R/mpMap Workshop

Part 3: QTL Mapping

Emma Huang

TAMU, 3 Sep. 2015

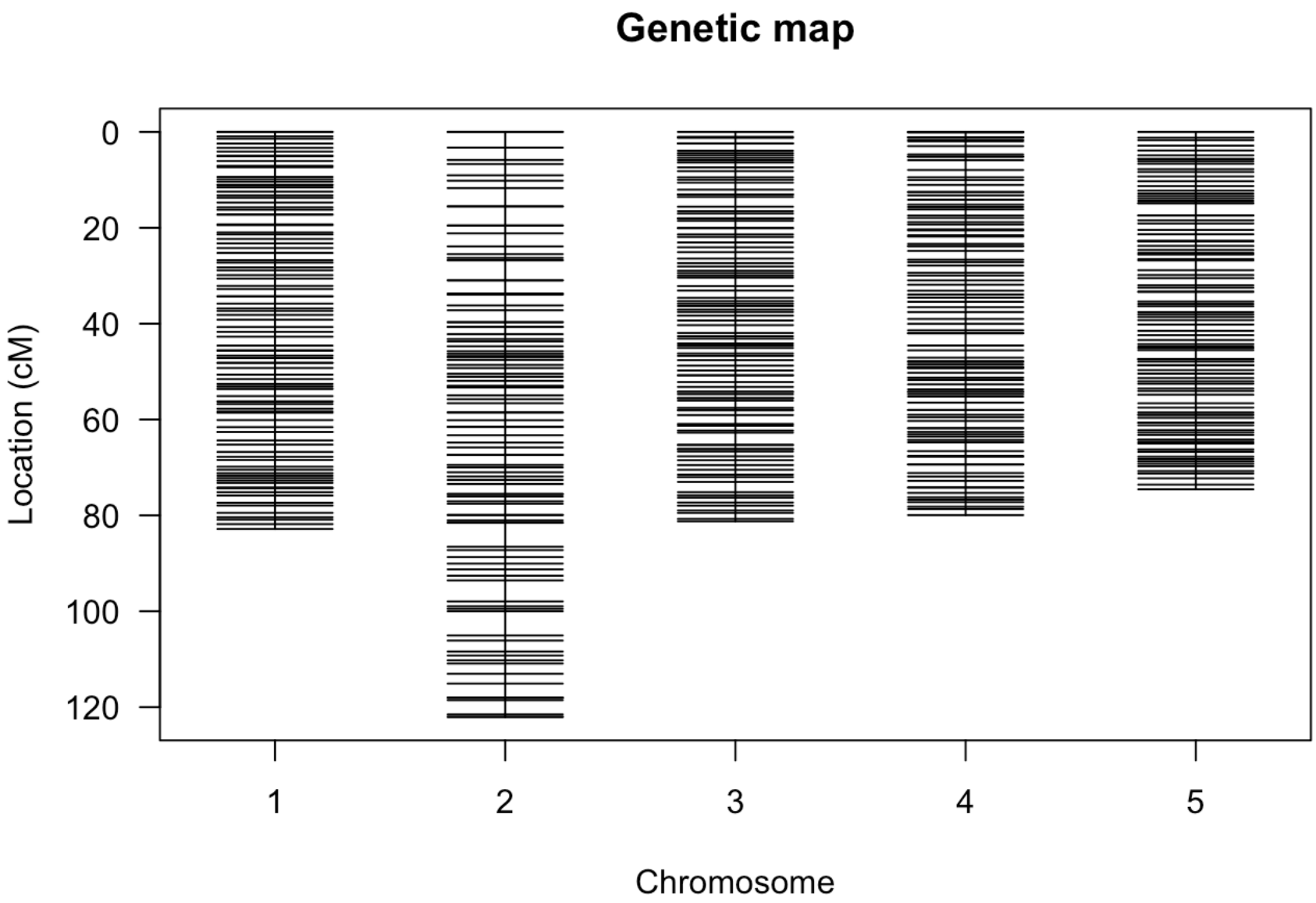
Plan

9:30-10:30

- Part 3: QTL Mapping (45 min)
    - Full Model
    - Association mapping
    - Meta-alleles
    - Mixed models
- Exercises (10 min)
- Break/Questions (5 min)

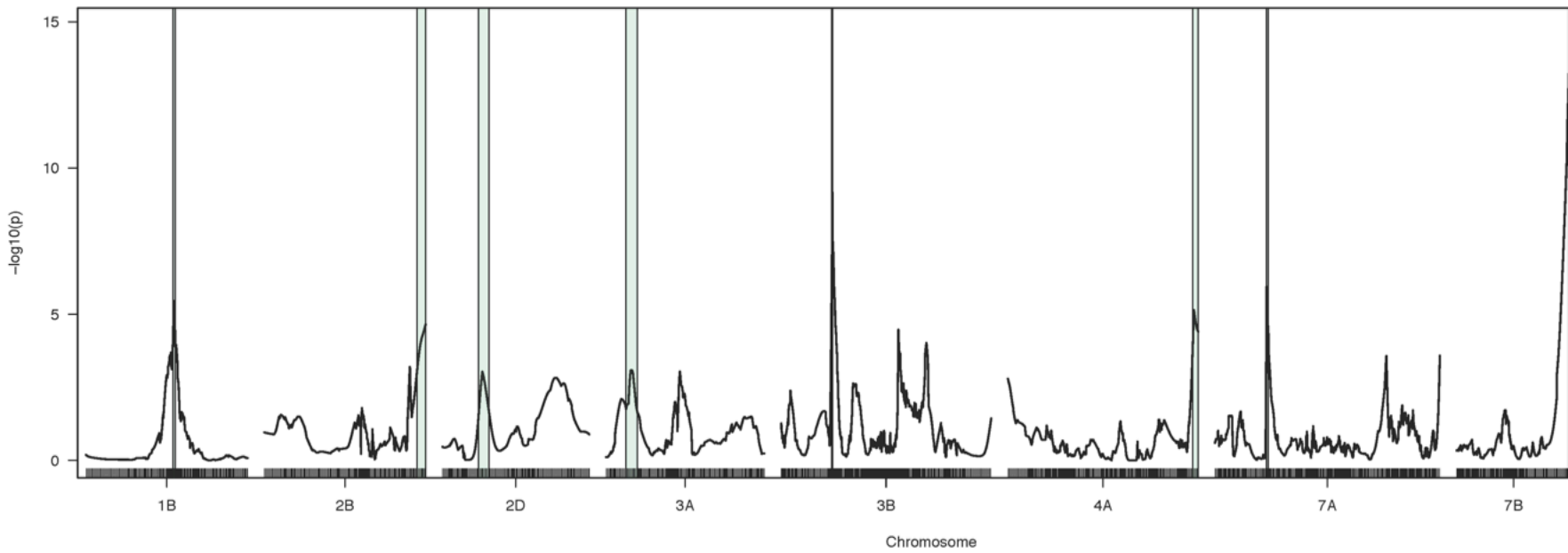
Starting Point

```
plot(datfinal$map)
```



Goal

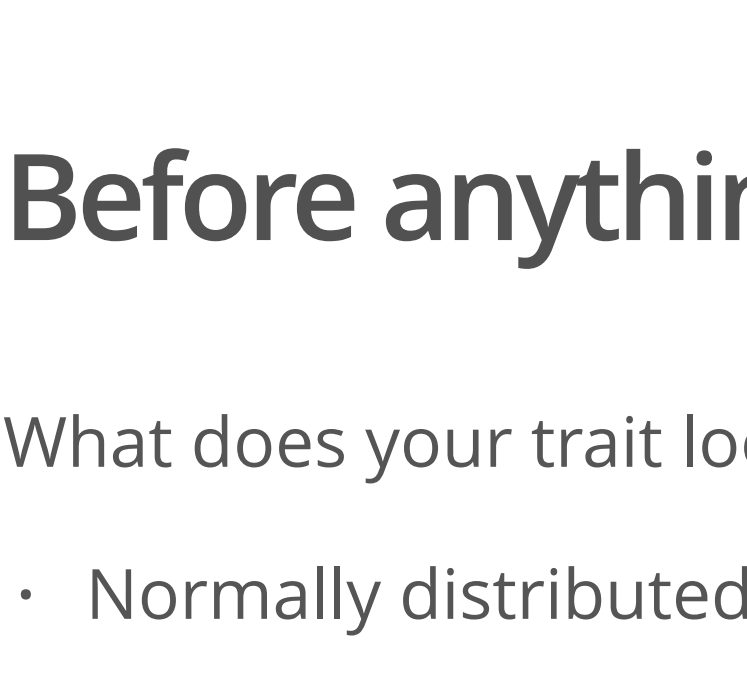
Associate phenotype with genotype - QTL mapping



Before anything else

What does your trait look like?

- Normally distributed?
- Transgressive segregation?
- Do you have more than one?
    - more than one field?
    - more than one year?



Beyond our scope:

- Epistasis
  - biallelic - EpiGPU (Hemani et al. 2011), GLIDE (Kam-Thong et al. 2012)
  - CrossTermsR package (in development) Josh.Bowden@csiro.au
- Binary traits
  - Not too difficult to modify for linear models
  - More complex for mixed models - Boden et al. 2015
- Multiple traits
  - In theory can fit with mpMap
  - May want to consider MPWGAIM (Verbyla et al. 2014)
  - or combinations of univariate analyses

Full Model (Linkage)

- e.g., HAPPY (Mott et al. 2000)
- Specify

$$y = \Sigma_{f=1}^{F} \beta_f X_f + \epsilon$$

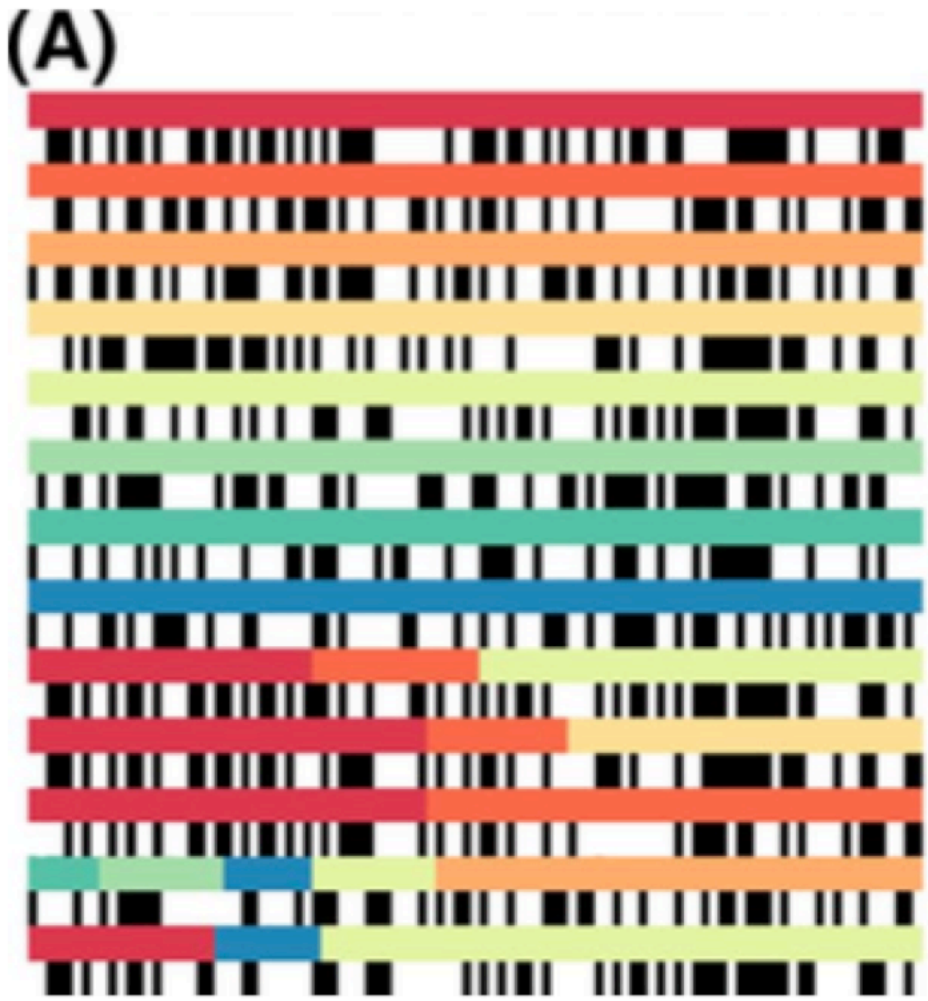
  where $\epsilon \sim N(0, \sigma^2)$ and f ranges over founders
- Advantages
  - Full flexibility
  - All founder effects estimated
- Drawbacks
  - Too many founder effects?
  - High df test

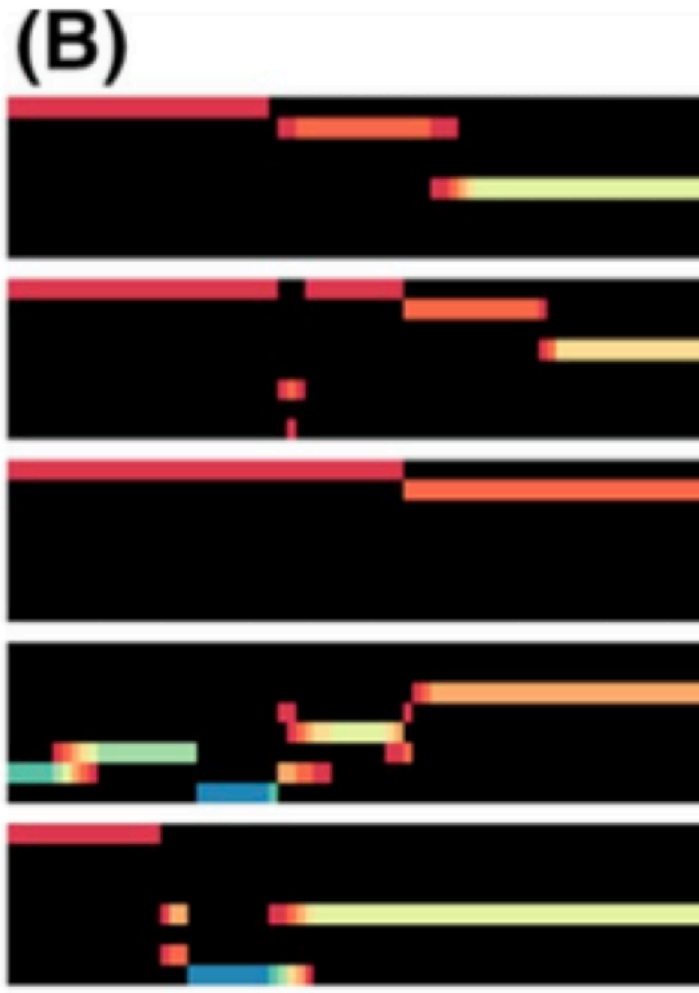
Digression: IBD Haplotypes

- Regions of the genome inherited as a chunk
- Full model is capturing effects of inheriting chunks from each parent
- Unobserved
    - Estimate which alleles inherited from different founders via Hidden Markov Model
    - Depends on distance between markers, genotypes of founders, genotypes of finals, population design

Haplotype Mosaics

How well can we reconstruct haplotypes?





Estimating haplotype mosaics

```
mpp <- mpprob(datfinal, program="qtl", step=2)
```

```
## [1] "No chromosomes specified, will default to all"
## Using map groupings for groups. Remove map object if you
##  --Read the following data:
##   1000  individuals
##   505  markers
##   2  phenotypes
```

Probabilities

```
round(mpp$prob[[1]][1:3, 1:8],2)
```

```
##    D1M1, Founder 1 D1M1, Founder 2 D1M1, Founder 3 D1M1,
## L1               0            0.99            0.01
## L2               1            0.00            0.00
## L3               1            0.00            0.00
##    D1M2, Founder 1 D1M2, Founder 2 D1M2, Founder 3 D1M2,
## L1               0            0.99            0.01
## L2               1            0.00            0.00
## L3               1            0.00            0.00
```

```
mpp$estfnd[[1]][1:3, 1:2]
```

```
##       D1M1 D1M2
## L1012    2    2
## L1018    1    1
## L1024    1    1
```

Biallelic Model (association)

- e.g., TASSEL (Bradbury et al. 2007)
- Utilized by Mackay et al., G3, 2014
- For each marker j,

$$y = \beta_j X_j + \epsilon$$

  where $\epsilon \sim N(0, \sigma^2)$
- Drawbacks
  - Reduced power when multiple founder effects
- Advantages
  - Computationally simpler and faster

Grouped Model (meta-alleles)

- e.g., ClustHaplo (Leroux et al., TAG, 2014)
- Presenting two alternate approaches at Eucarpia Biometrics in Plant Breeding on Sep. 11
- Drawbacks
  - Determining the best way to construct groups
  - Consistency with observed alleles; biologically meaningful
- Advantages
  - Intermediate to other models in both flexibility and computation

Arabidopsis MAGIC example - Days to Bolting

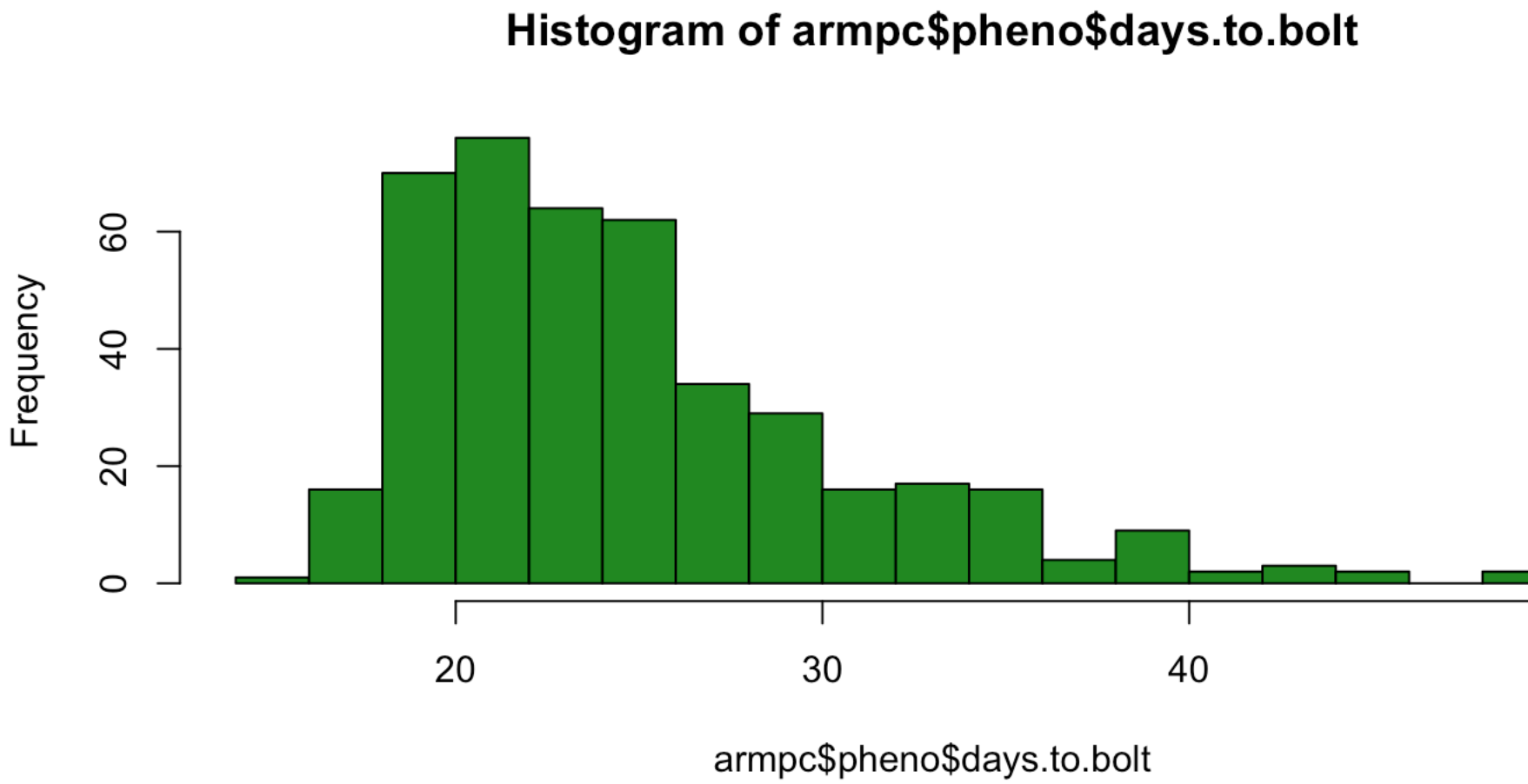
- Kover et al. 2009, 19 founders, 1254 SNPs

```
load('ArabidopsisExample.RData')
table(nall)
```

```
## nall
##   2   3   4   5   6   7   8
## 190 420 283 234 108  18   1
```

Distribution of days to bolting

```
hist(armpc$pheno$days.to.bolt, breaks=20, col="forestgreen")
```



QTL mapping Arabidopsis MAGIC

```
mpqfull <- mpIM(object=armpc, ncov=0, responsename="days.to.

mpqbi <- mpIM(object=armpc, ncov=0, responsename="days.to.bo
              foundergroups=armpc$founders)

mpqgr <- mpIM(object=armpc, ncov=0, responsename="days.to.bo
              foundergroups=gr)
```

Results of mapping

```
load('ResultsArabidopsisExample.RData')
dim(summary(mpqfull))
```

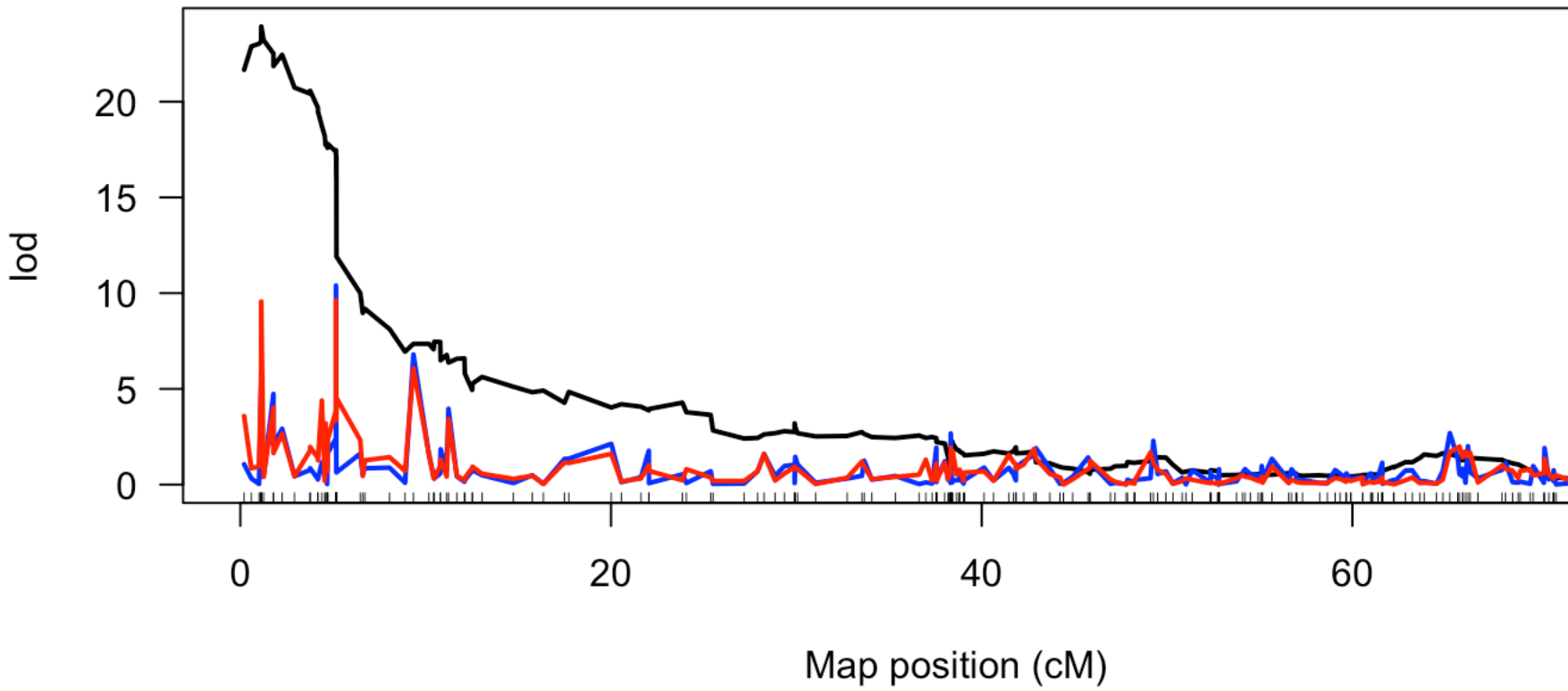
```
## [1] 43  7
```

- Refine results - look only at most significant

```
mpqfull2 <- findqtl(mpqfull, dwindow=5, threshold=6)
mpqbi2 <- findqtl(mpqbi, dwindow=5, threshold=6)
mpqgr2 <- findqtl(mpqgr, dwindow=5, threshold=6)
dim(summary(mpqfull2)) # findqtl is producing odd results, n
```

```
## [1] 16  7
```

Chr 4 QTL (FRI, GA)



Mixed models

- Can be important for including design information
  - variation across rows, columns in a field trial;
  - variation between experimenters in milling/baking;
  - variation over time
  - relationships between individuals
- Implementation in mpMap requires ASReml license
- mpIM(baseModel, pheno, idname, …)
  - one-stage
  - two-stage (no weights)

Composite interval mapping

mplM(…, ncov, …)

- Not particularly efficient
- Stepwise selection of covariates from total number of markers
- ncov is maximum number selected
- Can be very slow with many markers included

Simultaneous Model

- fit.mpqtl
  - Includes all QTL in the same model
  - Typically reduced significance once accounting for others
- qindex
  - Allows selection of certain QTL to fit
  - e.g., those still significant after accounting for others

Reducing number of QTL - either based on p-value or location

```
fit(mpqbi, qindex=which(fit(mpqbi)$table$pvalue<.1))$table
```

```
## Percent Phenotypic Variance explained by full model:  62.
## Percent Phenotypic Variance explained by full model:  22.
##         Chr    Pos      LeftMrk    RightMrk  Wald df   pva
## Chr13 Chr1  82.43 MN1_19782567   MASC03754 38.34 18 3.49e
## Chr14 Chr1 107.72  PERL0234167 PERL0235052 40.17 18 1.98e
## Chr52 Chr5   6.03  MN5_1446247 MN5_1586146 67.93 18 1.01e
```

Simultaneous vs. Individual

```
data.frame(fit(mpqbi2)$table[,-c(1,3)], Ind.pv=summary(mpqbi
```

```
## Percent Phenotypic Variance explained by full model:  36.
```

```
##          Pos    RightMrk  Wald df   pvalue PctVar   Ind.pv
## Chr41  5.16    GA1_7845 76.99 18 2.86e-09  20.08 3.81e-03
## Chr42  9.34 MN4_2241604 13.95 18 7.32e-01  10.91 1.60e-07
## Chr51 17.29 MN5_4179168 30.73 18 3.10e-02  13.33 4.47e-07
## Chr52 17.99 MN5_4318001 13.52 18 7.60e-01  13.32 4.80e-10
## Chr53 21.55     CO 1457 22.73 18 2.01e-01   8.75 9.51e-08
```

- Percent phenotypic variance explained computed from adjusted R^2

Support interval

- Calculates LOD-drop from peak of QTL

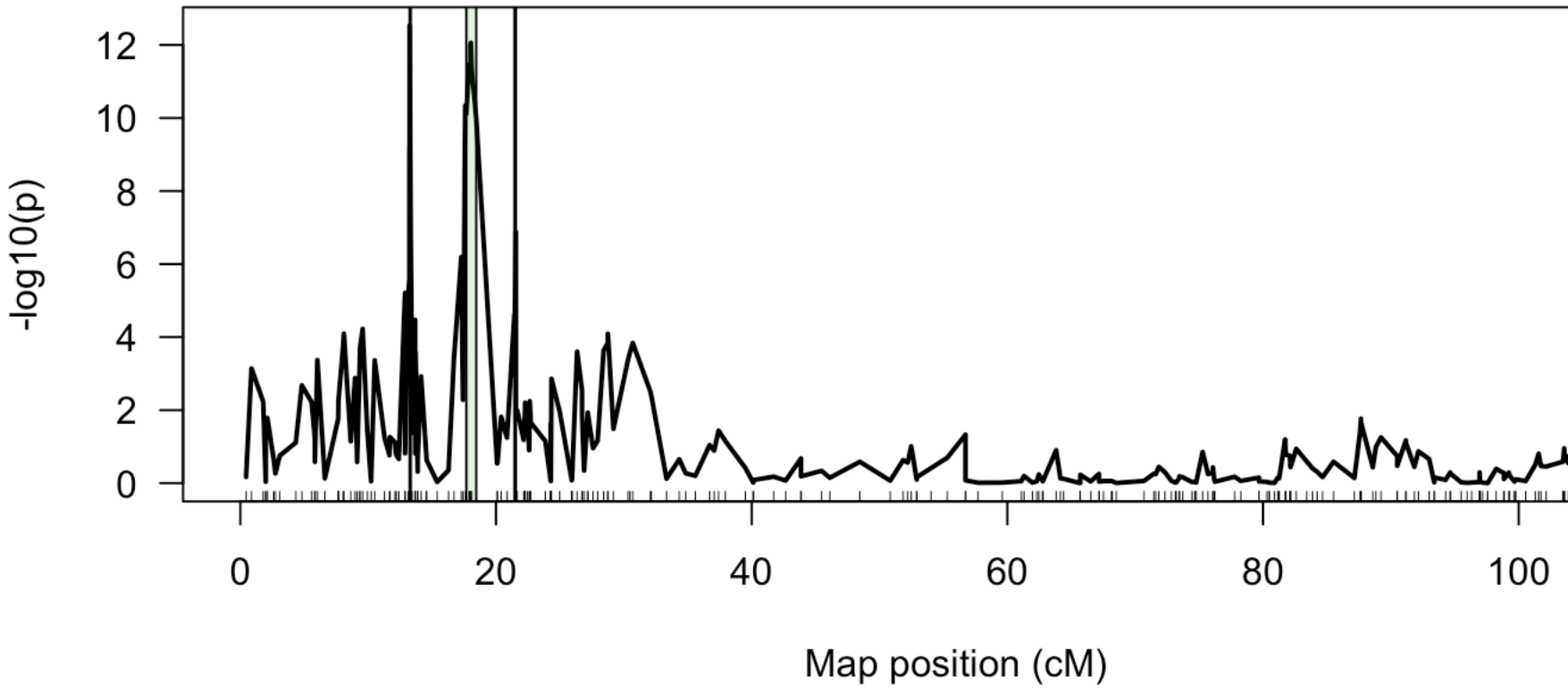
```
supportinterval(mpqgr2, lodsupport=2)$support
```

```
##              Chr4     Chr4       Chr4     Chr5     Chr5     C
## Lower 1.126700 4.703004  8.887708 13.23797 17.41320 21.43
## Upper 1.255542 6.605004 10.171371 13.33546 18.45359 21.55
```

- By default, included in plot of QTL profile

Plot of QTL profile

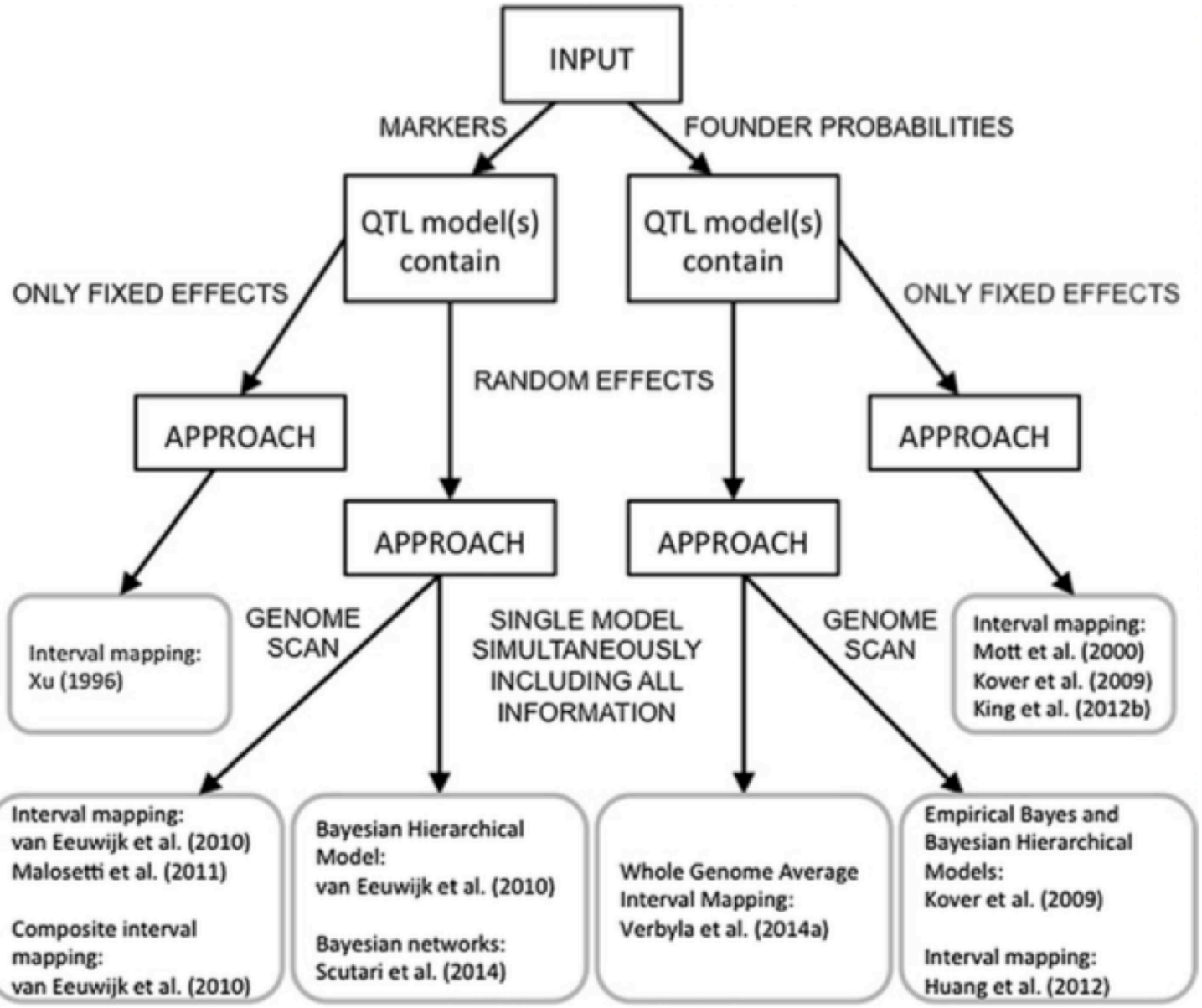
```
plot(mpqgr2, chr=5)
```



Significance thresholds

- Multiple testing correction
  - Bonferroni - divide by number of tests (conservative)
  - Empirical simulations - generate data under null
  - Permutations - break up associations
- Balancing time/computation vs. power

Other approaches



References (1/2)

Boden et al. 2015, Ppd-1 is a key regulator of inflorescence architecture and paired spikelet development in wheat. Nature Plants doi:10.1038/nplants.2014.16

Bradbury et al. 2007, TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635. doi: 10.1093/bioinformatics/btm308

Hemani et al. 2011, EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 27:1462-1465. doi: 10.1093/bioinformatics/btr172

Kam-Thong et al. 2012, GLIDE: GPU-based linear regression for detection of epistasis. Heredity 73:220-236.

References (2/2)

Kover et al. 2009, A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS Genetics doi: 10.1371/journal.pgen.1000551

Leroux et al. 2014, Clusthaplo: a plug-in for MCQTL to enhance QTL detection using ancestral alleles in multi-cross designs. TAG 127:921-33. doi: 10.1007/s00122-014-2267-1

Mackay et al. 2014, An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. G3 4: 1603-1610.

Mott et al. 2000, A new method for fine-mapping quantitative trait loci in outbred animal stocks. PNAS 97:12649-12654.

Verbyla et al. 2014, Whole-genome analysis of multienvironment or multitrait QTL in MAGIC. G3 4:1569-1584.

Exercises

Data sim3

- Plot haplotype mosaics for lines with 20 most recombinations
- Map QTL using
  - the full model
  - a biallelic model
  - composite interval mapping with 5 covariates
- How many do you detect?
- Which ones do you detect better with one model than another?
- Fit a model with all QTL in it simultaneously
  - How does this affect their significance?

Exercises - Data generation

```
map <- sim.map(len=rep(100, 5), n.mar=51, eq.spacing=T, incl
ped <- sim.mpped(4, 1, 500)
sim3 <- sim.mpcross(map=map, pedigree=ped, qtl=
                    rbind(c(1, 21, 0, .4, 0, 0),
                          c(1, 71, 0, 0, -.3, 0),
                          c(4, 35, .3, 0, .3, 0),
                          c(5, 59, 0, .4, 0, -.4)))
save(sim3, file="sim3.RData")
```

Exercises - haplotype mosaics

```
mpp <- mpprob(sim3, program="qtl")
plot(mpp, nlines=20)
```

Exercises - QTL mapping

```
mpq1 <- mpIM(object=mpp, responsename="pheno", ncov=0, dwind
mpq2 <- mpIM(object=mpp, responsename="pheno", ncov=0,
             foundergroups=mpp$founders, dwindow=50)
mpq3 <- mpIM(object=mpp, responsename="pheno", ncov=5, dwind
```