

QTL mapping exercise using r/qtl

BIO373 at the University of Zurich

Yasuhiro Sato (lecturer)¹ Katharina Jung (TA)
Naoto-Benjamin Hamaya (TA)

28 Sept. 2023

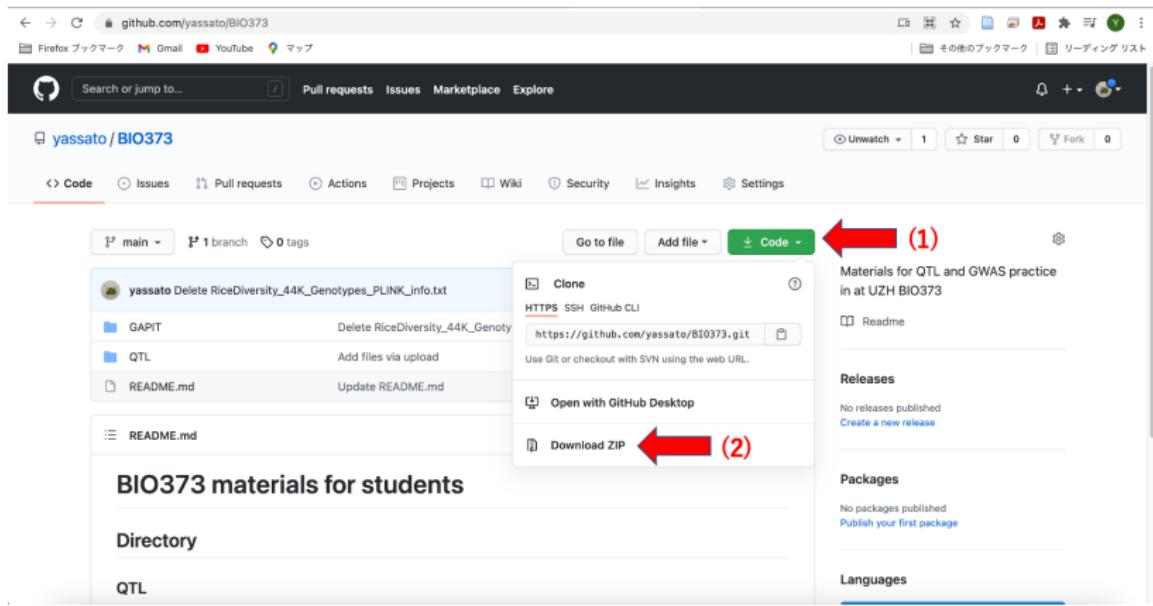
¹yasuhiro.sato@uzh.ch

What to do in this exercise

- ▶ QTL mapping of flowering time in *Arabidopsis thaliana*
- 1. Col x Kas Recombinant inbred lines (RILs) aim **to see how interval mapping works**
- 2. Multiparent Advanced Generation Intercross (MAGIC) lines aim **to see how fine mapping resolves the interval**
- ▶ This instruction PDF and input files are available at https://github.com/yassato/BIO373_YS2023/tree/main/QTL

Download materials

- ▶ Download .zip from
https://github.com/yassato/BIO373_YS2023 and unzip it
(Of course, “git clone URL” works if you are good at Git)



Preparation

Note: No support will be provided for your local environment (e.g., laptop)

1. Access to RStudio server (<https://fgcz-genomics.uzh.ch>) and log-in with your B-fabric username and Password
2. Make and change your working directory with `mkdir QTL` from Terminal; and then `setwd("./QTL")` from R Console
3. Upload input data ("ColKasFloweringPheno.csv" and "ColKasFloweringGeno.csv") to the directory you made

The screenshot shows the RStudio Server interface. The browser tab is 'RStudio Server' at 'fgcz-genomics.uzh.ch'. The R console window shows R code for correlation analysis. The 'Terminal' tab is selected. A red arrow points from the 'Terminal' tab to the 'Upload' button in the 'File' menu. Another red arrow points from the 'Upload' button in the 'File' menu to the 'Upload' button in the file tree sidebar. The sidebar shows a directory structure under 'QTL' with several files listed.

```
R 4.1.2 - (QTL)
5 090414-A09 5 1 5 1.8817708 18.49218 89.37386 91.25963
6 090105-A02 7 1 6 8.8639442 20.66431 89.37386 98.23788
> pred <- pred[order(pred$taxa),]
> y <- pOrder(p$taxID,]
> cor.test(pred$Prediction, y$Year06Flowering.time.at.Aransas)

Pearson's product-moment correlation

data: pred$Prediction and y$Year06Flowering.time.at.Aransas
t = 49.431, df = 335, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9234637 0.9494873
sample estimates:
  cor
0.9377791

>
> res <- lm(y$Year07Flowering.time.at.Aransas~pred$Prediction)
> plot(pred$Prediction,
+       y$Year07Flowering.time.at.Aransas,
+       ylab="flowering 2007", xlab="predicted",
+       main=paste("r =",round(sqrt(summary(res)$r.squared),2)))
> abline(res)
```

Name	Size	Modified
g_t_functions.txt	636.2 KB	Jun 8, 2022, 9:28 AM
RiceDiversity_44K_Genotypes_PLINK_info.txt.gz	250 KB	Jun 8, 2022, 9:28 AM
RiceDiversity_44K_Genotypes_PLINK_imputed.txt.gz	2 MB	Jun 8, 2022, 9:28 AM
CAPIT.Kin.VanRaden.pdf	783.9 KB	Jun 8, 2022, 9:35 AM
CAPIT.Kin.VanRaden.csv	3 MB	Jun 8, 2022, 9:35 AM
CAPIT.Heterozygosity.pdf	4.8 KB	Jun 8, 2022, 9:35 AM
CAPIT.Marker.Density.pdf	67.3 KB	Jun 8, 2022, 9:35 AM
CAPIT.Marker.LD.pdf	288.8 KB	Jun 8, 2022, 9:35 AM
CAPIT.GLM.Seed.length.phenotype_view.pdf	49.4 KB	Jun 8, 2022, 9:31 AM
CAPIT.GLM.Seed.length.PRED.csv	5 B	Jun 8, 2022, 9:32 AM
CAPIT.GLM.Seed.length.RDF.csv	516 B	Jun 8, 2022, 9:32 AM

1. Interval mapping using Col x Kas RILs

- ▶ Columbia (Col; early flowering) and Kashmir (Kas; late flowering) are crossed and selfed for several generations
- ▶ Days-to-bolting of 96 RILs were recorded under simulated Sweden or Spain climates (Li et al. 2006 below)



Figure 1. Side by side walk in Growth Chambers running Sweden (left) and Spain (right) simulated seasonal conditions. Days until flowering was recorded for 1123 plants in total seen growing in two blocks on the right and left side of each chamber.
doi:10.1371/journal.pone.0000105.g001

r/qtl provides useful tools for standard QTL mapping

```
# clean up your workplace
rm(list=ls())

# install r/qtl and its advanced package, r/qtl2
install.packages("qtl",repos="http://cran.us.r-project.org") # install the qtl package

## 
## The downloaded binary packages are in
## /var/folders/b1/3ml2xmv1381dfsh21rgw32sm0000gn/T//RtmpY9lt56 downloaded_packages

install.packages("qtl2",repos="http://cran.us.r-project.org") # install the qtl2 package

## 
## The downloaded binary packages are in
## /var/folders/b1/3ml2xmv1381dfsh21rgw32sm0000gn/T//RtmpY9lt56 downloaded_packages

# load r/qtl package
library(qtl) # load the qtl package
```

Load genotypes and phenotypes as a “cross” object

- ▶ Genotype and phenotype data from Li et al. (2006)
- ▶ Missing genotypes are specified by “na.strings”
- ▶ Import recombinant inbred lines (RILs) as “riself”

```
# load genotypes and phenotypes as a "cross" object
colkas <- read.cross(format="csvs", dir="./",
                      genfile="ColKasFloweringGeno.csv", # geno data
                      phefile = "ColKasFloweringPheno.csv", # pheno data
                      na.strings = c("-"), estimate.map=FALSE,
                      crosstype = "riself")

## --Read the following data:
##   96 individuals
##   119 markers
##   3 phenotypes
## --Cross type: riself
```

See summary of the cross object

- ▶ We can check the number of markers and individuals
- ▶ RILs are selfed and thus have AA or BB

```
## (colkas) # see summary of the cross object
```

```
##      RI strains via selfing
## 
##      No. individuals:     96
## 
##      No. phenotypes:      3
##      Percent phenotyped: 100 100 100
## 
##      No. chromosomes:    5
##      Autosomes:          1 2 3 4 5
## 
##      Total markers:       119
##      No. markers:         24 26 18 22 29
##      Percent genotyped:  90.2
##      Genotypes (%):      AA:46.5  BB:53.5
```

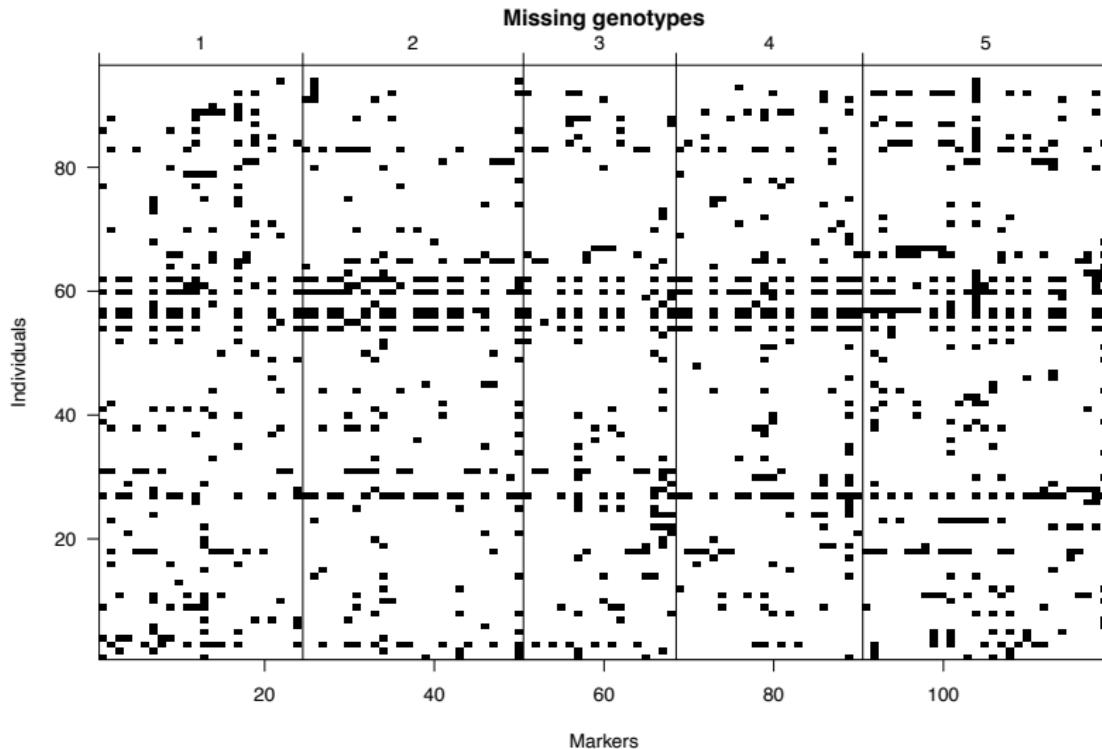
```
totmar(colkas) # total no. of genetic markers
```

```
## [1] 119
```

Check missing genotypes

- ▶ Missing genotype will be imputed by genotype probabilities

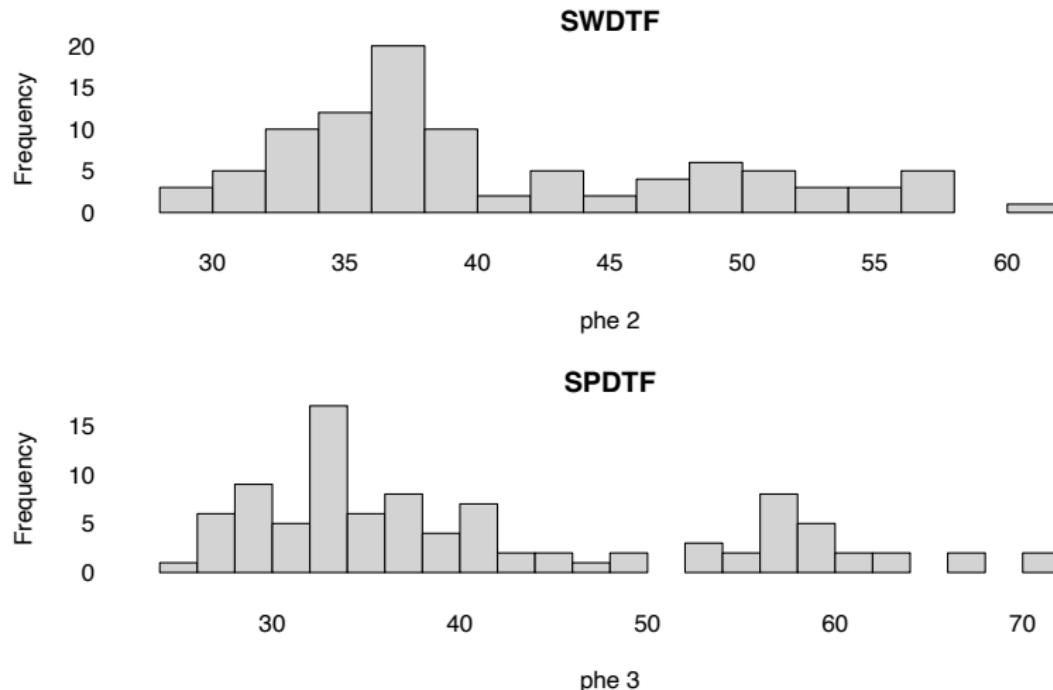
```
plotMissing(colkas) # check missing genotypes
```



Check phenotypic variation in flowering time

- Bi-modal but quantitative variation among Col x Kas RILs

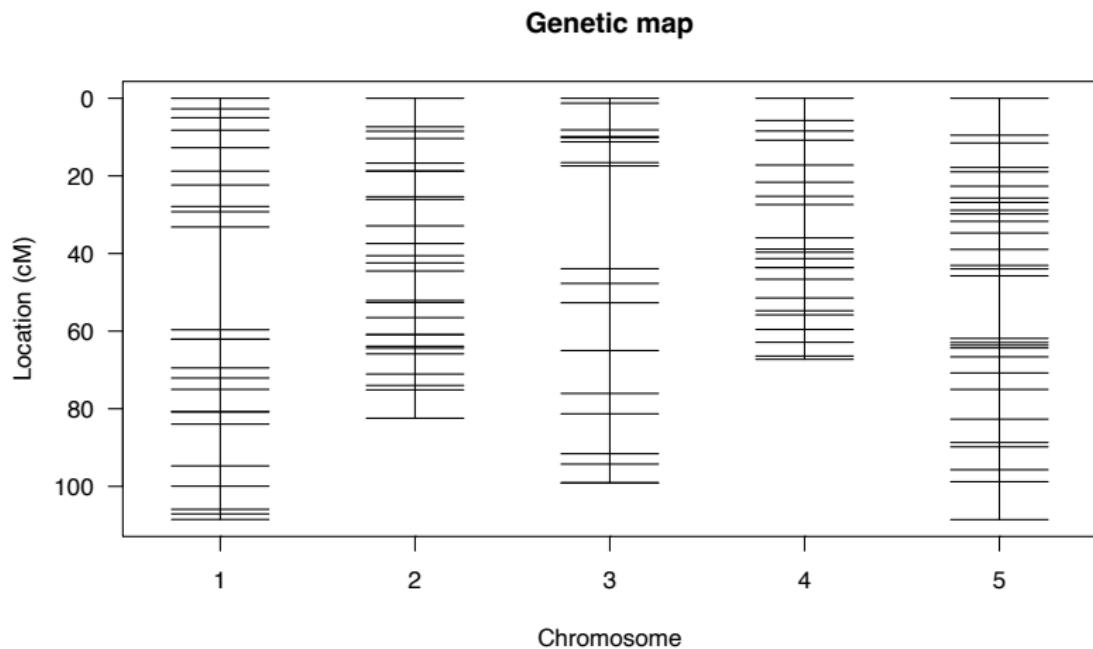
```
##(mfcol=c(2,1)); ##(mai=c(1,1,0.25,0.25)) # change the plot parameters  
plotPheno(colkas, pheno.col=2) # plot the results of Sweden days-to-bolting (SWDTF)  
plotPheno(colkas, pheno.col=3) # plot the results of Spain days-to-bolting (SPDTF)
```



Estimate and visualize a linkage map

- Distance in centimorgan (cM) is based on recombination rate

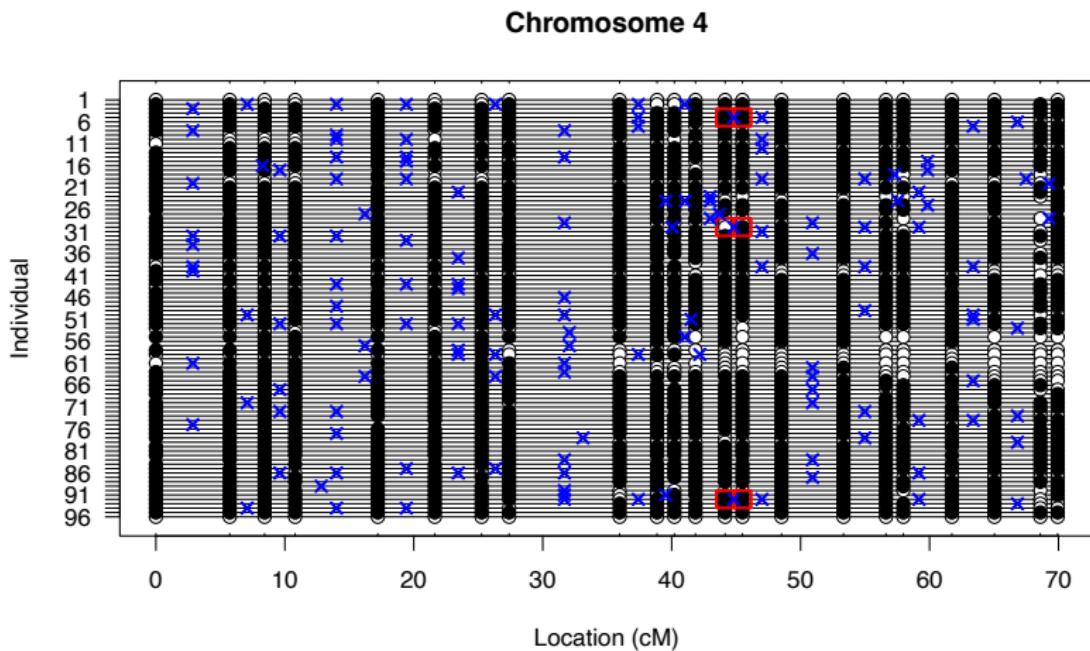
```
newmap <- est.map(colkas, error.prob=0.01)
colkas <- replace.map(colkas, newmap)
plotMap(colkas)
```



Display genome positions where recombination happened

- ▶ Cross marks indicate recombination points
- ▶ Plot on your own R to zoom in

```
colkas <- calc.errorlod(colkas, error.prob=0.01)
plotGeno(colkas, chr=4)
```



Calculate genotype probabilities

- A part of probabilities to observe AA is shown

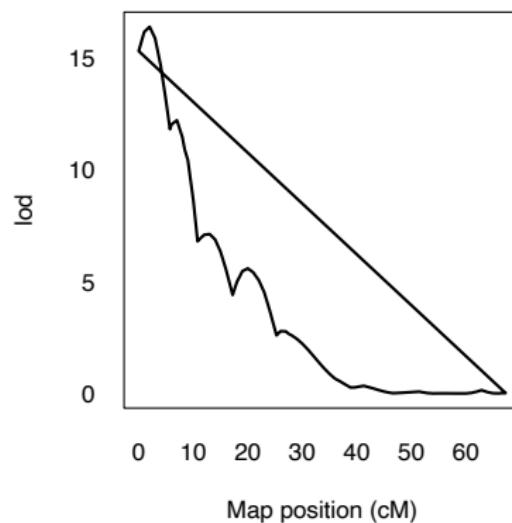
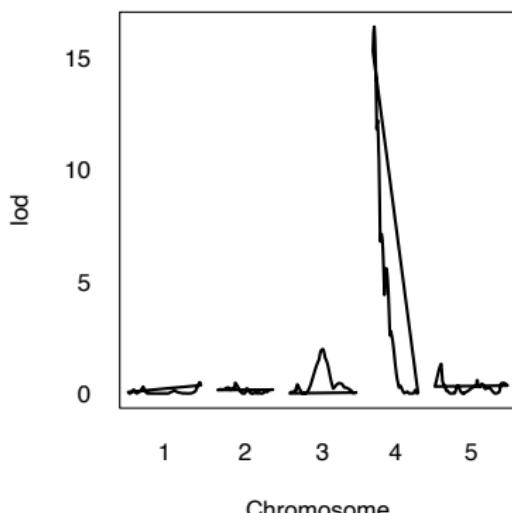
```
colkas_genoprob <- calc.genoprob(colkas, step=1)
colkas_genoprob$geno$"4"$prob
```

```
## , , AA
##
##      chr4.89659      loc1      loc2      loc3      loc4      loc5
## [1,] 9.999887e-01 0.998145398 0.997083676 0.996801870 0.997299540 0.998577465
## [2,] 1.148028e-05 0.001891173 0.002989341 0.003307706 0.002846768 0.001605804
## [3,] 9.991154e-01 0.824293513 0.649980446 0.475902697 0.301787139 0.127360588
## [4,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [5,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [6,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [7,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [8,] 8.846085e-04 0.175706487 0.350019554 0.524097303 0.698212861 0.872639412
## [9,] 9.999887e-01 0.998145398 0.997083676 0.996801870 0.997299540 0.998577465
## [10,] 9.999887e-01 0.998145398 0.997083676 0.996801870 0.997299540 0.998577465
## [11,] 9.999887e-01 0.998145398 0.997083676 0.996801870 0.997299540 0.998577465
## [12,] 1.014916e-01 0.085390792 0.068639469 0.051211348 0.033079083 0.014214227
## [13,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [14,] 9.999887e-01 0.998145398 0.997083676 0.996801870 0.997299540 0.998577465
## [15,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [16,] 9.999886e-01 0.998125899 0.997044745 0.996743446 0.997221531 0.998479749
## [17,] 1.129659e-05 0.001854606 0.002916332 0.003198142 0.002700477 0.001422556
## [18,] 3.285399e-05 0.006146079 0.011484455 0.016056356 0.019868958 0.022928241
## [19,] 9.999887e-01 0.998145398 0.997083676 0.996801870 0.997299540 0.998577465
## [20,] 8.846085e-04 0.175706487 0.350019554 0.524097303 0.698212861 0.872639412
## [21,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [22,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [23,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [24,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [25,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
## [26,] 1.129657e-05 0.001854602 0.002916324 0.003198130 0.002700460 0.001422535
```

QTL mapping for flowering time under Sweden climates

- ▶ Vertical whiskers indicate the position of observed markers
- ▶ QTL at an interval between the 1st and 2nd marker on chr. 4

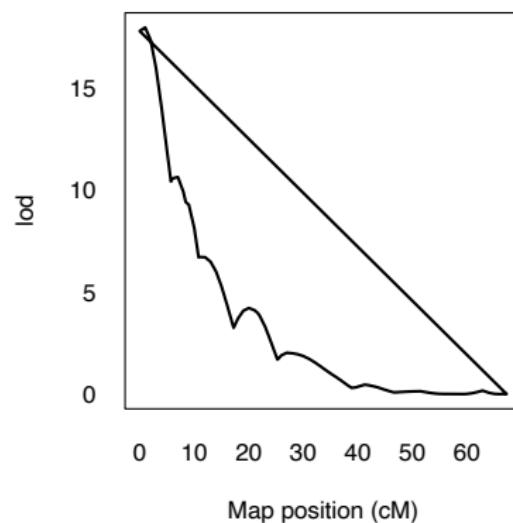
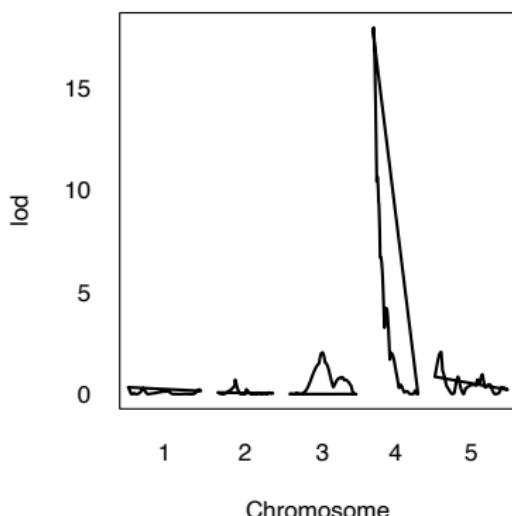
```
scanSWDTF <- scanone(colkas_genoprob, pheno.col=colkas$pheno$SWDTF,  
method="hk") # hk = Haley-Knott regression  
##(mfcol=c(1,2)); ##(mai=c(1,1,0.25,0.25)) # change the plot parameters  
##(scanSWDTF); ##(scanSWDTF, chr=4) # plot the results
```



QTL mapping for flowering time under Spain climates

- ▶ QTL at an interval between the 1st and 2nd marker on chr. 4,
- ▶ suggesting a stable QTL between the two environments.

```
scanSPDTF <- scanone(colkas_genoprob, pheno.col=colkas$pheno$SPDTF,  
                      method="hk") # hk = Haley-Knott regression  
## (mfcol=c(1,2)); ## (mai=c(1,1,0.25,0.25)) # change the plot parameters  
## (scanSPDTF); ## (scanSPDTF, chr=4) # plot the results
```



2. Fine QTL mapping using multiparental intercross lines

- ▶ Then let us use multiparental lines to see within the interval
- ▶ Kover et al. (2009) made MAGIC lines from 19 accessions
- ▶ Days-to-bolting of 703 lines were reported by Gnan et al. (2014)

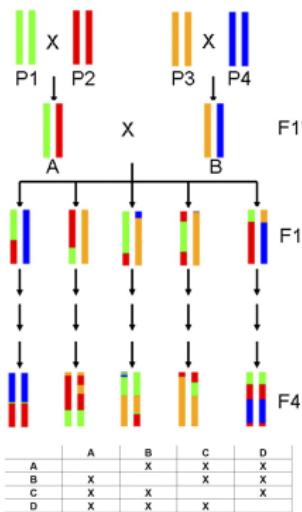


Figure 1: Image of multiparental intercross from Huang et al. 2011

Advanced package, r/qtl2, can handle multiparental lines

- ▶ The r/qtl2 package reads .zip set of input files via online
- ▶ The data include 8 phenotypes

```
[1] > library(qtl2) # load the qtl2 package
```

```
##  
## Attaching package: 'qtl2'  
  
## The following object is masked from 'package:qtl':  
##  
##     clean  
  
# download dataset from the r/qtl2 website  
file <- paste0("https://raw.githubusercontent.com/rqtl/",  
               "qtl2data/master/ArabMAGIC/arabmagic_tair9.zip")  
magic <- read_cross2(file)  
head(magic$pheno) # see phenotypes
```

```
##          bolting_days seed_weight seed_area ttl_seedsfruit branches height  
## MAGIC.1        15.33      17.15      0.64        45.11    10.50    NA  
## MAGIC.2        22.00      22.71      0.75        49.11     4.33   42.33  
## MAGIC.3        23.00      21.03      0.68        57.00     4.67   50.00  
## MAGIC.4        18.67      22.45      0.74        54.33     6.33    NA  
## MAGIC.5        18.67      25.36      0.82        38.33     5.67   42.25  
## MAGIC.6        25.00      21.53      0.71        52.00     4.33    NA  
##          pc_seeds_aborted fruit_length  
## MAGIC.1            0.00       14.95  
## MAGIC.2            1.09       13.27  
## MAGIC.3            0.00       13.90  
## MAGIC.4            0.23       15.93  
## MAGIC.5            0.00       12.81  
## MAGIC.6            0.00       12.27
```

See summary and a part of marker information

- ▶ Check the number of markers and individuals
- ▶ Markers indeed exist on the top of chr 4

```
XXXXXX(magic) # see summary of the cross2 object
```

```
## Object of class cross2 (crosstype "magic19")
##
## Total individuals          703
## No. genotyped individuals 703
## No. phenotyped individuals 677
## No. with both geno & pheno 677
##
## No. phenotypes             8
## No. covariates            0
## No. phenotype covariates  0
##
## No. chromosomes            5
## Total markers              1251
##
## No. markers by chr:
##   1   2   3   4   5
## 274 211 248 228 290
```

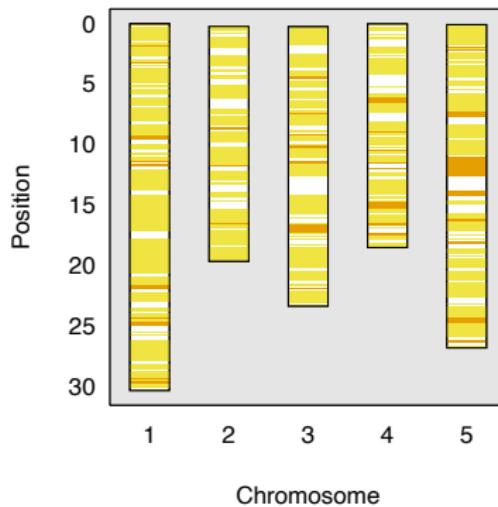
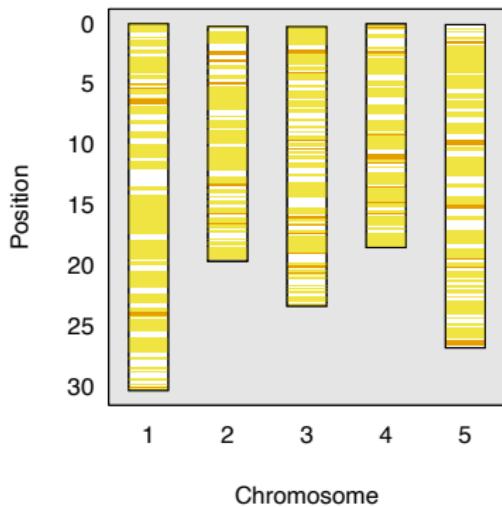
```
magic$gmap$"4"[1:50] # markers at the top of chr. 4
```

	MN4_488120K	MN4_142943	MN4_241821	FRI_725	FRI_927
##	0.048813	0.142943	0.241821	0.269260	0.269462
##	NMSNP4_270272	FRI_1888	FRI_2343	MASC04123	MN4_428535
##	0.270272	0.270407	0.270862	0.301329	0.428535
##	MN4_428737	MN4_541323	MASC02680	MASC04651	MASC02790
##	0.428737	0.541323	0.701440	0.903543	0.903867
##	MASC02819	MASC05116	MN4_1055425	MASC04725	MASC04658

Visualize fine recombination for each individual

- ▶ Colors indicate different parental alleles at each locus
- ▶ Use “ind=” option to change individuals

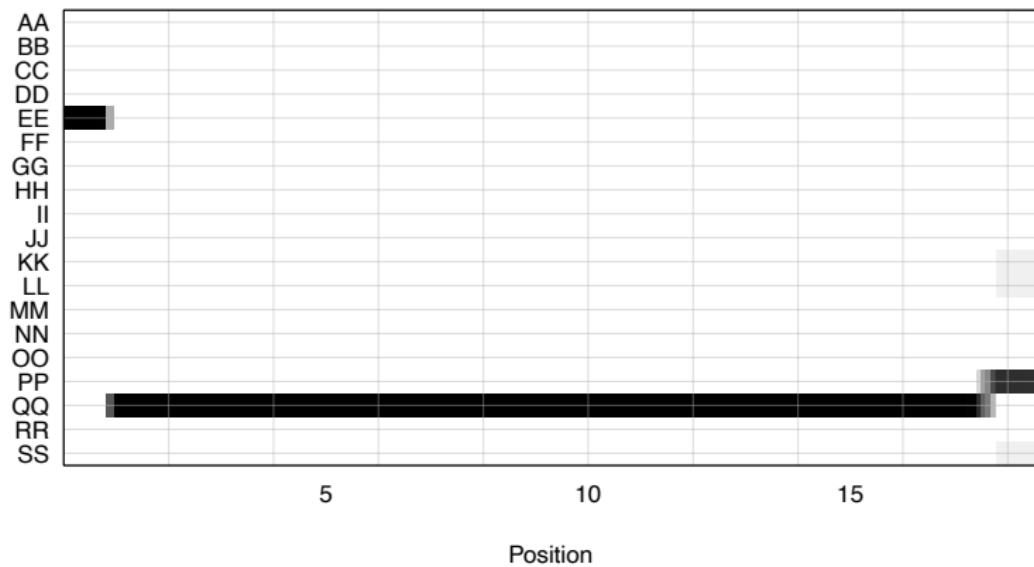
```
mfcol=c(1,2); mai=c(1,1,0.25,0.25) # change the plot parameters  
plot_onegeno(magic$geno, map=magic$gmap, ind=3) # Position in Mbp  
plot_onegeno(magic$geno, map=magic$gmap, ind=4) # use "ind=" to change individuals
```



Calculate genotype probabilities

- ▶ Display which parental alleles an individual more likely has
- ▶ Use “ind=” option to change individuals

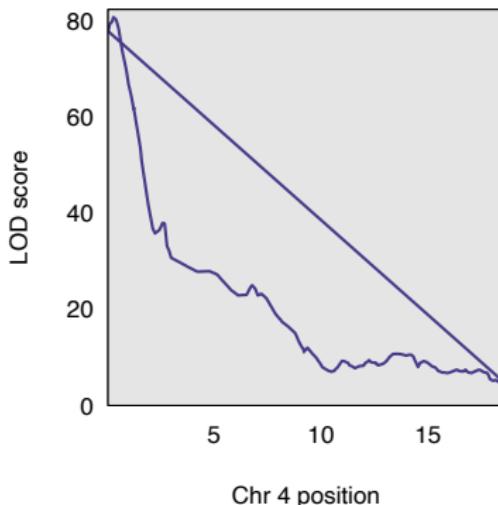
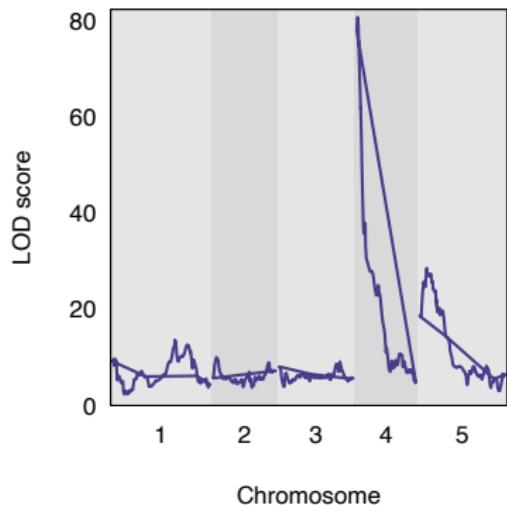
```
map2 <- insert_pseudomarkers(magic$gmap, step=1)
magic_p <- calc_genoprob(magic, map=map2)
plot_genoprob(magic_p, map=map2, chr=4, ind=3) # use "ind=" to change individuals
```



QTL mapping for flowering time in MAGIC lines

- Fine mapping also found QTL on the top of chr 4

```
res_scan2 <- scan1(magic_p, pheno=magic$pheno[,1])  
## mfcol=c(1,2); mai=c(1,1,0.25,0.25) # change the plot parameters  
## (res_scan2, map=map2); ## (res_scan2, map=map2, chr=4) # plot the results
```



What genes located near the flowering-related QTL?

- ▶ *FRIGIDA (FRI)* is located on the top of chromosome 4 (Aranzana et al. 2005).
- ▶ Many lab accessions carry loss of function mutations in *FRI* and thus are early-flowering.
- ▶ The same QTL was reported by GWAS of flowering time (Atwell et al. 2010).

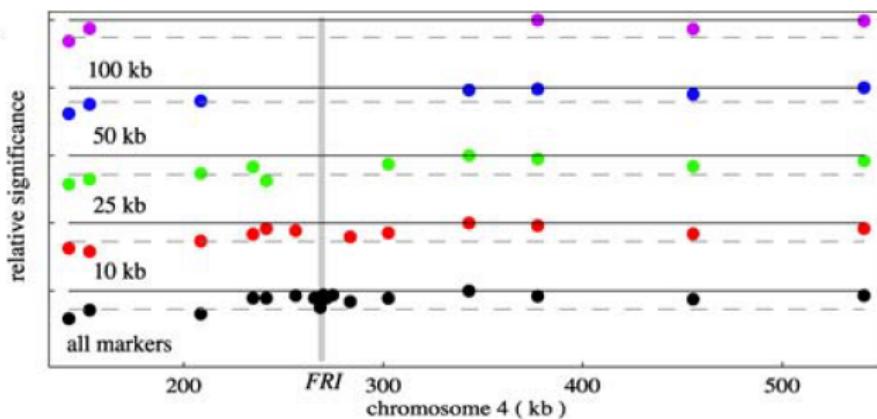


Figure 2: Single-marker analysis of flowering time by Aranzana et al. 2005. X-axis focuses on the top of chr 4.

Summary

1. Interval mapping can estimate the location of QTL even when markers are sparse
2. Multiparental intercross lines enable fine mapping with dense markers
3. We found a significant QTL on the top of chromosome 4 in *A. thaliana*

Exercise

- Q1. The cross2 object named “magic” includes 8 traits. Other than the flowering time (=‘bolting_days’), which traits can you find a sharp peak?
- Q2. ... and on which chromosome can you find the peak?
- Q3. Please read the last paragraph of ‘Results’ section of Gnan et al. (2014), <https://doi.org/10.1534/genetics.114.170746>. You will find a plausible answer there.

References

- ▶ Aranzana, María José, Sung Kim, Keyan Zhao, Erica Bakker, Matthew Horton, Katrin Jakob, Clare Lister, et al. (2005). Genome-Wide Association Mapping in *Arabidopsis* Identifies Previously Known Flowering Time and Pathogen Resistance Genes. PLoS Genetics 1(5):e60. <https://doi.org/10.1371/journal.pgen.0010060>.
- ▶ Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. (2010). Genome-Wide Association Study of 107 Phenotypes in *Arabidopsis thaliana* Inbred Lines. Nature 465:627–31. <https://doi.org/10.1038/nature08800>.
- ▶ Broman KW et al. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890. <https://doi.org/10.1093/bioinformatics/btg112>
- ▶ Broman, K. W., & Sen, S. (2009). A Guide to QTL Mapping with R/qtl. New York: Springer.
- ▶ Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Sen S, Yandell BS, Churchill GA. (2018). R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multi-parent populations. Genetics 211:495–502. <https://doi.org/10.1534/genetics.118.301595>
- ▶ Gnan, Sebastian, Anne Priest, and Paula X Kover. (2014). The Genetic Basis of Natural Variation in Seed Size and Seed Number and Their Trade-Off Using *Arabidopsis thaliana* MAGIC Lines. Genetics 198(4): 1751–58. <https://doi.org/10.1534/genetics.114.170746>.
- ▶ Huang, X., M.-J. Paulo, M. Boer, S. Effgen, P. Keizer, M. Koornneef, and F. A. van Eeuwijk. (2011). Analysis of Natural Allelic Variation in *Arabidopsis* Using a Multiparent Recombinant Inbred Line Population. Proceedings of the National Academy of Sciences USA 108(11):4488–93. <https://doi.org/10.1073/pnas.1100465108>.
- ▶ Kover, Paula X., William Valdar, Joseph Trakalo, Nora Scarcelli, Ian M. Ehrenreich, Michael D. Purugganan, Caroline Durrant, and Richard Mott. (2009). A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. PLoS Genetics 5(7): e1000551. <https://doi.org/10.1371/journal.pgen.1000551>.
- ▶ Li, Yan, Peter Roycewicz, Evadne Smith, and Justin O. Borevitz. (2006). Genetics of Local Adaptation in the Laboratory: Flowering Time Quantitative Trait Loci under Geographic and Seasonal Conditions in *Arabidopsis*." PLoS ONE 1(1):e105. <https://doi.org/10.1371/journal.pone.0000105>.