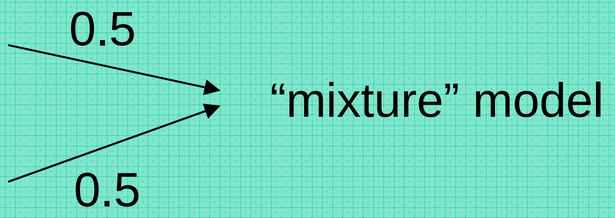
qQ genotype → 20 $\Pr(qQ) = \frac{1}{2}$

Which is the true situation?

qq $y = 10 + \varepsilon$

qQ $y = 20 + \varepsilon$

Fit both situations and then weight them



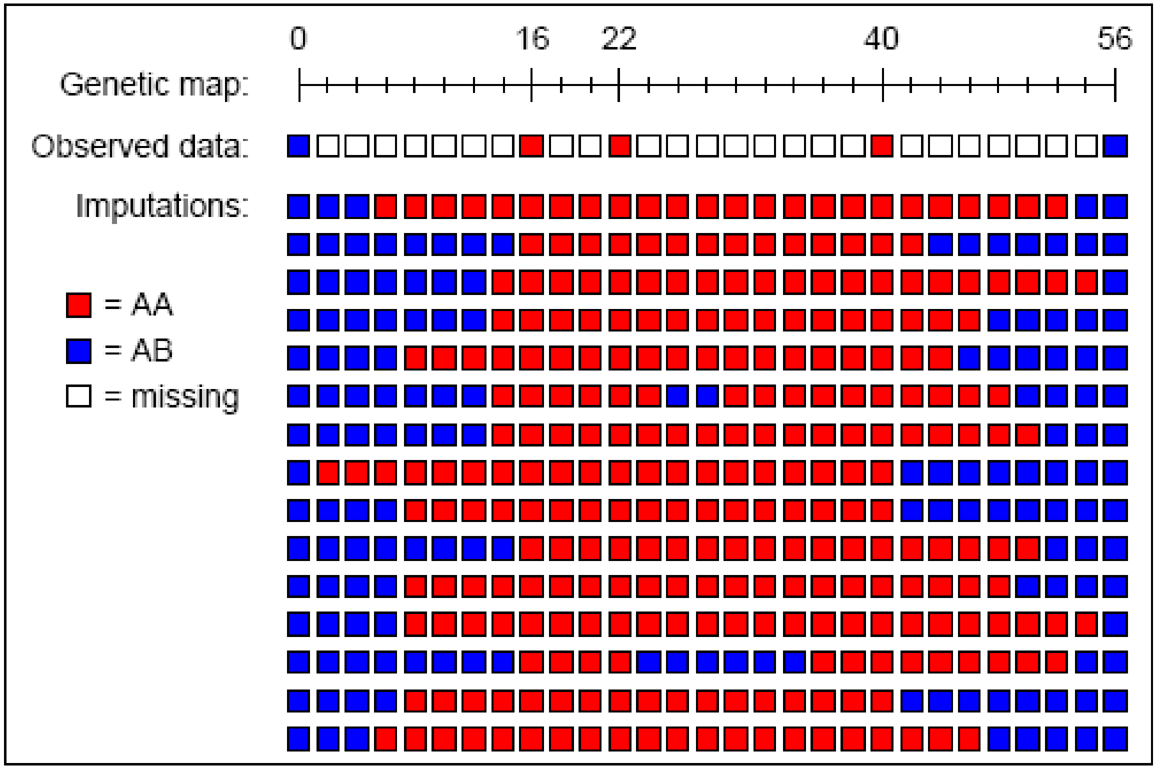
ML interval mapping

Fit the "average" situation
(which is technically false, but quicker)

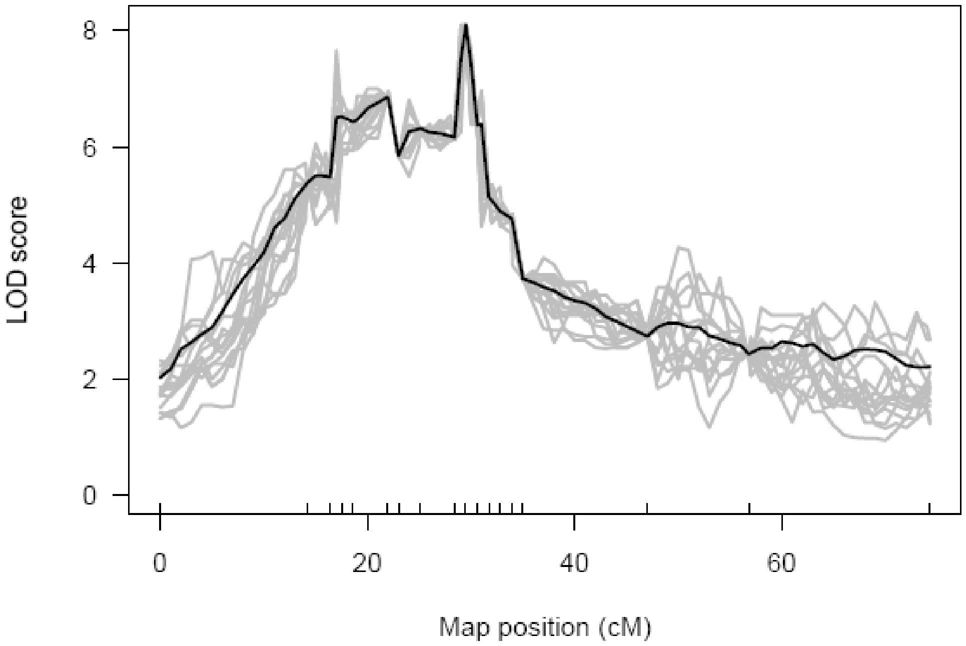
$$y = 15 + \varepsilon$$

Haley-Knott regression

Imputation


0
16
22
40
56
Genetic map:
Observed data:
Imputations:
= AA
= AB
= missing

Imputation

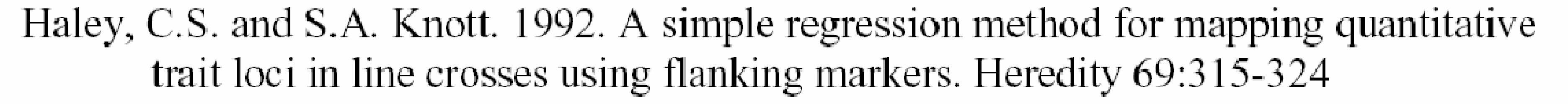


Key references

Maximum likelihood methods

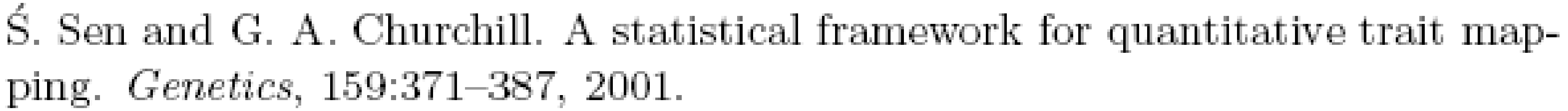
Lander, E.S. and D. Botstein.1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.

Linear regression



Haley, C.S. and S.A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324

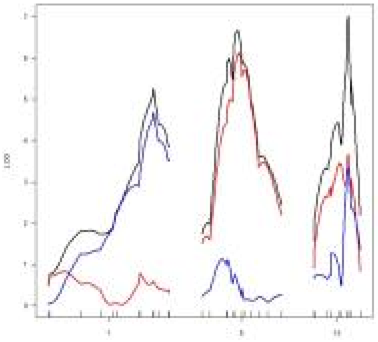
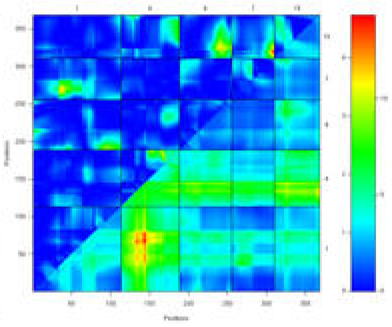
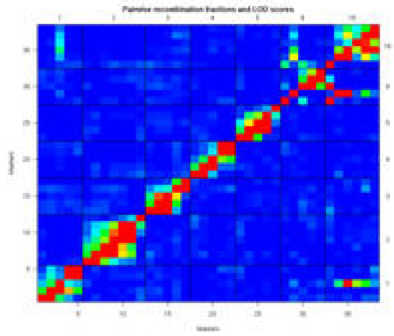
Imputation

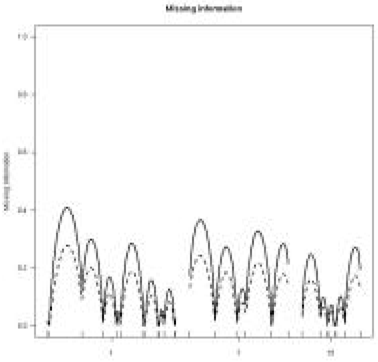


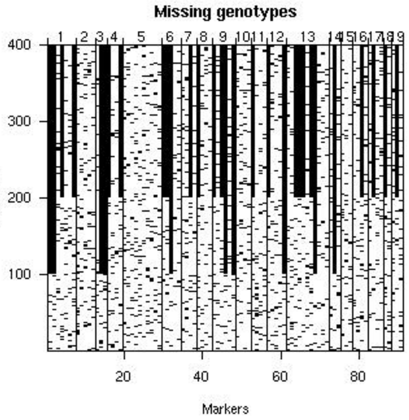
Ś. Sen and G. A. Churchill. A statistical framework for quantitative trait mapping. *Genetics*, 159:371–387, 2001.

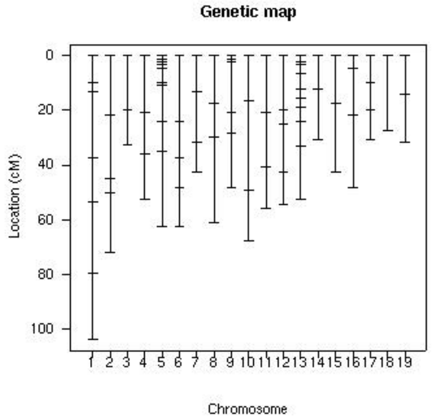
r/qtl

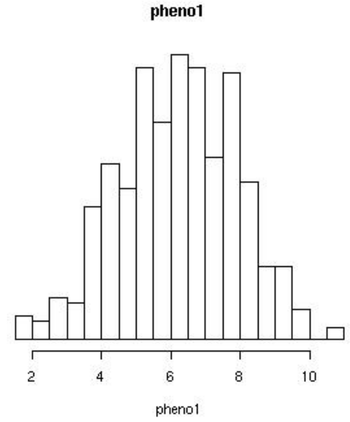
http://www.rqtl.org/
Broman, Sen & Churchill

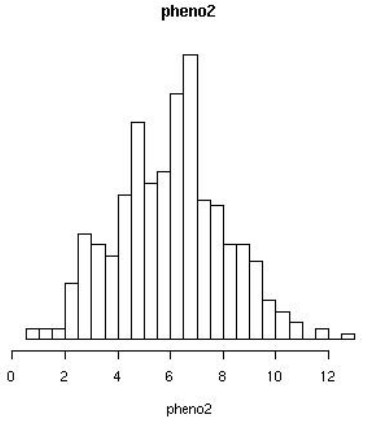






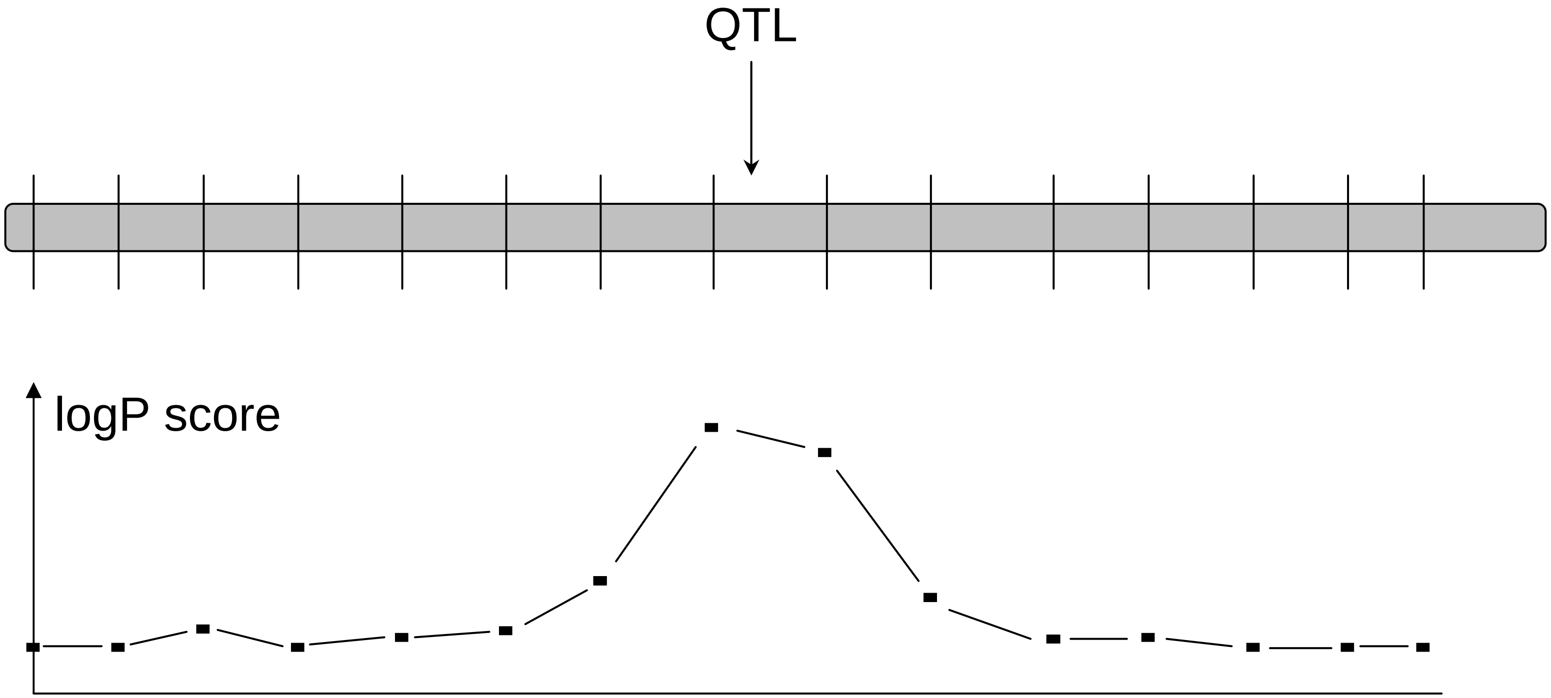


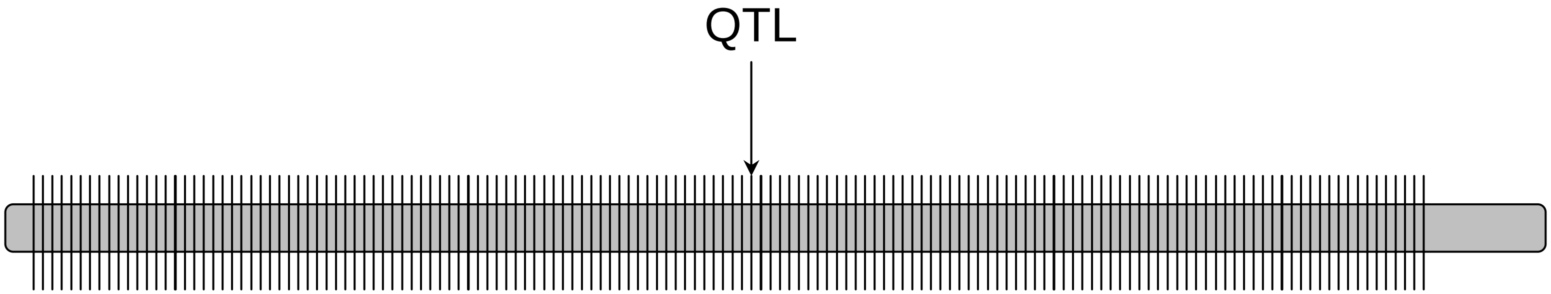


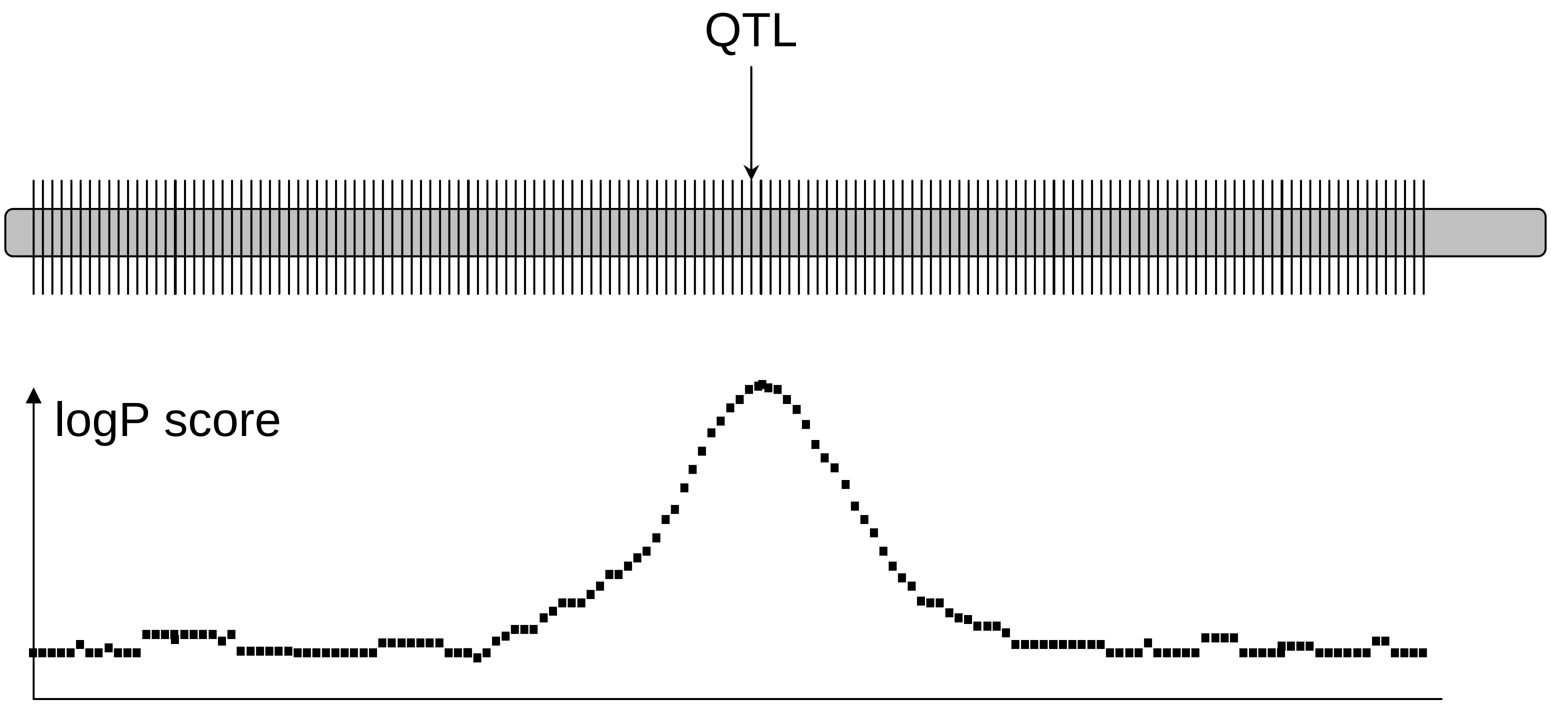




Is interval mapping necessary?


QTL
logP score


QTL


QTL
logP score

Significance Thresholds

Significance Thresholds

Mapping method	**Suggestive**		**Significant**	
	P	LOD	P	LOD
Backcross	3.40E-03	1.9	1.00E-04	3.3
Intercross (2 df)	1.60E-03	2.8	5.20E-05	4.3

Lander, E. Kruglyak, L. *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results* **Nature Genetics.** 11, 241-7, 1995

Thresholds

Permutation test

```
SUBJECT.NAME Sex Phenotype m1 m2 m3 m4
      F2$798   F -0.738004 -1  1  1 -1
      F2$364   F  0.413330  0  0  0  0
      F2$367   F  1.417480 -1  1  1 -1
      F2$287   F  0.811208  1 -1 -1  1
      F2$205   M  1.198270  0  0  0  0
```

Thresholds

Permutation test

```
 SUBJECT.NAME Sex Phenotype m1 m2 m3 m4
       F2$798   F -0.738004 -1  1  1 -1
       F2$364   F  0.413330  0  0  0  0
       F2$367   F  1.417480 -1  1  1 -1
       F2$287   F  0.811208  1 -1 -1  1
       F2$205   M  1.198270  0  0  0  0
```

shuffle

```
 SUBJECT.NAME Sex Phenotype m1 m2 m3 m4
       F2$798   F  0.413330 -1  1  1 -1
       F2$364   F  1.417480  0  0  0  0
       F2$367   F  1.198270 -1  1  1 -1
       F2$287   F -0.738004  1 -1 -1  1
       F2$205   M  0.811208  0  0  0  0
```

Permutation tests to establish thresholds

Empirical threshold values for quantitative trait mapping
GA Churchill and RW Doerge
Genetics, 138, 963-971 1994

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.

PRACTICAL: significance thresholds by permutation

Two problems in QTL analysis

Missing genotype problem

Model selection problem

The model problem

How QTL genotypes combine to produce the phenotype

The model problem

Linked QTL corrupt the position estimates

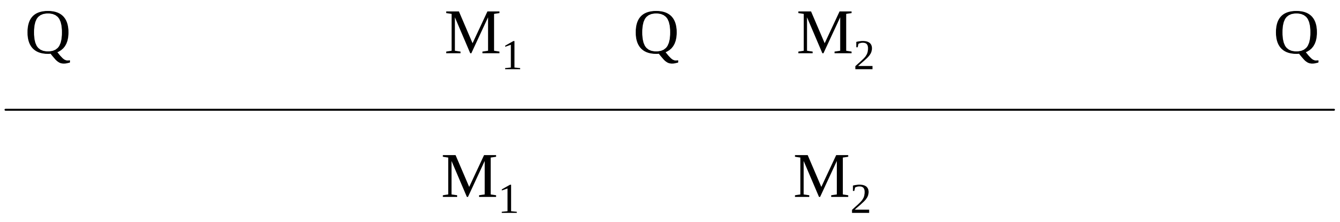
Unlinked QTL decreases the power of QTL detection

Composite interval mapping

ZB Zeng *Precision mapping of quantitative trait loci*
Genetics, Vol 136, 1457-1468, 1994

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

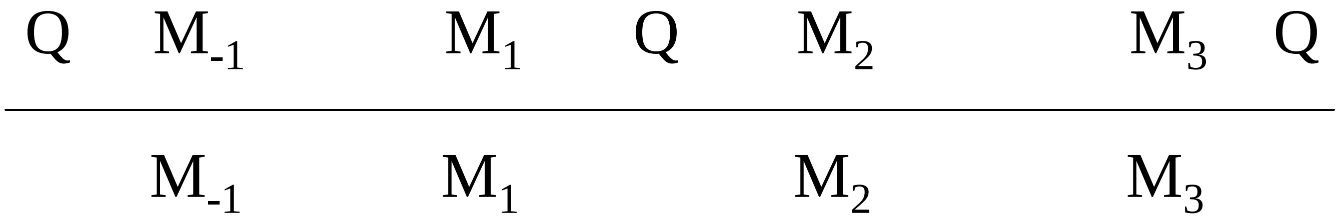
Composite interval mapping



Q M_1 Q M_2 Q

M_1 M_2

Composite interval mapping



Q M_{-1} M_1 Q M_2 M_3 Q

M_{-1} M_1 M_2 M_3

Model selection

Inclusion of covariates: gender, environment and other things too many too enumerate here

Inclusion of covariates

H_0: phenotype ~ covariates

H_1: phenotype ~ covariates + **LocusX**

Inclusion of covariates

H_0: phenotype ~ covariates

H_1: phenotype ~ covariates + **LocusX**

H_1 vs H_0 : how much extra does **LocusX** explain?

Inclusion of covariates

H_0: phenotype ~ covariates

H_1: phenotype ~ covariates + **LocusX**

H_1 vs H_0 : how much extra does **LocusX** explain?

H_0: startle ~ Sex + BodyWeight + TestChamber + Age

H_1: startle ~ Sex + BodyWeight + TestChamber + Age + **Locus432**

PRACTICAL: Inclusion of gender effects in a genome scan

To start:
In Firefox, then File -> Open File, and open “gxe.R”

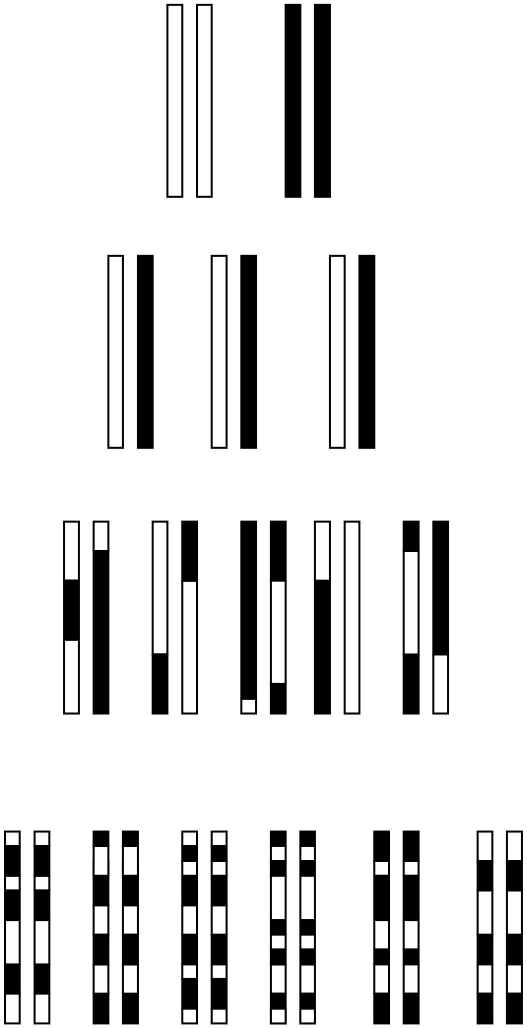
Experimental crosses

Inbred strain crosses

Recombinant inbreds

Alternatives

Recombinant Inbreds



F_0 Parental Generation

F_1 Generation

F_2 Generation

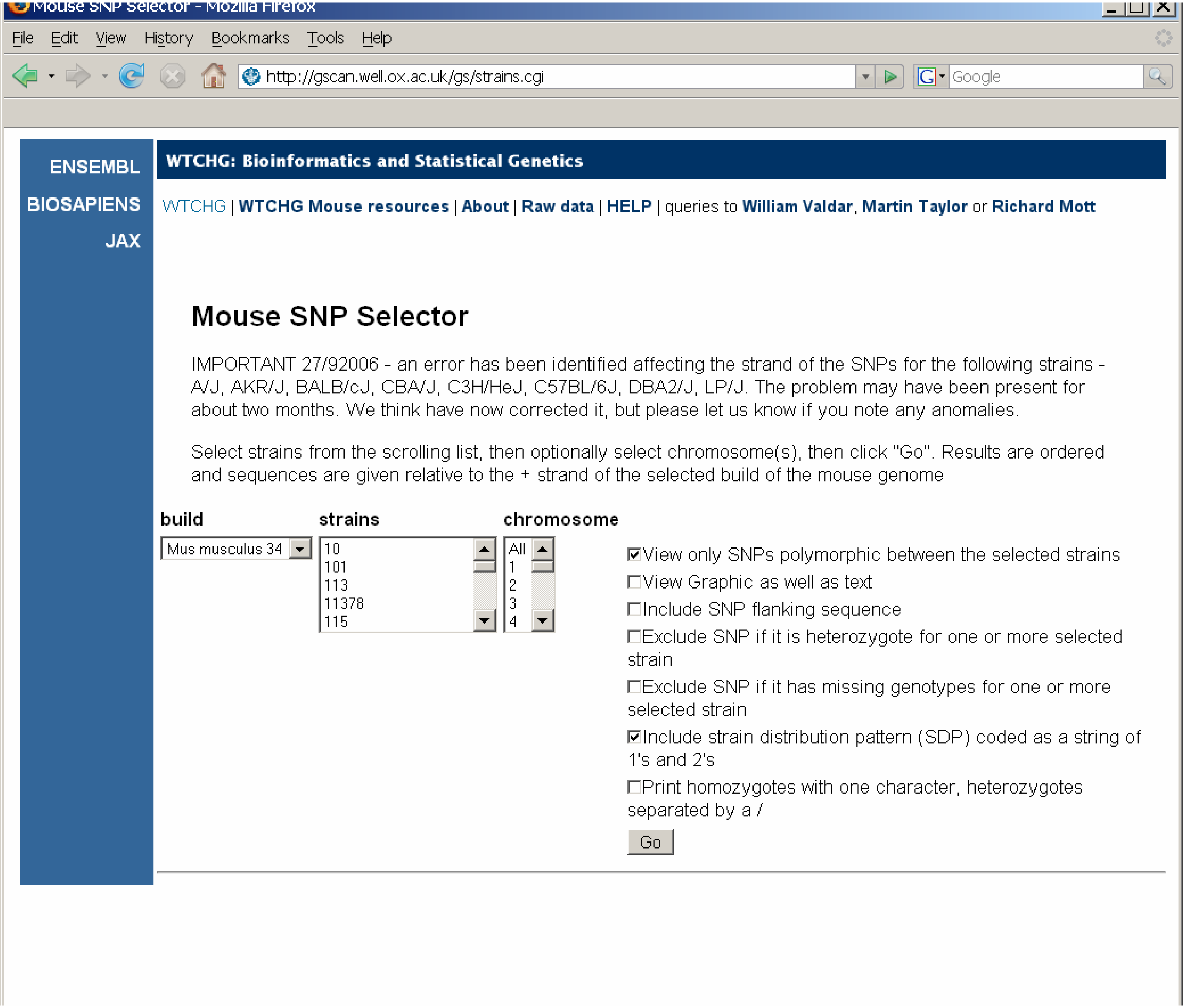
Interbreeding for approximately 20 generations to produce recombinant inbreds

RI strain genotypes

http://www.well.ox.ac.uk/mouse/INBREDS

SNP SELECTOR

http://gscan.well.ox.ac.uk/gs/strains.cgi



WTCHG: Bioinformatics and Statistical Genetics

WTCHG | **WTCHG Mouse resources** | **About** | **Raw data** | **HELP** | queries to **William Valdar**, **Martin Taylor** or **Richard Mott**

Mouse SNP Selector

IMPORTANT 27/92006 - an error has been identified affecting the strand of the SNPs for the following strains - A/J, AKR/J, BALB/cJ, CBA/J, C3H/HeJ, C57BL/6J, DBA2/J, LP/J. The problem may have been present for about two months. We think have now corrected it, but please let us know if you note any anomalies.

Select strains from the scrolling list, then optionally select chromosome(s), then click "Go". Results are ordered and sequences are given relative to the + strand of the selected build of the mouse genome

build
Mus musculus 34

strains
10
101
113
11378
115

chromosome
All
1
2
3
4

☑View only SNPs polymorphic between the selected strains
☐View Graphic as well as text
☐Include SNP flanking sequence
☐Exclude SNP if it is heterozygote for one or more selected strain
☐Exclude SNP if it has missing genotypes for one or more selected strain
☑Include strain distribution pattern (SDP) coded as a string of 1's and 2's
☐Print homozygotes with one character, heterozygotes separated by a /

RI strain phenotypes

RI analysis

Power of RIs

Number	**Power**	**QTL %Var$_{exp}$**
24	90	55
	50	45
37	90	30
	50	35

Effect size of a QTL that can be detected with RI strain sets, at P= 0.00013

Experimental crosses

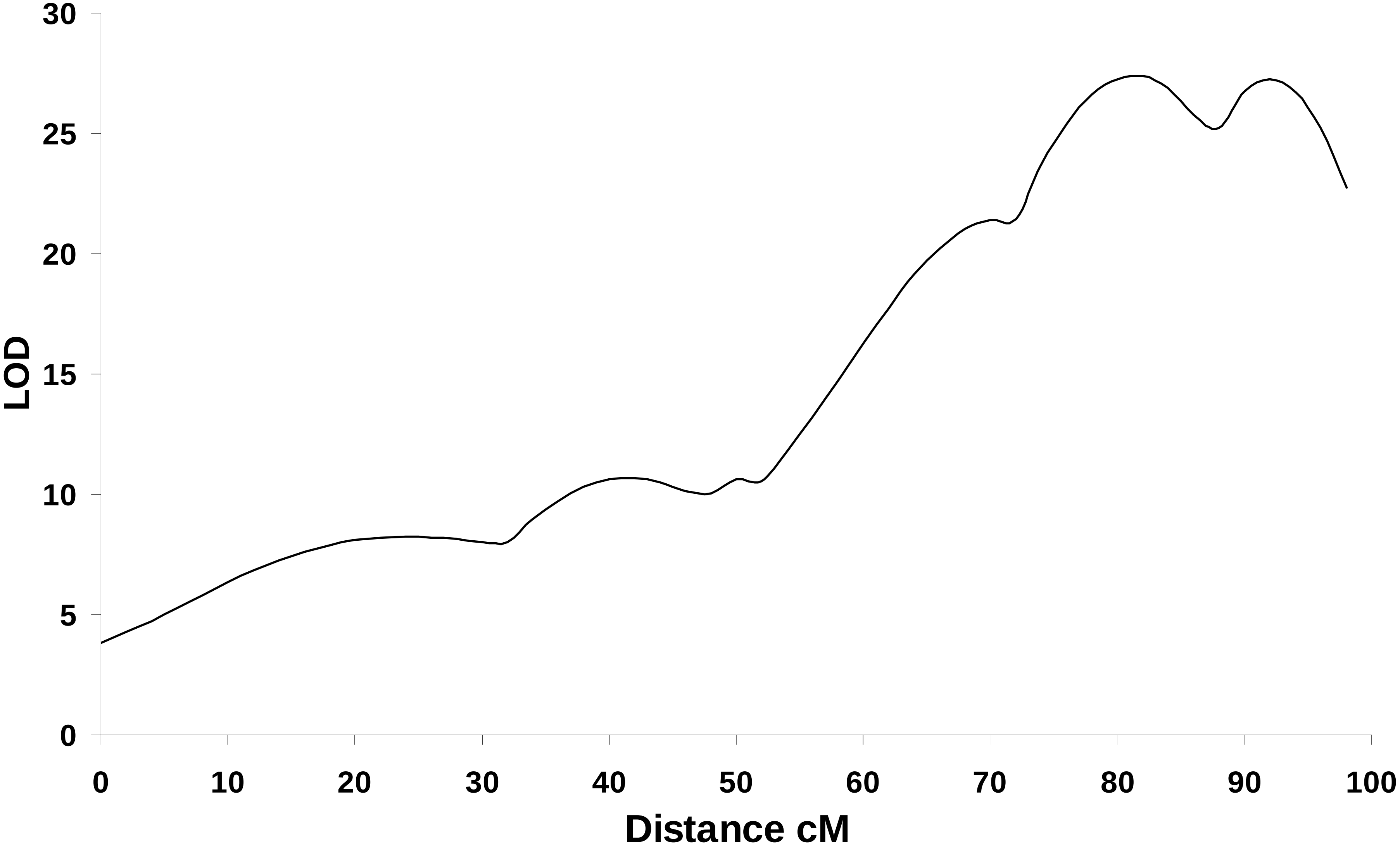
Inbred strain crosses

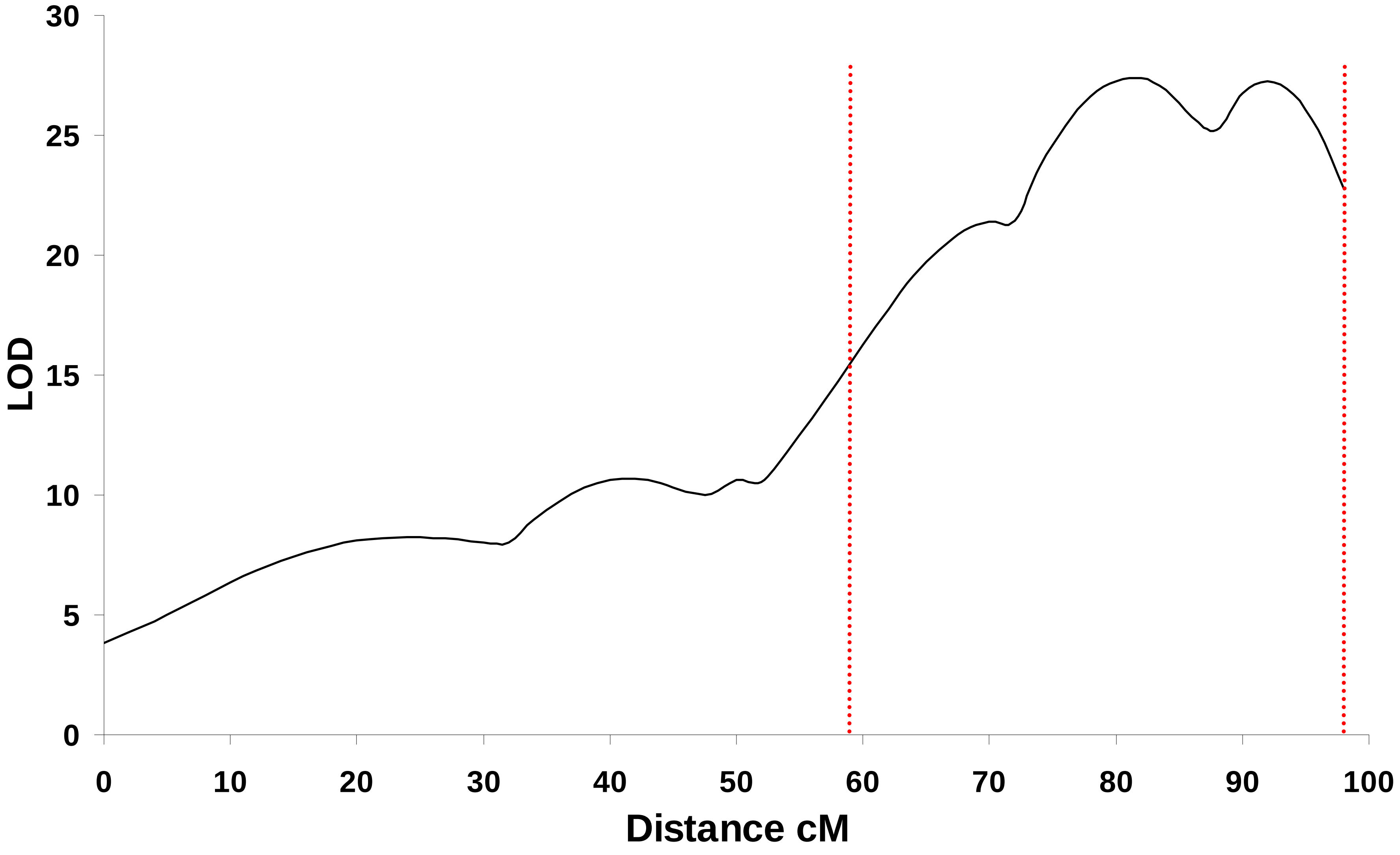
Recombinant inbreds

Alternatives

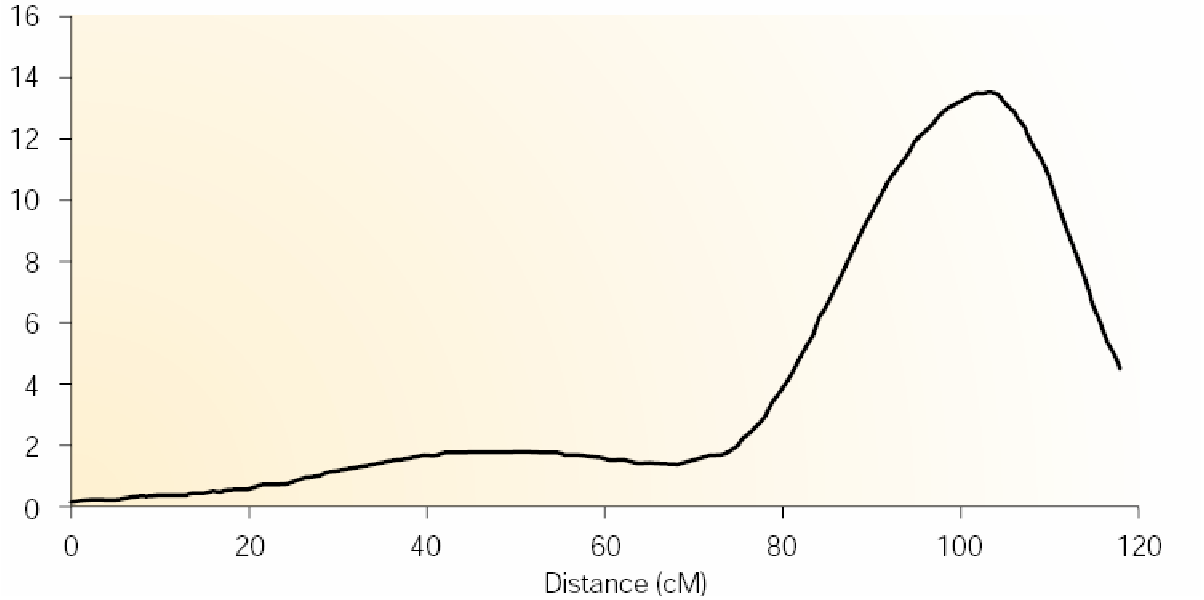
Why do we need alternatives?

Classical strategies don’t find genes because of poor resolution

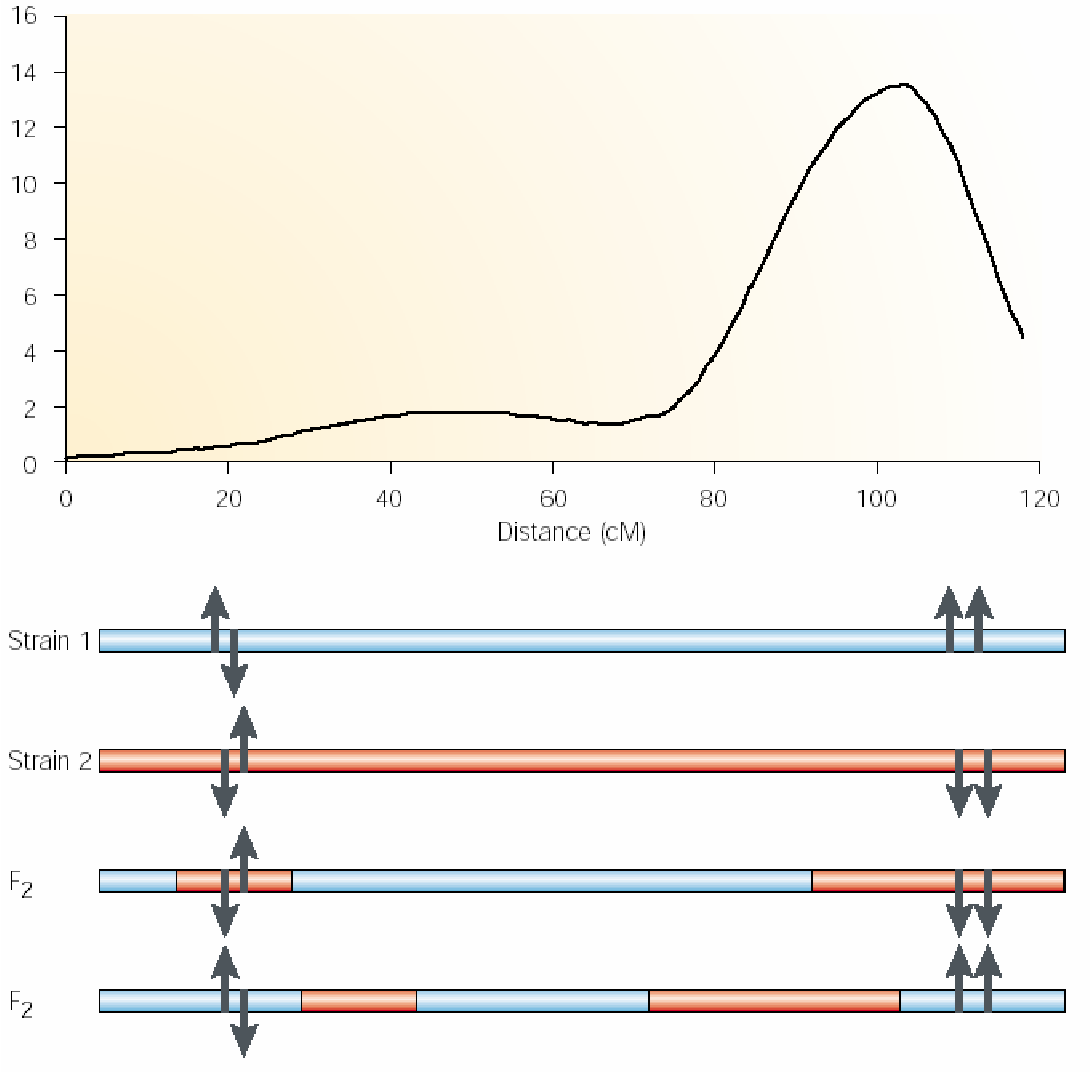

LOD
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
80
90
100
Distance cM



LOD
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
80
90
100
Distance cM

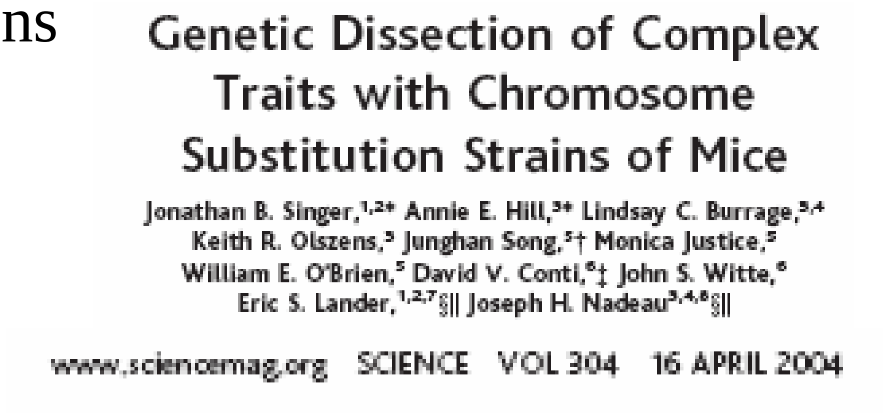

16
14
12
10
8
6
4
2
0
0
20
40
60
80
100
120
Distance (cM)

One locus may contain many QTL



New approaches

Chromosome substitution strains



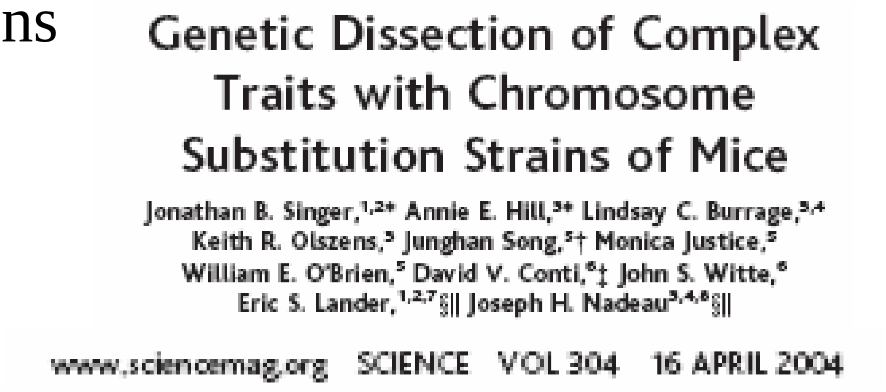
Genetic Dissection of Complex Traits with Chromosome Substitution Strains of Mice

Jonathan B. Singer,[1,2]* Annie E. Hill,[3]* Lindsay C. Burrage,[3,4]
Keith R. Olszens,[5] Junghan Song,[5]† Monica Justice,[5]
William E. O'Brien,[5] David V. Conti,[6]‡ John S. Witte,[6]
Eric S. Lander,[1,2,7]§|| Joseph H. Nadeau[3,4,8]§||

www.sciencemag.org SCIENCE VOL 304 16 APRIL 2004

New approaches

Chromosome substitution strains

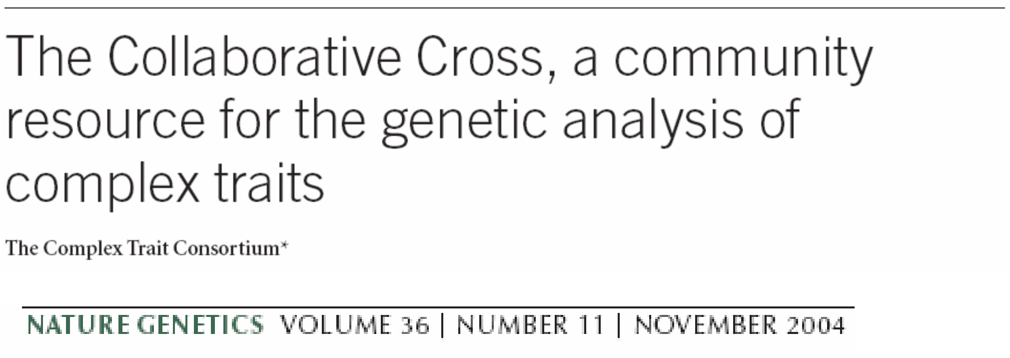


Genetic Dissection of Complex Traits with Chromosome Substitution Strains of Mice

Jonathan B. Singer,[1,2]* Annie E. Hill,[3]* Lindsay C. Burrage,[3,4]
Keith R. Olszens,[5] Junghan Song,[5]† Monica Justice,[5]
William E. O'Brien,[5] David V. Conti,[6]‡ John S. Witte,[6]
Eric S. Lander,[1,2,7]§|| Joseph H. Nadeau[3,4,8]§||

www.sciencemag.org SCIENCE VOL 304 16 APRIL 2004

Collaborative cross



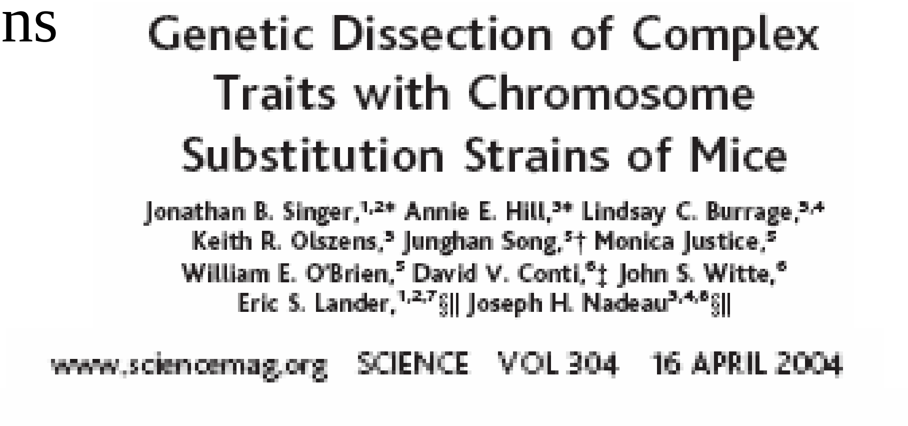
The Collaborative Cross, a community resource for the genetic analysis of complex traits

The Complex Trait Consortium*

NATURE GENETICS VOLUME 36 | NUMBER 11 | NOVEMBER 2004

New approaches

Chromosome substitution strains

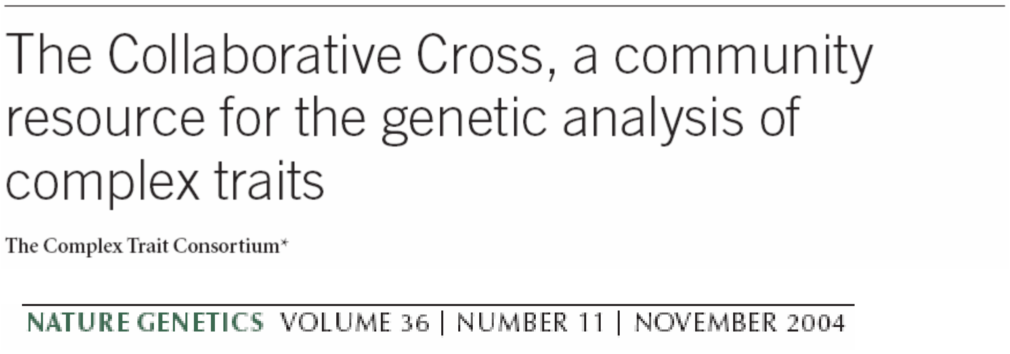


Genetic Dissection of Complex Traits with Chromosome Substitution Strains of Mice

Jonathan B. Singer,[1,2]* Annie E. Hill,[3]* Lindsay C. Burrage,[3,4]
Keith R. Olszens,[3] Junghan Song,[5]† Monica Justice,[5]
William E. O'Brien,[5] David V. Conti,[6]‡ John S. Witte,[6]
Eric S. Lander,[1,2,7]§|| Joseph H. Nadeau[3,4,8]§||

www.sciencemag.org SCIENCE VOL 304 16 APRIL 2004

Collaborative cross

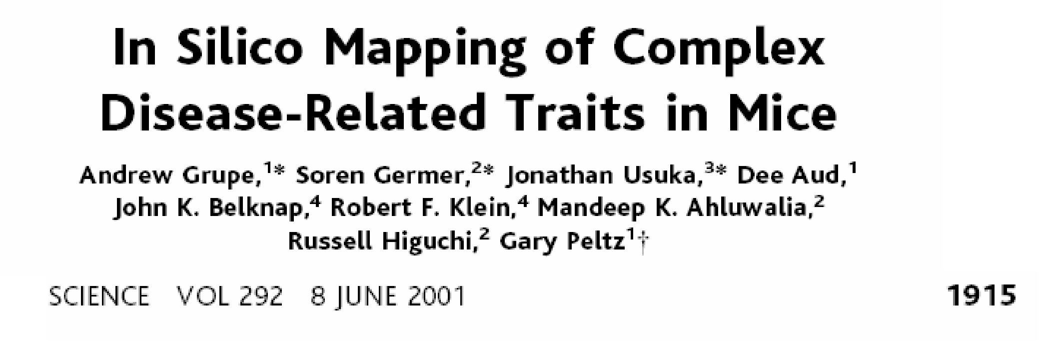


The Collaborative Cross, a community resource for the genetic analysis of complex traits

The Complex Trait Consortium*

NATURE GENETICS VOLUME 36 | NUMBER 11 | NOVEMBER 2004

In silico mapping



In Silico Mapping of Complex Disease-Related Traits in Mice

Andrew Grupe,[1]* Soren Germer,[2]* Jonathan Usuka,[3]* Dee Aud,[1]
John K. Belknap,[4] Robert F. Klein,[4] Mandeep K. Ahluwalia,[2]
Russell Higuchi,[2] Gary Peltz[1]†

SCIENCE VOL 292 8 JUNE 2001 1915

Resources

R	http://www.r-project.org/
R help	http://news.gmane.org/gmane.comp.lang.r.general
R/qtl	http://www.rqtl.org
Composite interval mapping (QTL Cartographer)	http://statgen.ncsu.edu/qtlcart/index.php
Markers	http://www.well.ox.ac.uk/mouse/inbreds

Gscan (HAPPY and associated analyses) http://gscan.well.ox.ac.uk

General reading

Lynch & Walsh (1998) Genetics and analysis of quantitative traits (Sinauer).

Dalgaard (2002) Introductory statistics with R (Springer-Verlag).

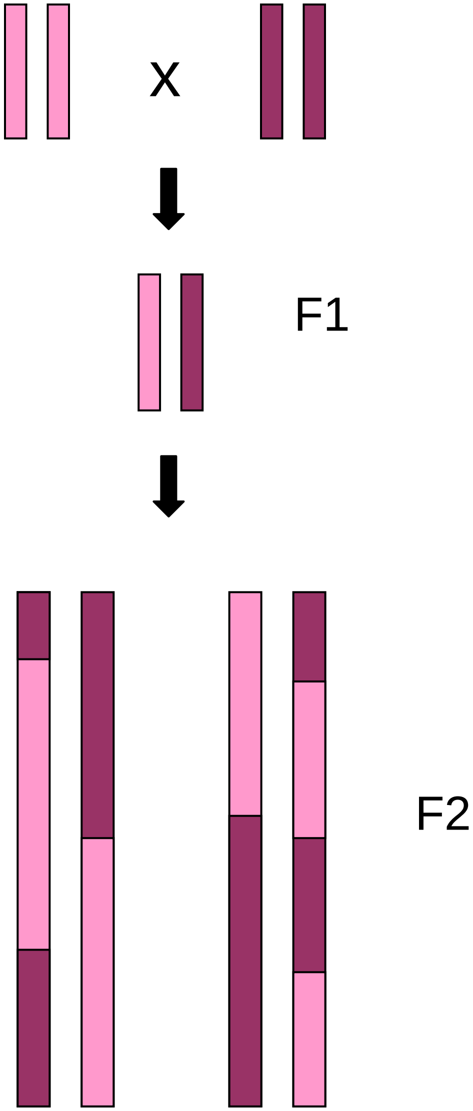
END SECTION

New approaches

Advanced intercross lines

Genetically heterogeneous stocks

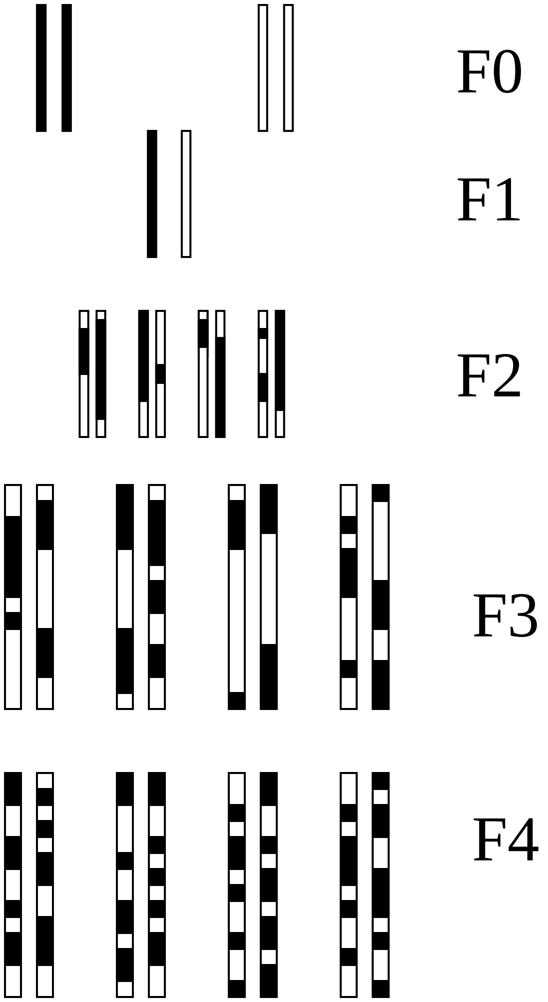
F2 Intercross



Avg. Distance Between
Recombinations

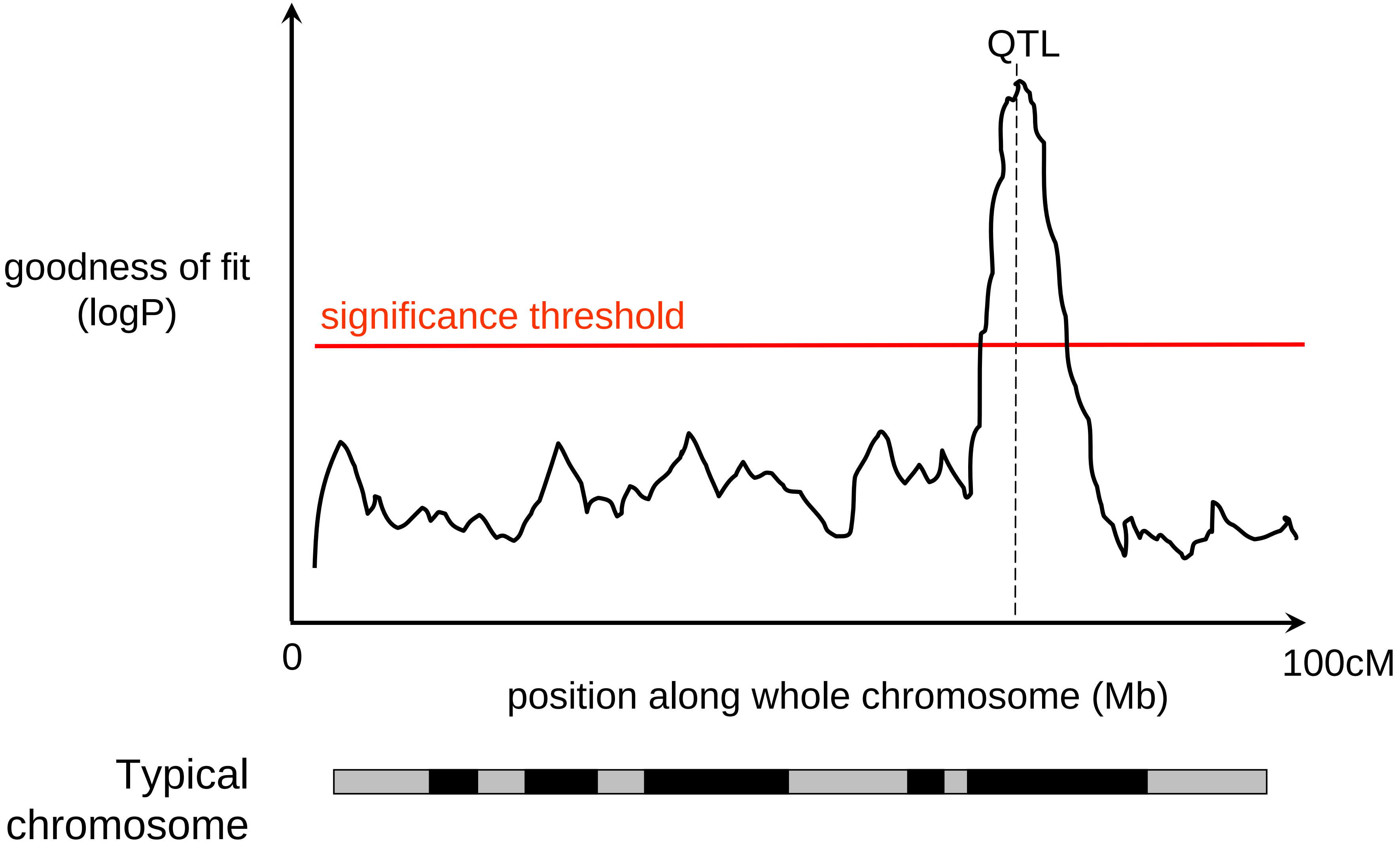
F2 intercross
~30 cM

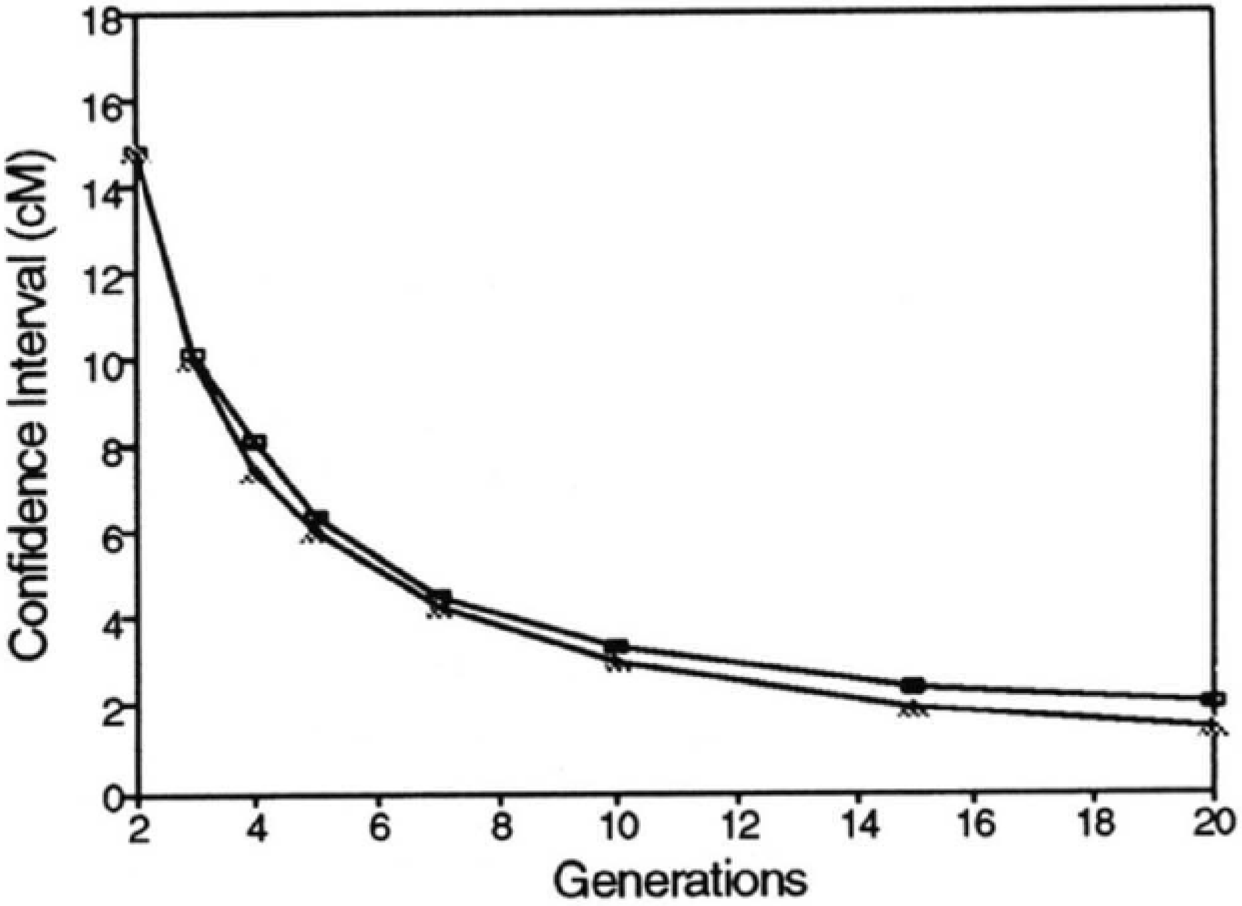
Advanced intercross lines (AILs)



Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141: 1199-1207.

Chromosome scan for F12

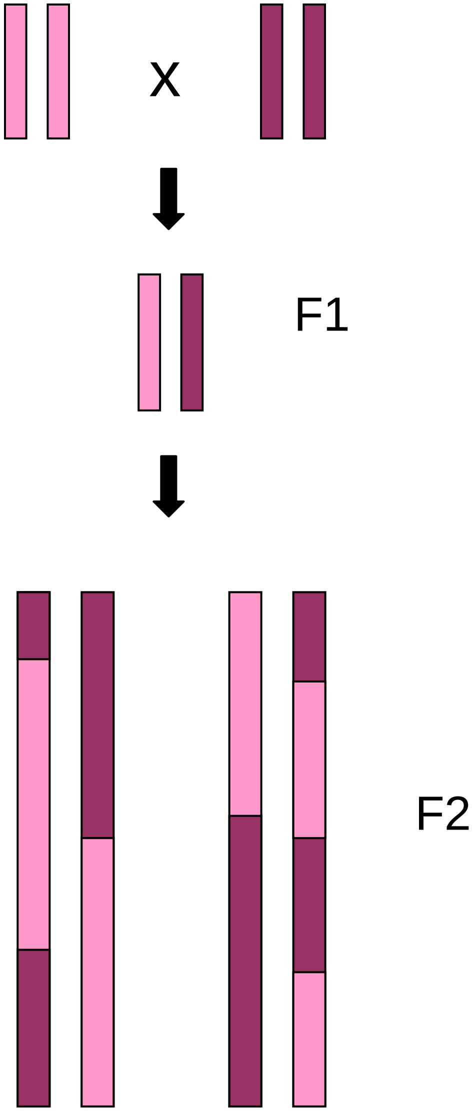



Confidence Interval (cM)
0
2
4
6
8
10
12
14
16
18
Generations
2
4
6
8
10
12
14
16
18
20

PRACTICAL: AILs

Genetically Heterogeneous Mice

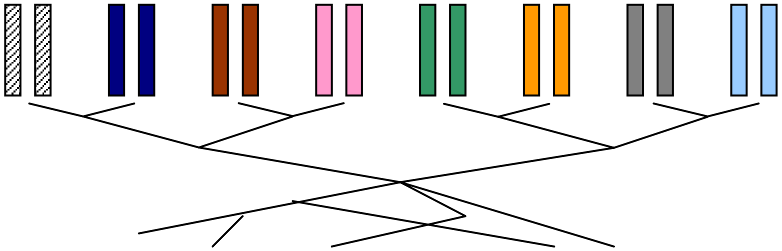
F2 Intercross



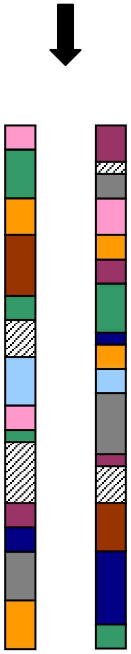
Avg. Distance Between
Recombinations

F2 intercross
~30 cM

Heterogeneous Stock



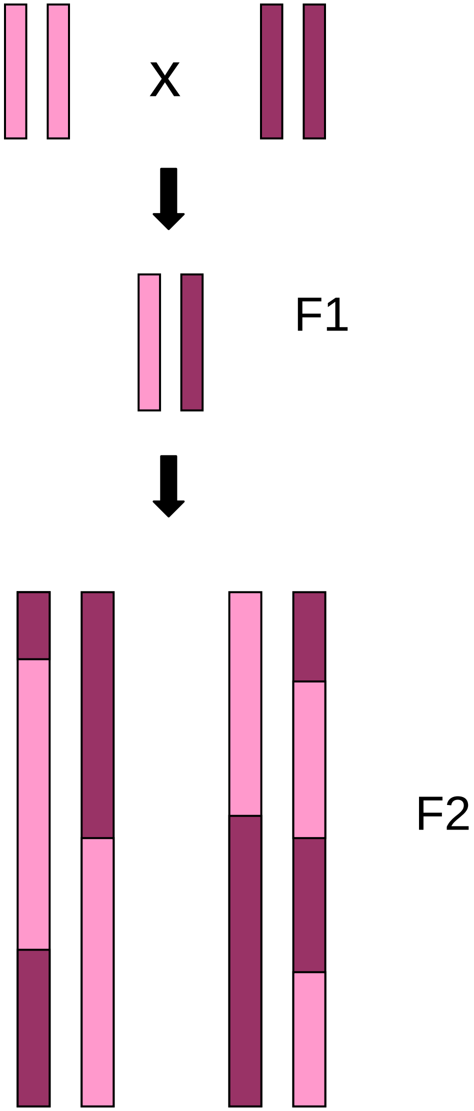
Pseudo-random mating
for 50 generations



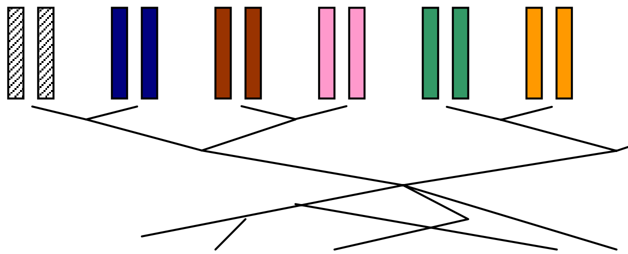
Avg. Distance Between
Recombinations:

HS	F2 intercross
~2 cM	~30 cM

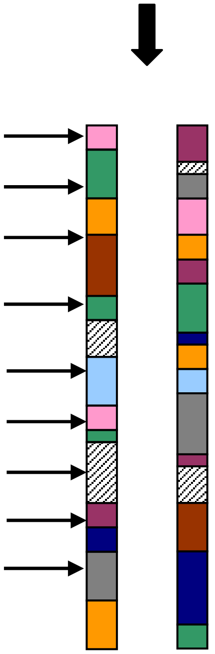
F2 Intercross



Heterogeneous Stock



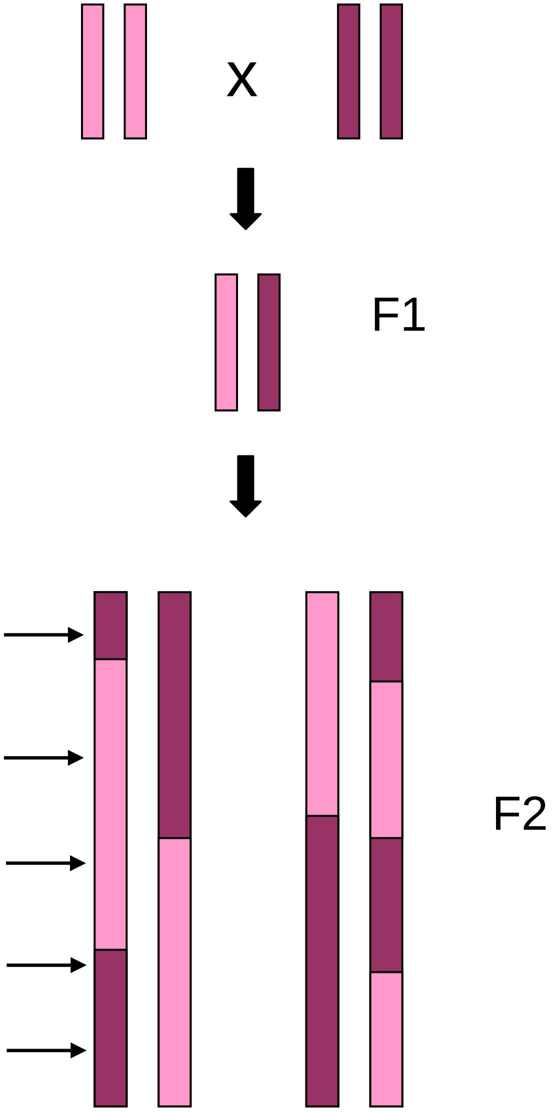
Pseudo-random mating
for 50 generations



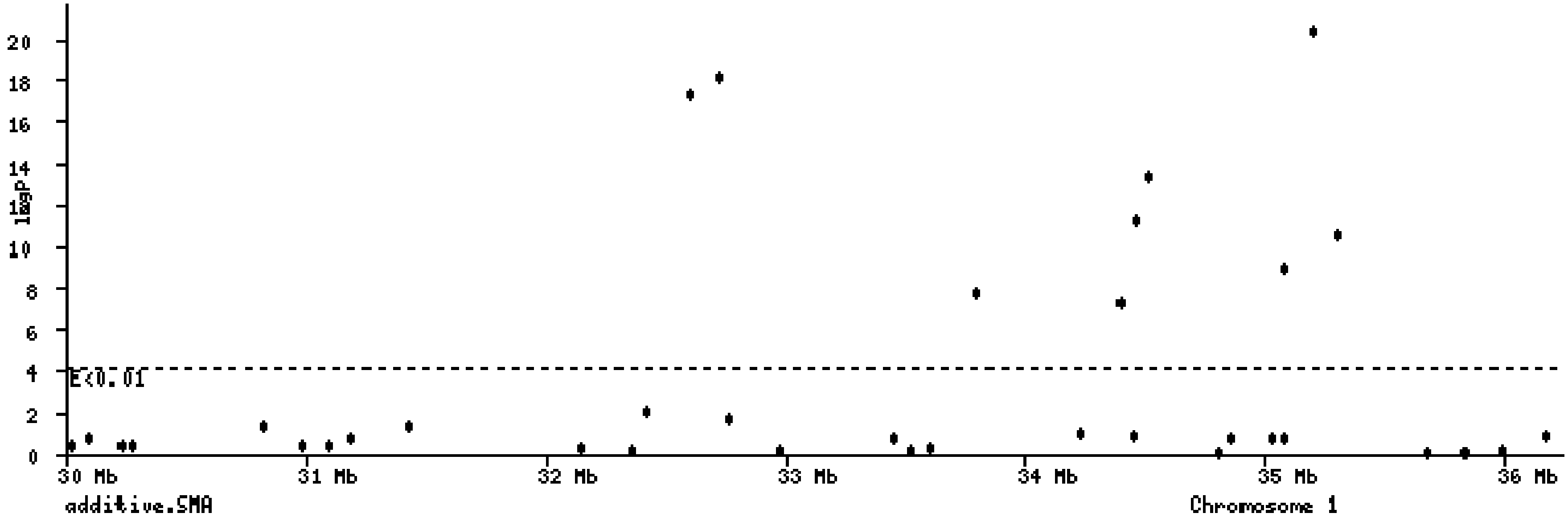
Avg. Distance Between Recombinations:

HS	F2 intercross
~2 cM	~30 cM

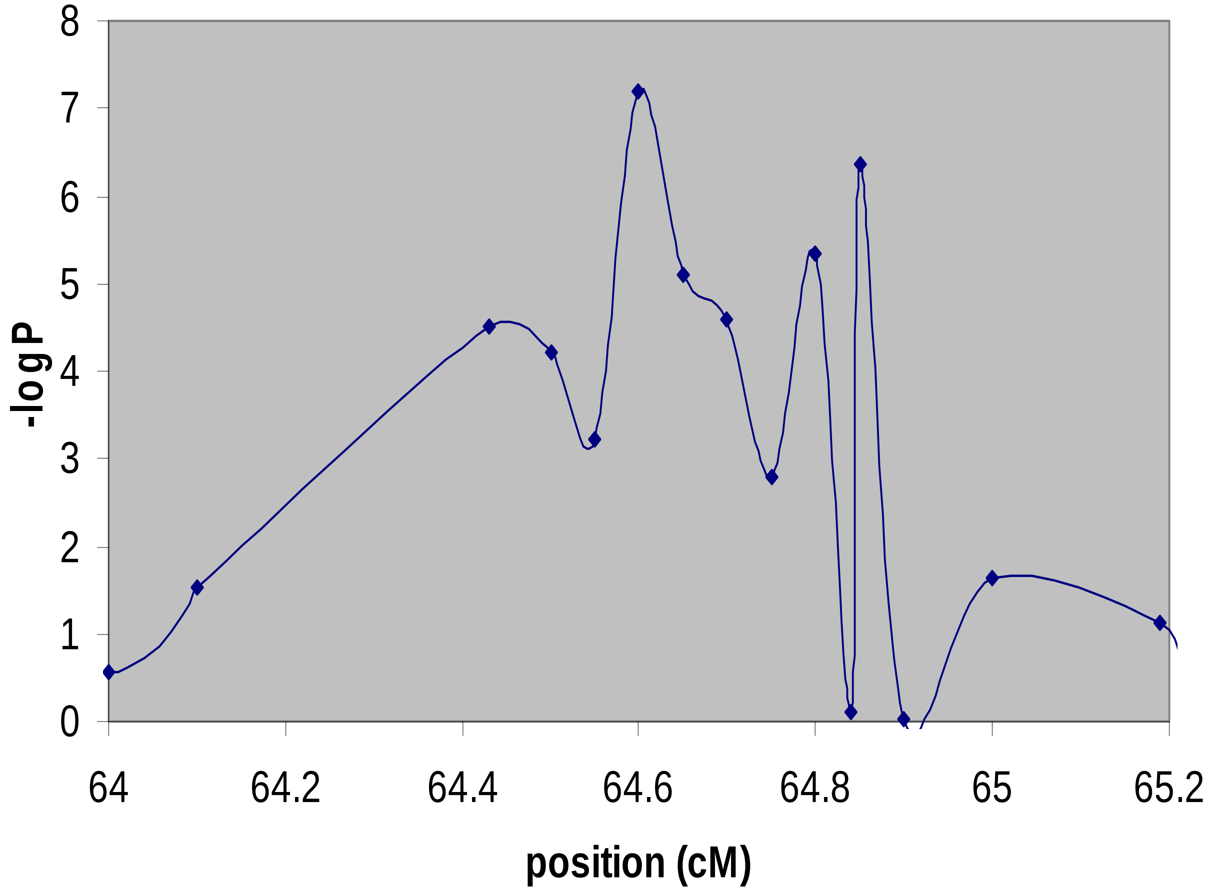
F2 Intercross



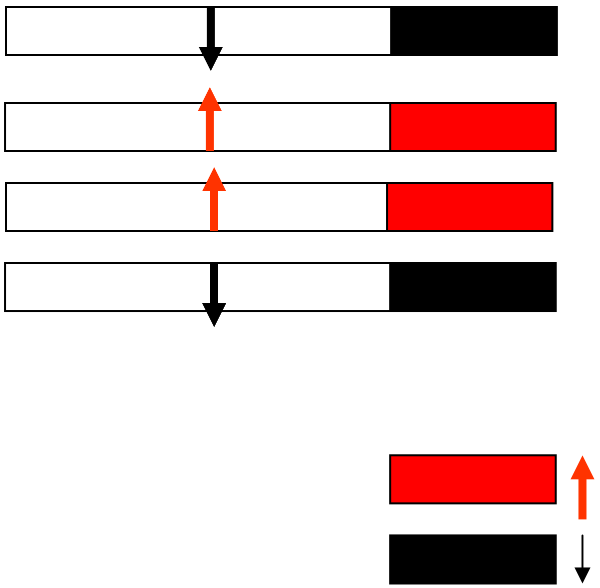
Genome scans with single marker association



High resolution mapping



Relation Between Marker and Genetic Effect

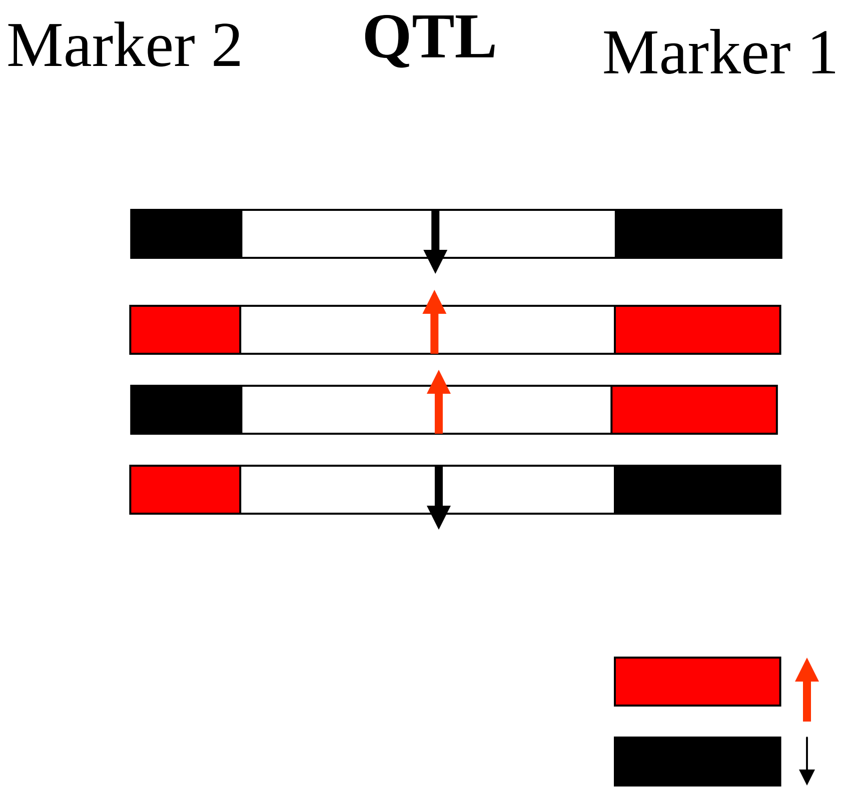


QTL

Marker 1

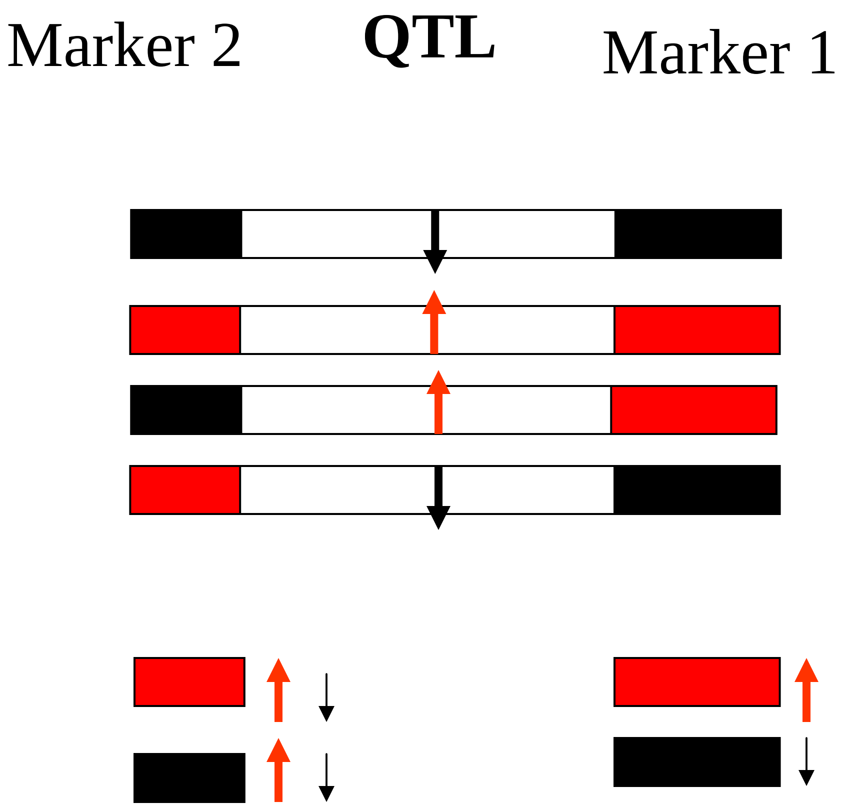
Observable effect

Relation Between Marker and Genetic Effect



Observable effect

Relation Between Marker and Genetic Effect



Marker 2

QTL

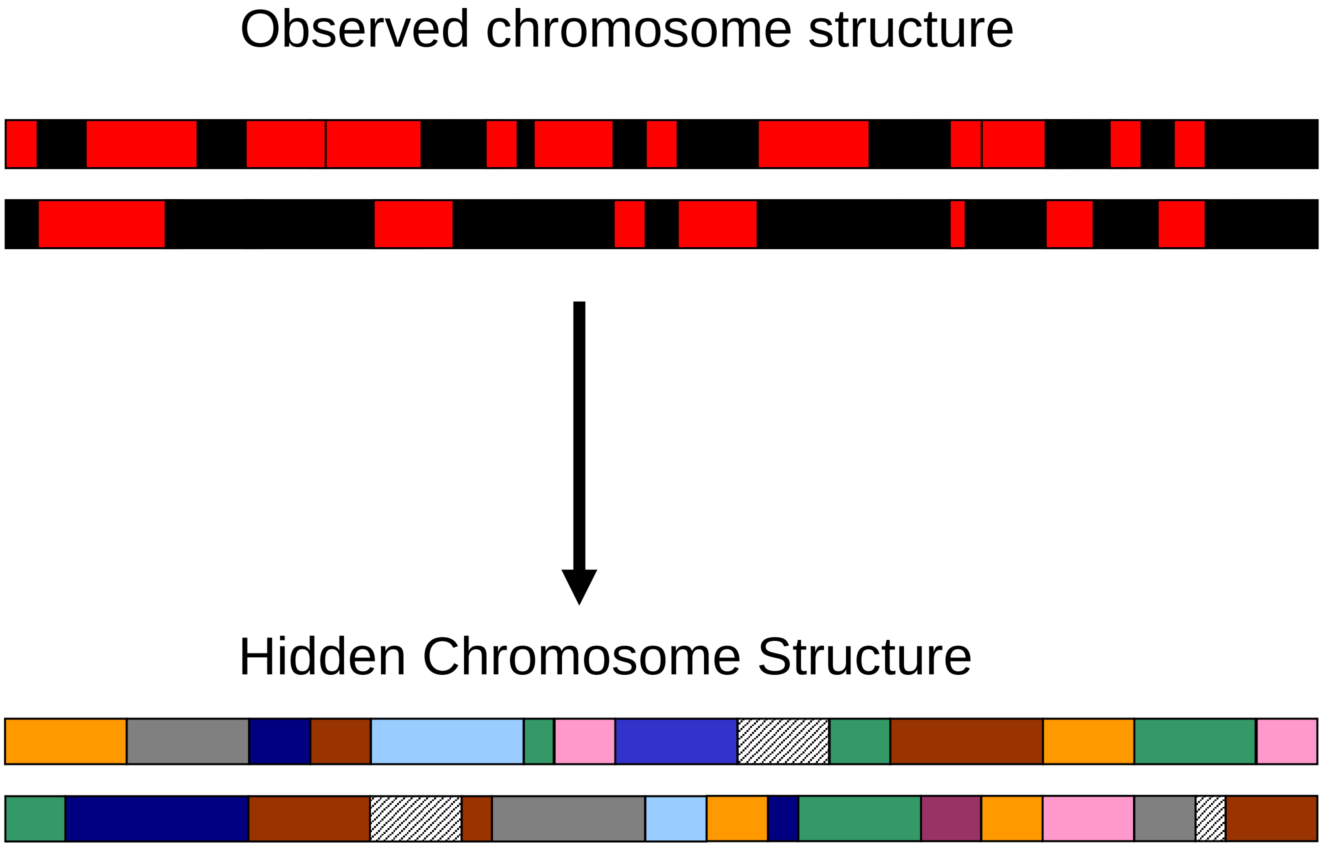
Marker 1

No effect

observable

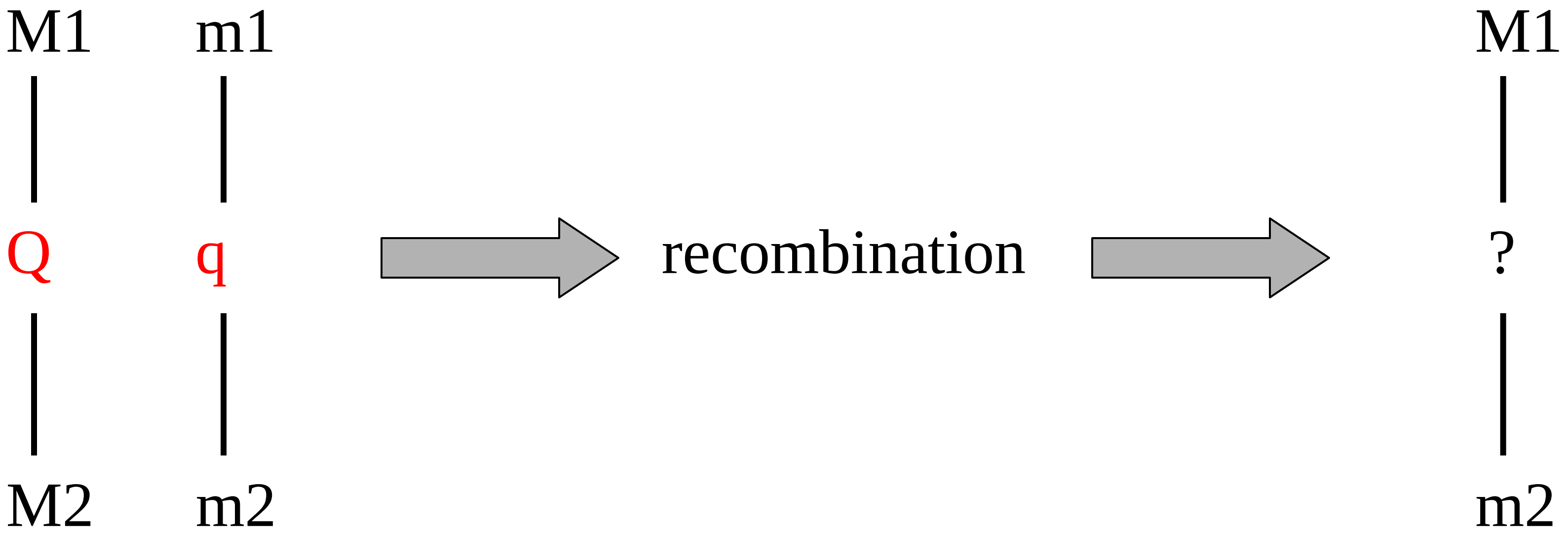
Observable
effect

Multipoint method (HAPPY) calculates the probability that an allele descends from a founder using multiple markers



Observed chromosome structure

Hidden Chromosome Structure



M1 m1

Q q

M2 m2

recombination

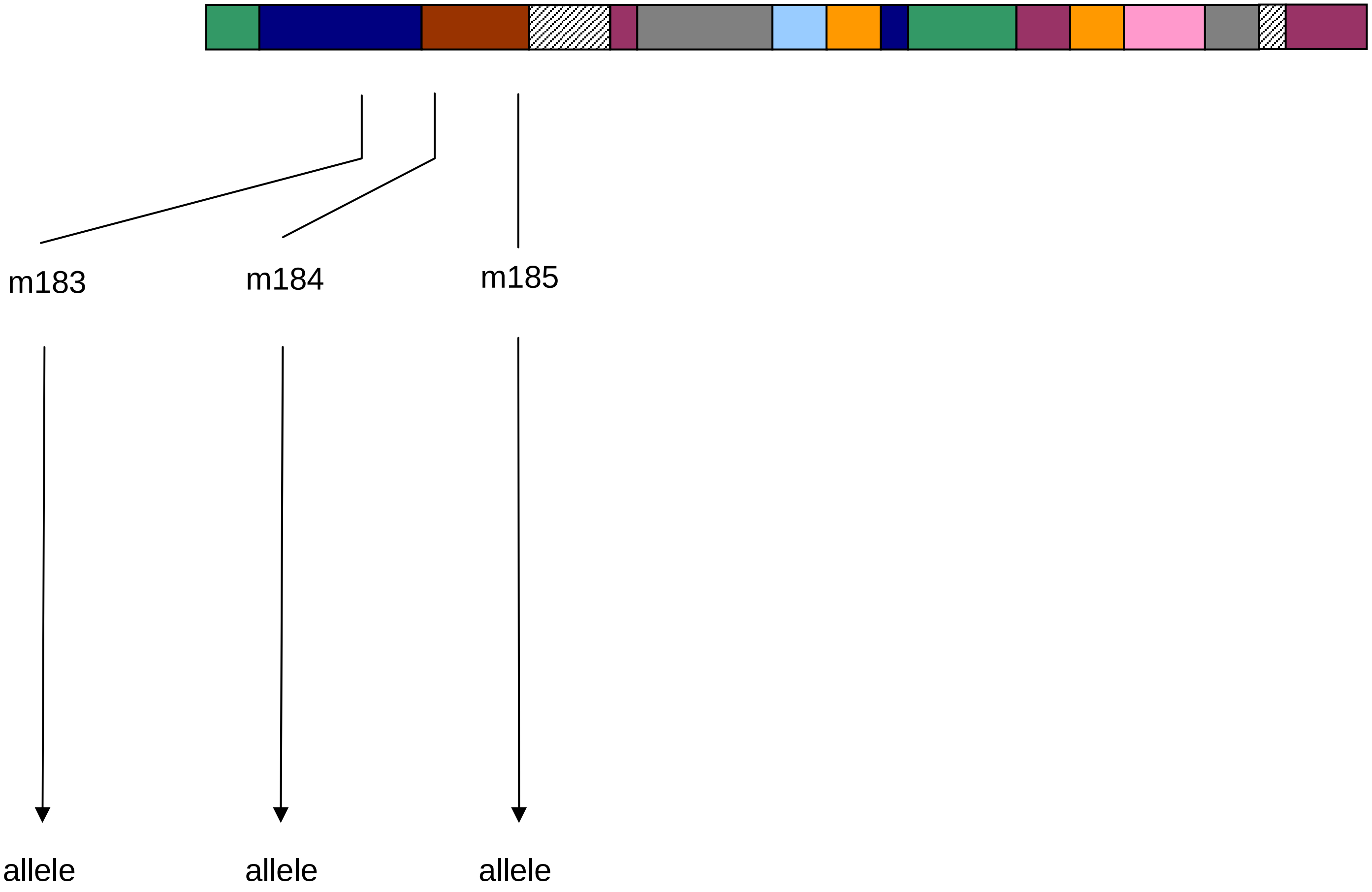
M1

?

m2

Haplotype reconstruction using HAPPY

A typical chromosome from an HS mouse

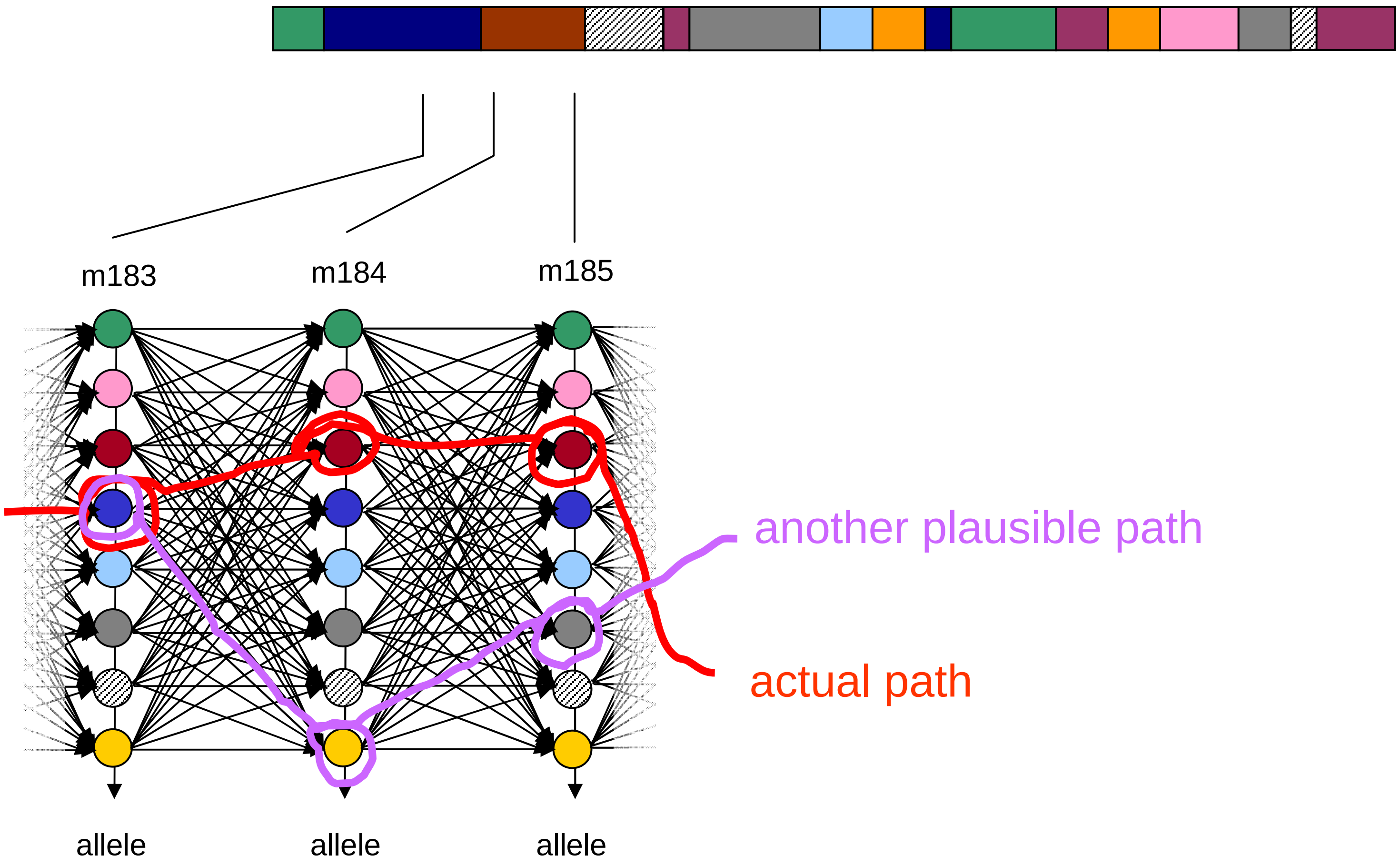


m183 m184 m185

allele allele allele

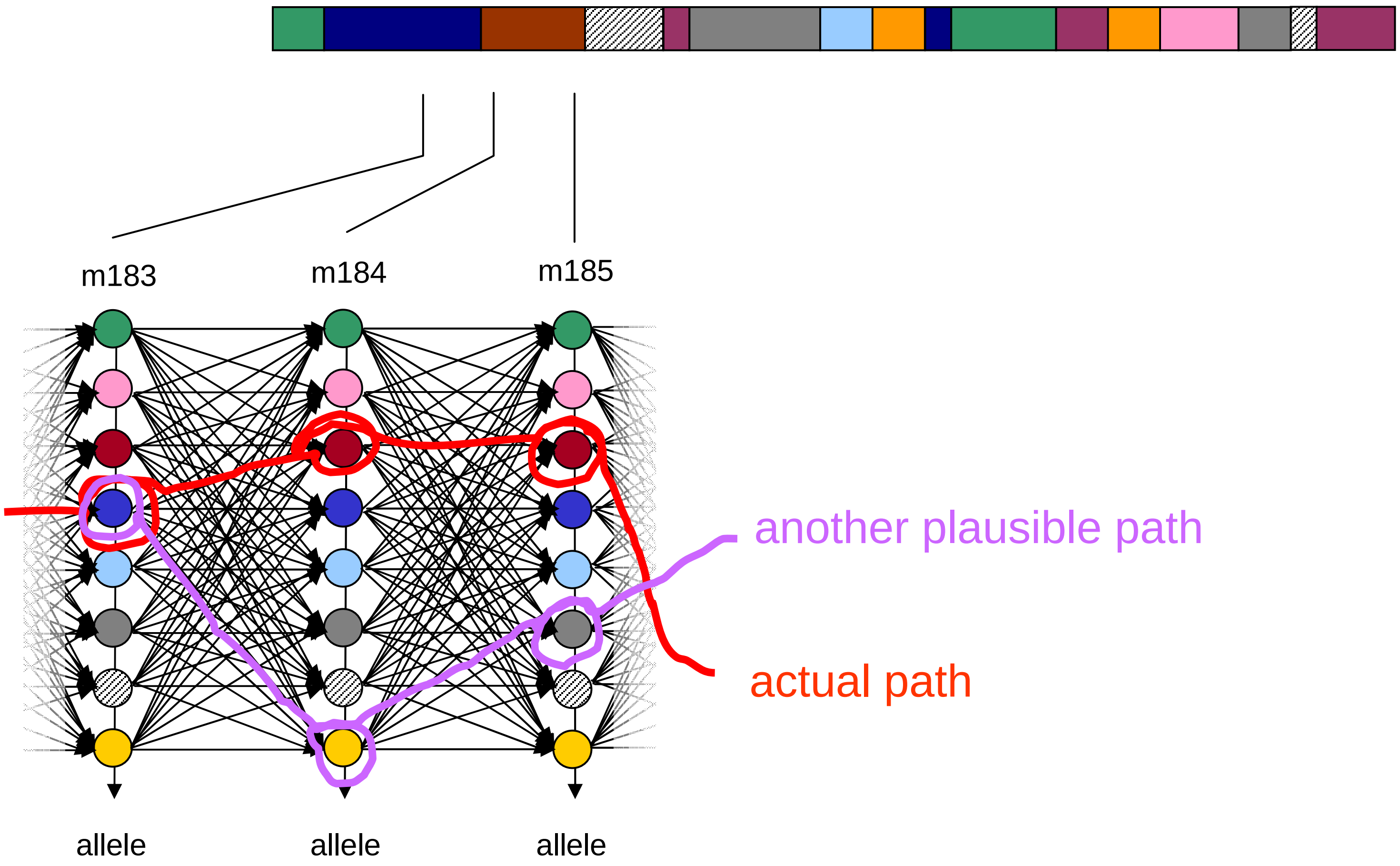
Haplotype reconstruction using HAPPY

A typical chromosome from an HS mouse



Haplotype reconstruction using HAPPY

A typical chromosome from an HS mouse

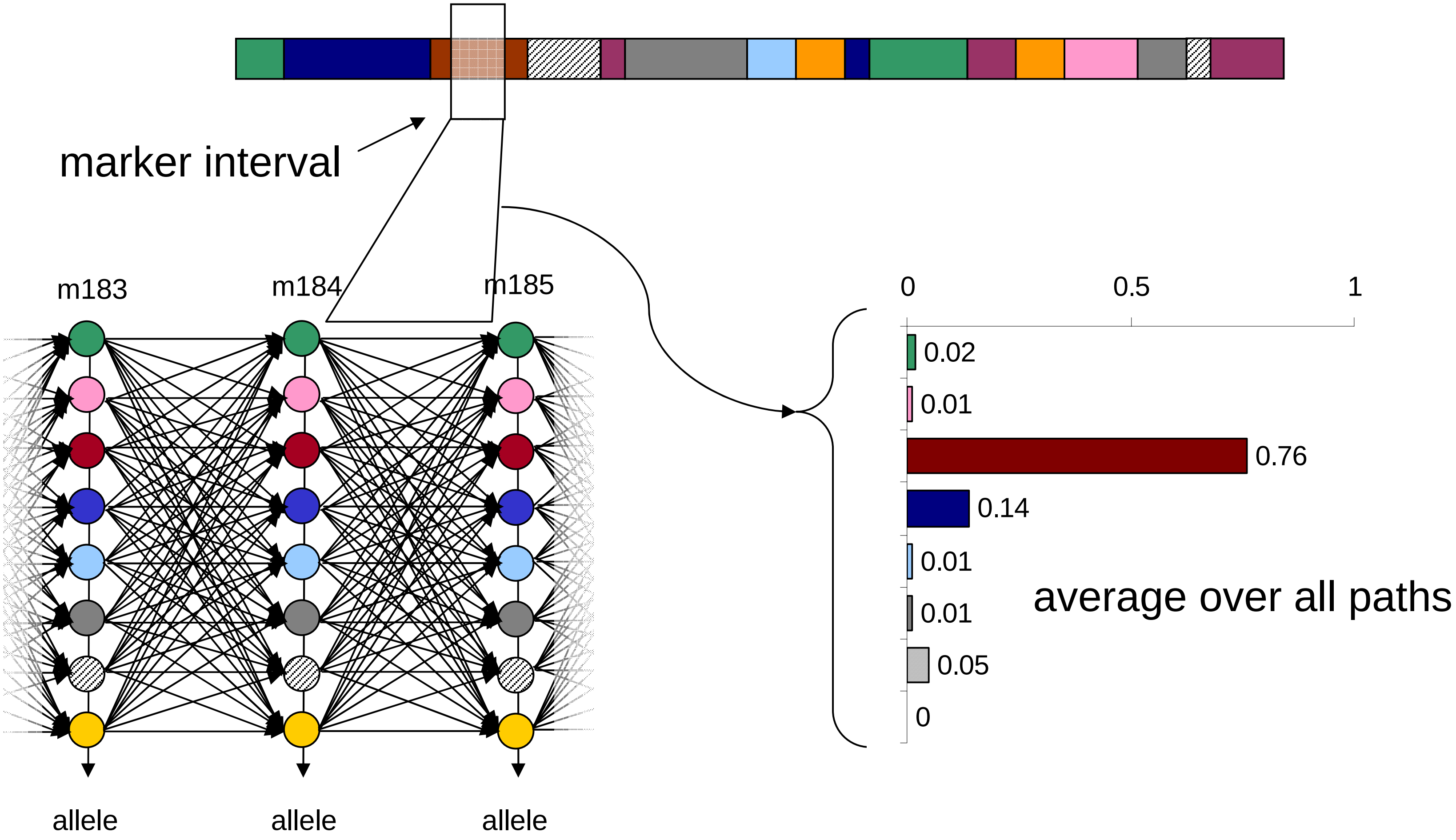


another plausible path

actual path

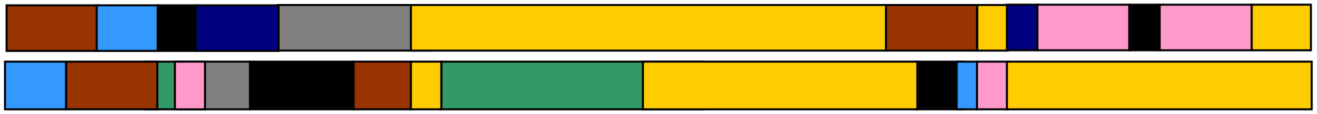
Haplotype reconstruction using HAPPY

A typical chromosome from an HS mouse

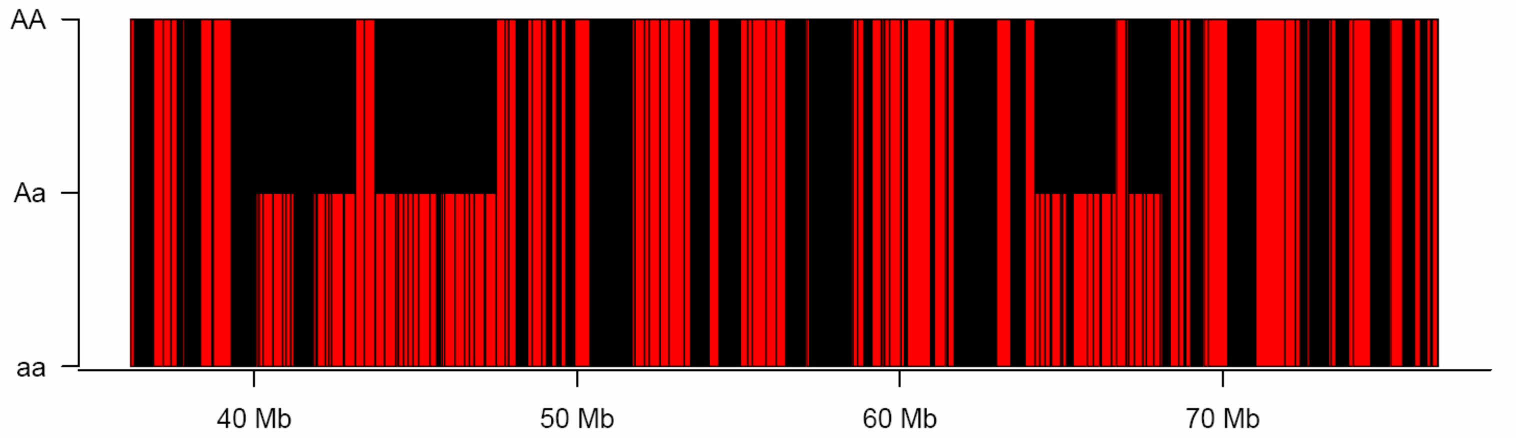


Haplotype reconstruction using HAPPY

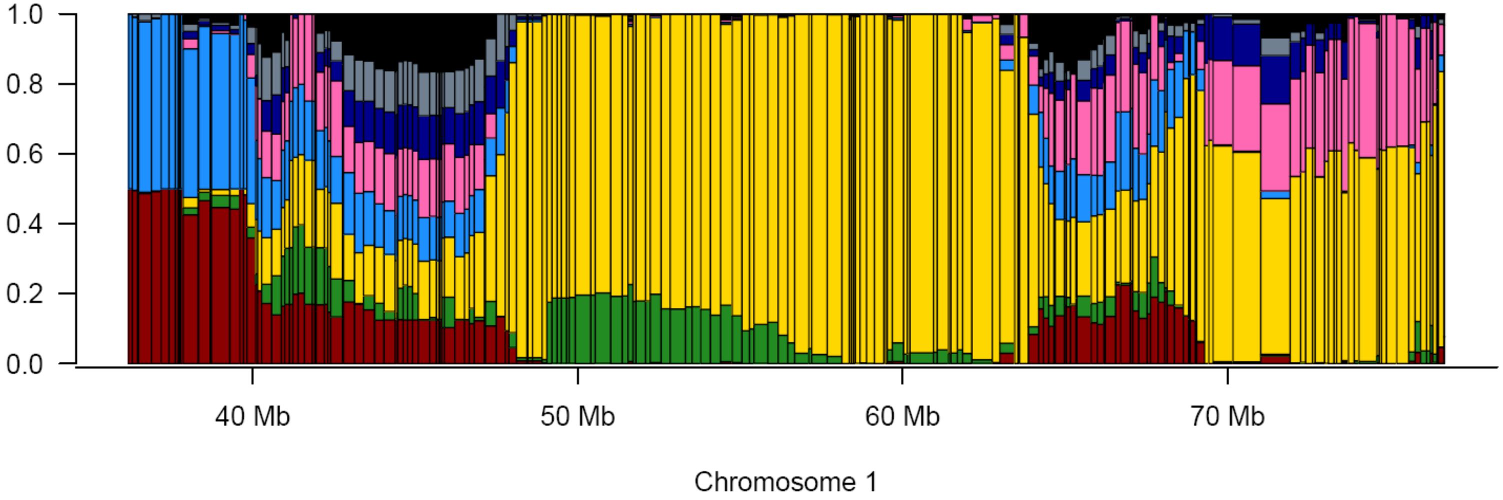
chromosome



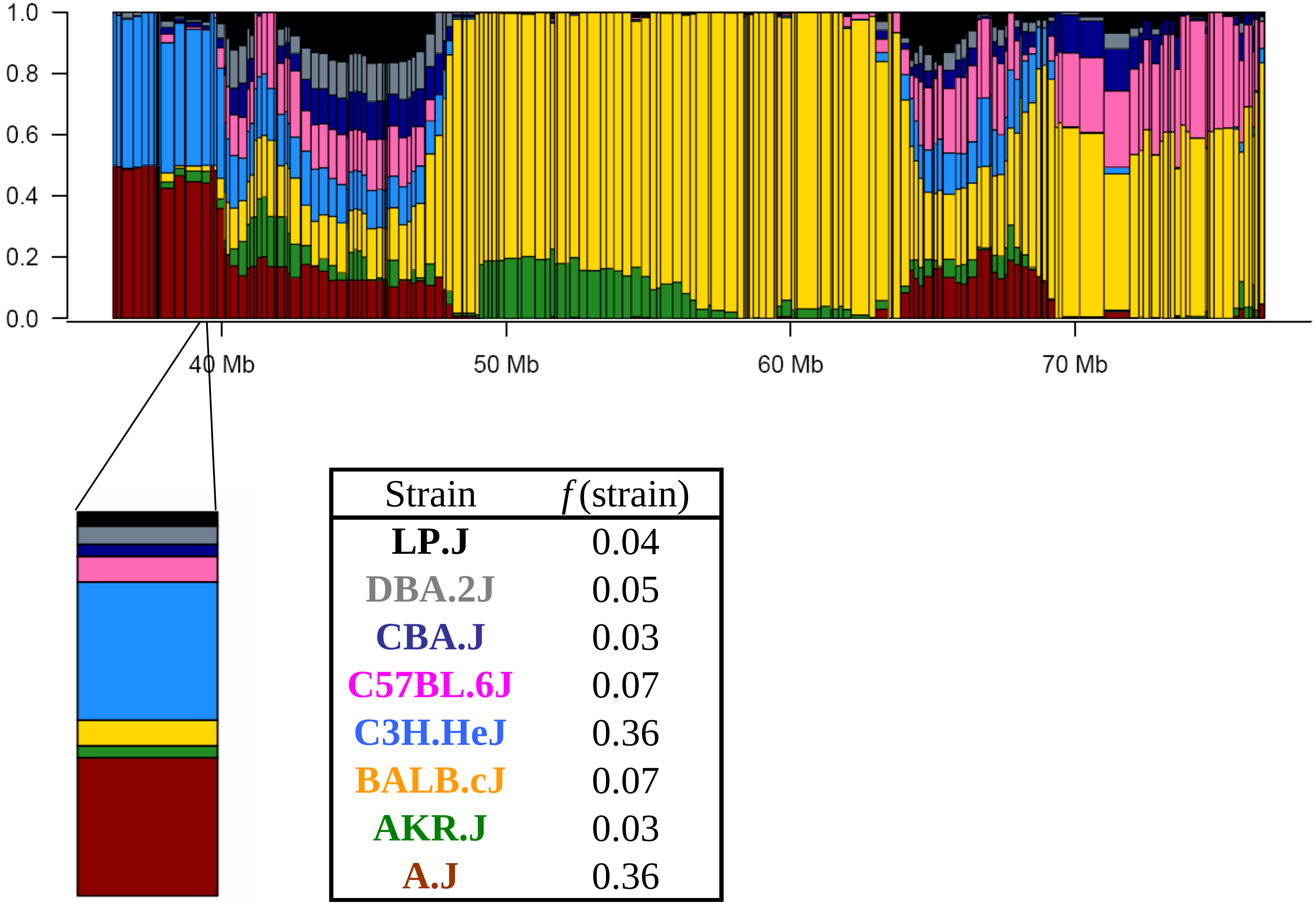
genotypes



haplotype
proportions
predicted by
HAPPY

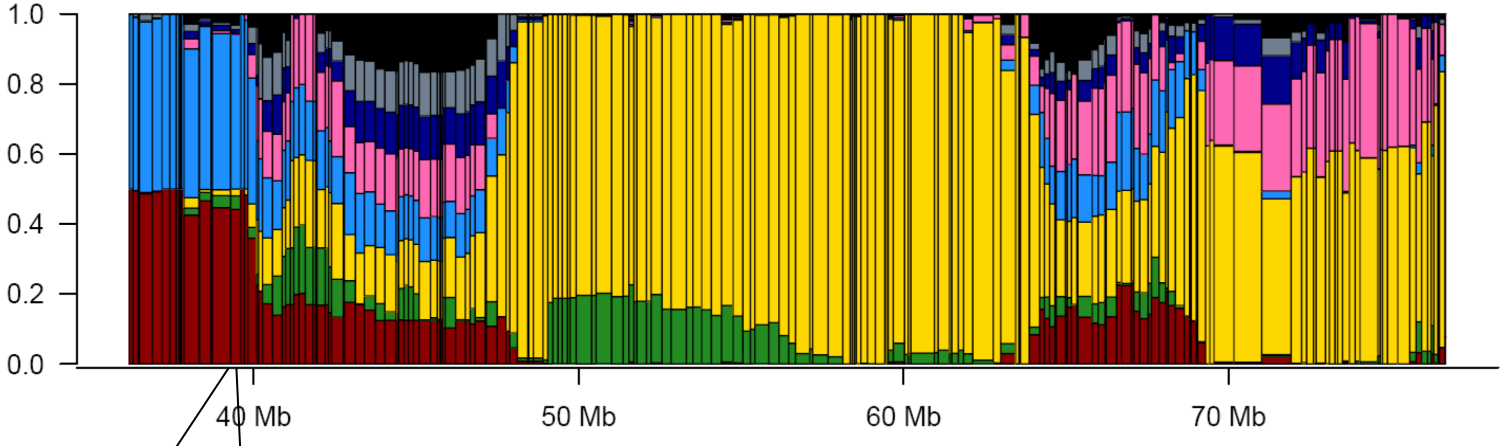


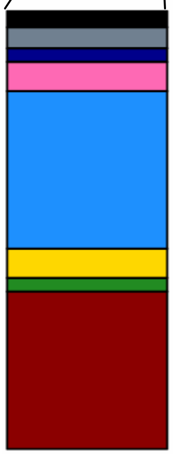
HAPPY model for additive effects



Strain	*f* (strain)
LP.J	0.04
DBA.2J	0.05
CBA.J	0.03
C57BL.6J	0.07
C3H.HeJ	0.36
BALB.cJ	0.07
AKR.J	0.03
A.J	0.36

HAPPY model for additive effects





Strain	*f* (strain)
LP.J	0.04
DBA.2J	0.05
CBA.J	0.03
C57BL.6J	0.07
C3H.HeJ	0.36
BALB.cJ	0.07
AKR.J	0.03
A.J	0.36

Phenotype *y* is modeled as

$$y = \sum_{s=1}^{8} \tau(s) f(s) + \varepsilon$$

$\tau(s)$ is effect of strain *s*

HAPPY effects models

Additive model

$$y = \sum_{s=1}^{8} \tau(s) f(s) + \varepsilon$$

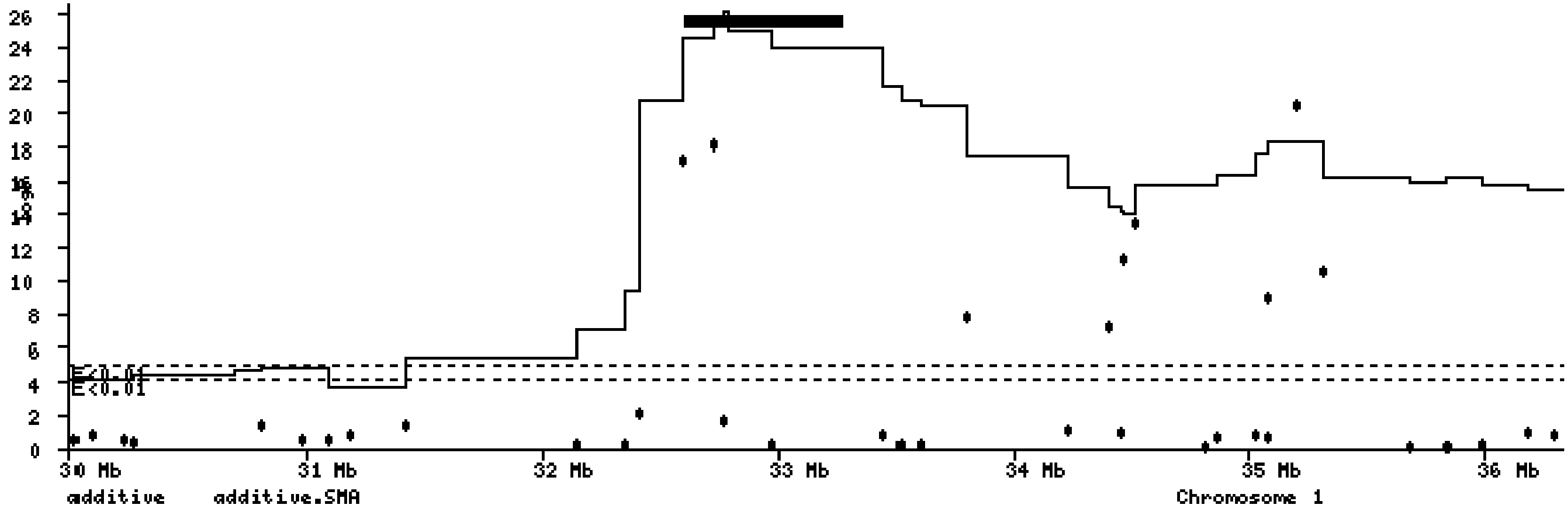
Additive model with covariate effects

$$y = \mu + \sum_{j \in \text{covariates}} \alpha_j + \sum_{s=1}^{8} \tau(s) f(s) + \varepsilon$$

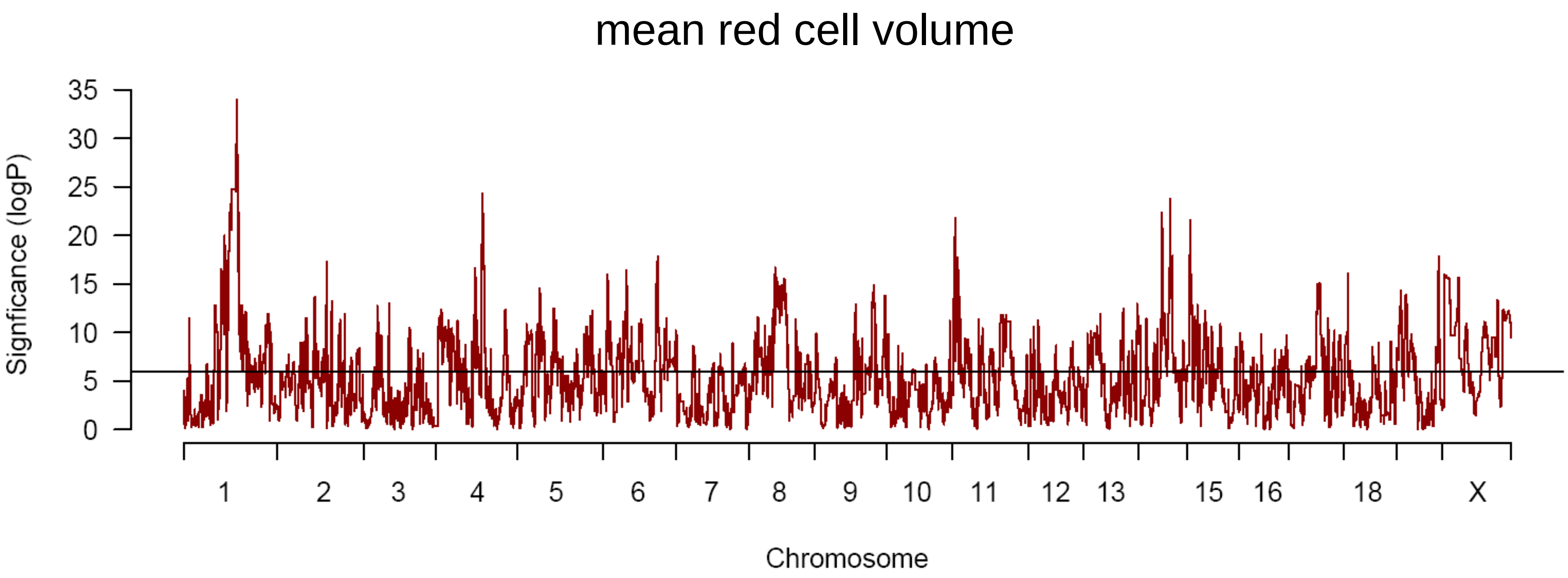
Full (ie, additive & dominance) model with covariate effects

$$y = \mu + \sum_{j \in \text{covariates}} \alpha_j + \sum_{s,t} \tau(s,t) f(s,t) + \varepsilon$$

Genome scans with HAPPY



Many peaks

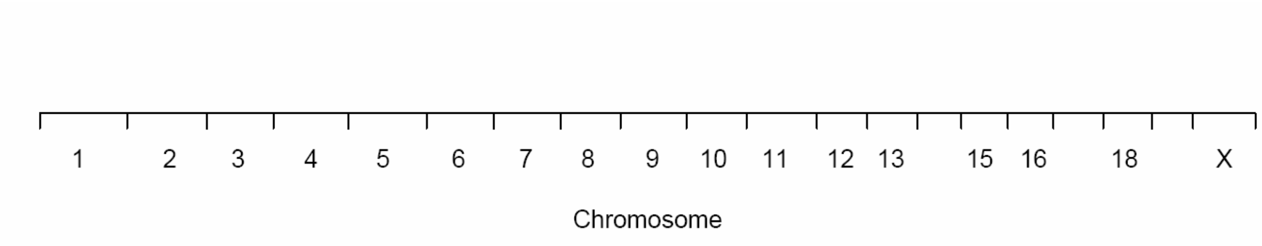

mean red cell volume
Signficance (logP)
0
5
10
15
20
25
30
35
1
2
3
4
5
6
7
8
9
10
11
12
13
15
16
18
X
Chromosome

Ghost peaks

family effects, cage effects, odd breeding

…complex pattern of linkage disequilibrium

How to select peaks: a simulated example



How to select peaks: a simulated example

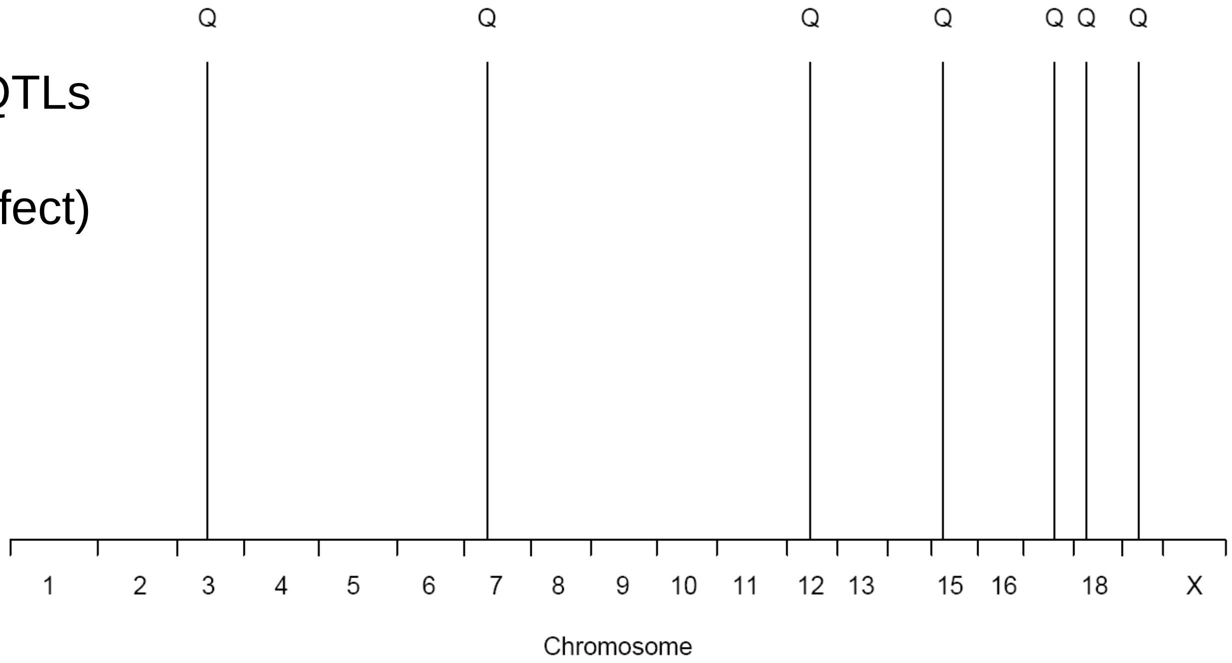
Simulate 7 x 5% QTLs

(ie, 35% genetic effect)

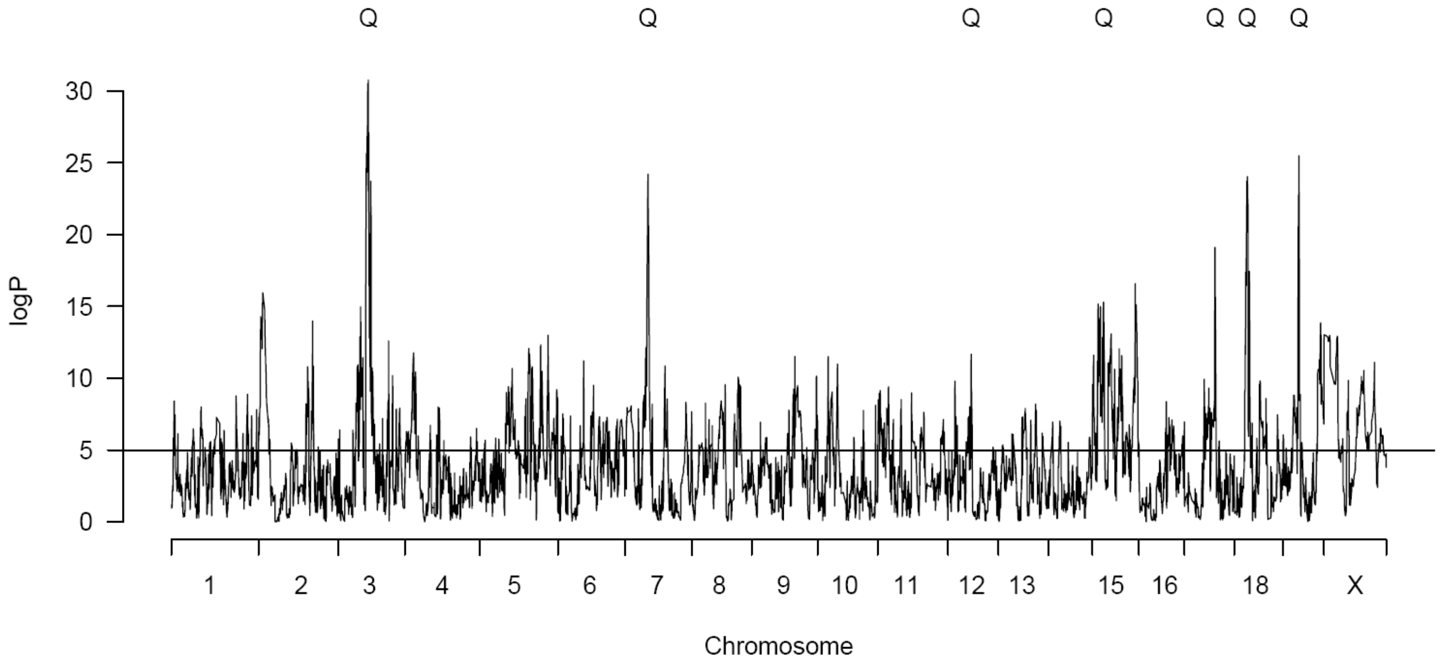
+ 20% shared environment effect

+ 45% noise

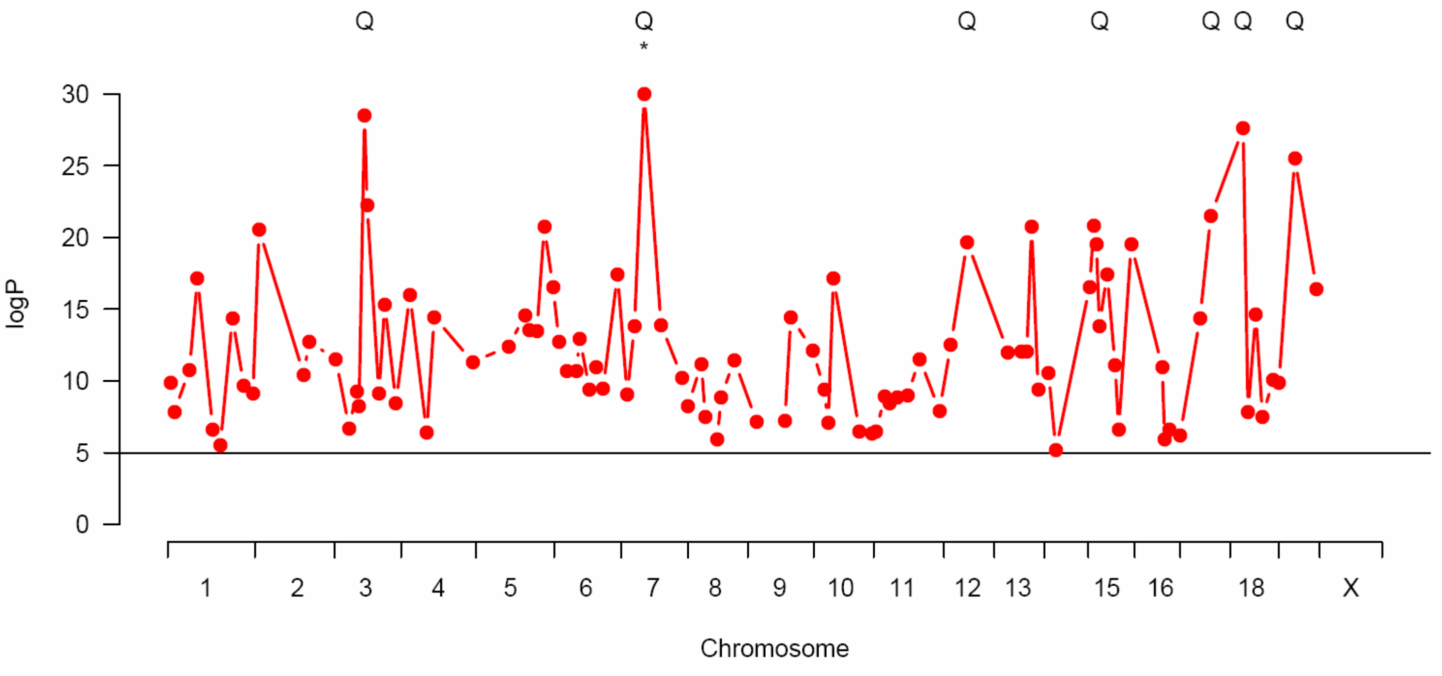
= 100% variance



Simulated example: 1D scan

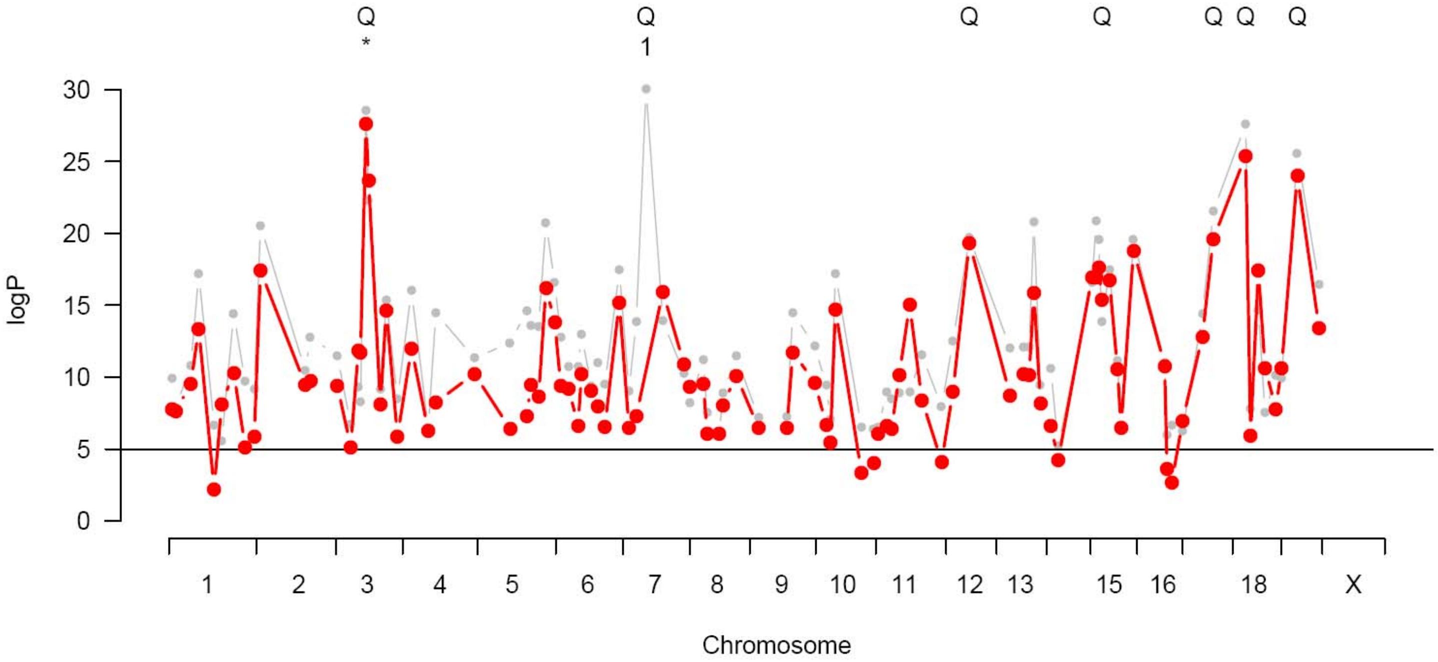


Peaks from 1D scan



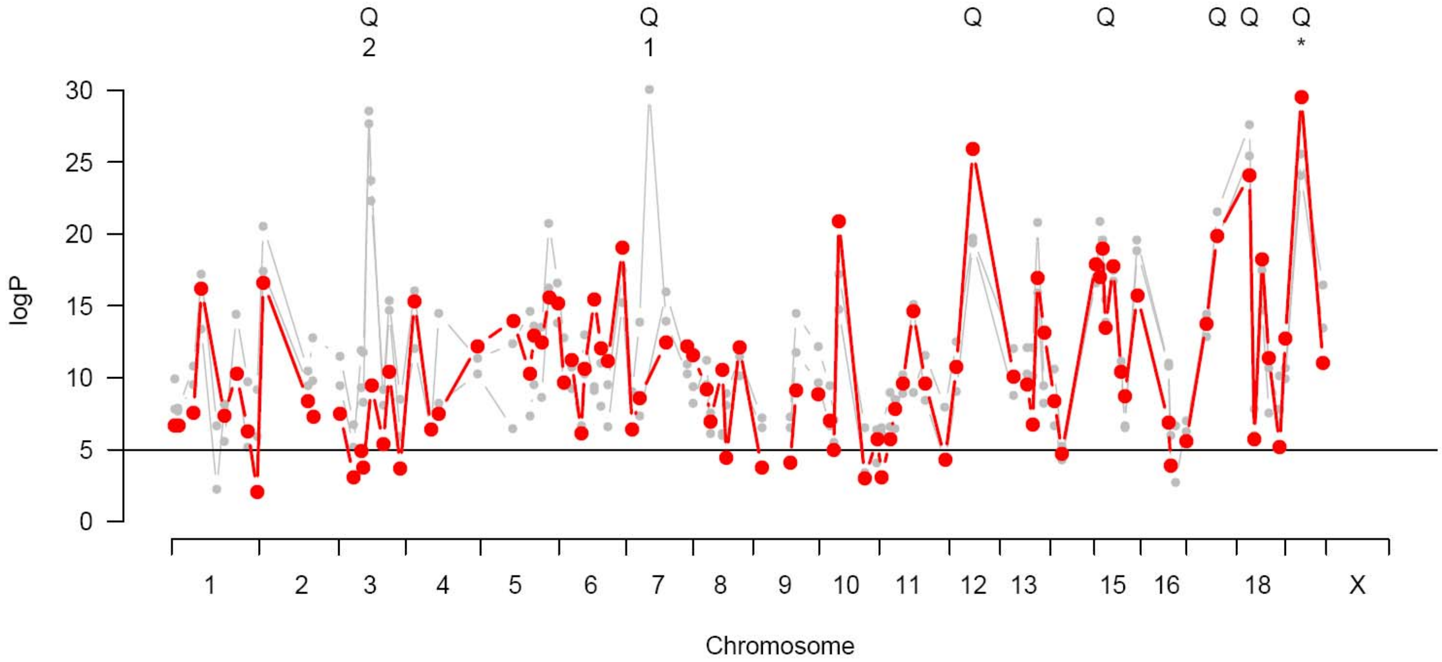
phenotype ~ covariates + ?

1D scan: condition on 1 peak



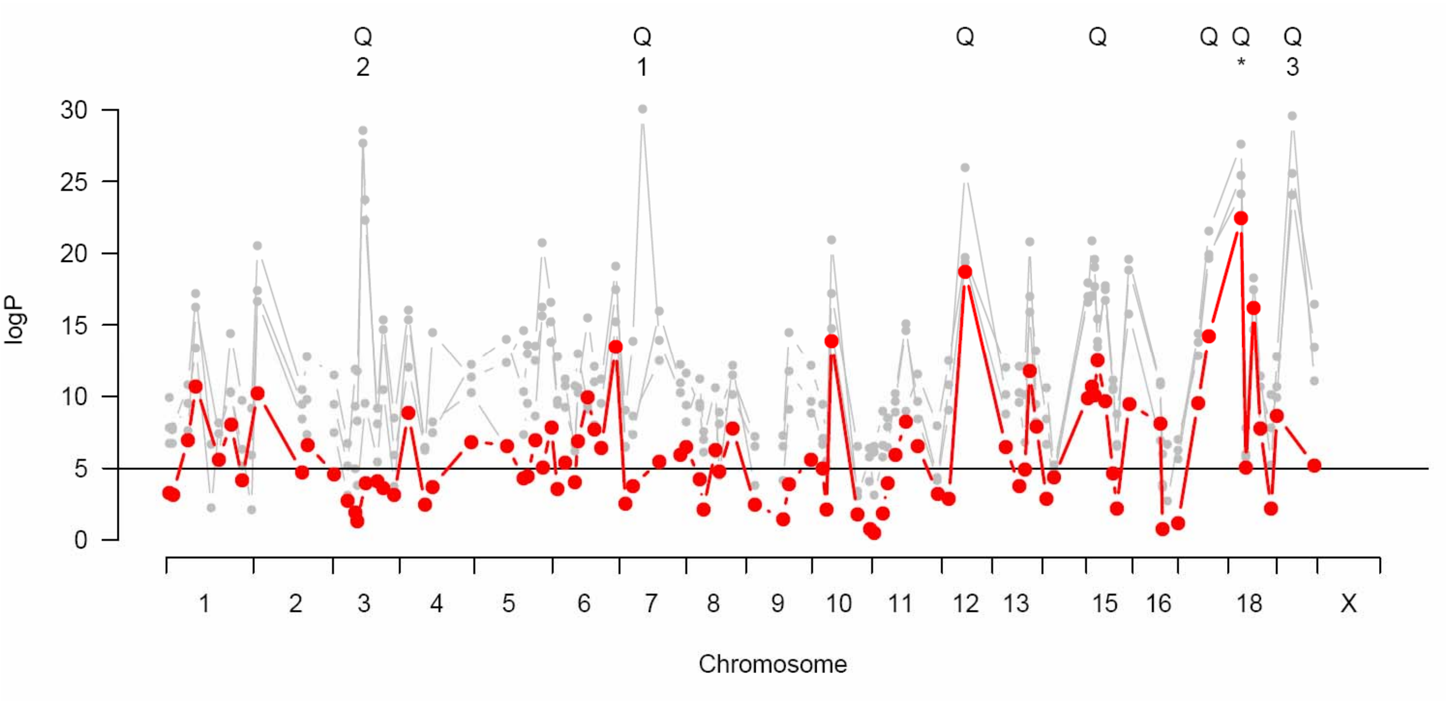
phenotype ~ covariates + peak 1 + ?

1D scan: condition on 2 peaks



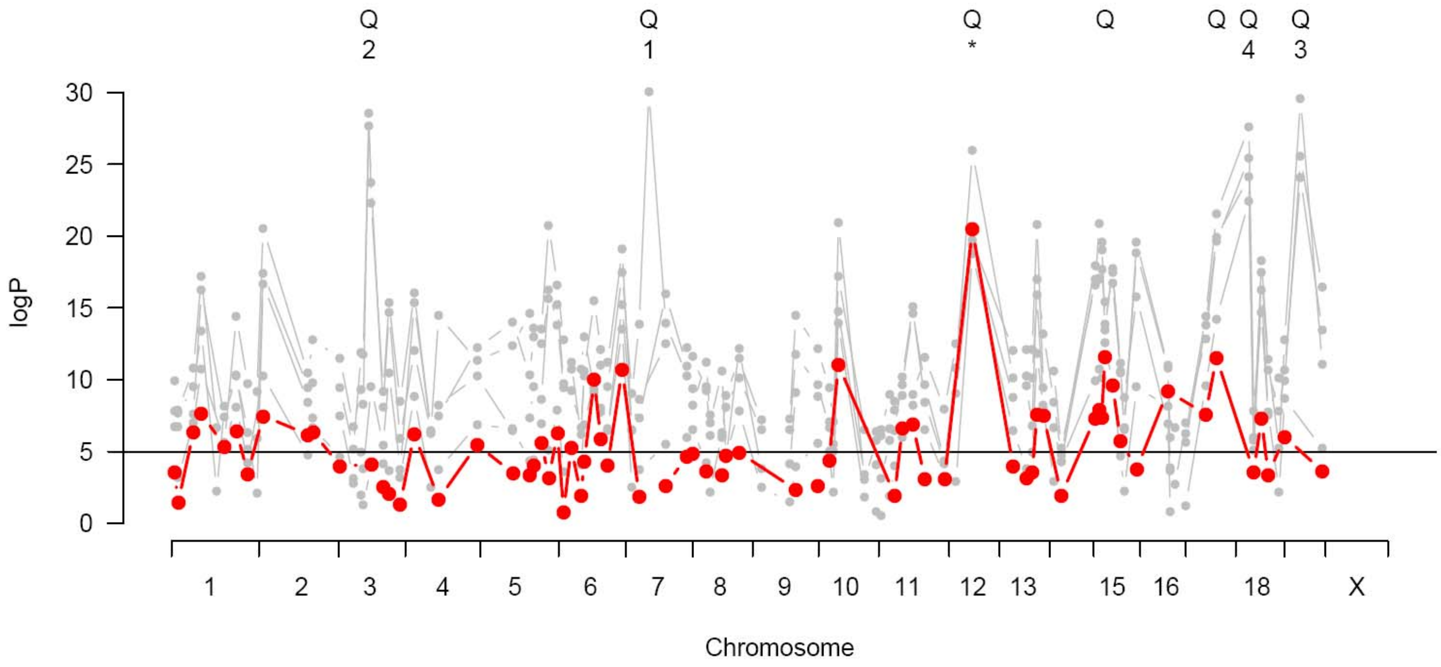
phenotype ~ covariates + peak 1 + peak 2 + ?

1D scan: condition on 3 peaks



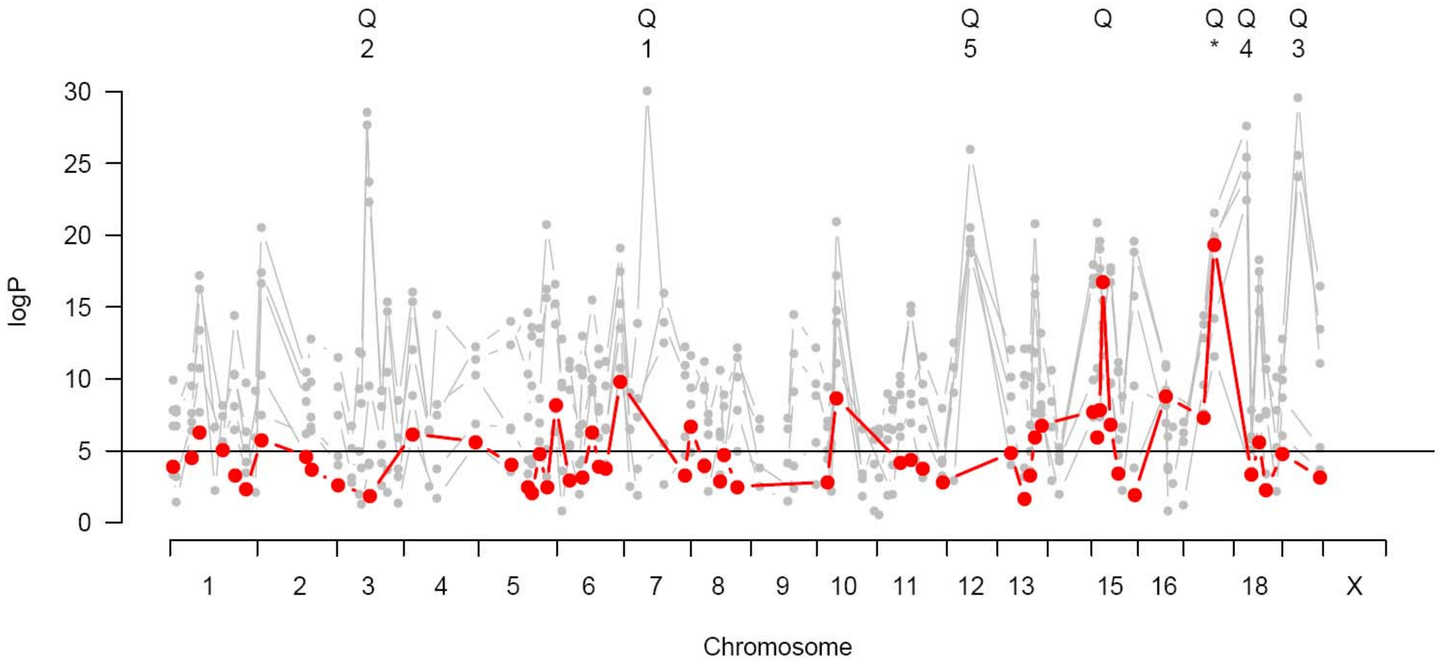
phenotype ~ covariates + peak 1 + peak 2 + peak 3 + ?

1D scan: condition on 4 peaks



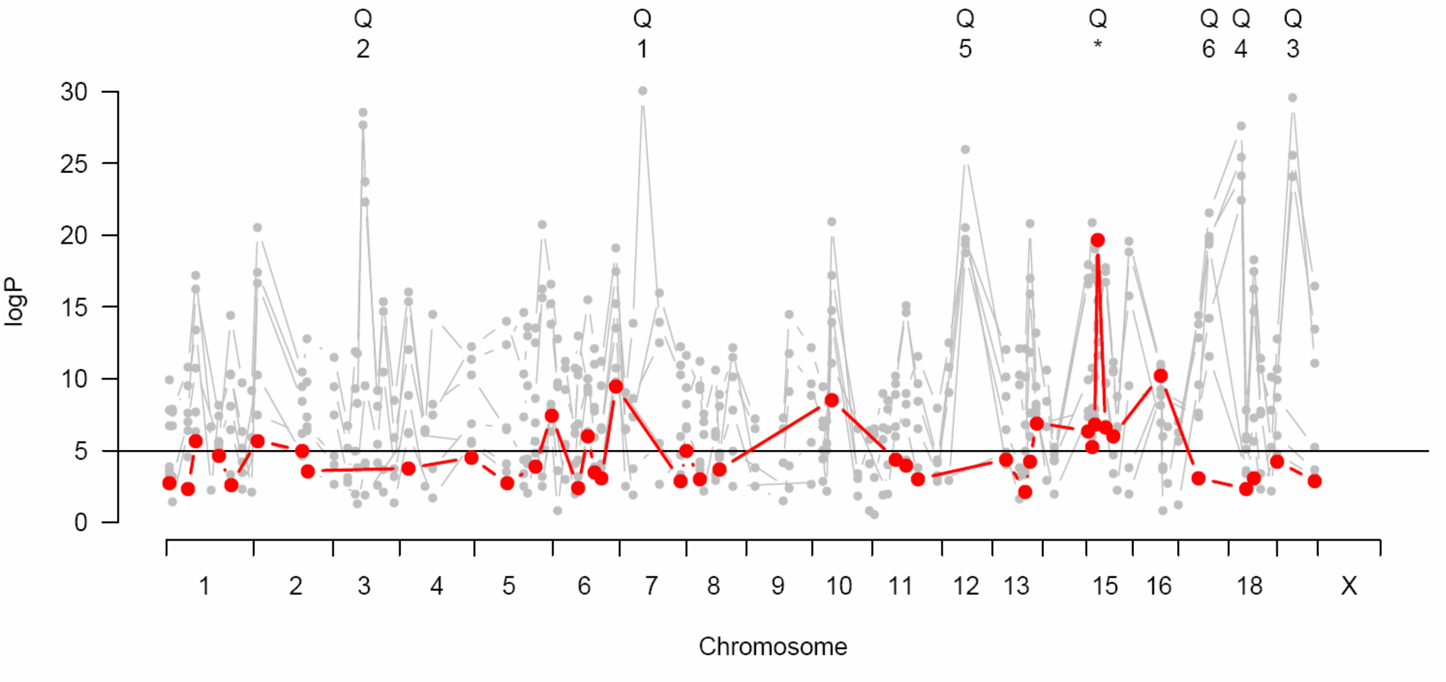
phenotype ~ covariates + peak 1 + peak 2 + peak 3 +
peak 4 + ?

1D scan: condition on 5 peaks



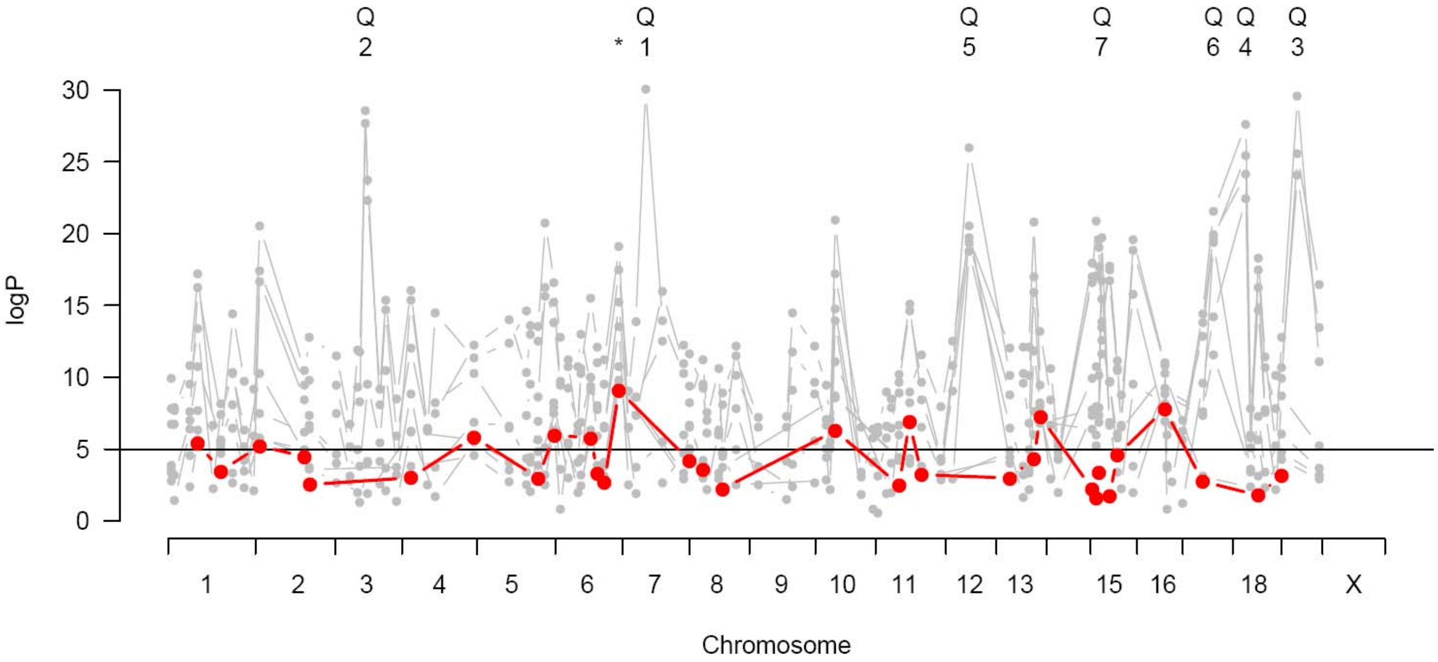
phenotype ~ covariates + peak 1 + peak 2 + peak 3 + peak 4 + peak 5 + ?

1D scan: condition on 6 peaks



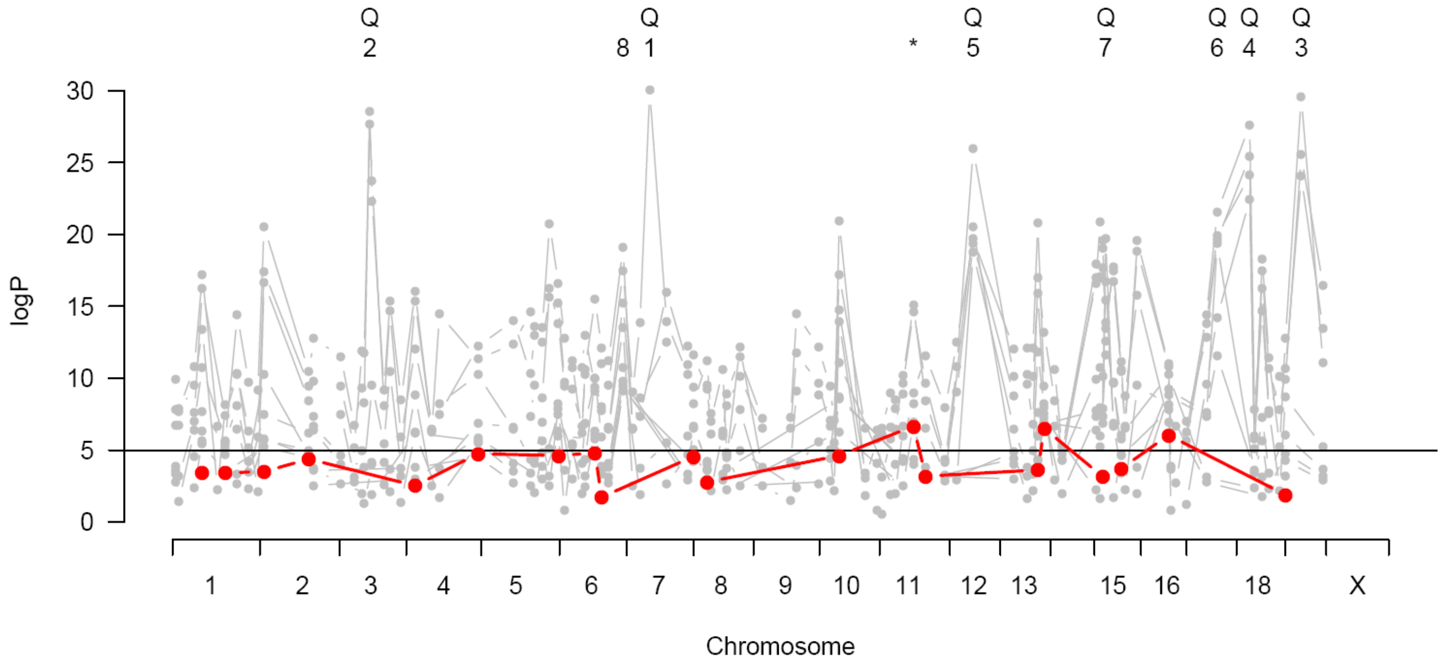
phenotype ~ covariates + peak 1 + peak 2 + peak 3 +
peak 4 + peak 5 + peak 6 + ?

1D scan: condition on 7 peaks



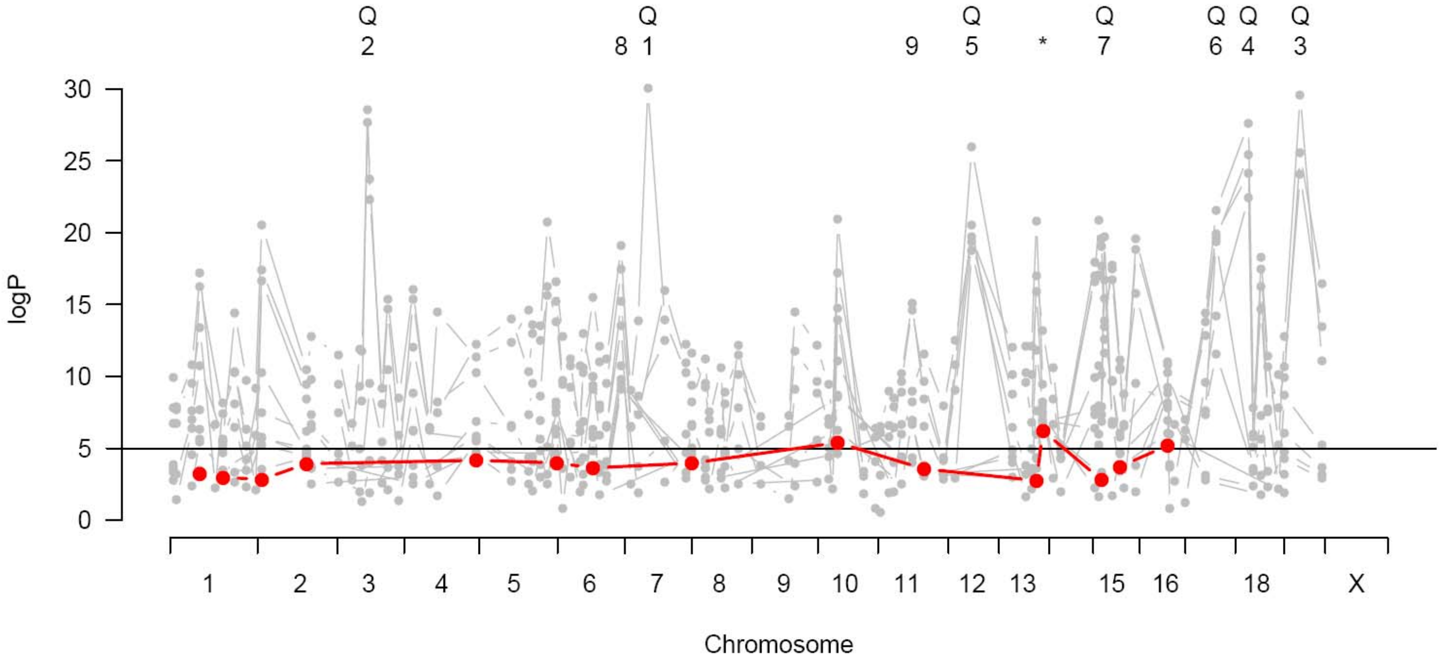
phenotype ~ covariates + peak 1 + peak 2 + peak 3 + peak 4 + peak 5 + peak 6 + peak 7 + ?

1D scan: condition on 8 peaks



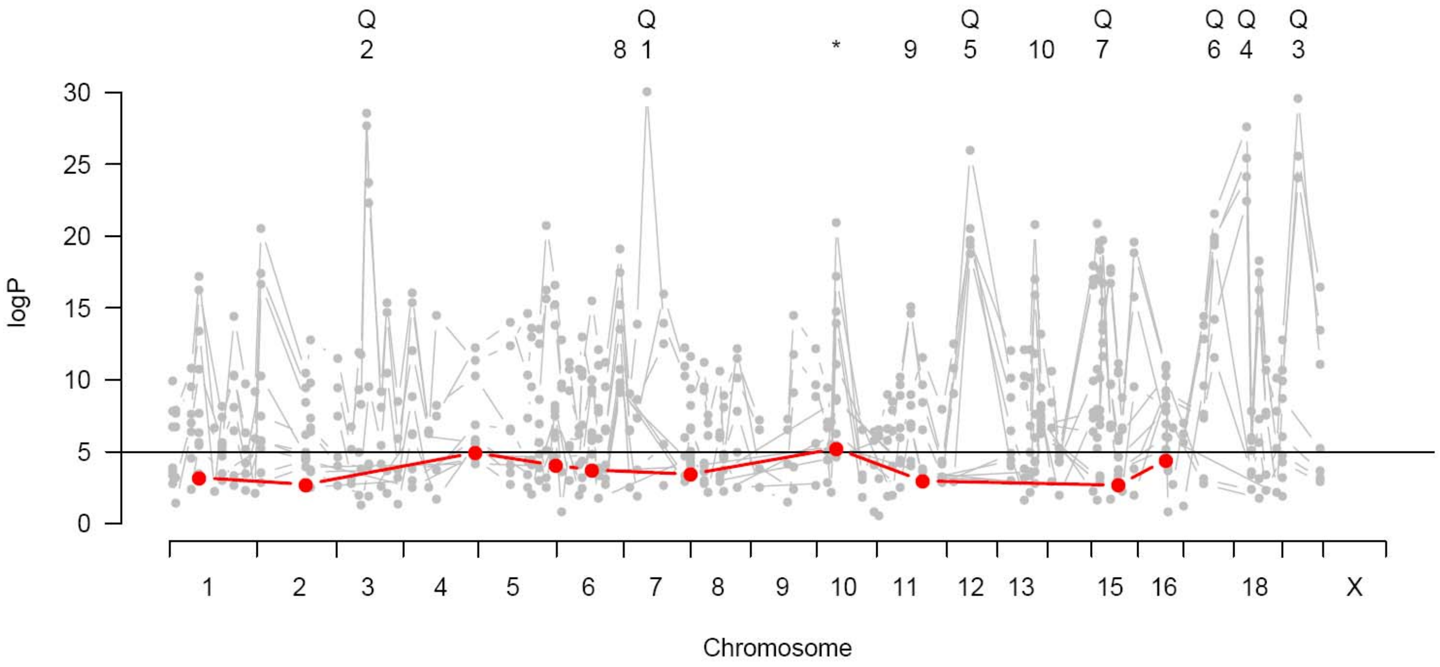
phenotype ~ covariates + peak 1 + peak 2 + peak 3 + peak 4 + peak 5 + peak 6 + peak 7 + peak 8 + ?

1D scan: condition on 9 peaks



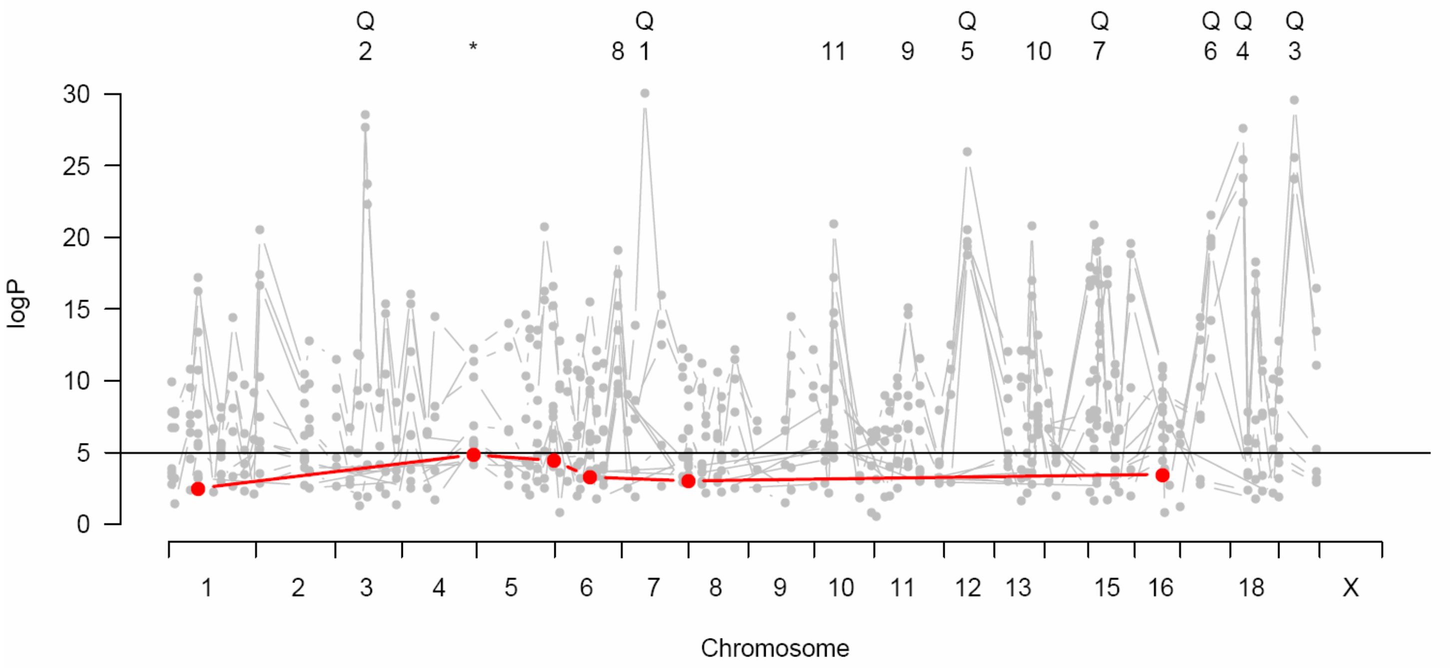
phenotype ~ covariates + peak 1 + peak 2 + peak 3 + peak 4 + peak 5 + peak 6 + peak 7 + peak 8 + peak 9 + ?

1D scan: condition on 10 peaks



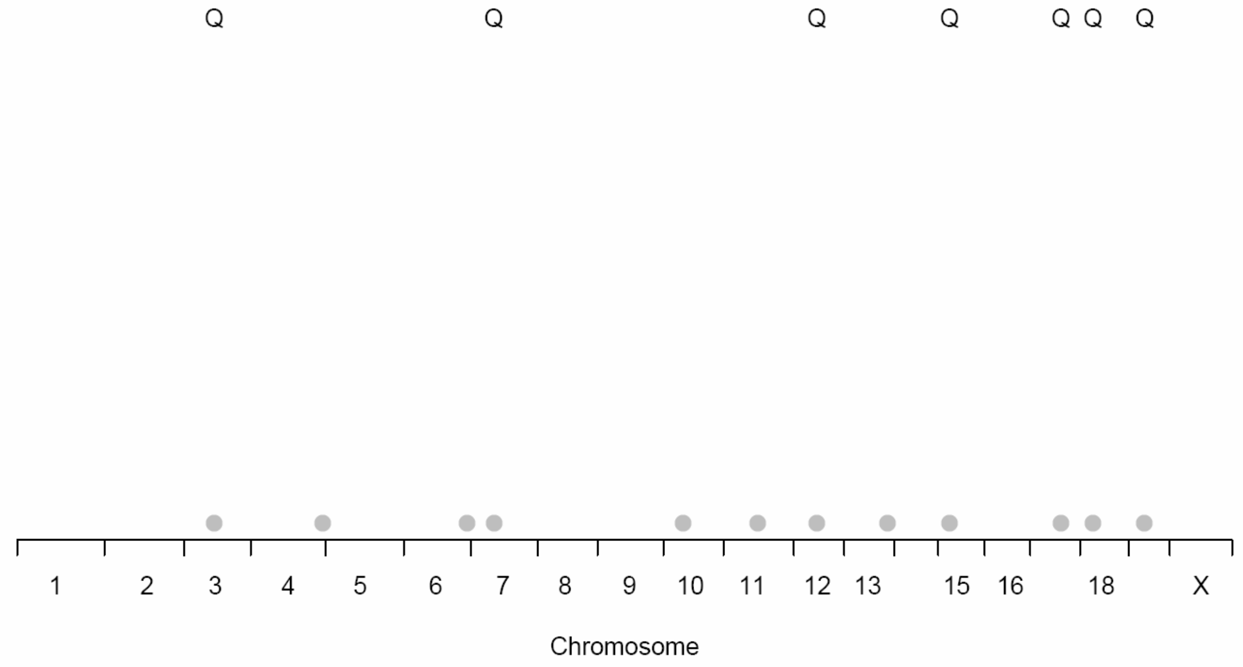
phenotype ~ covariates + peak 1 + peak 2 + peak 3 + peak 4 + peak 5 + peak 6 + peak 7 + peak 8 + peak 9 + peak 10 + ?

1D scan: condition on 11 peaks



phenotype ~ covariates + peak 1 + peak 2 + peak 3 + peak 4 + peak 5 + peak 6 + peak 7 + peak 8 + peak 9 + peak 10 + peak 11 + ?

Peaks chosen by forward selection



Bootstrap sampling

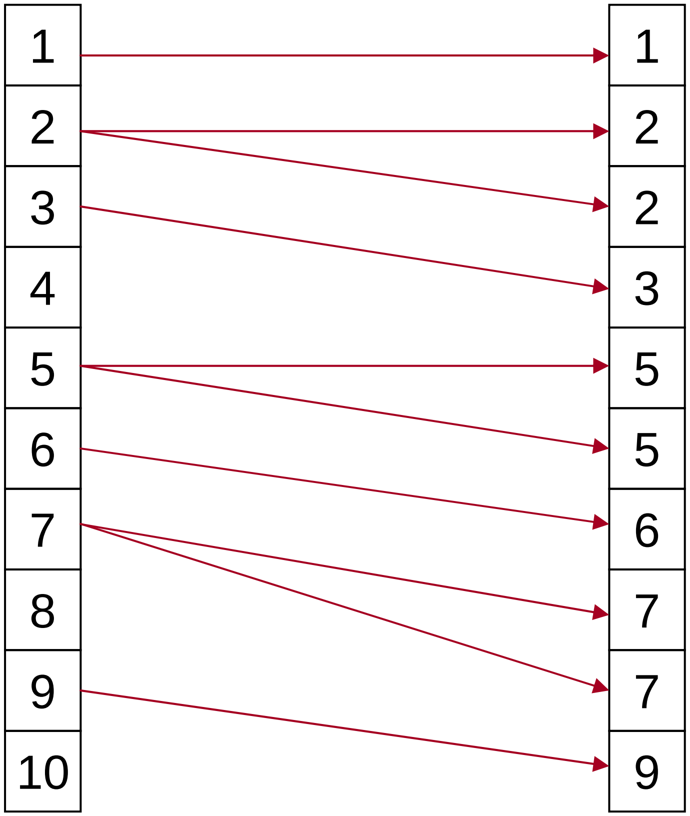
10 subjects

1
2
3
4
5
6
7
8
9
10

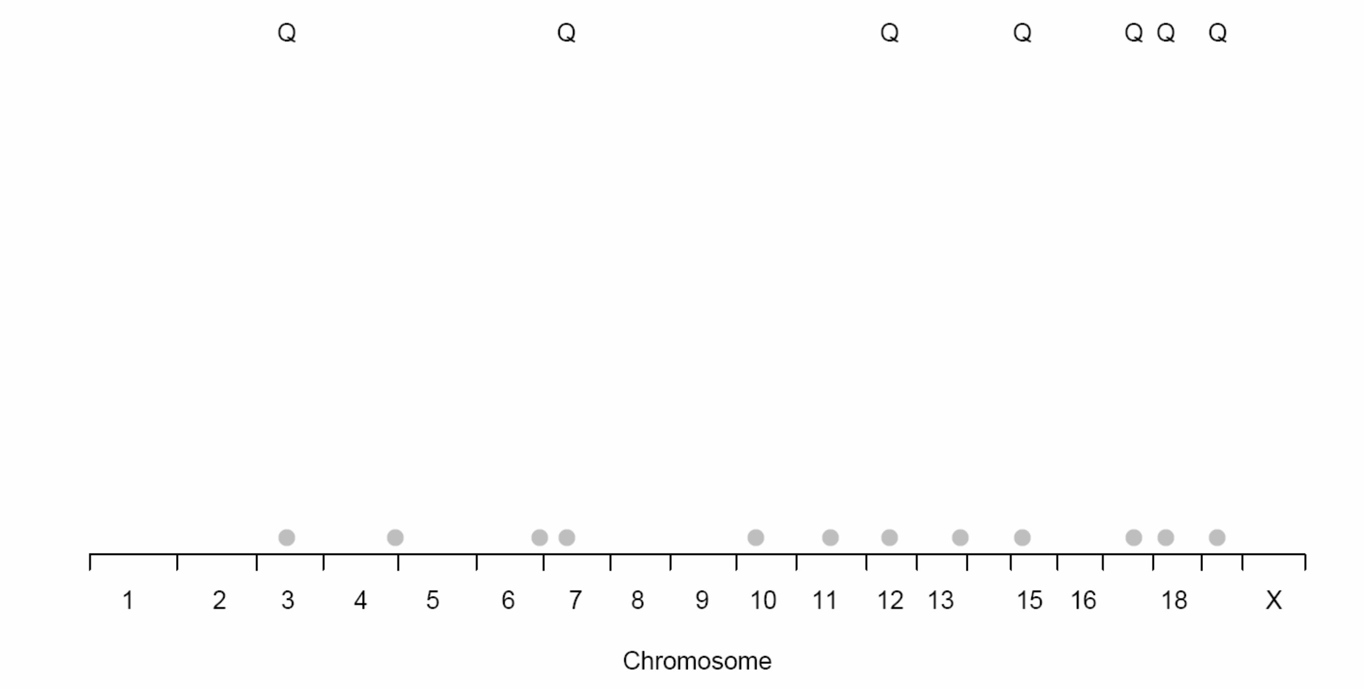
Bootstrap sampling

sample with
replacement

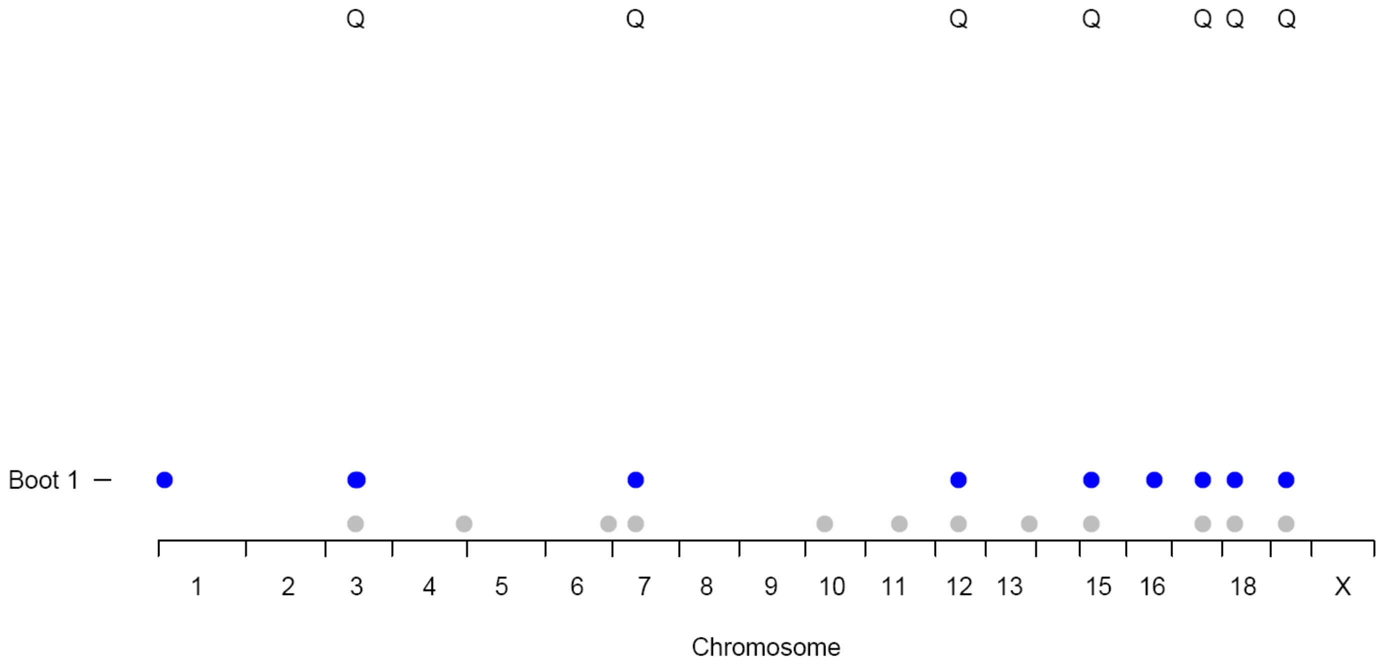
10 subjects



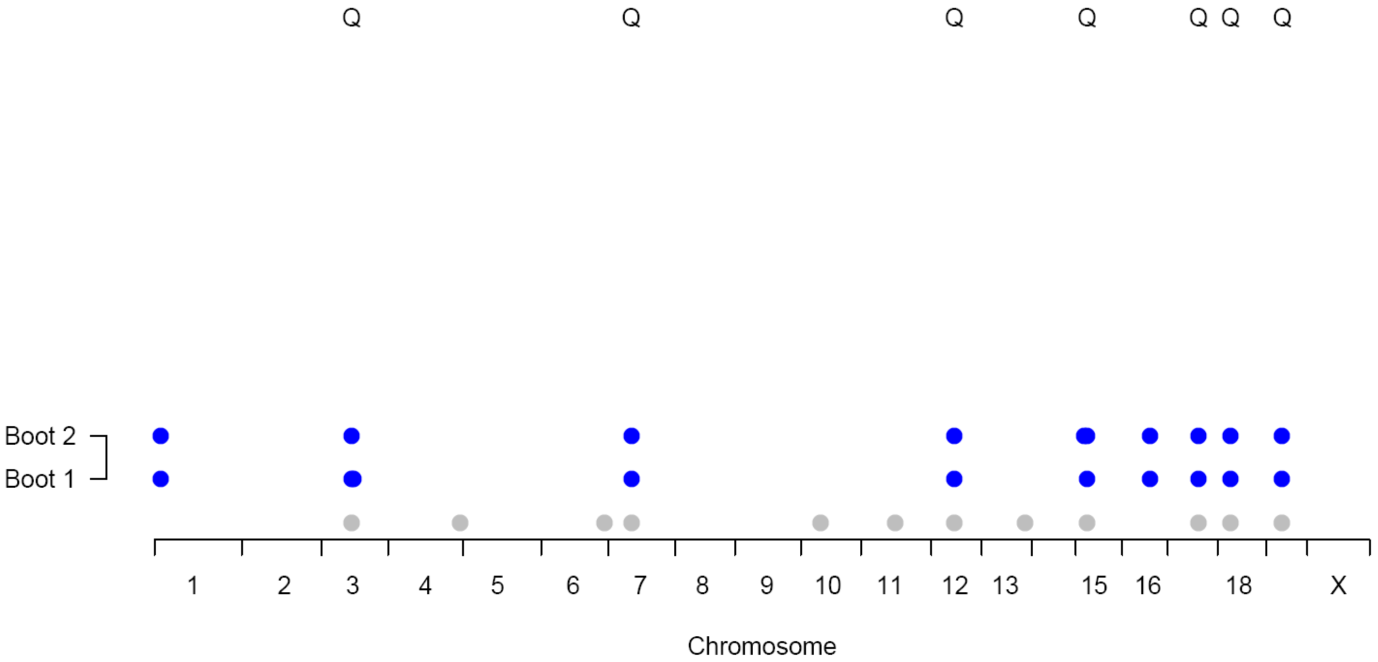
bootstrap sample
from
10 subjects


Q
Q
Q
Q
Q
Q
Q
1
2
3
4
5
6
7
8
9
10
11
12
13
15
16
18
X
Chromosome

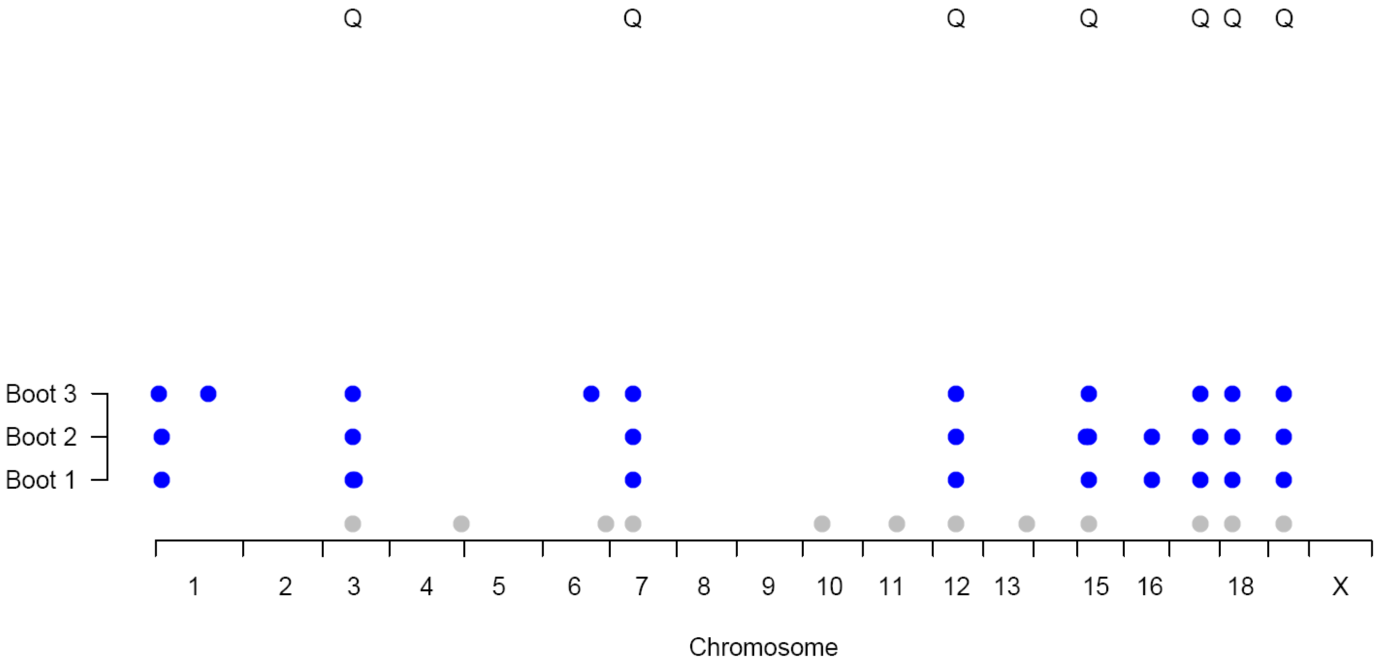
Forward selection on a bootstrap sample



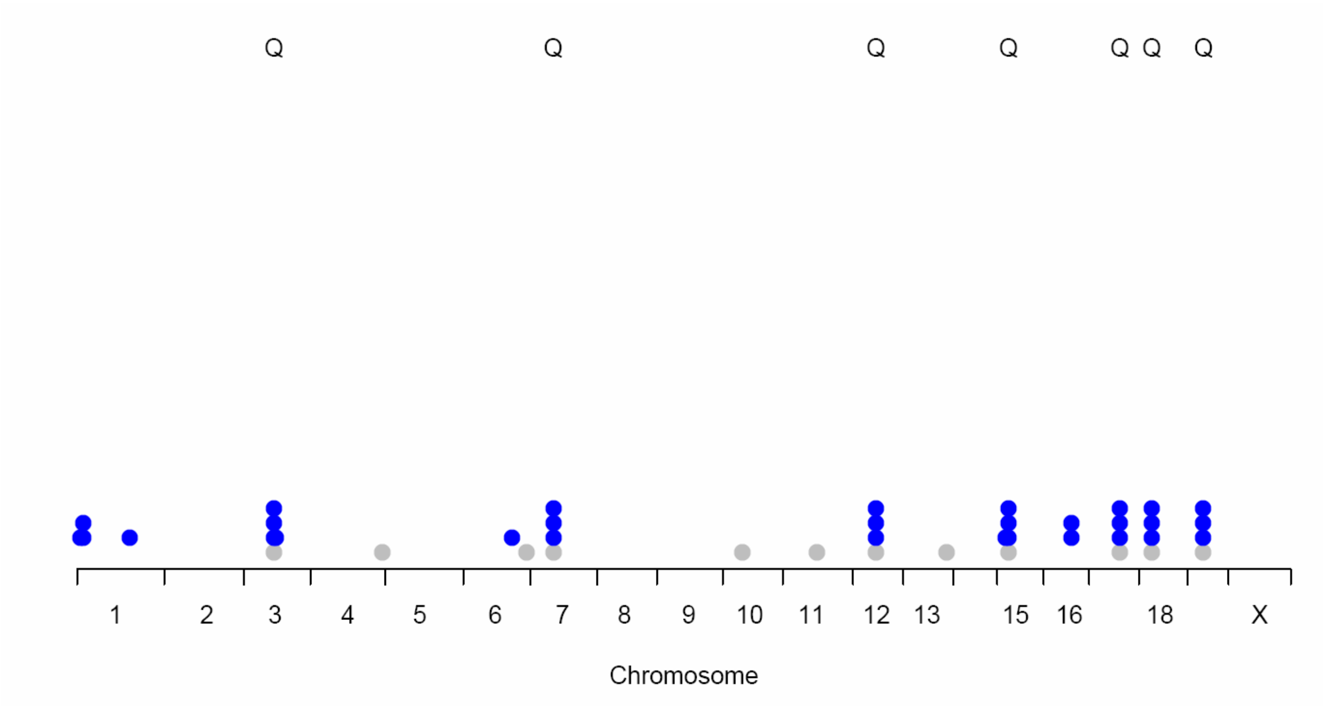
Forward selection on a bootstrap sample



Forward selection on a bootstrap sample

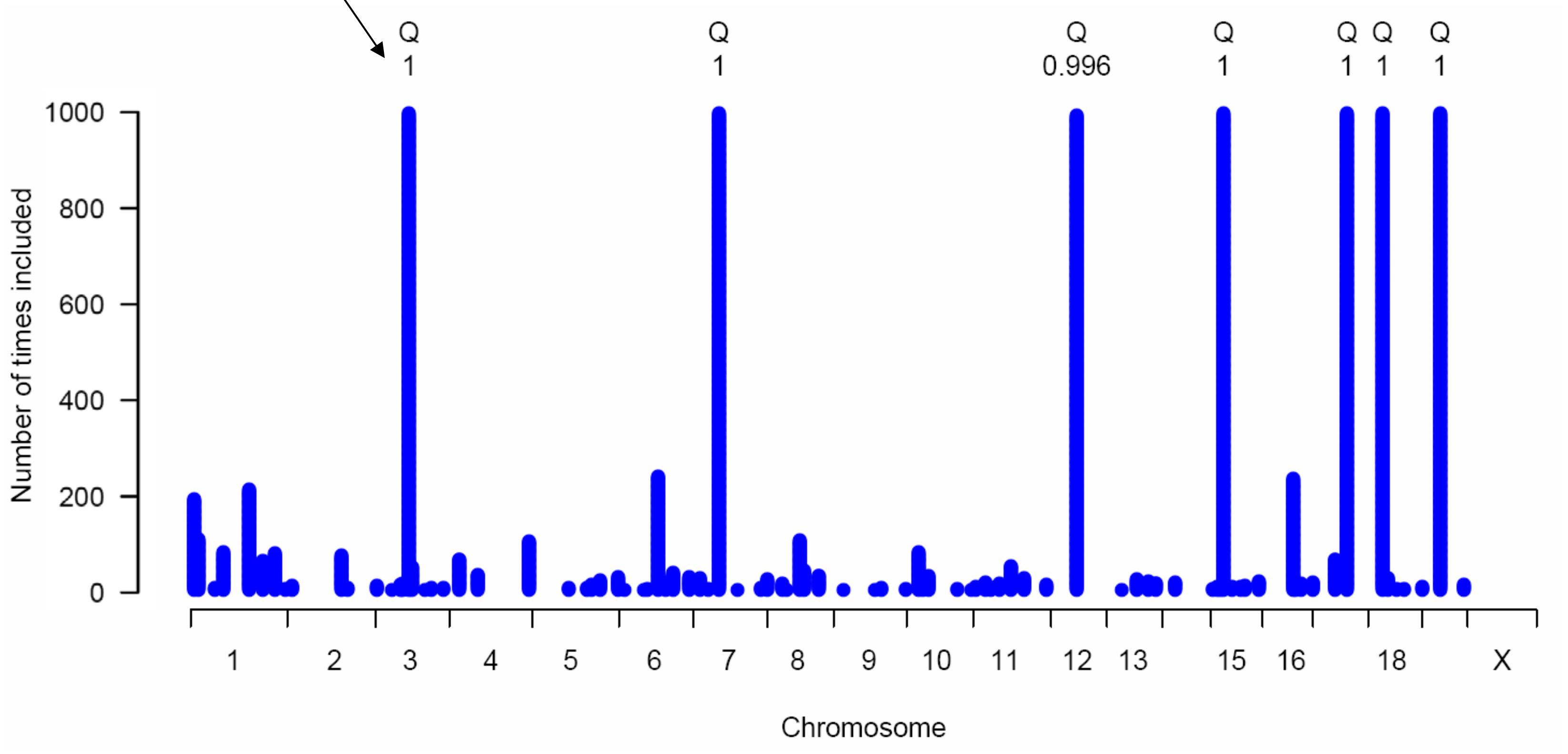


Bootstrap evidence mounts up…



In 1000 bootstraps…

Bootstrap Posterior Probability
(BPP)



Model averaging by bootstrap aggregation

Choosing only one model:
- very data-dependent, arbitrary
- can’t get all the true QTLs in one model

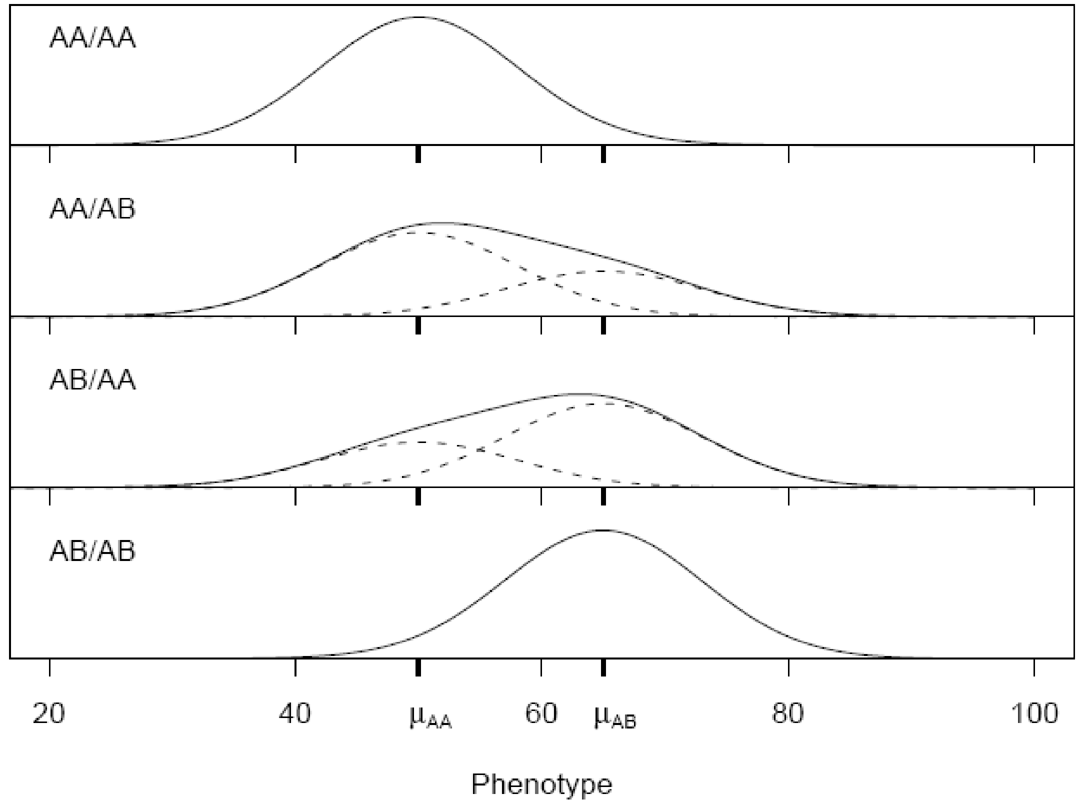
Bootstrap aggregation averages over models
- true QTLs get included more often than false ones

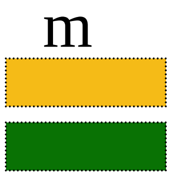
References:
- Broman & Speed (2002)
- Hackett et al (2001)

PRACTICAL: http://gscan.well.ox.ac.uk

ADDITIONAL SLIDES FROM HERE

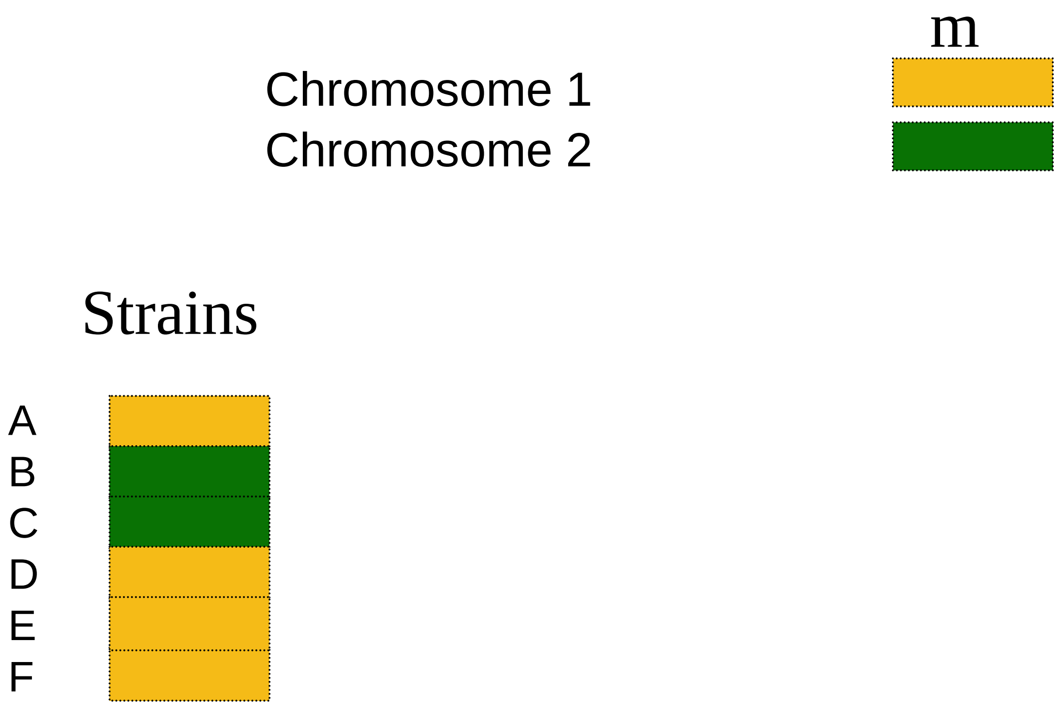
An individual's phenotype follows a mixture of normal distributions





Paternal chromosome

Maternal chromosome

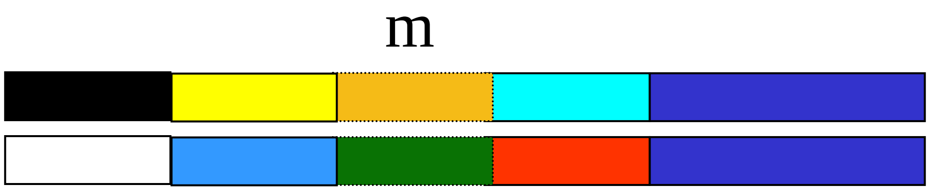


	m
Chromosome 1	🟨
Chromosome 2	🟩

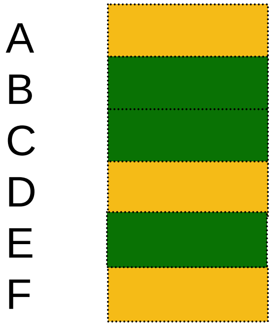
Strains

A	🟨
B	🟩
C	🟩
D	🟨
E	🟨
F	🟨

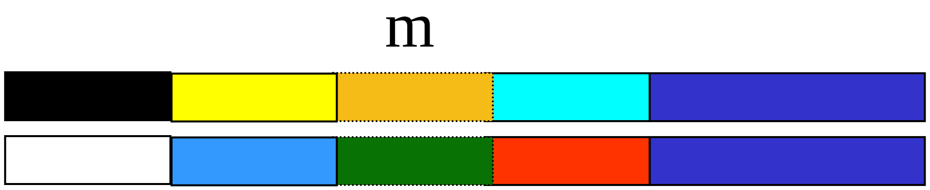
Markers



Strains

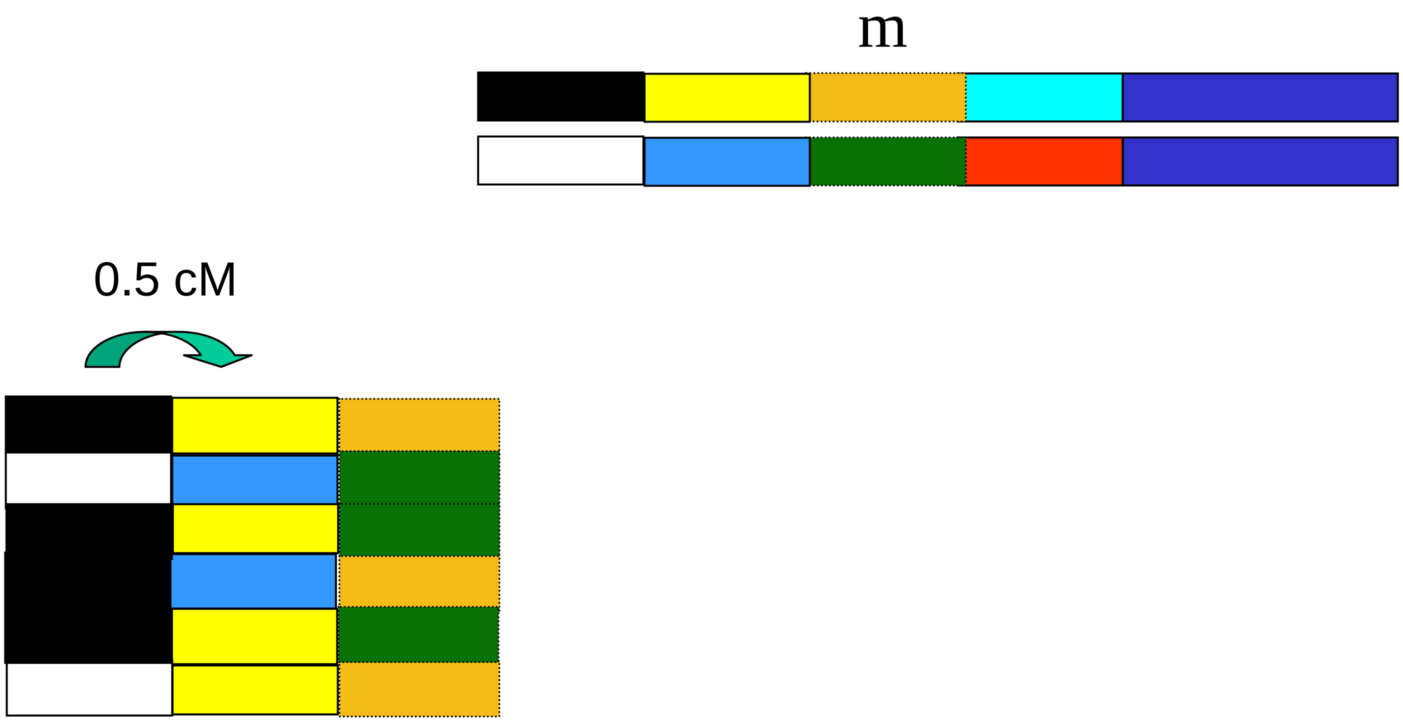


Markers

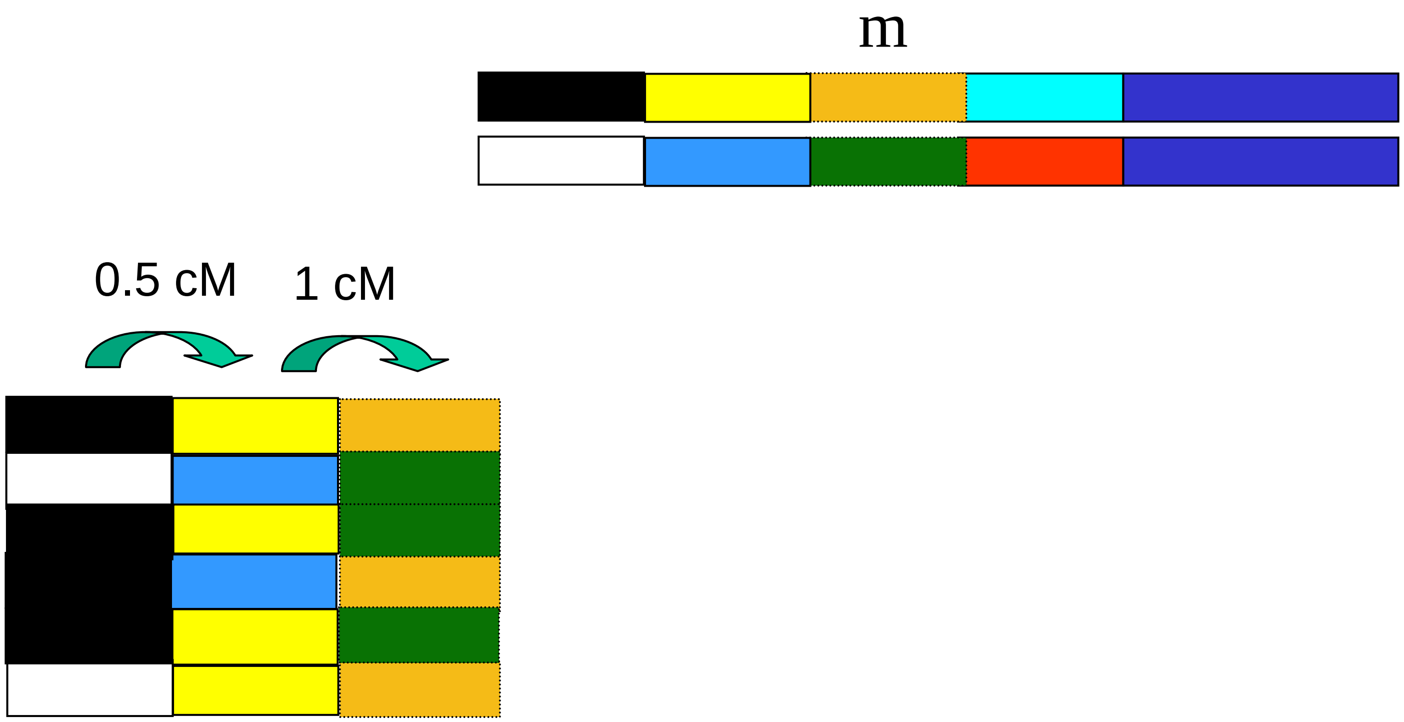


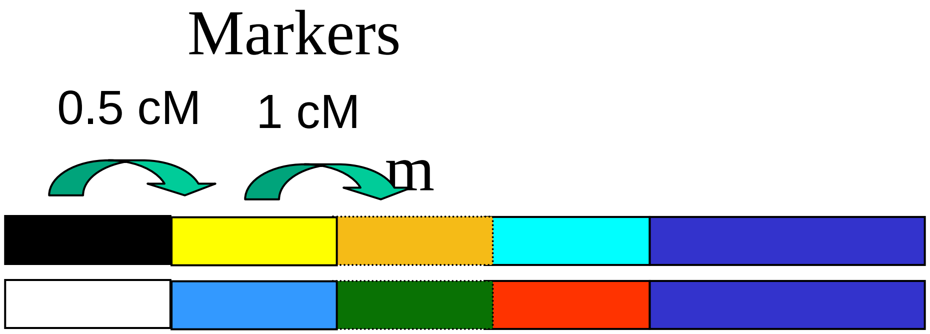
Strains

Markers



Markers

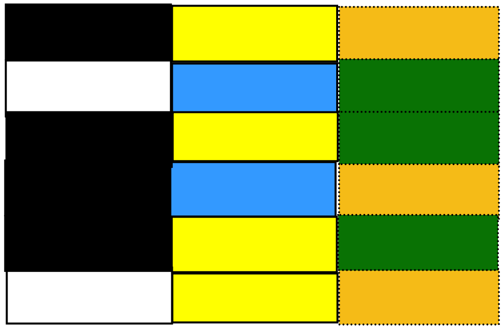




Markers

0.5 cM

1 cM

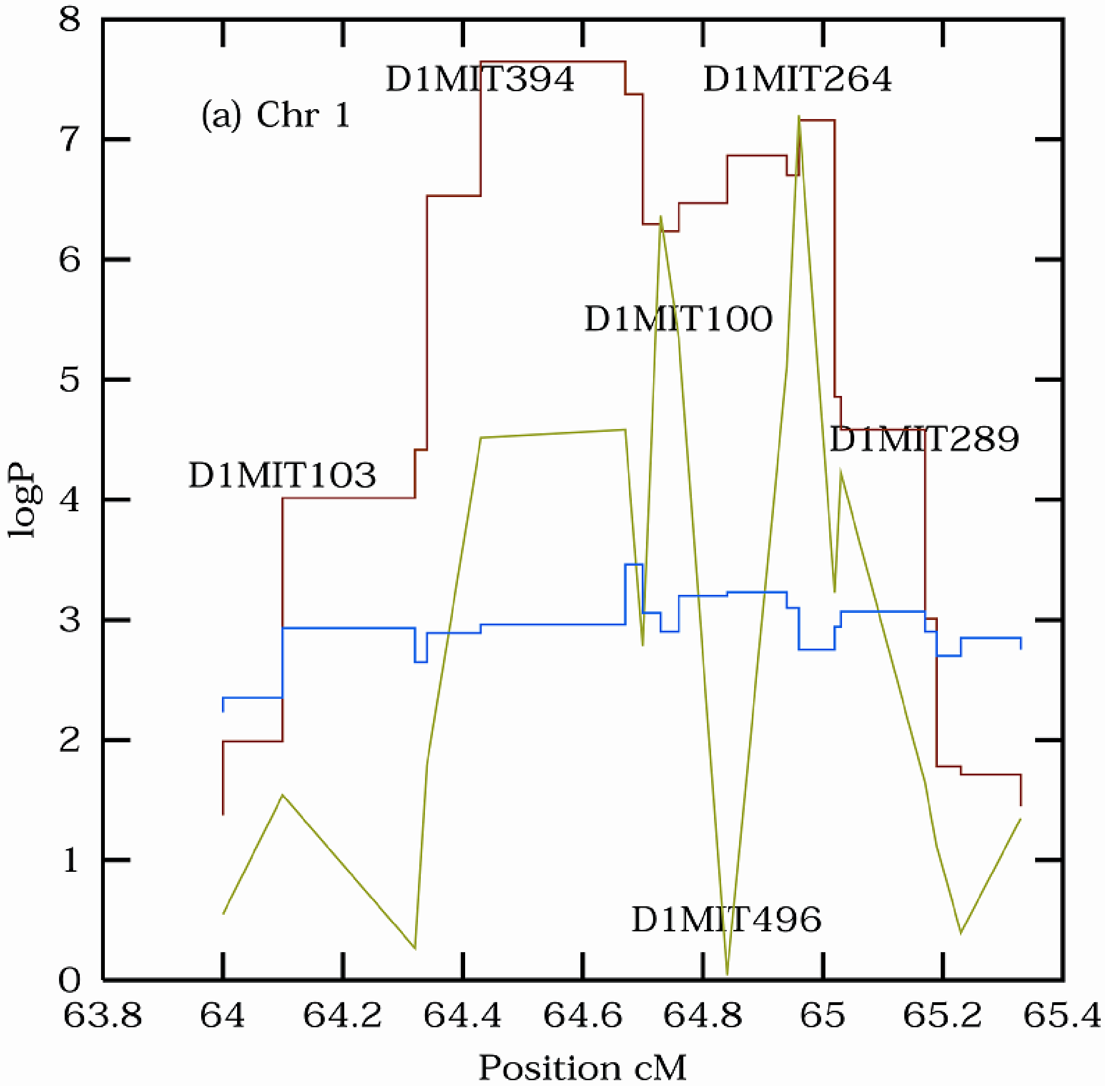


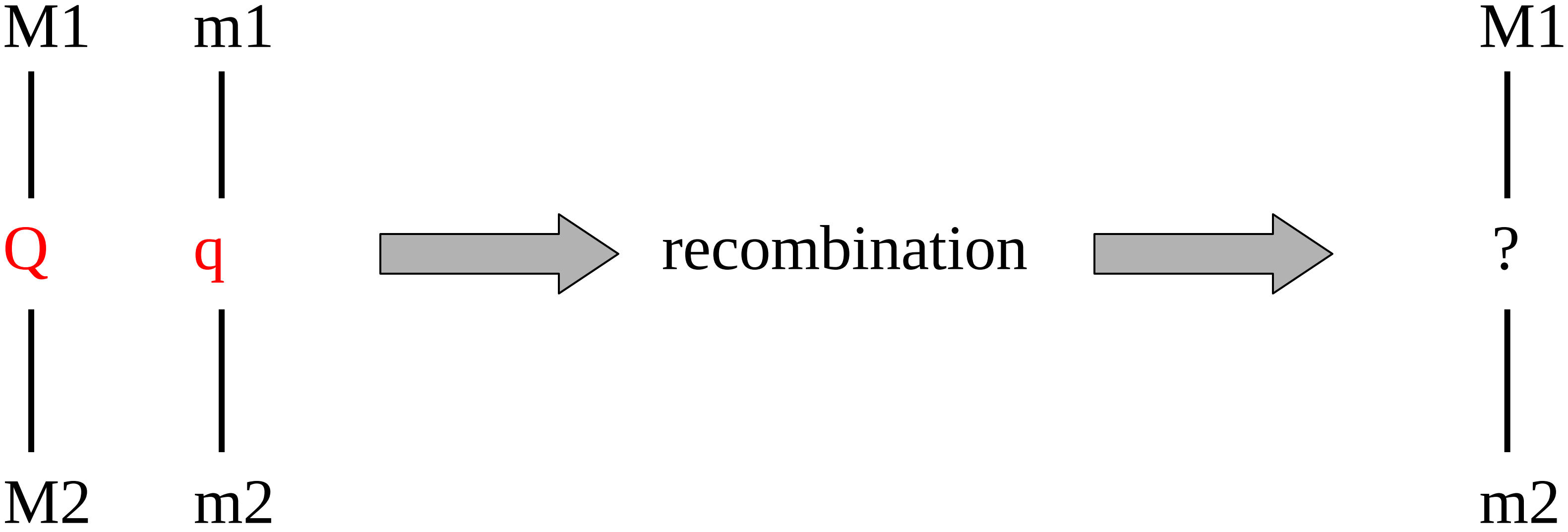
Analysis

$$x_m(s,t) = \sum_{s',t'} x_{m-1}(s',t')\psi_m(s,t|s',t')$$

Probabilistic Ancestral Haplotype Reconstruction (descent mapping): implemented in HAPPY

http://www.well.ox.ac.uk/~rmott/happy.html


(a) Chr 1
D1MIT394
D1MIT264
D1MIT100
D1MIT289
D1MIT103
D1MIT496
logP
Position cM
0
1
2
3
4
5
6
7
8
63.8
64
64.2
64.4
64.6
64.8
65
65.2
65.4



M1 m1

Q q

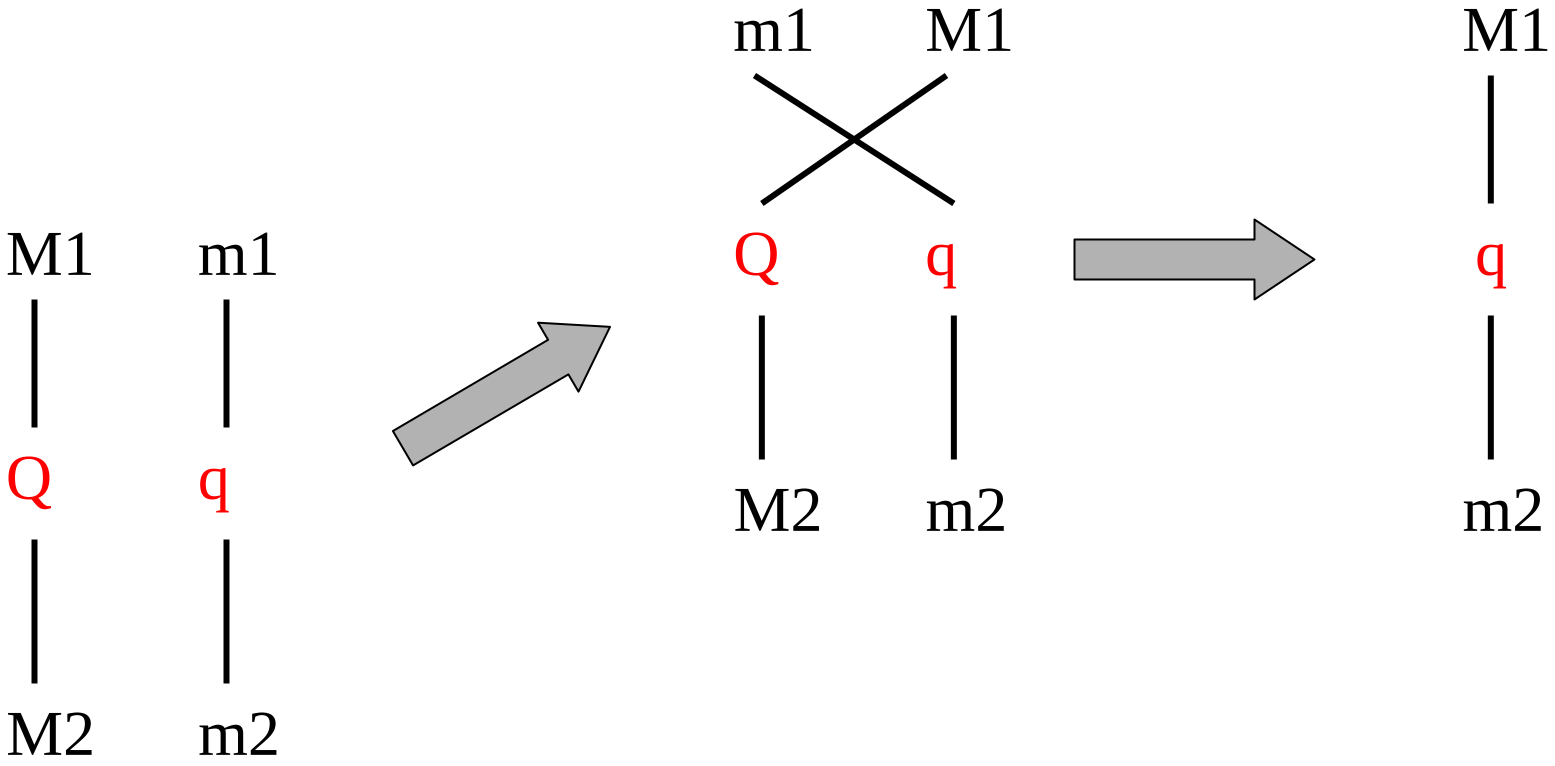
M2 m2

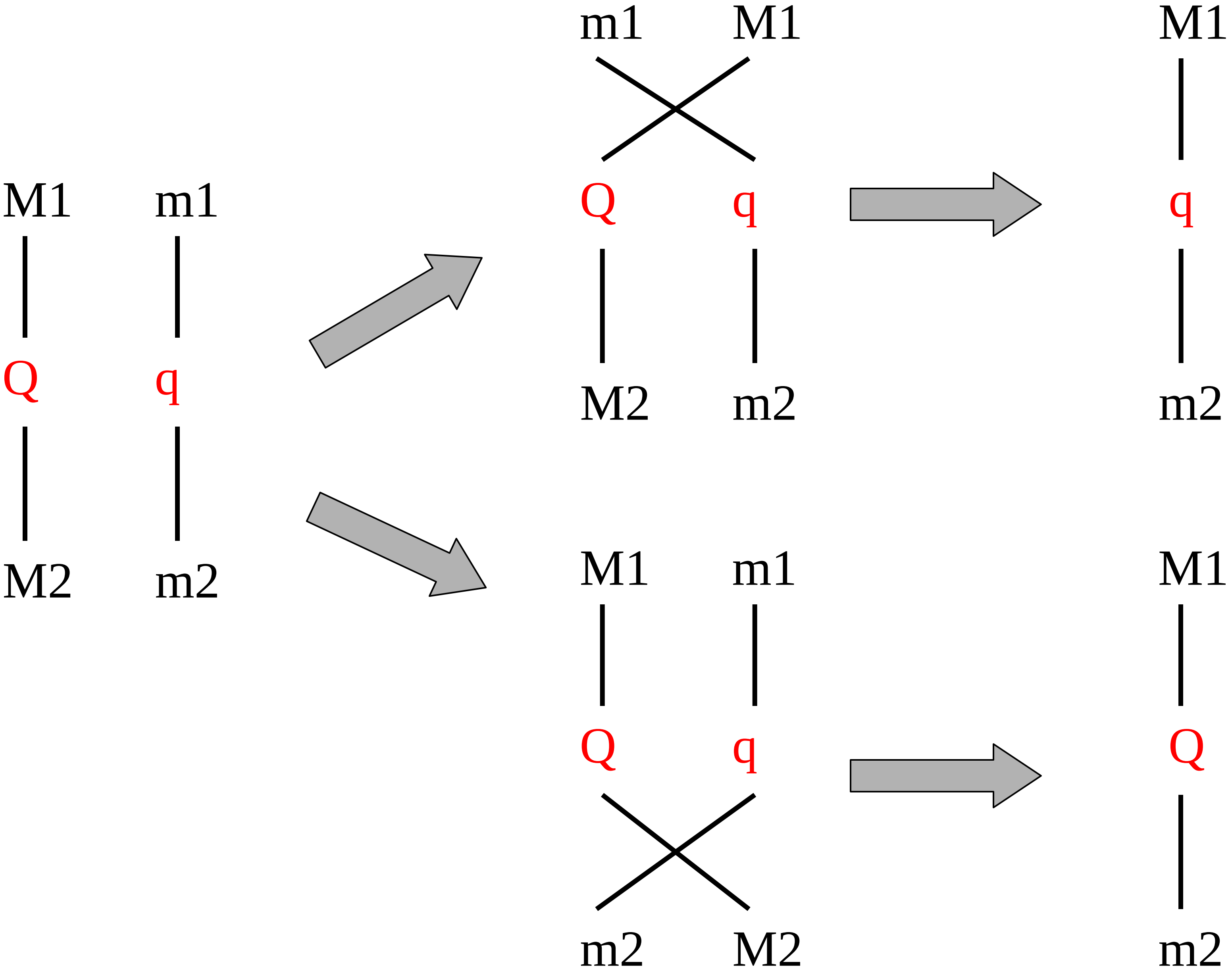
recombination

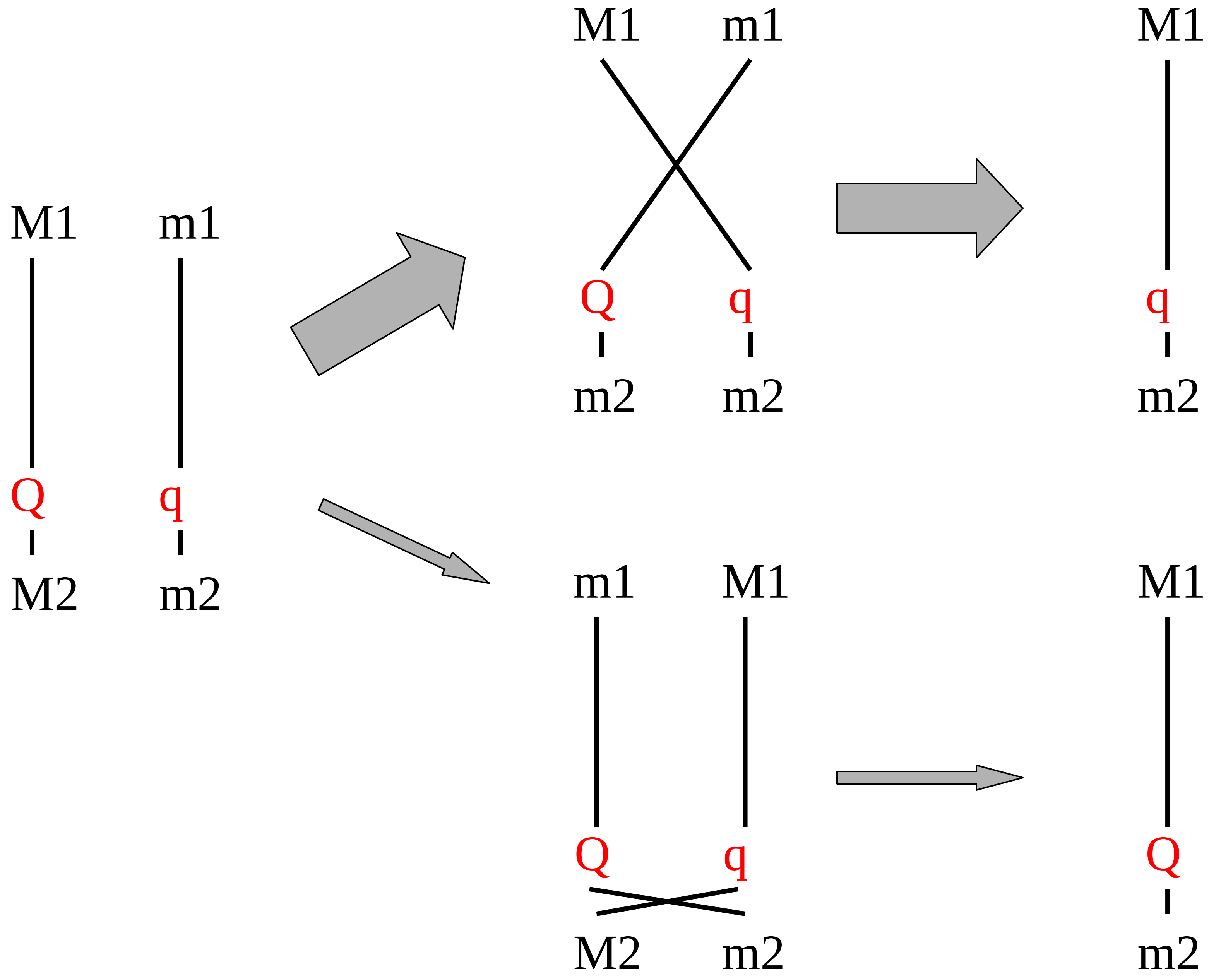
M1

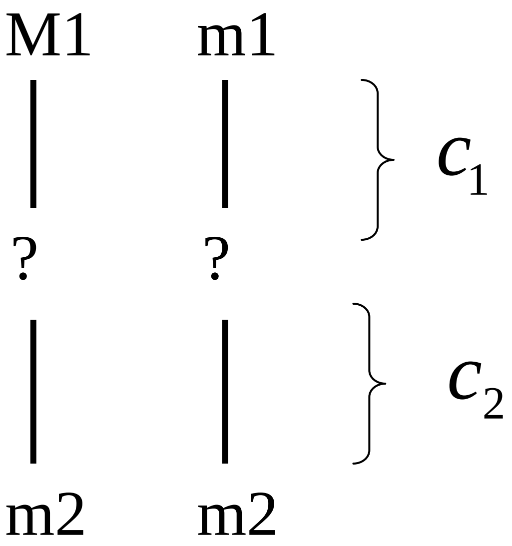
?

m2

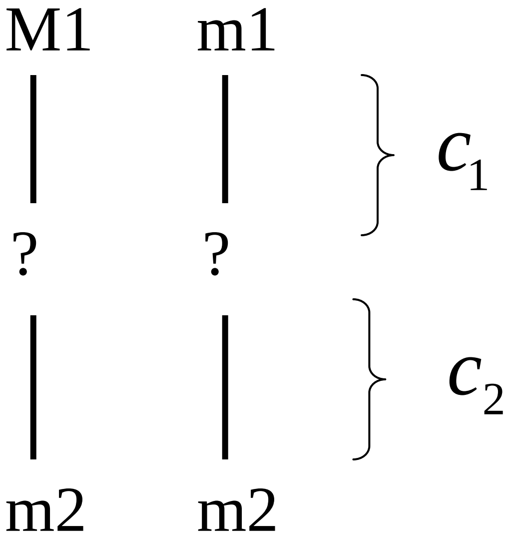

M1
m1
Q
q
M2
m2
m1
M1
Q
q
M2
m2
M1
q
m2


M1
m1
Q
q
M2
m2
m1
M1
Q
q
M2
m2
M1
q
m2
M1
m1
Q
q
m2
M2
M1
Q
m2


M1
m1
Q
q
M2
m2
M1
m1
Q
q
m2
m2
M1
q
m2
m1
M1
Q
q
M2
m2
M1
Q
m2



cM distances
determine
probabilities



cM distances
determine
probabilities

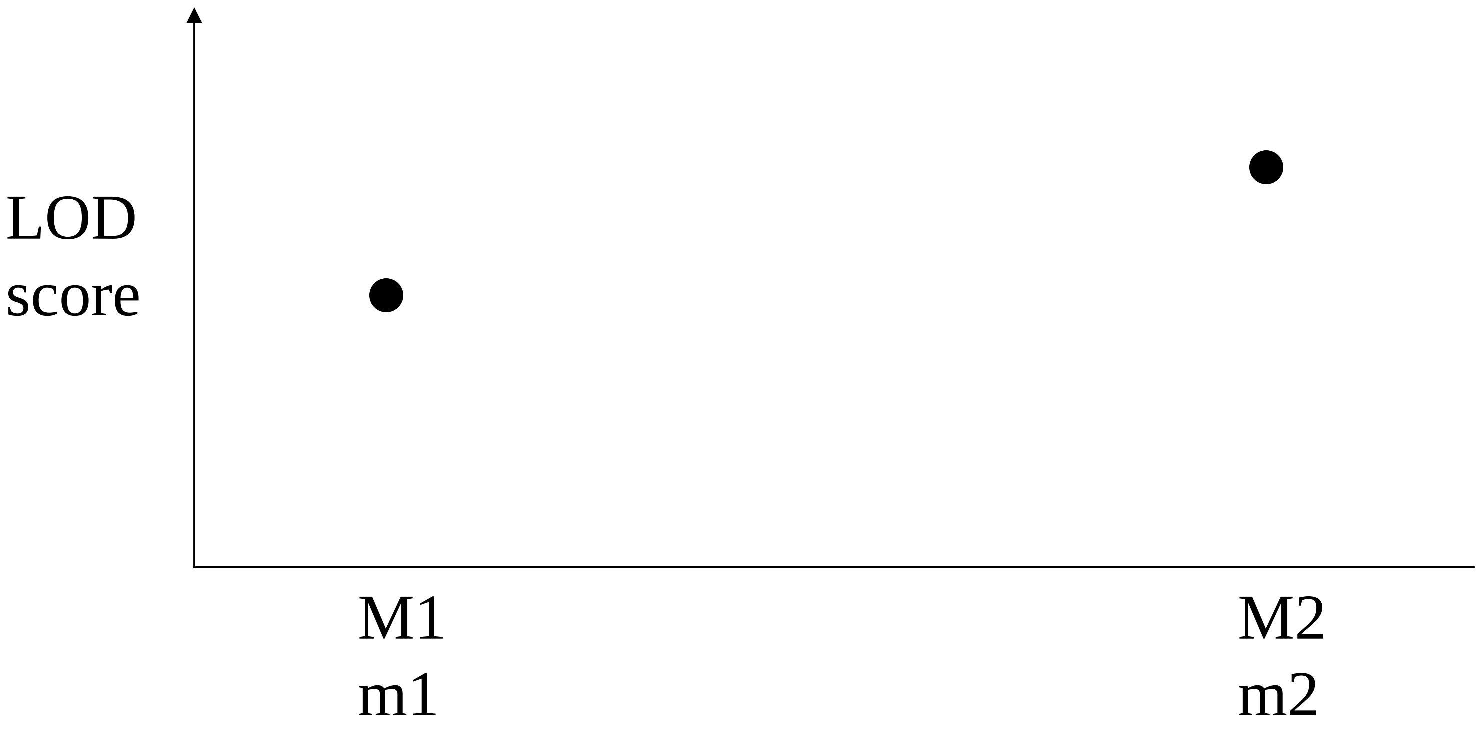
Eg,

$$\Pr\left(\,qq\ \mid M_1 m_2 m_1 m_2\right) = 0.5$$

$$\Pr\left(\,qQ\ \mid M_1 m_2 m_1 m_2\right) = 0.5$$

$$\Pr\left(\,QQ\ \mid M_1 m_2 m_1 m_2\right) = 0$$

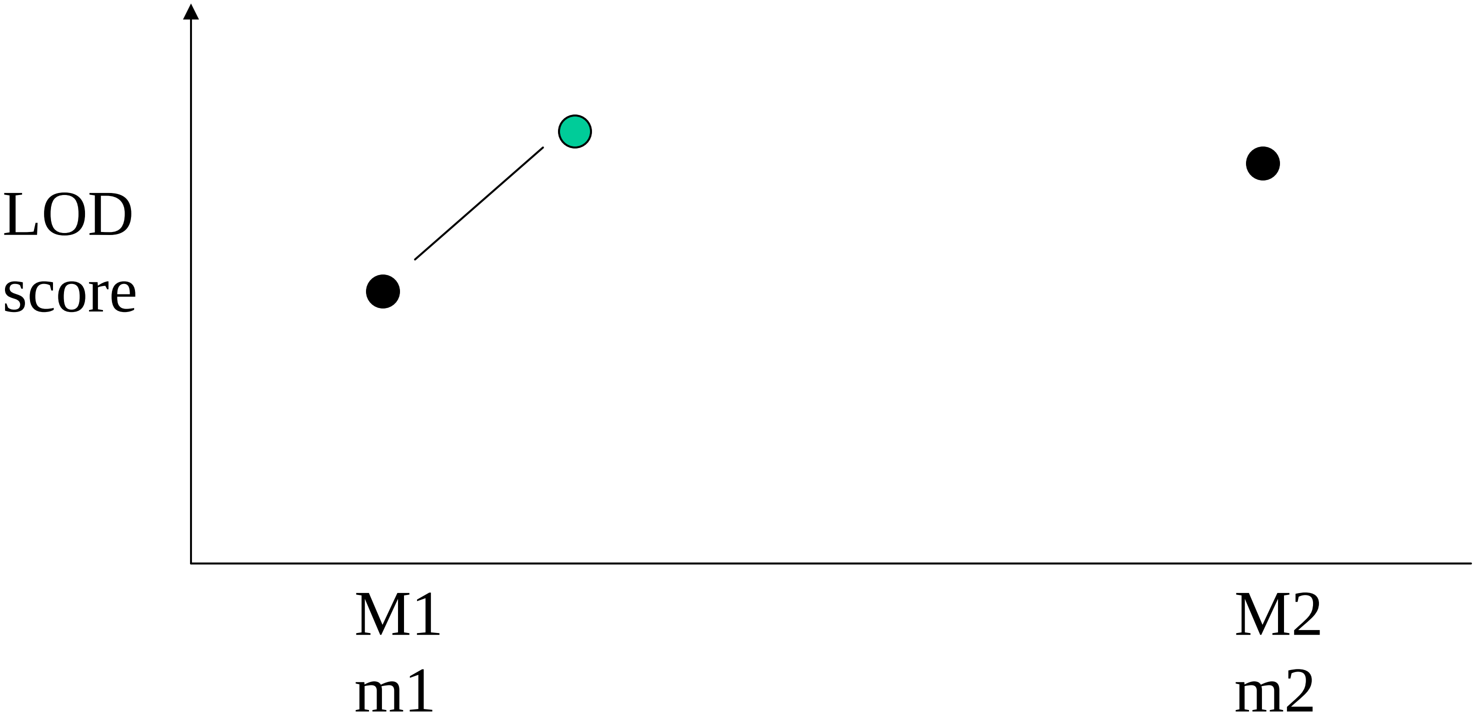
Interval mapping



Interval mapping

M1 – Q ---- M2

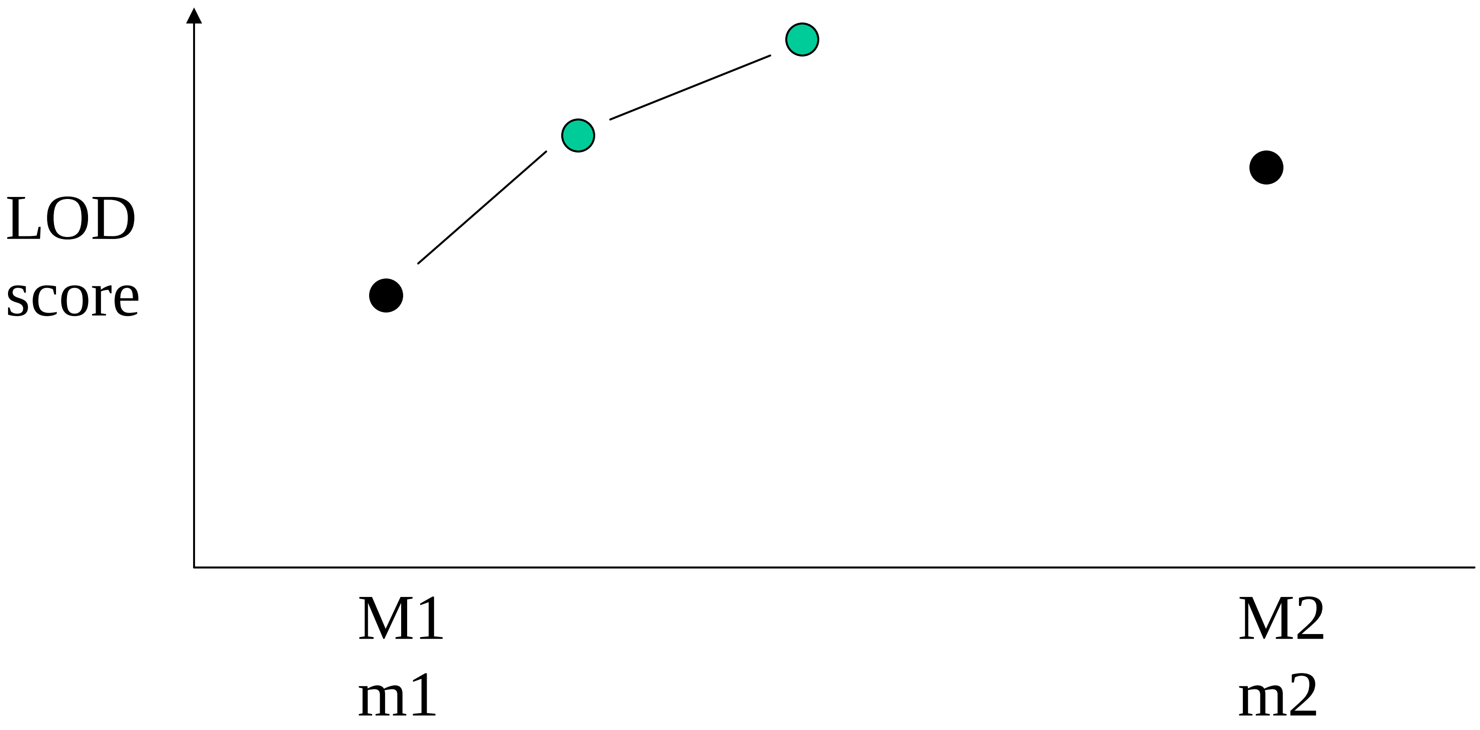
m1 – q ---- m2



Interval mapping

M1 —— Q —— M2

m1 —— q —— m2



Interval mapping