

GWAS/GS exercise using GAPIT

BIO373 at the University of Zurich

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

29 Sept. 2023

¹yasuhiro.sato@uzh.ch

Dataset: Rice 44k genomes

- ▶ Data from Zhao et al. (2011) Nature Communications 2:467
- ▶ Data available at <http://www.ricediversity.org/data/>
- ▶ 34 agronomic traits were examined for 413 rice accessions



ARTICLE

Received 14 Jan 2011 | Accepted 2 Aug 2011 | Published 13 Sep 2011

DOI: 10.1038/ncomms1467

Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*

Keyan Zhao^{1,2}, Chih-Wei Tung³, Georgia C. Eizenga⁴, Mark H. Wright¹, M. Liakat Ali⁵, Adam H. Price⁶, Gareth J. Norton⁶, M. Rafiqul Islam⁷, Andy Reynolds¹, Jason Mezey¹, Anna M. McClung⁴, Carlos D. Bustamante^{1,2} & Susan R. McCouch³

What to do in this exercise

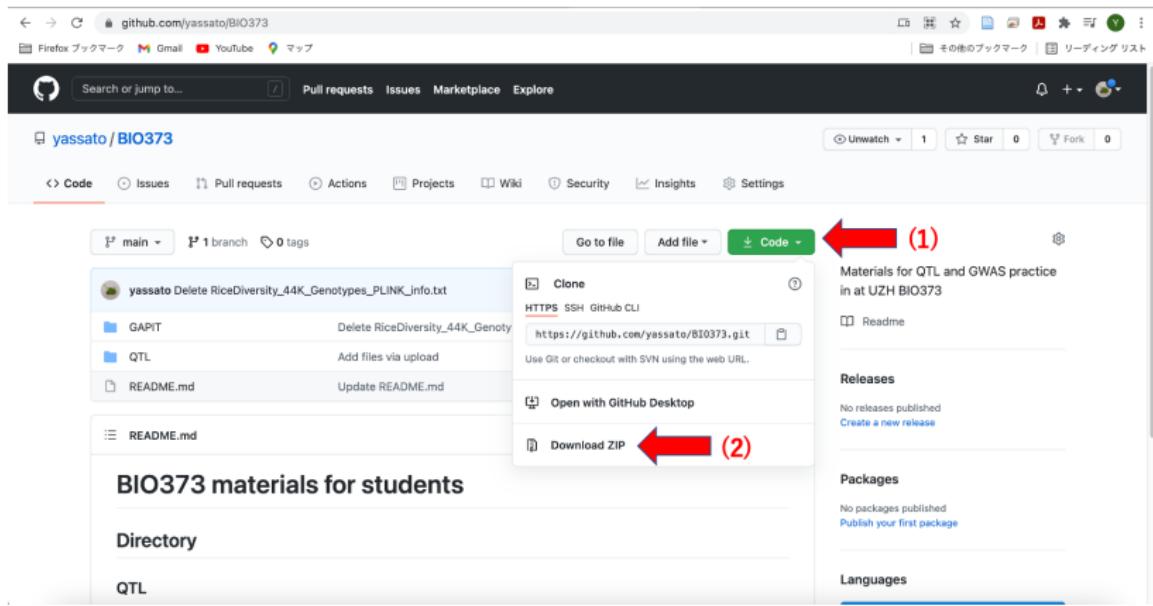
- ▶ GWAS example: Find genomic region associated with the seed length
- ▶ GS example: Predict flowering time between different study years
- ▶ Finally, you will be able to try GWAS/GS for traits of your interests!

Source codes and input data

- ▶ RiceDiversity_44K_Genotypes_PLINK_imputed.txt.gz
- ▶ RiceDiversity_44K_Genotypes_PLINK_info.txt
- ▶ *This instruction PDF*
- ▶ All available at
https://github.com/yassato/BIO373_YS2023/GAPIT

Download materials (same as the QTL exercise)

- ▶ Download .zip from
<https://github.com/yassato/BIO373> and unzip it
(Of course, “git clone URL” works if you are good at Git)



Set up the working directory

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 GAPIT` from Terminal; and then `setwd("./GAPIT")` from R Console
3. Upload RiceDiversity_44K_Genotypes_PLINK_imputed.txt.gz and RiceDiversity_44K_Genotypes_PLINK_info.txt to the directory you made

The screenshot shows the RStudio Server interface. The browser tab is 'BIO373/GAPIT at main · yasato' on 'fgcz-genomics.uzh.ch'. The RStudio console shows the following R code:

```
R 4.1.2 --> /GAPIT/
5 090414-A09   5   1   5   1.8817708 18.49218 89.37386 91.25563
6 090105-A02   7   1   6   8.8639442 20.66431 89.37386 98.23780
> pred <- predOrder(pred$Taxo,]
> y <- pOrder($HgID,]
> cor.test(pred$Prediction, y$Year06Flowering.time.at.Arkansas)

Pearson's product-moment correlation

data: pred$Prediction and y$Year06Flowering.time.at.Arkansas
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$Year07Flowering.time.at.Arkansas ~ pred$Prediction)
> plot(pred$Prediction,
+       y$Year07Flowering.time.at.Arkansas)
```

The RStudio environment pane shows the global environment with 'res' and 'y' loaded. The file tree pane shows the directory structure under 'Home / GAPT' with files like 'GAPT.R', 'GAPT.Rnw', 'GAPT.kin.R', 'GAPT.kin.Rnw', 'GAPT.Heterozygosity.pdf', etc.

First of all, load Genomic Association and Prediction Integrated Tool (GAPIT)

- ▶ Install GAPIT source code and its dependency
- ▶ Wait ca. 15 min. to install everything
- ▶ Some packages are not installed but they are negligible

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

# select "1:All" if this asks something about dependent packages
install.packages("devtools")
BiocManager::install("snpStats")
devtools::install_github("SFUStatgen/LDheatmap")
devtools::install_github("jiabowang/GAPIT3@078fe28",force=TRUE)
# load GAPIT3 package
```

Load and see phenotype data

```
pheno_url <- "http://www.ricediversity.org/data/sets/44kgwas/RiceDiversity_44K_Phenotypes_34traits_PLINK.RData"
p <- read.table(pheno_url, sep="\t", header=TRUE)
nrow(p) # no. of plants
```

```
## [1] 413
```

```
head(p)
```

```
##      HybID NSFTVID Flowering.time.at.Arkansas Flowering.time.at.Faridpur
## 1 081215-A05      1          75.08333            64
## 2 081215-A06      3          89.50000            66
## 3 081215-A07      4          94.50000            67
## 4 081215-A08      5          87.50000            70
## 5 090414-A09      6          89.08333            73
## 6 090414-A10      7          105.00000           NA
##   Flowering.time.at.Aberdeen FT.ratio.of.Arkansas.Aberdeen
## 1                      81          0.9269547
## 2                      83          1.0783133
## 3                      93          1.0161290
## 4                     108          0.8101852
## 5                     101          0.8820132
## 6                     158          0.6645570
##   FT.ratio.of.Faridpur.Aberdeen Culm.habit Leaf.pubescence Flag.leaf.length
## 1          0.7901235       4.0             1        28.37500
## 2          0.7951807       7.5             0        39.00833
## 3          0.7204301       6.0             1        27.68333
## 4          0.6481481       3.5             1        30.41667
## 5          0.7227723       6.0             1        36.90833
## 6             NA            3.0             1        36.99000
##   Flag.leaf.width Awn.presence Panicle.number.per.plant Plant.height
## 1     1.2833333         0            3.068053       110.9167
## 2     1.0000000         0            4.051785       143.5000
## 3     1.5166667         0            3.124565       128.0833
## 4     0.8916667         0            3.697178       153.7500
```

Read genotype data and marker information

```
g <- read.table("RiceDiversity_44K_Genotypes_PLINK_imputed.txt.gz",
                 header=TRUE, sep="\t")
gm <- read.table("RiceDiversity_44K_Genotypes_PLINK_info.txt.gz",
                 header=TRUE, sep="\t")
nrow(g) # no. of plants
```

```
## [1] 413
```

```
ncol(g[,-1]) # no. of SNPs
```

```
## [1] 36901
```

```
head(gm) # marker info
```

```
##           ID CHROM   POS
## 1 id1000001     1 13147
## 2 id1000003     1 73192
## 3 id1000005     1 74969
## 4 id1000007     1 75852
## 5 id1000008     1 75953
## 6 id1000011     1 91016
```

(1) Genome-wide association study (GWAS)

Aim: Looking for genomic region underlying the length of rice grains

- ▶ Indica cultivars have long grains, while Japonica have round-shaped grains
- ▶ *Grain Size 3 (GS3)* is known to regulate the seed length in rice (Wang et al. 2011)
- ▶ Can we detect the known loci with GWAS?

Run GWAS with a general linear model (GLM) or mixed linear model (MLM)

- ▶ It takes several minutes. Wait.
- ▶ When finished, output files appear in the current directory
- ▶ Warning messages occur but the program still works
- ▶ **Note: When you run GAPIT twice, the second run may not work. In such a case, log-out once and retry from data loading.**

```
myGAPIT <- GAPIT( # warnings occur but it still works
  Y=p[,c("HybID", "Seed.length")],
  GD=g,
  GM=gm,
  SNP.MAF=0.05, # cut-off minor alleles at 0.05
  Inter.Plot=TRUE, # option to make interactive plots
  model=c("GLM", "MLM"),
  kinship.algorithm="VanRaden",
  Multiple_analysis=TRUE)
```

GWAS is done. Let us see a trait diagnosis first

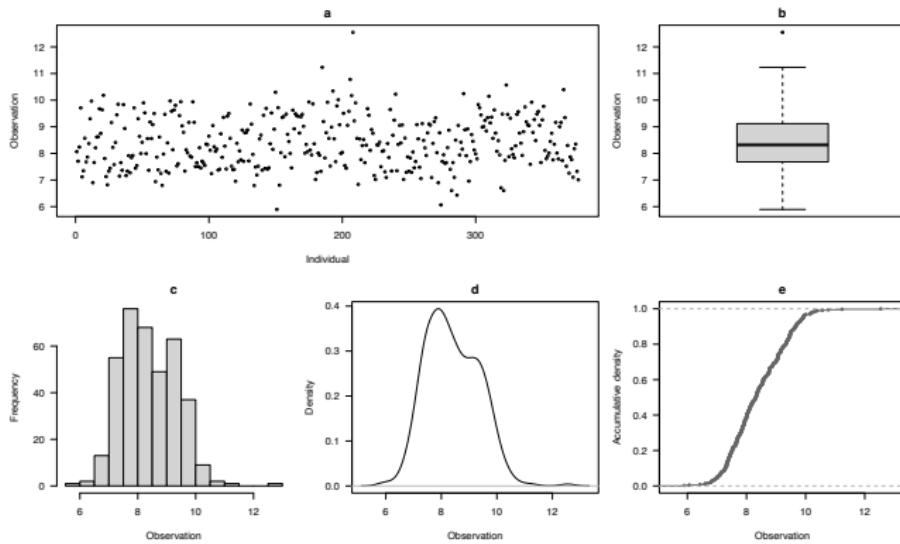


Figure 1: “GAPIT.Phenotype.View.Seed.length.pdf”

- ▶ The seed length looks normally distributed

... and also check heritability in the seed length

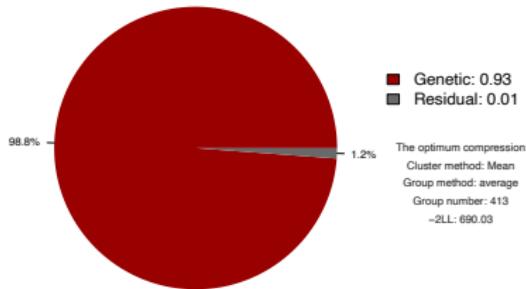


Figure 2: "GAPIT.Association.Optimum.MLM.Seed.length.pdf"

- ▶ Heritability, $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$,
- ▶ where σ_g^2 : genetic variance; σ_e^2 : residual variance
- ▶ h^2 (%) is calculated as
$$100 * (\sigma_g^2 / (\sigma_g^2 + \sigma_e^2)) = 100 * (0.93 / (0.93 + 0.01)) = 98.8\%$$

Check LD to see what kbp we should refer around the SNPs.

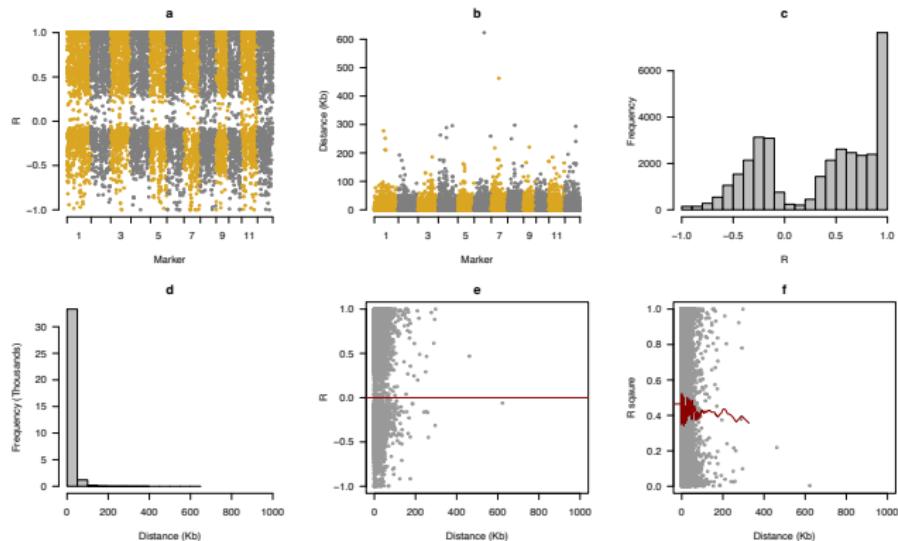


Figure 3: “GAPIT.Genotype.Density_R_sqaure.pdf”

- (b) The length of linkage disequilibrium is at most 600 kbp.

Compare the marker density with the LD length

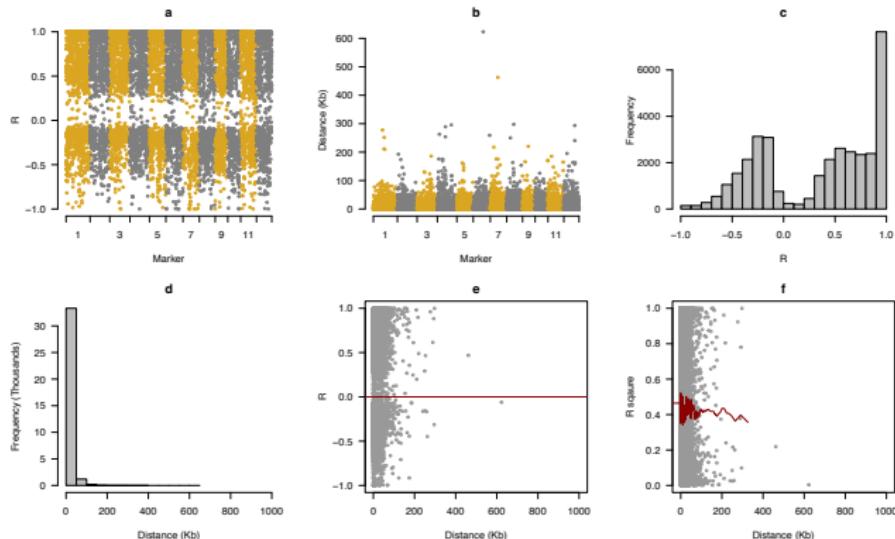


Figure 4: “GAPIT.Genotype.Density_R_sqare.pdf”

- ▶ (d) The marker intervals are much shorter than the length of LD,
- ▶ indicating that marker density was enough

Manhattan plot of the general linear model (GLM)

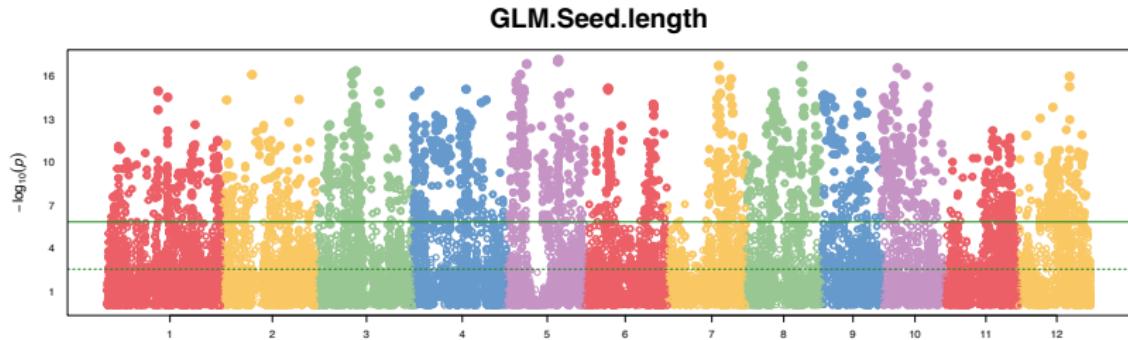


Figure 5: The file name
“GAPIT.Association.Manhattan_Geno.GLM.Seed.length.pdf”

- ▶ Is everything significant?? Very difficult to find key variants. . .

Quantile-quantile (QQ) plot also shows inflated p -values

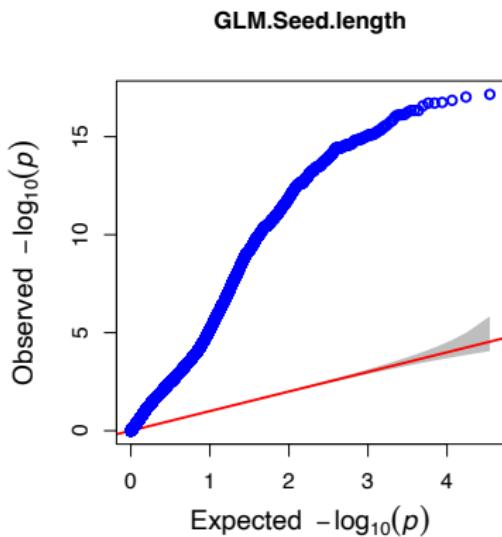


Figure 6: "GAPIT.Association.QQ.GLM.Seed.length.pdf"

- ▶ Blue dots: Observed $-\log_{10}(p\text{-values})$
- ▶ Red line: Expected $-\log_{10}(p\text{-values})$ when they are random

Heatmap of the kinship matrix shows two clusters

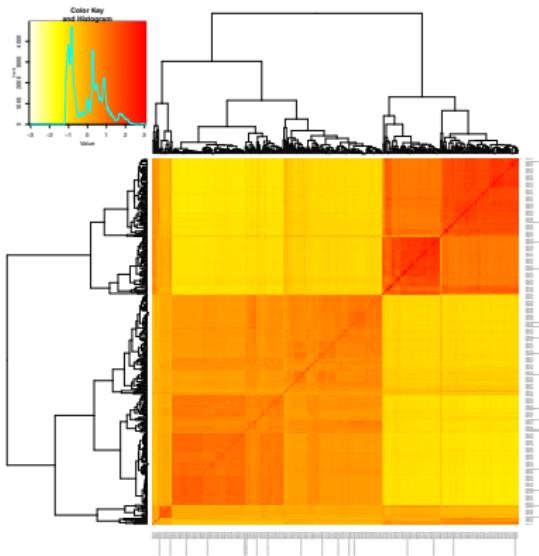


Figure 7: “GAPIT.Genotype.Kin_VanRaden.pdf”

- ▶ Here we see a complex kinship structure
- ▶ A mixed linear model is worth trying to correct it

Manhattan plot of the mixed linear model (MLM)

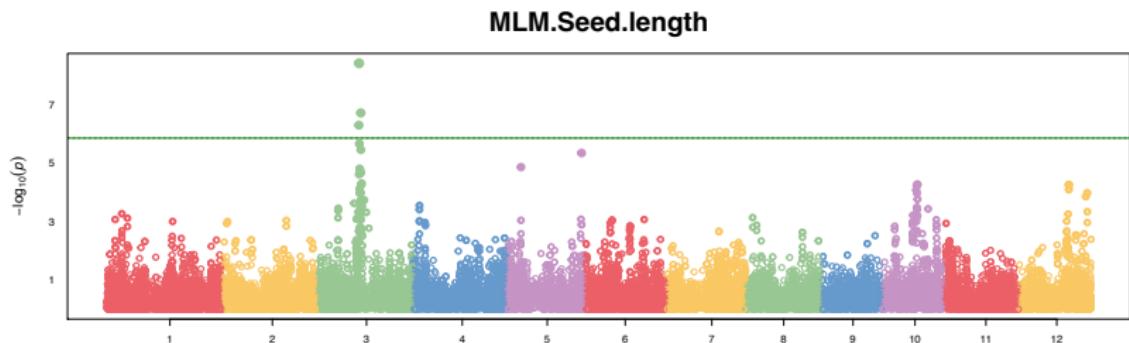


Figure 8: “GAPIT.Association.Manhattan_Geno.MLM.Seed.length.pdf”

- ▶ We can find a peak on the chromosome 3!

QQ plot of the mixed linear model

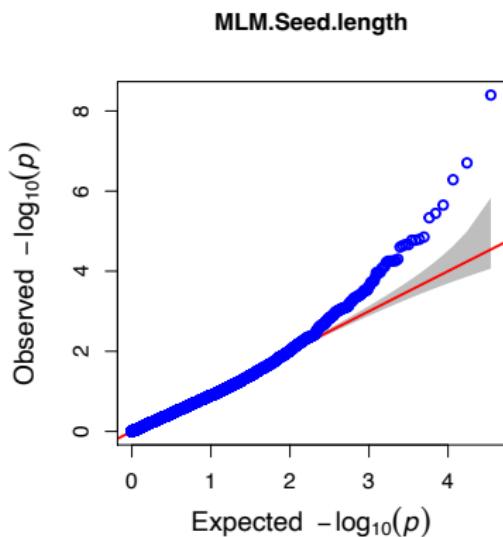


Figure 9: “GAPIT.Association.QQ.MLM.Seed.length.pdf”

- ▶ Only for top-scoring SNPs, $-\log_{10}(p\text{-values})$ are higher than expected

GWAS works well. Check the position of top-scoring SNPs

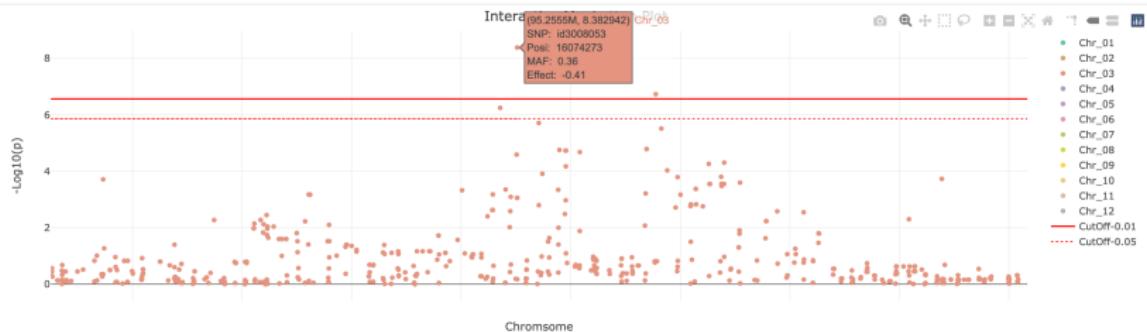
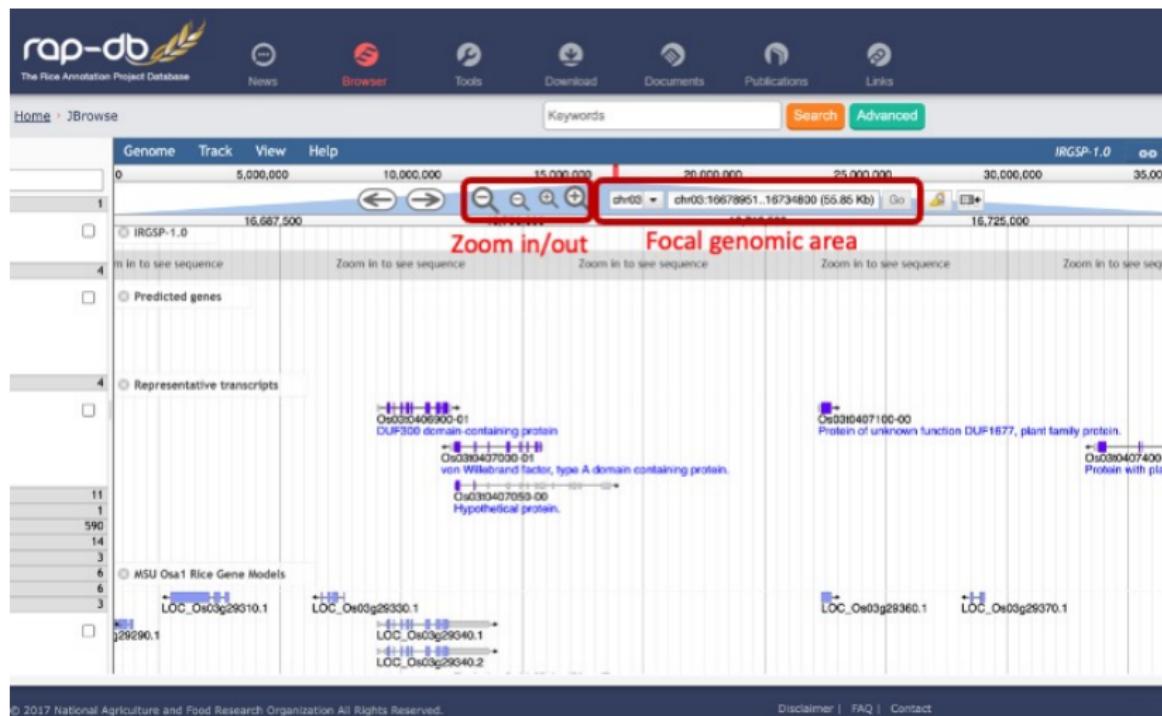


Figure 10:
“GAPIT.Association.Interactive_Manhattan.MLM.Seed.length.html”

- ▶ Open “Interactive.Manhattan.MLM.Seed.length.html”
- ▶ You can find 2 significant SNPs on the chromosome 3
- ▶ The position info appears when you put your cursor on a SNP

What genes are located nearby? Check the database

- ▶ Access RAP-DB website at <https://rapdb.dna.affrc.go.jp/>
- ▶ Look for and click “JBrowse”



Short exercise: Let's find the *GS3* locus.

By using RAP-DB,

1. Input significant SNP positions \pm 3 bp as a focal genomic area and “Go”
e.g., chr03:16706777..16706779
2. Zoom in/out \pm the average LD length near the SNP
3. Find the locus ID “Os03t0407400-01” (= *GS3*) and click!

Point of (biological) interpretation

Which SNPs can you find *GS3* locus nearby?

How far is the *GS3* from the significant SNP?

What family of proteins does the *GS3* encode?

(2) Genomic selection (GS)

Aim: Prediction of the flowering time in rice cultivars

- ▶ Flowering time, or heading date in rice, was recorded at Arkansas on 2006 and 2007 (Zhao et al. 2011)
- ▶ Genotypes were same but environment should be different between years
- ▶ The flowering time of some accessions were unavailable
- ▶ Can we predict the flowering time only on the basis of genotypes?

Estimate a trait value of each plant with gBLUP

- When finished, results are stored in “myGAPIT_BLUP” object.

```
# gBLUP for the flowering time 2006 at Arkansas
myGAPIT_BLUP <- GAPIT( # warnings occur but it still works
  Y=p[,c("HybID", "Year06Flowering.time.at.Arkansas")],
  GD=g,
  GM=gm,
  SNP.MAF=0.05,
  model="gBLUP",
  kinship.algorithm="VanRaden",
  file.output=FALSE)
```

```
## [1] "----- Welcome to GAPIT -----"
## [1] "gBLUP"
## [1] "-----Processing traits-----"
## [1] "Phenotype provided!"
## [1] "The 1 model in all."
## [1] "MLM"
## [1] "GAPIT.DP in process..."
## [1] "GAPIT will filter marker with MAF setting !!"
## [1] "The markers will be filtered by SNP.MAF: 0.05"
## maf_index
## FALSE TRUE
## 2150 34751
## [1] "Calculating kinship..."
## [1] "Number of individuals and SNPs are 413 and 34751"
## [1] "Calculating kinship with VanRaden method..."
## [1] "subtracting P..."
## [1] "Getting X'X..."
## [1] "Adjusting..."
## [1] "Calculating kinship with VanRaden method: done"
## [1] "kinship calculated"
```

We get BLUP, PEV, BLUE, and predicted trait values

- ▶ BLUP: Best Linear Unbiased Predictor shows trait variance around mean
- ▶ PEV: prediction error variance of BLUP
- ▶ BLUE: Best Linear Unbiased Estimator shows mean differences of traits
- ▶ BLUP + BLUE = Prediction

```
# load results of genomic prediction
pred <- myGAPIT_BLUP$Pred
head(pred)
```

##	Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction
## 1	081215-A05		1	1	-10.3304078	11.01517	89.3738566605352	79.04345
## 2	081215-A06		2	1	0.7679973	21.42044	89.3738566605352	90.14185
## 3	081215-A07		3	1	-1.9286126	20.80244	89.3738566605352	87.44524
## 4	081215-A08		4	1	0.4137206	15.23512	89.3738566605352	89.78758
## 5	090414-A09		5	1	1.8817708	18.49218	89.3738566605352	91.25563
## 6	090105-A02		7	1	8.8639442	20.66431	89.3738566605352	98.23780

Of course, predicted flowering time is well correlated with observed values

```
# align predicted and observed traits following the taxa name
pred <- pred[order(pred$Taxa),]
y <- p[order(p$HybID),]

# calculate Pearson's correlation between predicted and observed flowering
cor(pred$Prediction, y$Year06Flowering.time.at.Arkansas)

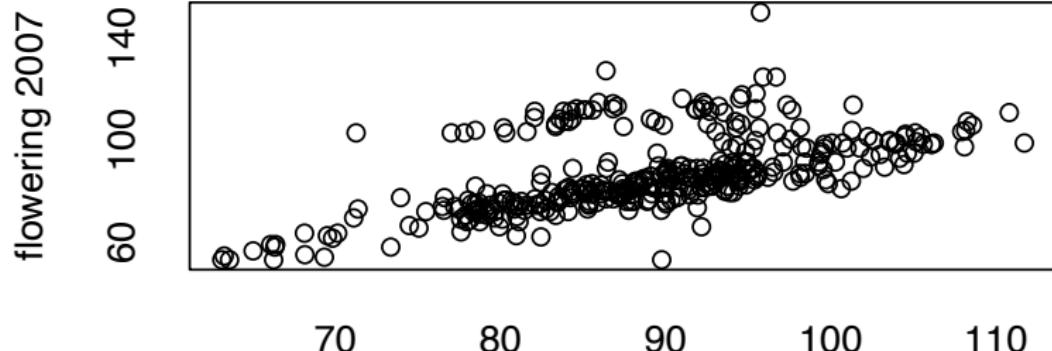
## 
## Pearson's product-moment correlation
##
## data: pred$Prediction and y$Year06Flowering.time.at.Arkansas
## t = 49.431, df = 335, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.9234637 0.9494873
## sample estimates:
## cor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

Predicted flowering time is correlated with those observed on 2007

```
# perform a linear regression to estimate the slope and intercept
res <- lm(y$Year07Flowering.time.at.Arkansas~pred$Prediction)

# plot the results
plot(pred$Prediction,
      y$Year07Flowering.time.at.Arkansas,
      ylab="flowering 2007", xlab="predicted",
      main=paste("r =", round(sqrt(summary(res)$r.squared), 2)))
abline(res)
```

$$r = 0.58$$



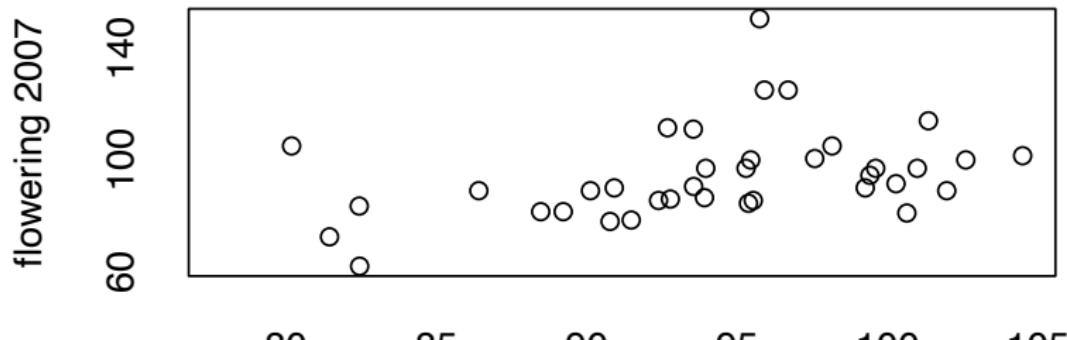
Predicted flowering time is correlated with those of missing accessions

```
NA06 <- is.na(y$Year06Flowering.time.at.Arkansas)

# perform a linear regression to estimate the slope and intercept
res <- lm(y$Year07Flowering.time.at.Arkansas[NA06] ~ pred$Prediction[NA06])

# plot the results
plot(pred$Prediction[NA06],
      y$Year07Flowering.time.at.Arkansas[NA06],
      ylab="flowering 2007", xlab="predicted",
      main=paste("r =", round(sqrt(summary(res)$r.squared), 2)))
abline(res)
```

$$r = 0.37$$



(3) Exercise

- ▶ Q1. Try GWAS of the flowering time at Aberdeen. How high is the heritability of this trait? At which chromosome can you find a peak?
- ▶ Q2. Find *HEADING DATE1* (*Hd1*: locus ID “Os06g0275000”). How distant is this gene from the top-scoring SNP? What is the ortholog of *Hd1* in *Arabidopsis thaliana*?
- ▶ Q3. Try gBLUP of the flowering time at Aberdeen. How large is the correlation between the predicted flowering time and observed one at Arkansas?
- ▶ Q4. More? You can test any traits of your interests!

(4) GWAS group work (30 min. incl. a break)

- ▶ Select 1 interesting trait for 1 group
- 1. Report its heritability,
- 2. perform MLM, and report $-\log_{10}(p)$ of the most significant SNP,
- 3. and list up 2 interesting candidate genes (specified by the code like Os03txxxxx) <200 kb near the most significant SNP.
- ▶ Send the result of 2., 3., and 4. to me (yasuhiro.sato@uzh.ch). 1 email from a representative is ok.
Please add your group no. to the email title.

References

- ▶ Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J. et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397-2399. <https://zzlab.net/GAPIT/>
- ▶ Rice Diversity website: <http://www.ricediversity.org/>
- ▶ Sakai, Hiroaki, Sung Shin Lee, Tsuyoshi Tanaka, Hisataka Numa, Jungsok Kim, Yoshihiro Kawahara, Hironobu Wakimoto, et al. (2013). Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics. *Plant and Cell Physiology* 54(2):e6. <https://rapdb.dna.affrc.go.jp/>
- ▶ Wang, Chongrong, Sheng Chen, and Sibin Yu. (2011). Functional Markers Developed from Multiple Loci in GS3 for Fine Marker-Assisted Selection of Grain Length in Rice. *Theoretical and Applied Genetics* 122(5):905-13. <https://doi.org/10.1007/s00122-010-1497-0>.
- ▶ Wang, Jiabo, and Zhiwu Zhang. (2021) GAPIT version 3: boosting power and accuracy for genomic association and prediction. *Genomics, Proteomics & Bioinformatics* 19(4):629-640. <https://doi.org/10.1016/j.gpb.2021.08.005>
- ▶ Zhao, Keyan, Chih-Wei Tung, Georgia C. Eizenga, Mark H. Wright, M. Liakat Ali, Adam H. Price, Gareth J. Norton, et al. (2011). Genome-Wide Association Mapping Reveals a Rich Genetic Architecture of Complex Traits in *Oryza sativa*. *Nature Communications* 2(1):467. <https://doi.org/10.1038/ncomms1467>.