Gene Hunting, Bread Making, and GAPIT Engineering

Zhiwu Zhang

WASHINGTON STATE UNIVERSITY

Zhiwu Zhang Laboratory

for Statistical Genomics

Home People Publication Research Teaching Software Outreach Jobs

Five ingredients to succeed: CS-VMV

Culture: Trying to understand.

Strategy: Solve biological problems with analytical and computational challenges.

Vision: Genomic and phenomic stream data is stationary water for organisms.

Mission: You get data, we help with our analytical methods, tools, and expertise.

Value: Every idea makes sense.

zzlab.net/share

Zhiwu Zhang Laboratory
for Statistical Genomics

GAPIT

Blink

GRID

GridFree

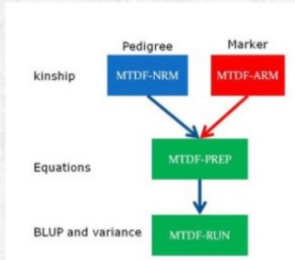
iPat

FarmCPU

Rooster

LADDER

mMAP

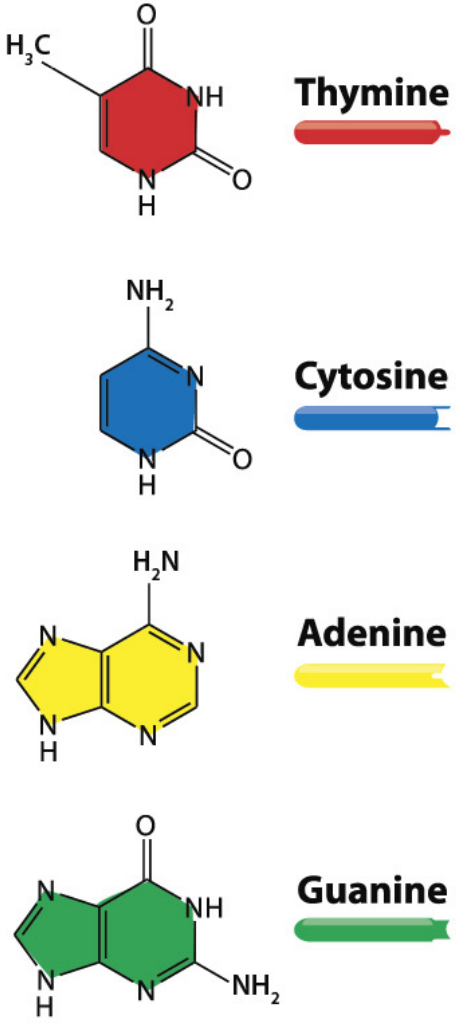


MTDFREML

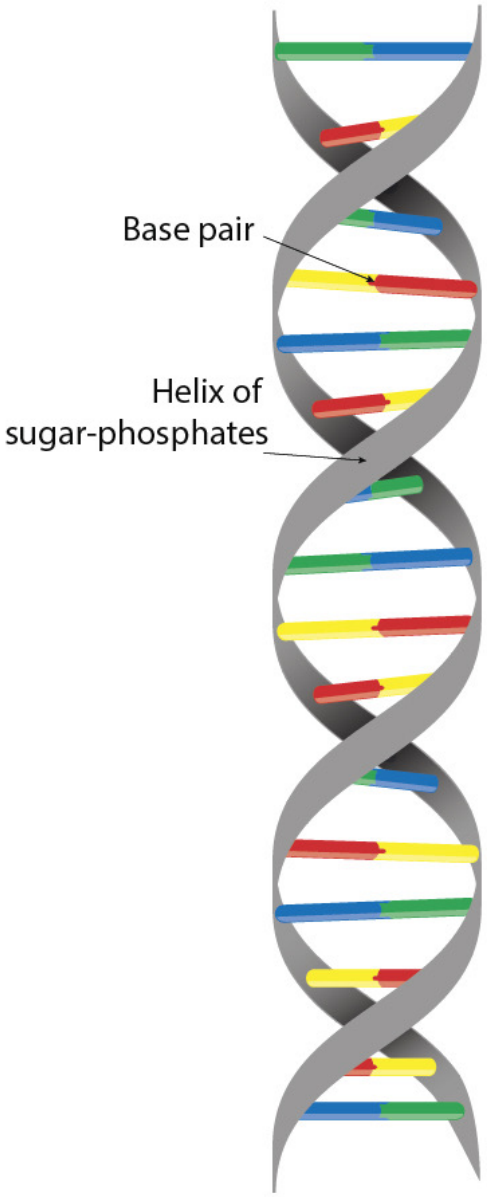
Audio4EDU

AI4EVER

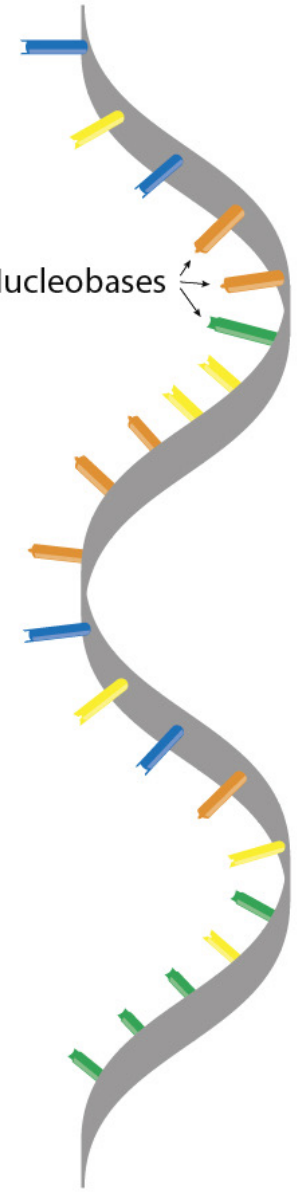
James Watson and Francis Crick DNA Model (Nobel Prize, 1962)



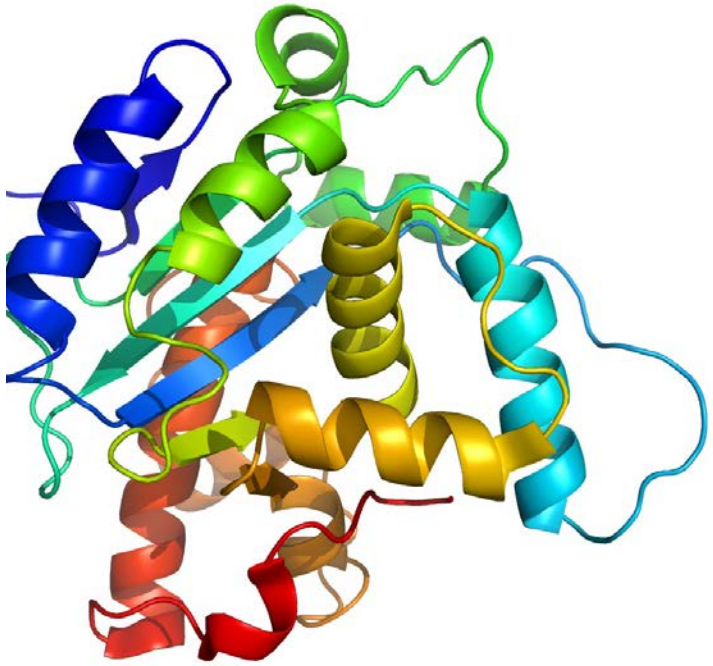
Nucleobases of DNA



DNA
Deoxyribonucleic acid



RNA
Ribonucleic Acid



Genotypes

AA: 0

AT: 1 ★

TT: 2 ★★

Genomic study

❖Explanation
- Candidate gene
- Cloning
- Linkage analysis
- GWAS

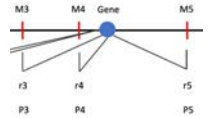
} Backward

} Forward

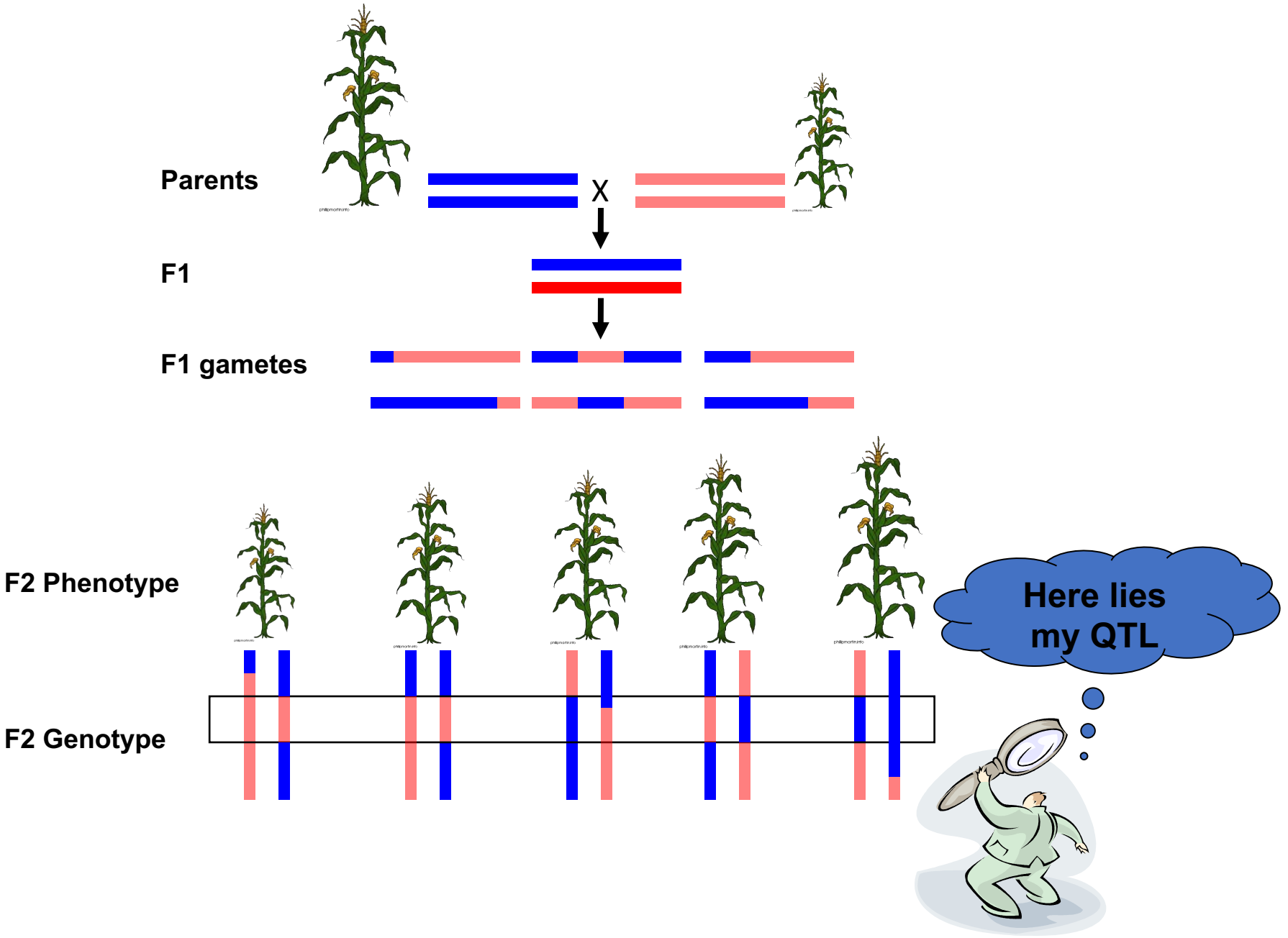
❖Prediction
- MAS
- GS
- GWAS+GS
- AI

Outline

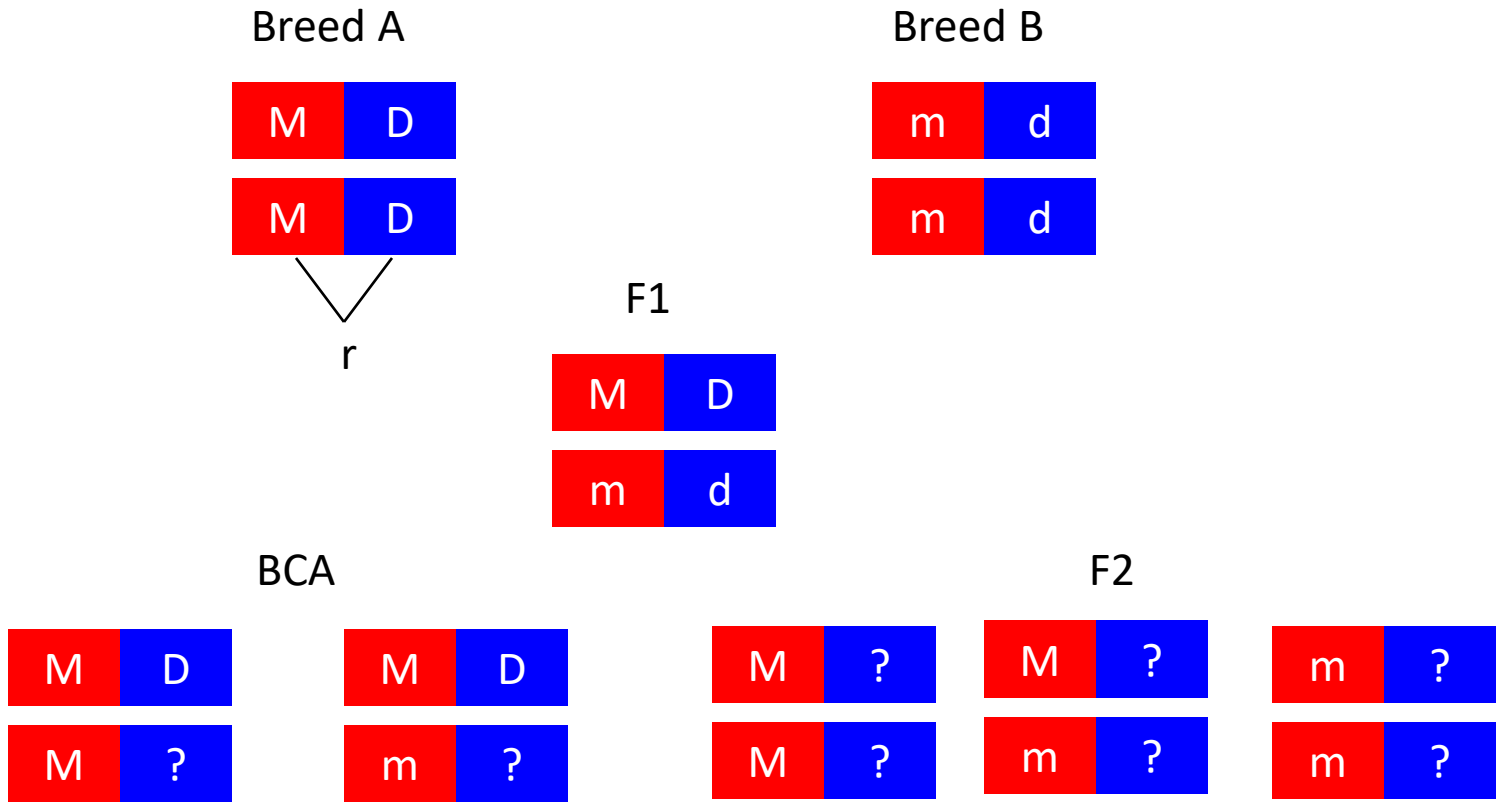
- Linkage analysis
- Association study
- Population structure and GLM
- Kinship and MLM
- BLINK



Linkage analysis

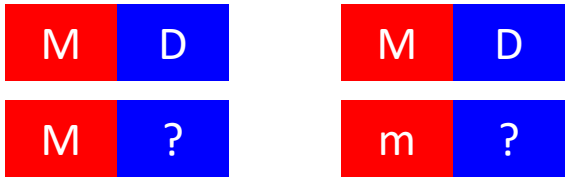


Crosses



Probability

BCA



P(?=D | MM)=1-r P(?=D | Mm)=r

P(?=d | MM)=r P(?=d | Mm)=1-r

	DD	Dd
MM	n1	n2
Mm	n3	n4

Recombine Unrecombine

$$P = r^{(n2+n3)} (1-r)^{(n1+n4)}$$

Mapping: vary r to maximize P

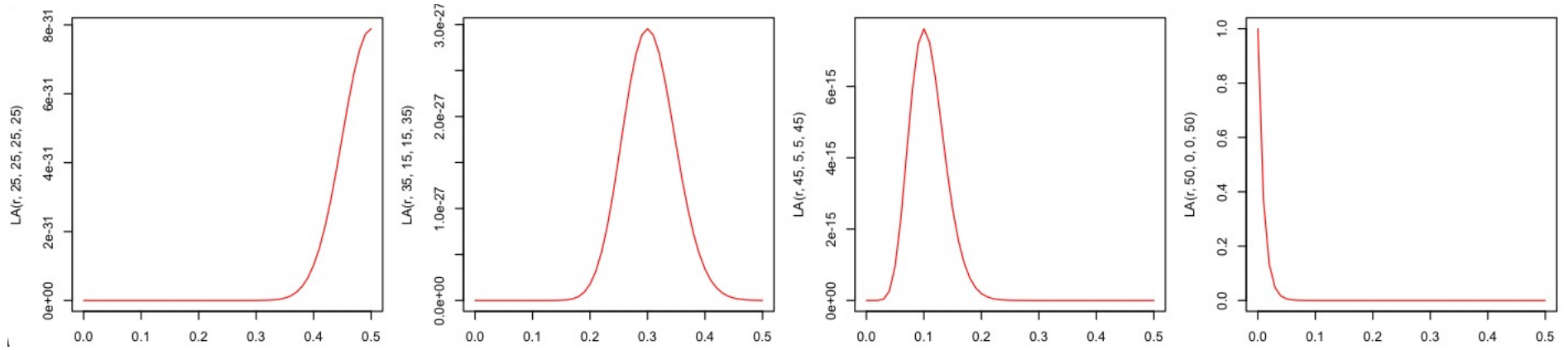
$$P = r^{(n2+n3)} (1-r)^{(n1+n4)}$$

	D	d
MM	25	25
Mm	25	25

	D	d
MM	35	15
Mm	15	35

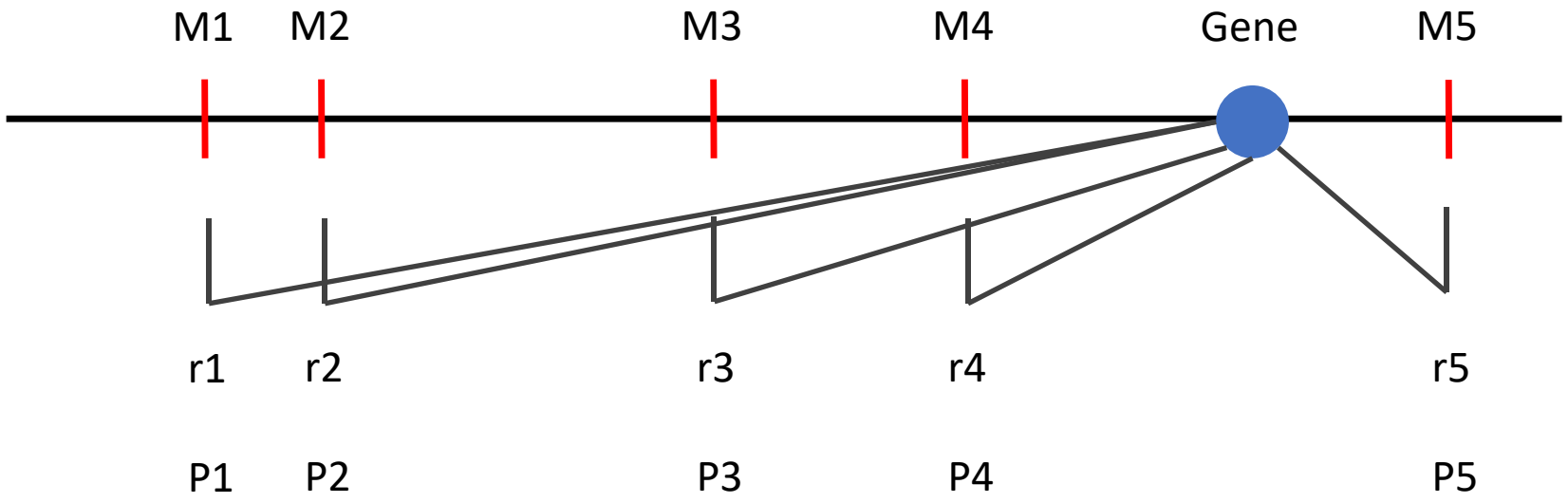
	D	d
MM	45	5
Mm	5	45

	D	d
MM	50	0
Mm	0	50



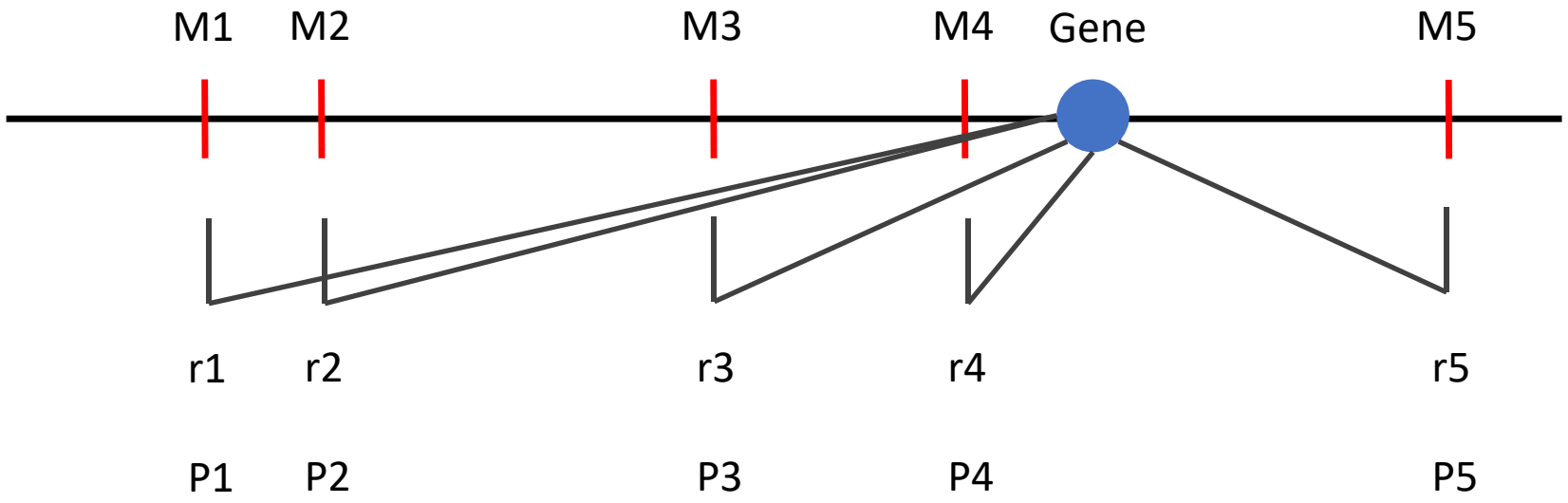

```
r=seq(0,.5,.01)
LA=function(r,n1,n2,n3,n4){return(r^(n2+n3)*(1-r)^(n1+n4))}
par(mfrow=c(1,4),mar = c(3,4,1,1))
plot(r,LA(r,25,25,25,25),type="l",col="red")
plot(r,LA(r,35,15,15,35),type="l",col="red")
plot(r,LA(r,45,5,5,45),type="l",col="red")
plot(r,LA(r,50,0,0,50),type="l",col="red")
```

Multiple markers

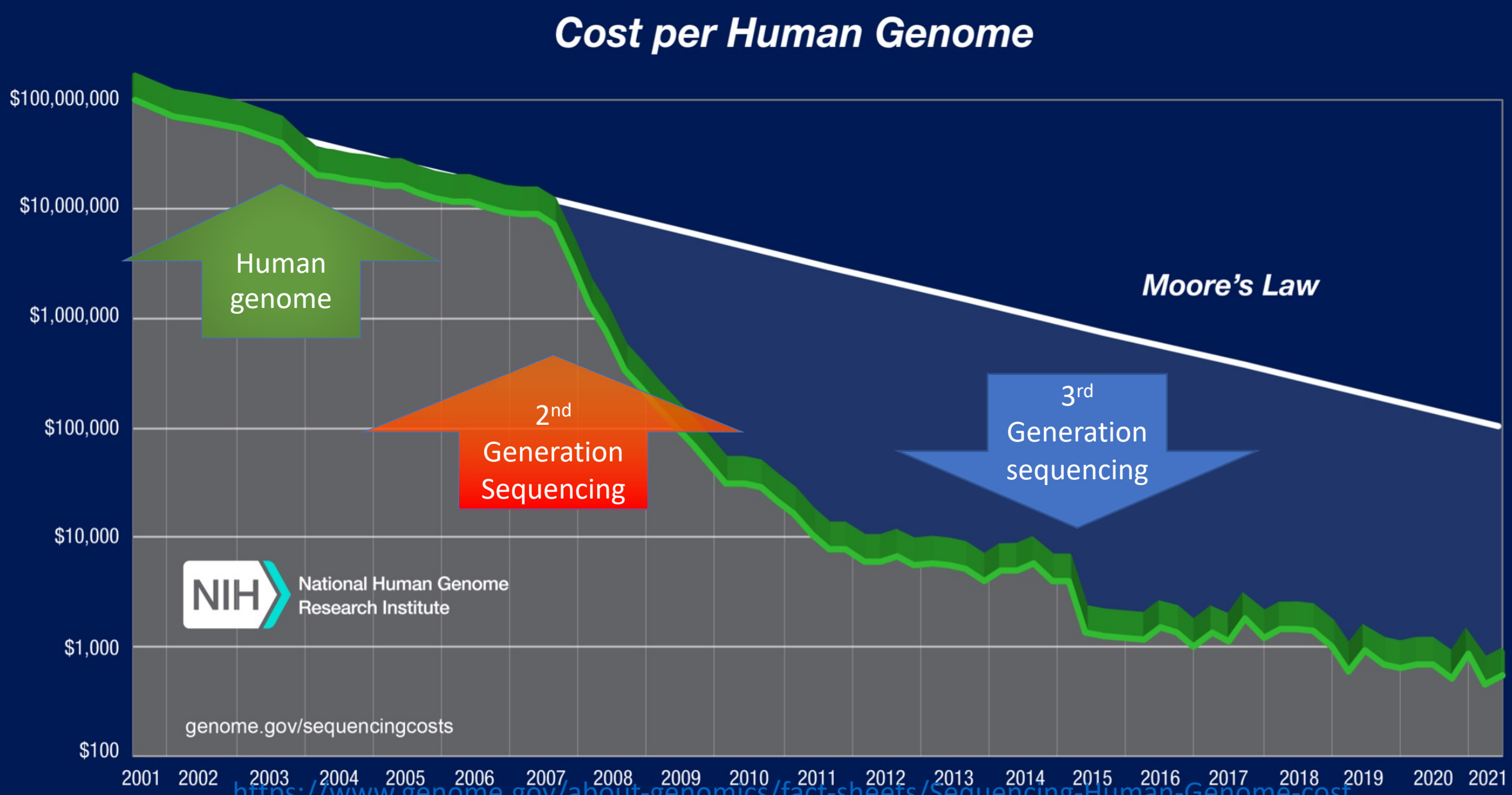


$P = P1*P2*P3*P4*P5$

Multiple markers

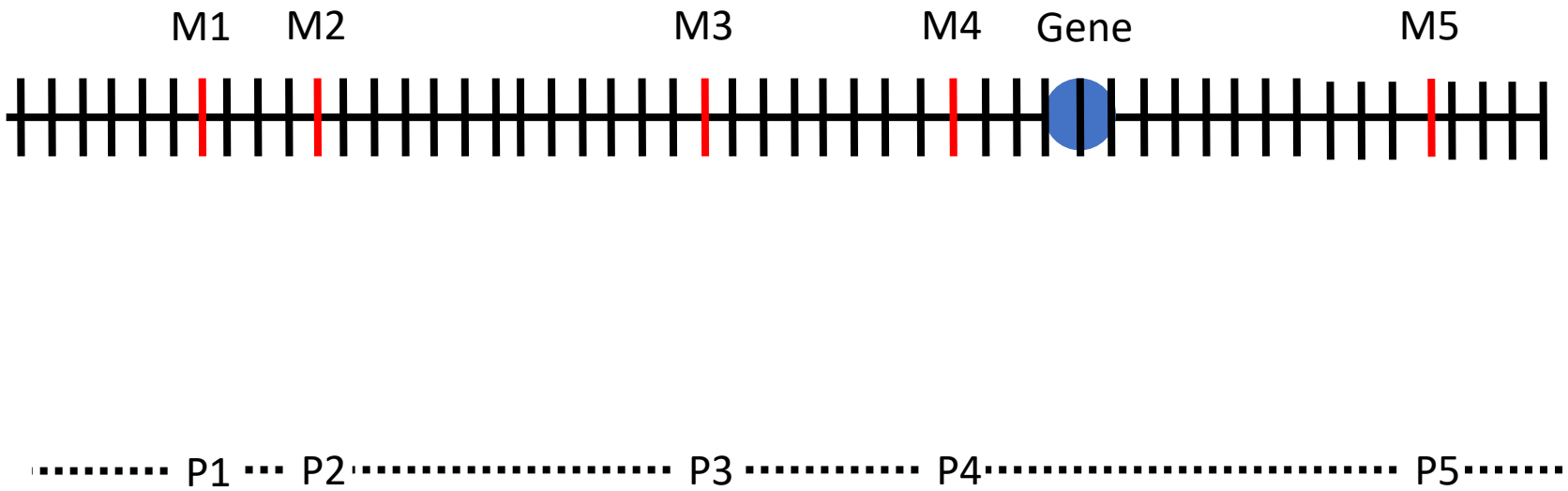


$P = P1*P2*P3*P4*P5$



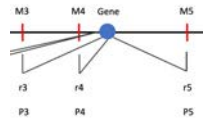
https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Dense markers (GWAS)

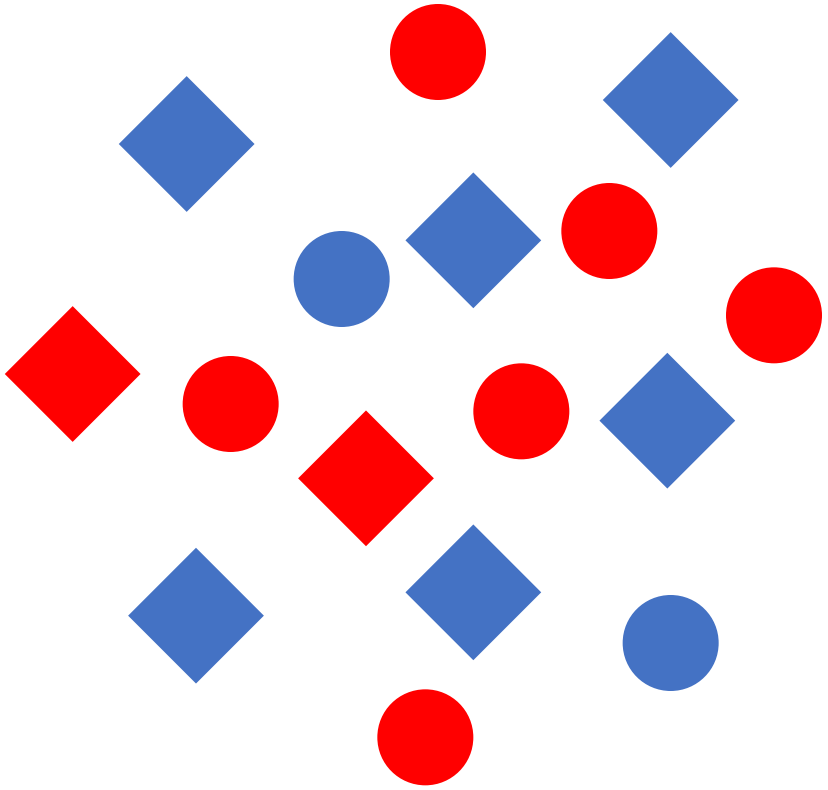


Outline

- Linkage analysis
- Association study
- Population structure and GLM
- Kinship and MLM
- BLINK



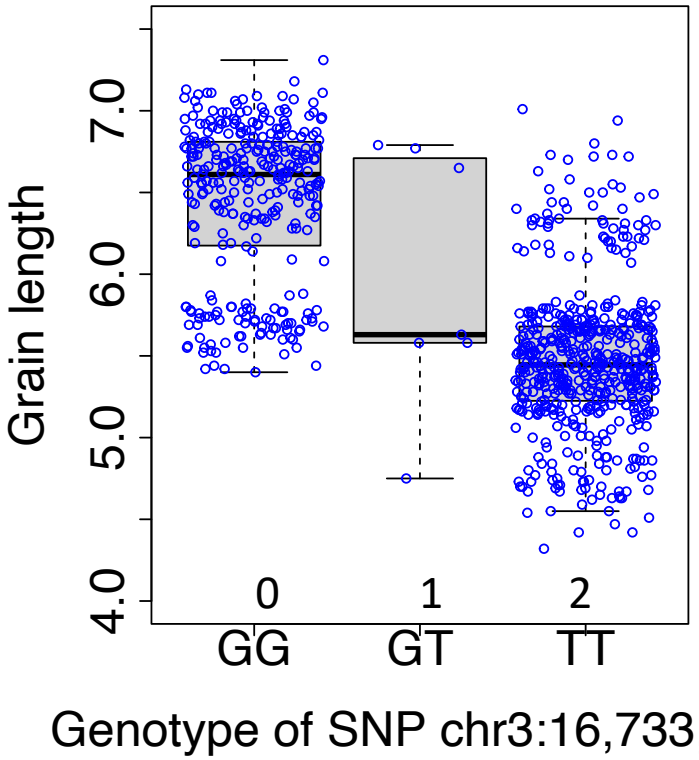
Association study



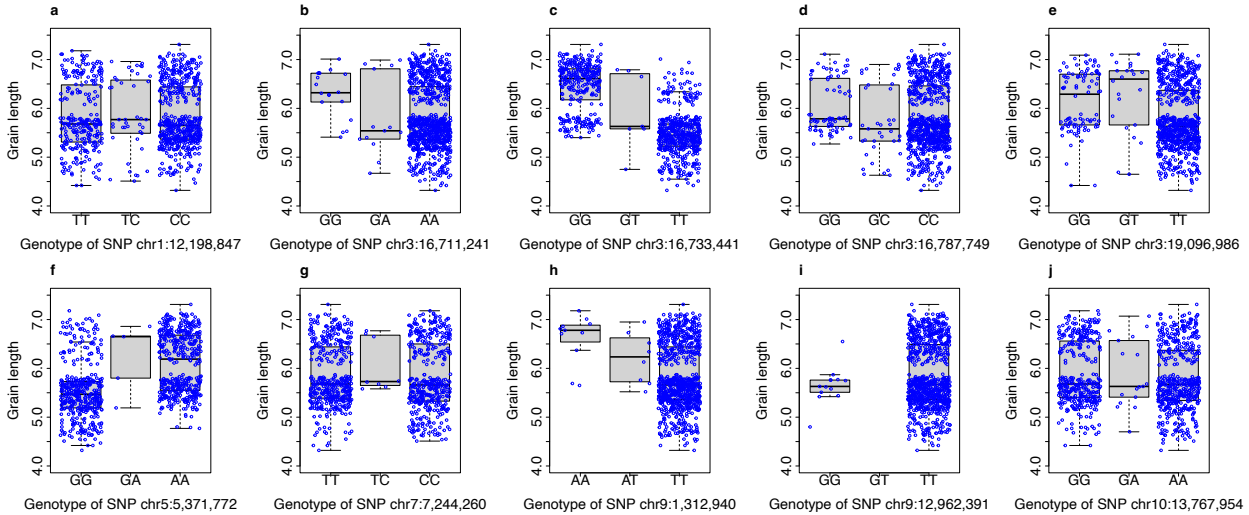
Marker	Control	Case
◇	6	2
○	2	6

$X^2=4(2*2/4)=4$, df=1,
P=4.5%

Correlation



Genotype of SNP chr3:16,733,441

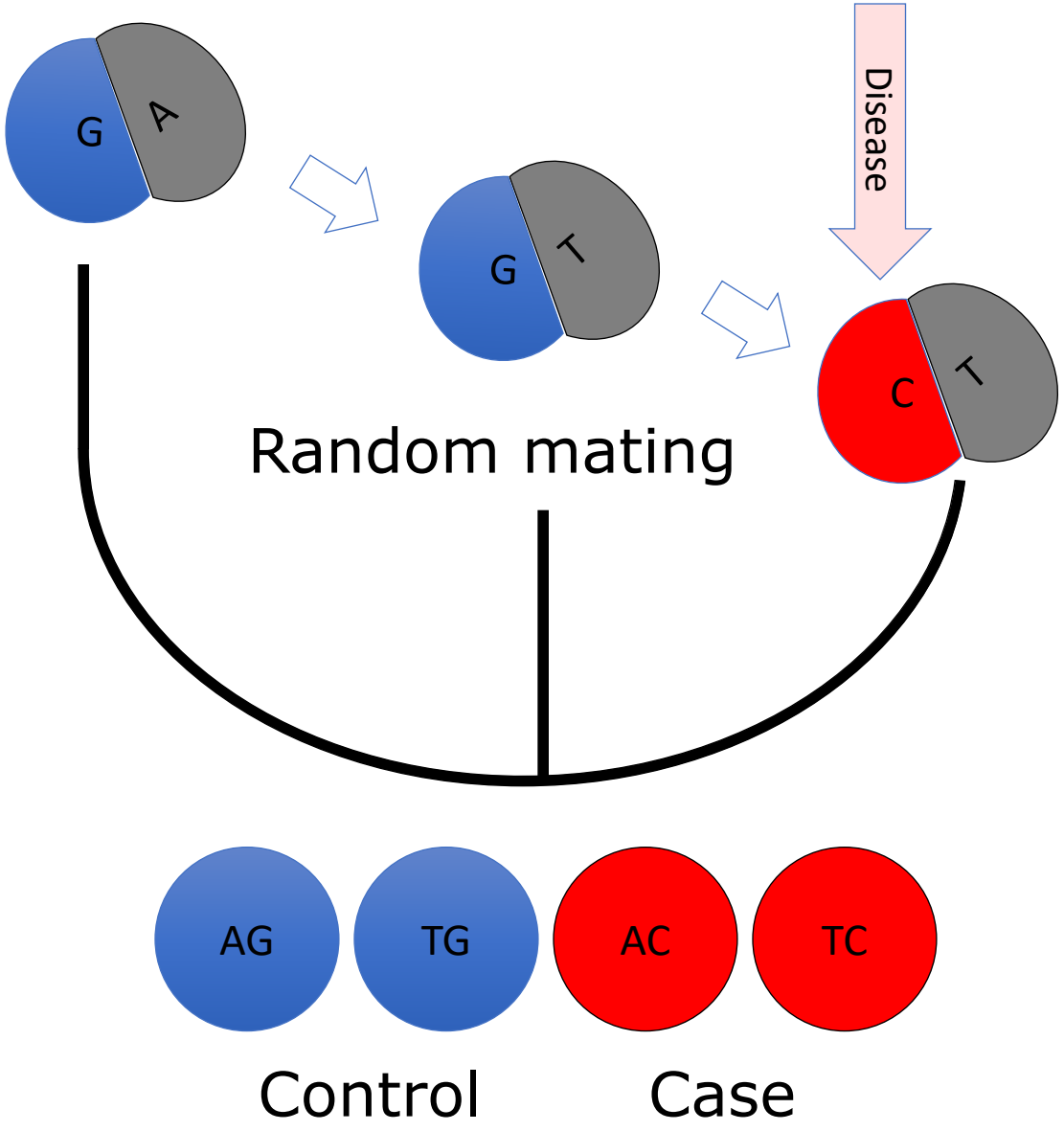


Genotype of SNP chr1:12,198,847
Genotype of SNP chr3:16,711,241
Genotype of SNP chr3:16,733,441
Genotype of SNP chr3:16,787,749
Genotype of SNP chr3:19,096,986
Genotype of SNP chr5:5,371,772
Genotype of SNP chr7:7,244,260
Genotype of SNP chr9:1,312,940
Genotype of SNP chr9:12,962,391
Genotype of SNP chr10:13,767,954

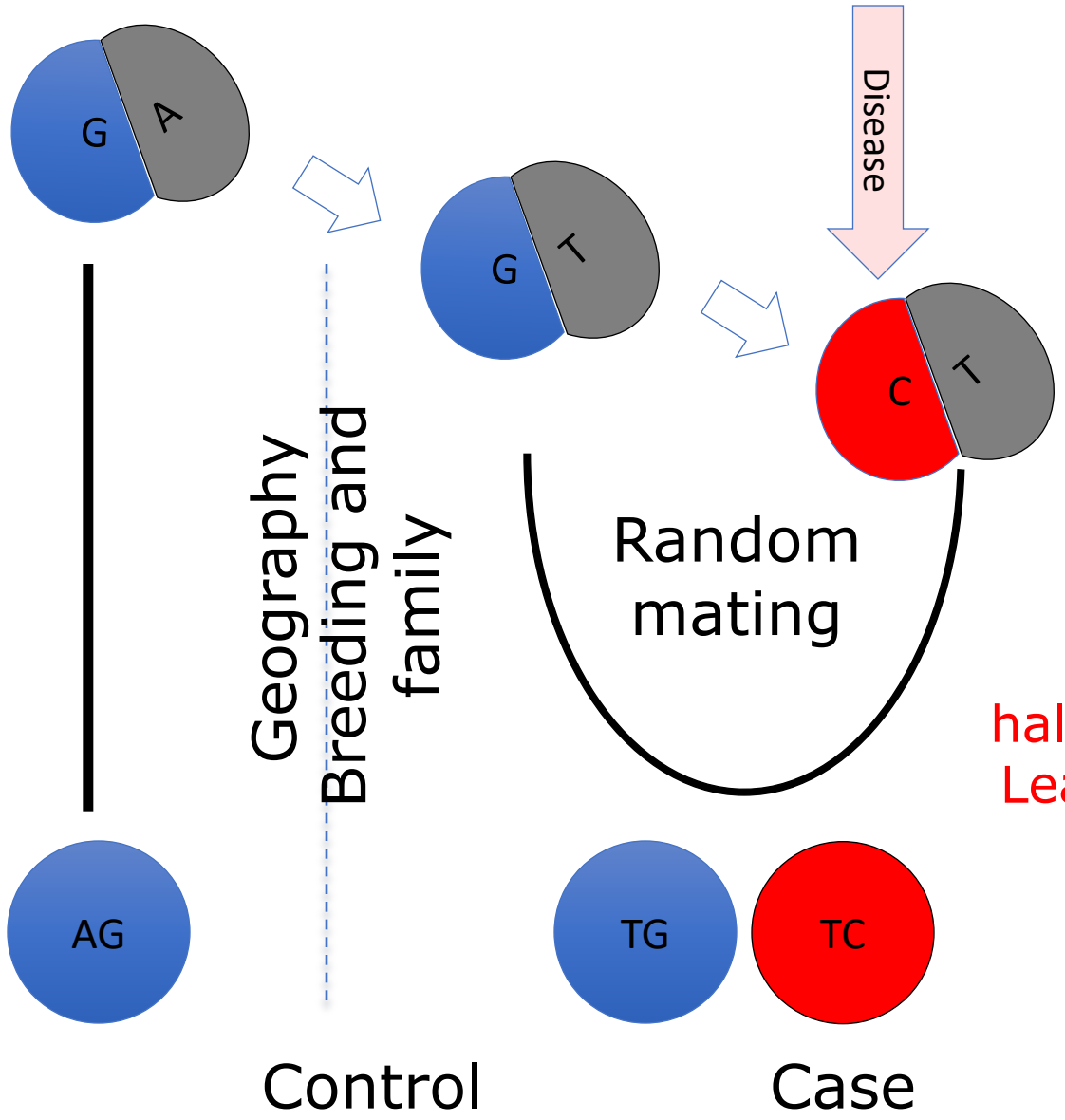
Comparison between linkage analysis and GWAS

Property	Linkage analysis	GWAS
Resolution		
Generation		
Genetic base		

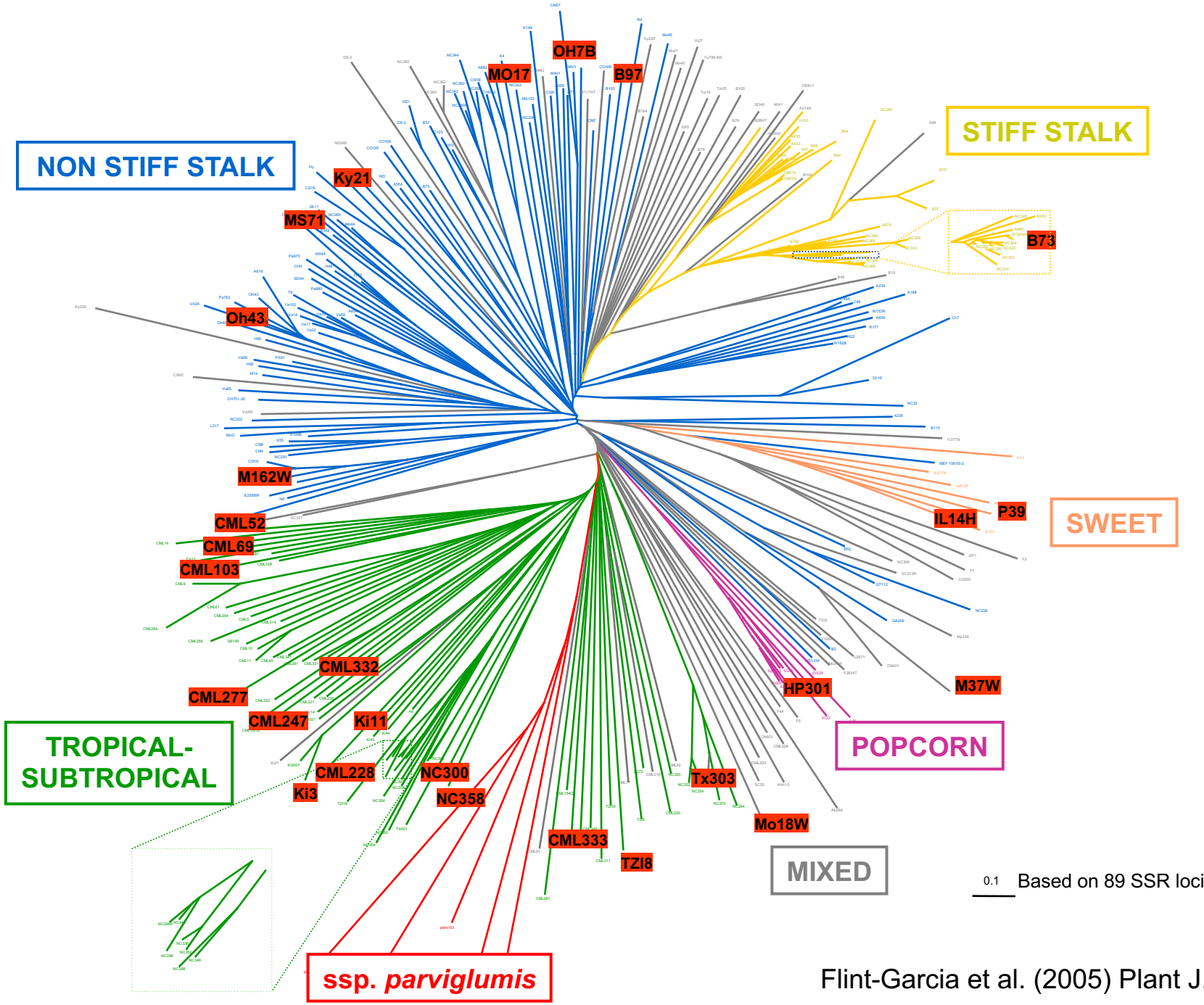
Linkage equilibrium



Linkage disequilibrium (LD)



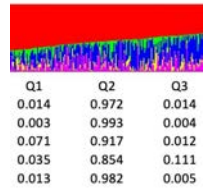
Association
half T as case, none for A
Lead to mistake if G/C is
not a marker



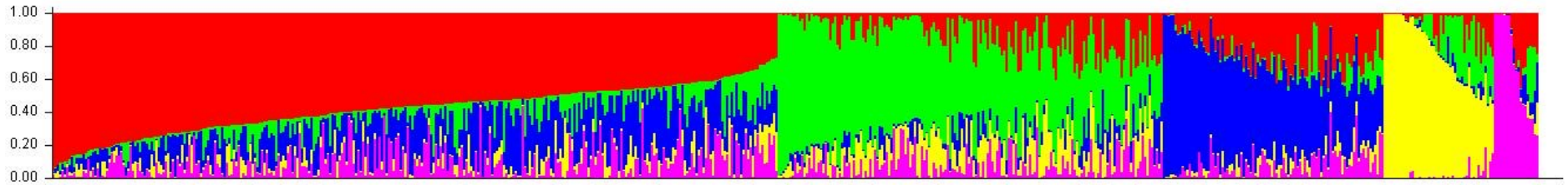
Flint-Garcia et al. (2005) Plant J. 44: 1054

Outline

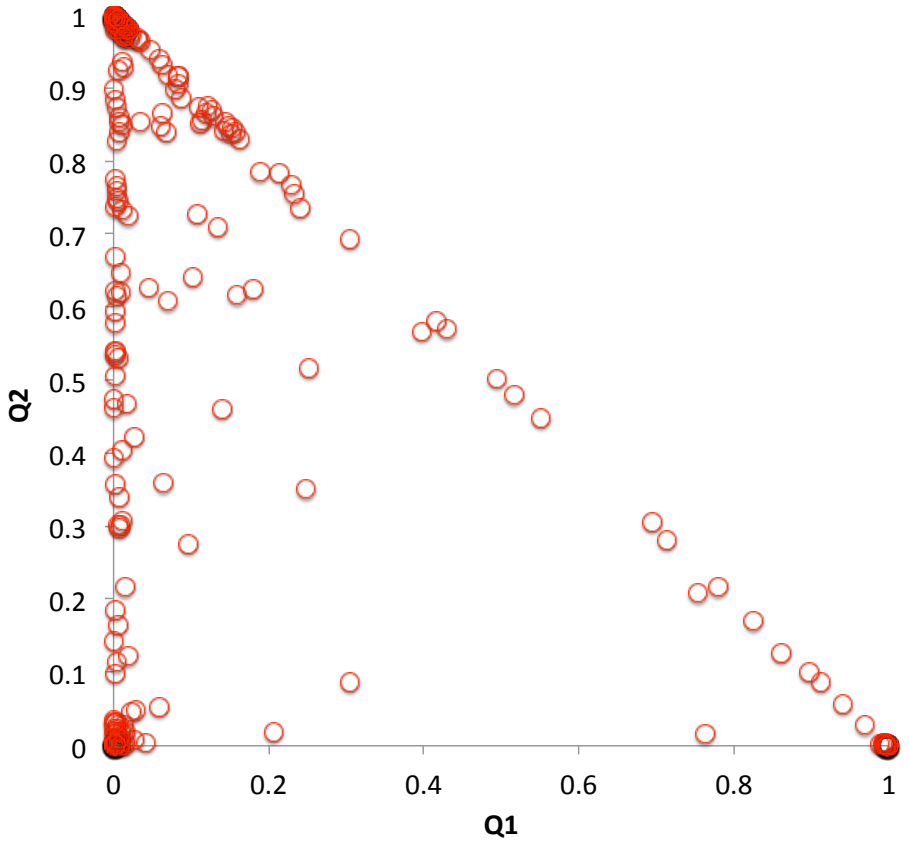
- Linkage analysis
- Association study
- Population structure and GLM
- Kinship and MLM
- BLINK



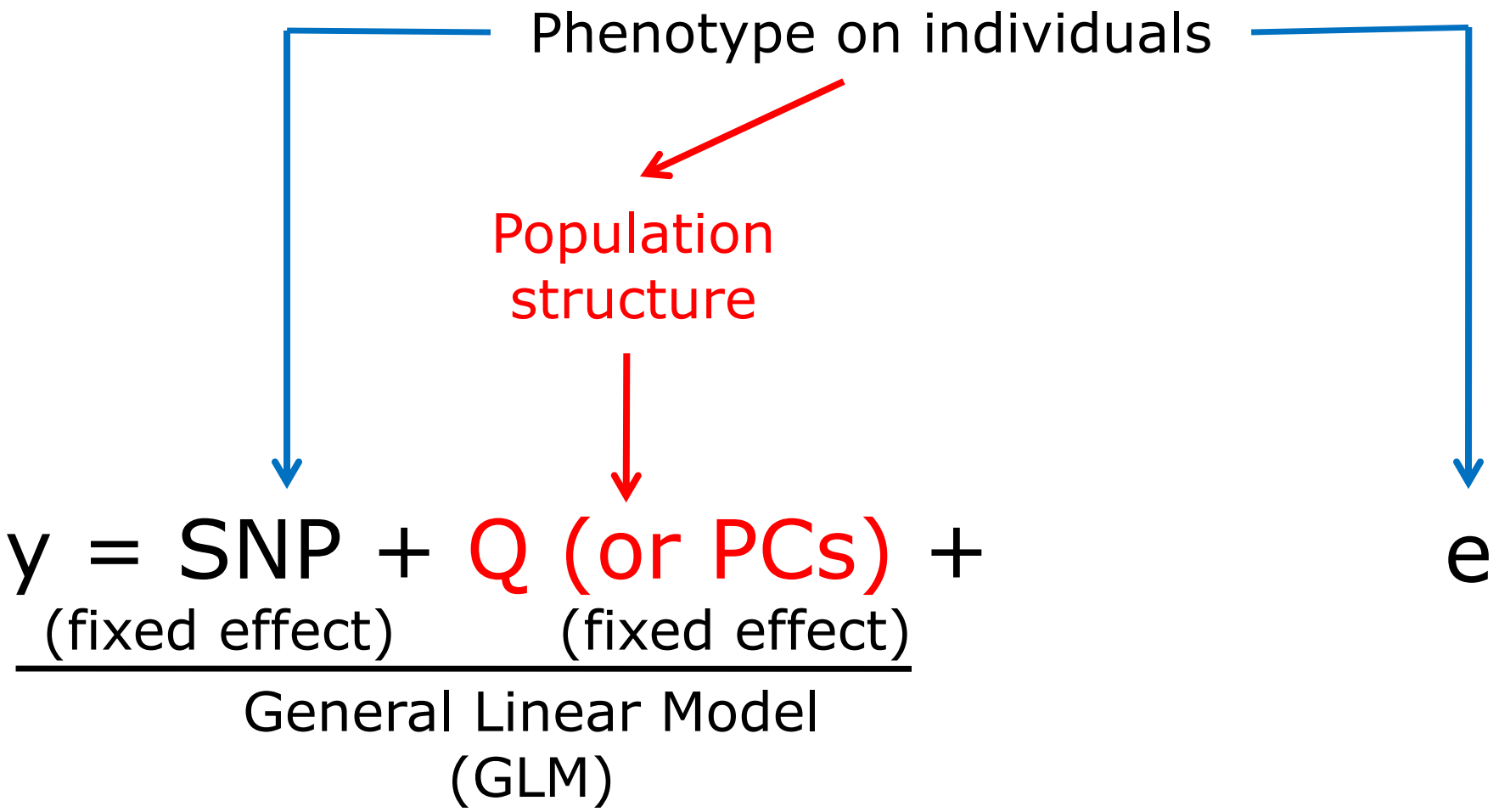
Population structure of maize



Taxa	Q1	Q2	Q3
33-16	0.014	0.972	0.014
38-11	0.003	0.993	0.004
4226	0.071	0.917	0.012
4722	0.035	0.854	0.111
A188	0.013	0.982	0.005
B73	0.999	0.001	1.10E-16
B73HTRHM	0.999	0.001	1.10E-16
B75	0.005	0.993	0.002
WD	0.014	0.97	0.016
WF9	0.005	0.994	0.001
YU796NS	0.189	0.785	0.026



GLM (Conceptual)



GLM on individuals

observation	mean	PC2	SNP
b=	bo	b1	b2
-4.709379	1	8.040247	2
-5.103188	1	4.824156	2
-2.782490	1	4.750749	2
-3.835722	1	-5.773005	2
-9.195871	1	-14.023364	2
-3.283042	1	-7.073483	2
-5.659523	1	8.636867	2
-4.264048	1	-15.491325	2
-10.486154	1	13.363734	0
-3.057630	1	-2.142841	2

	Ind1	Ind2	...	Ind9	Ind10
u=	u1	u2	...	u9	u10
	1	0	...	0	0
	0	1	...	0	0
	0	0	...	1	0
	0	0	...	0	1

y [1 x1 x2] =X Z

$$y = Xb + Zu + e$$

General linear model

$y=b_0 + x_1b_1 + x_2b_2 + \ldots + x_pb_p + e$

y: observation, dependent variable

x: Explainary/independent variables

e: Residuals/errors

$\Delta = e_1^2 + e_2^2 + \ldots + e_n^2$

$= e'e$

$= (y-Xb)'(y-Xb)$

Optimization to minimize residual

$$\Delta = e'e$$

$$= e^2 = (y - Xb)^2$$

$$\partial\Delta/\partial b$$

$$= 2X'(y - Xb)$$

$$= 2X'y - 2X'Xb = 0$$

$$X'Xb = X'y$$

$$b = [X'X]^{-1}[X'Y]$$

Statistical test

$$\hat{y} = X'\hat{b}$$

$$\sigma_e^2 = (y - \hat{y})'(y - \hat{y})/n$$

$$Var(\hat{b}) = [X'X]^{-1}\sigma_e^2$$

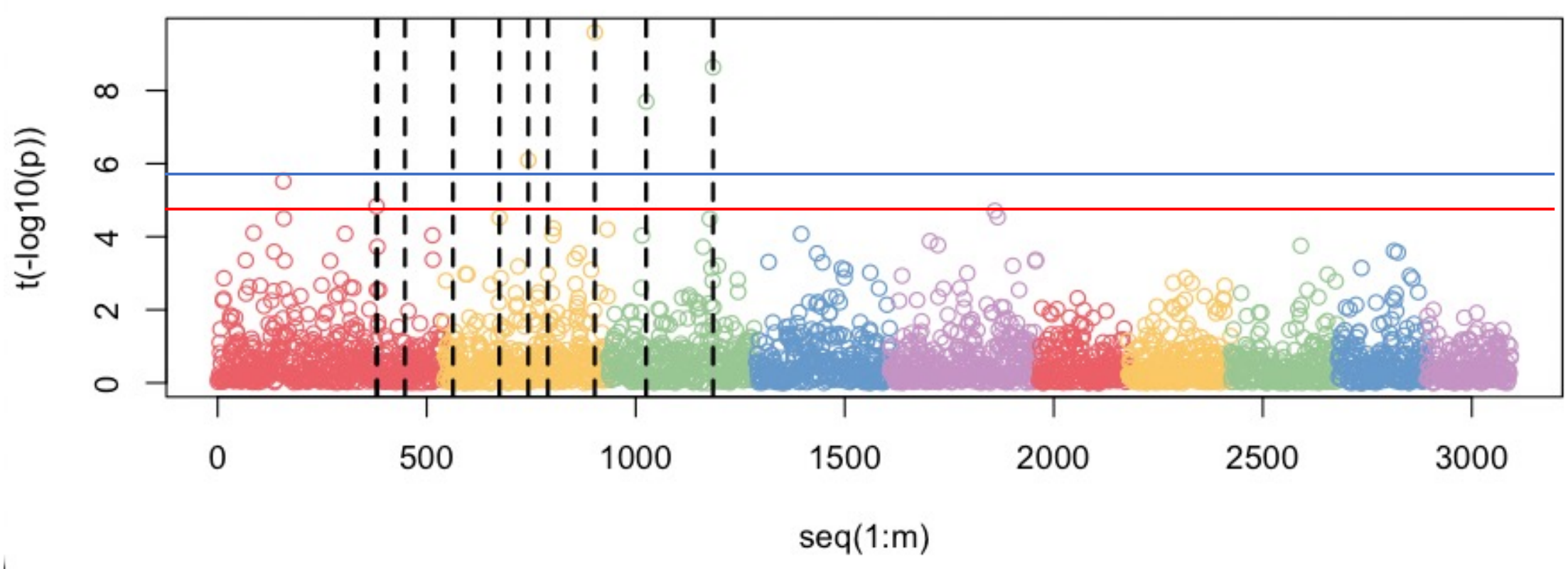
$$t = \hat{b}/\sqrt{Var(\hat{b})} \quad \sim t(n-1)$$

QTNs On CHR 1-5, leave 6-10 empty

```
myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)
source("http://zzlab.net/StaGen/2020/R/G2P.R")
source("http://zzlab.net/StaGen/2020/R/GWASbyCor.R")
X=myGD[,-1]
index1to5=myGM[,2]<6
X1to5 = X[,index1to5]
set.seed(99164)
mySim=G2P(X= X1to5,h2=.75,alpha=1,NQTN=10,distribution="norm")
p= GWASbyCor(X=X,y=mySim$y)
```

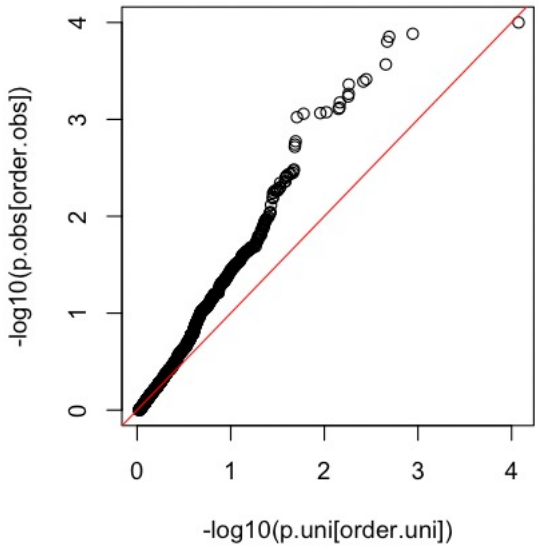
False positives

```
color.vector <- rep(c('#EC5f67', '#FAC863', '#99C794', '#6699CC', '#C594C5'),10)
m=nrow(myGM)
plot(t(-log10(p))~seq(1:m),col=color.vector[myGM[,2]])
abline(v=mySim$QTN.position, lty = 2, lwd=2, col = "black")
```



QQ plot

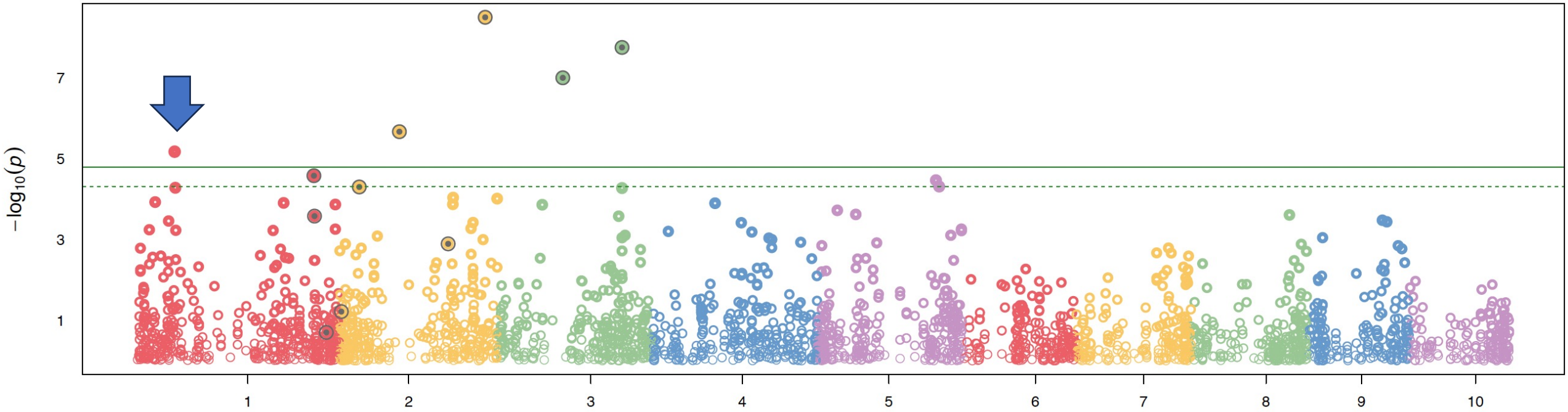
```
p.obs=p[!index1to5]
m2=length(p.obs)
p.uni=runif(m2,0,1)
order.obs=order(p.obs)
order.uni=order(p.uni)

plot(-log10(p.uni[order.uni]),-log10(p.obs[order.obs]))
abline(a = 0, b = 1, col = "red")
```



T Test

```
setwd("~/Desktop/temp")
myY=(cbind(myGD[,1], as.data.frame(mySim$y)))
source("http://zzlab.net/GAPIT/gapit_functions.txt")
```

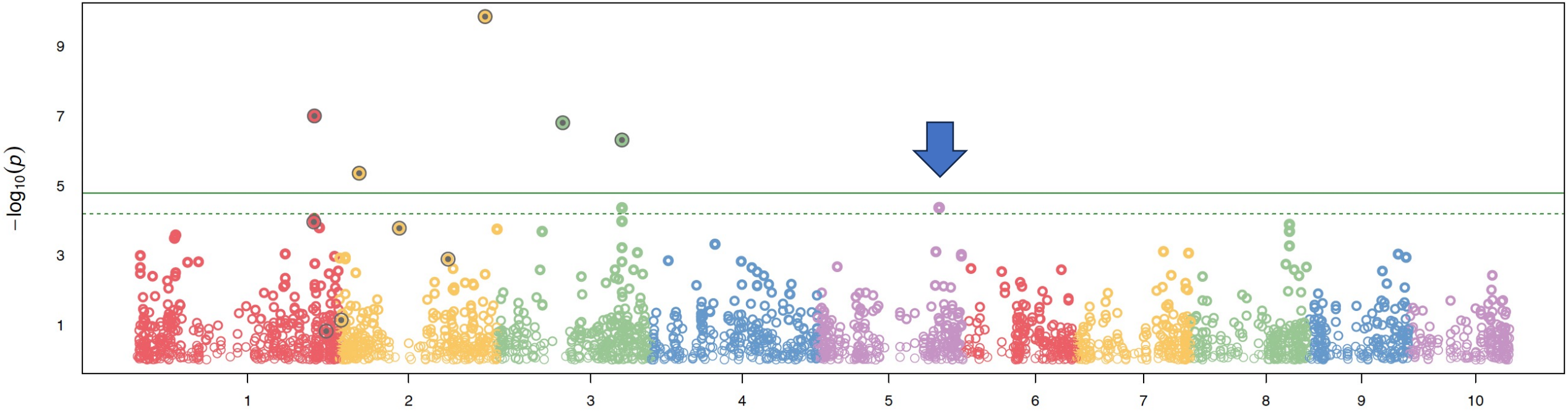
```
#GWAS by GAPIT
myGAPIT=GAPIT(
  Y=myY,
  GD=myGD,
  GM=myGM,
  QTN.position=mySim$QTN.position,
  PCA.total=0,
model="GLM",
memo="tTest")
```



```
setwd("~/Desktop/temp")
myY=(cbind(myGD[,1], as.data.frame(mySim$y)))
source("http://zzlab.net/GAPIT/gapit_functions.txt")
```

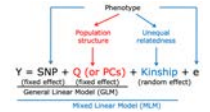
GLM

```
#GWAS by GAPIT
myGAPIT=GAPIT(
  Y=myY,
  GD=myGD,
  GM=myGM,
  QTN.position=mySim$QTN.position,
  PCA.total=3,
model="GLM",
memo="GLM")
```



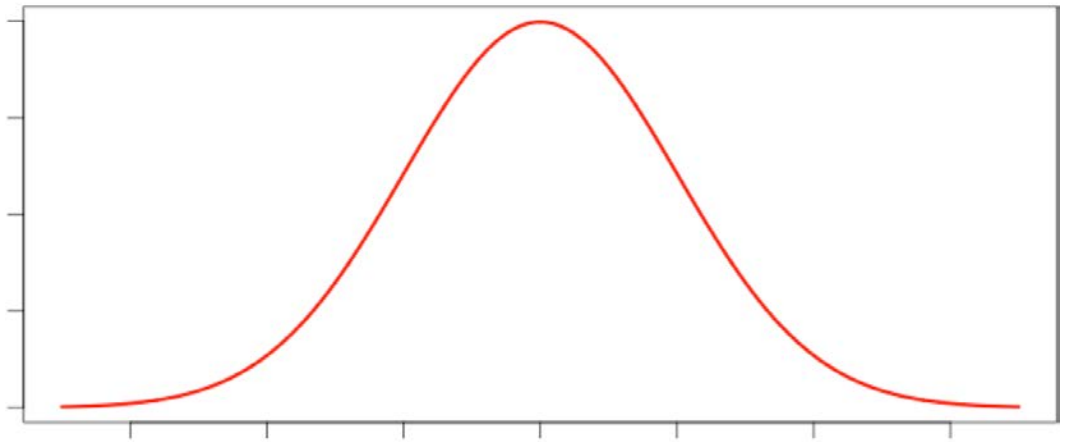
Outline

- Linkage analysis
- Association study
- Population structure and GLM
- Kinship and MLM
- BLINK



Minimize residual does not work for adding individuals' effects

- More parameters than observations
- Residuals are always zero due to over model fitting
- Needs new rules

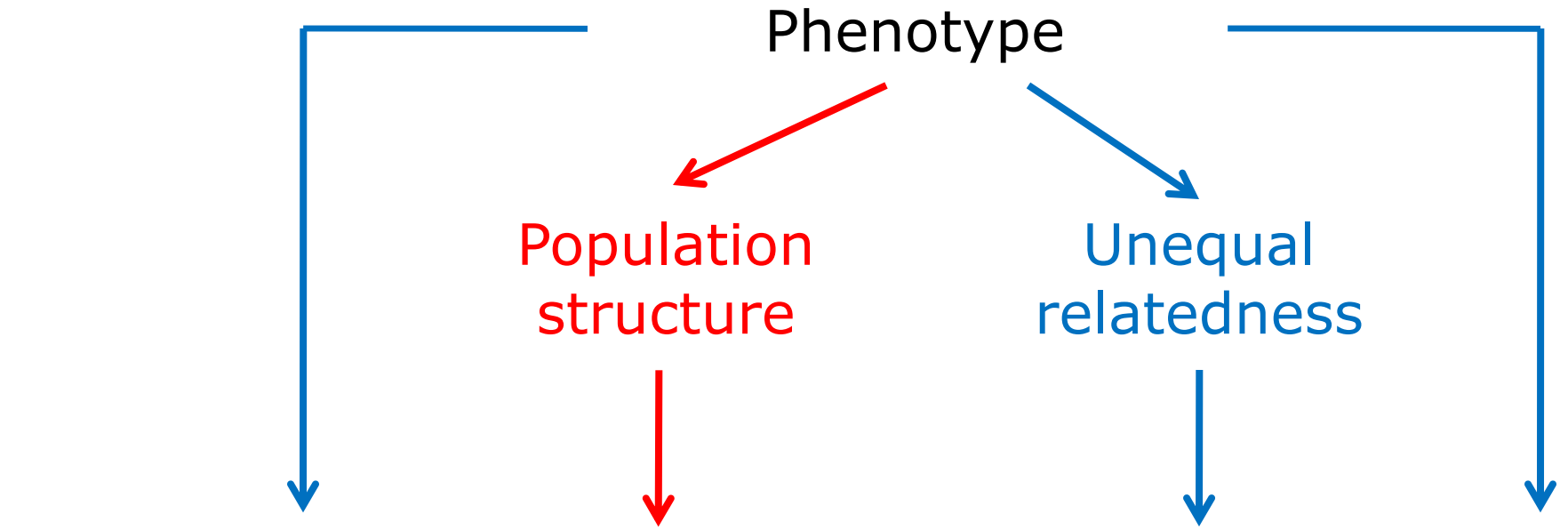


X

New rules

- Free individuals effects
- Only regulate the features for the population where they from: means (0) and variances
- Optimize variances according to kinship and observation to maximize the likelihood

MLM for GWAS



(Yu et al. 2005, Nature Genetics)

Mixed Linear Model (MLM)

$y = Xb + Zu + e$

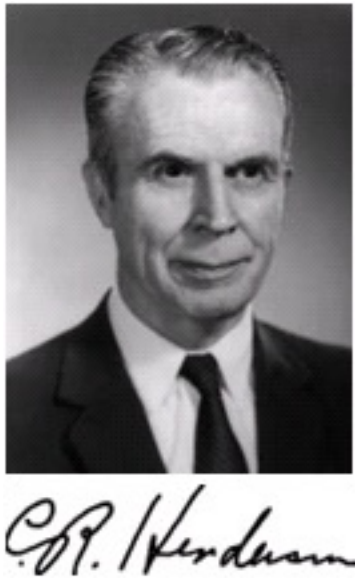
$Var(y)=V=Var(u)+Var(e)$

$Var(u)=G=2K\sigma_a^2$

$Var(e)=R=I\sigma_e^2$

u prediction: Best Linear Unbiased Prediction, BLUP)

b prediction: Best Linear Unbiased Estimate, BLUE)



Mixed Model Equation

$$y = Xb + Zu + e$$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_e^2}{\sigma_a^2} A^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ z'y \end{bmatrix}$$

$$\begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_e^2}{\sigma_a^2} A^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

$$\text{Var}(\begin{bmatrix} b \\ u \end{bmatrix}) = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_e^2}{\sigma_a^2} A^{-1} \end{bmatrix}^{-1} \sigma_e^2$$

SPAGeDi

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* 2: 618-620.

- Kinship coefficient
  - Loiselle et al. (1995)
  - Ritland (1996)
- Relationship coefficient
  - Queller & Goodnight (1989)
  - Hardy & Vekemans (1999)
  - Lynch & Ritland (1999)
  - Wang (2002);
- Genetic distance: Rousset (2000)

Identical by status

	AA	AT	TT
AA	1	.5	0
AT	.5	.5	.5
TT	0	.5	1

Proportion of shared alleles

	-1	0	1
-1	1	0	-1
0	0	0	0
1	-1	0	1

Genotype coding

Identical by status

Efficient algorithm

- M: n individual by m SNPs
- M: -1, 0 and 1
- Pi: frequency of 2nd allele for SNP i
- P: Column of i is 2(pi-.5)
- Z=M-P

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i(1-p_i)}$$

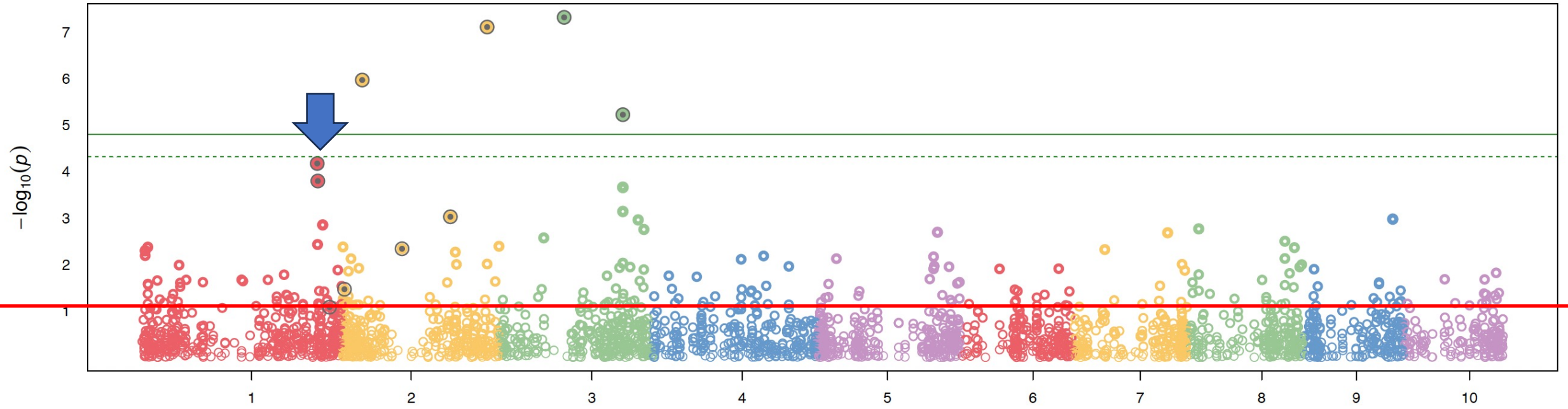
J. Dairy Sci. 2008. 91 (11) 4414-4423. Efficient Methods to Compute Genomic Predictions P. M. VanRaden

Paul VanRaden: Image Number K7168-6

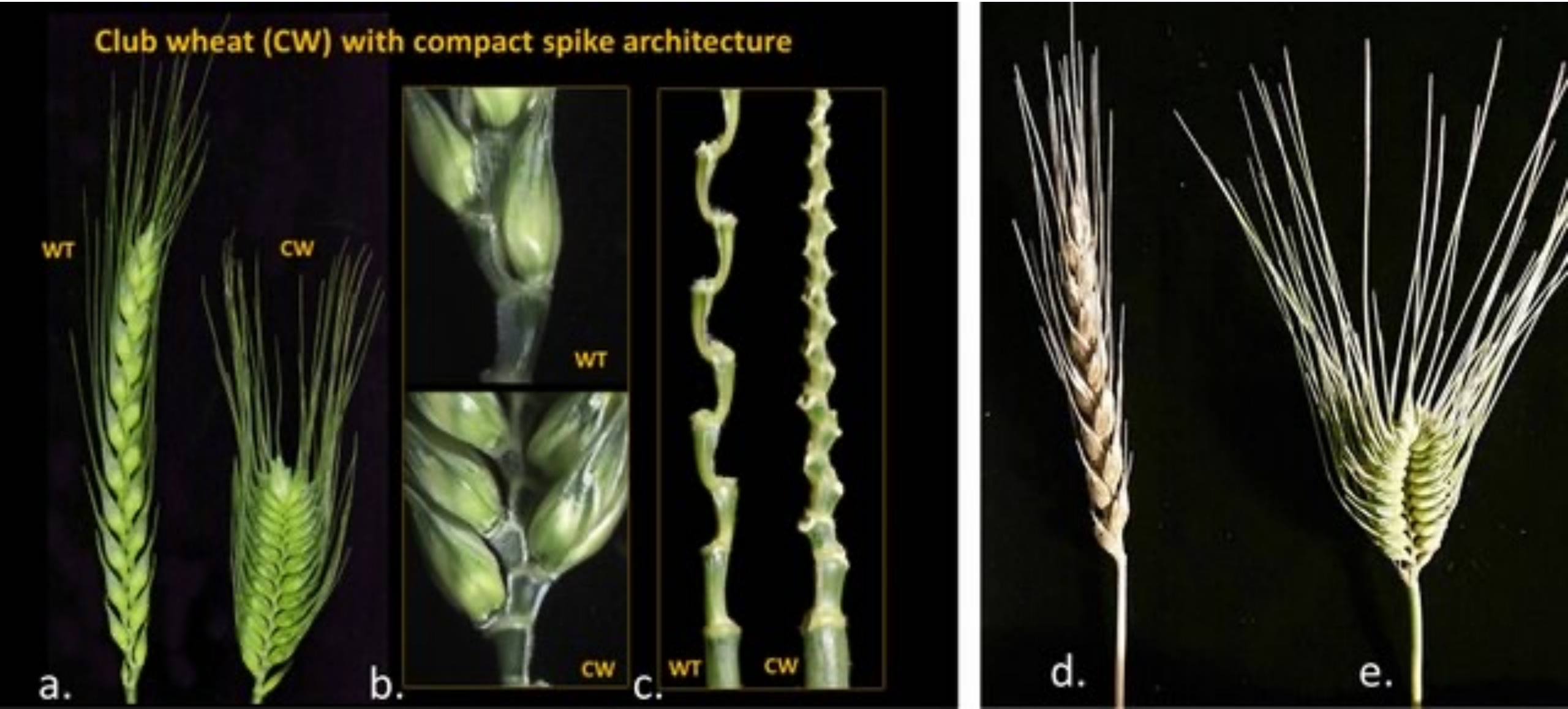
```
setwd("~/Desktop/temp")
myY=(cbind(myGD[,1], as.data.frame(mySim$y)))
source("http://zzlab.net/GAPIT/gapit_functions.txt")
```

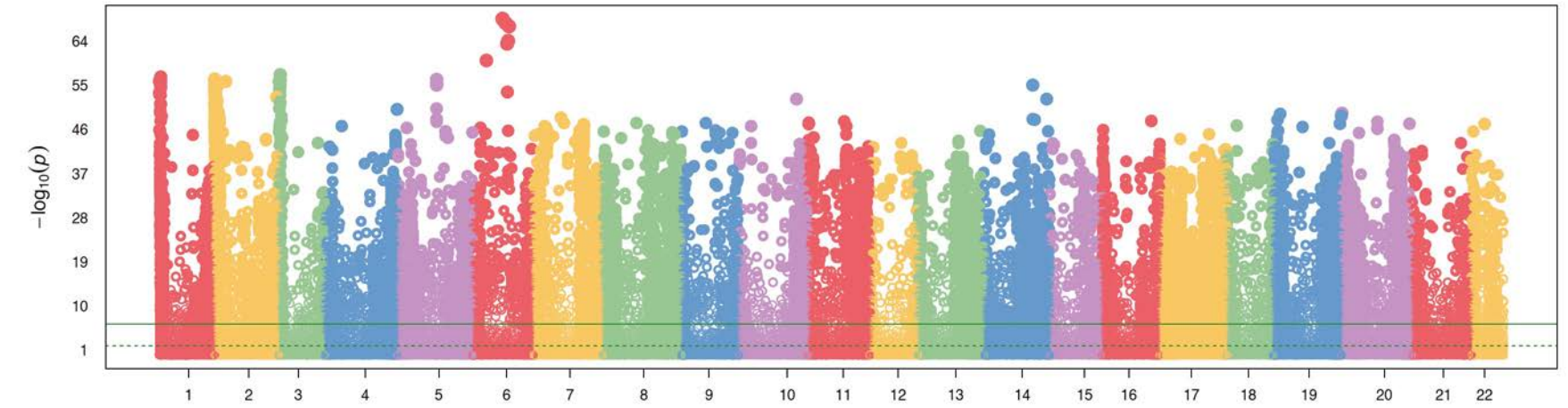
MLM

```
#GWAS by GAPIT
myGAPIT=GAPIT(
  Y=myY,
  GD=myGD,
  GM=myGM,
  QTN.position=mySim$QTN.position,
  PCA.total=3,
model="MLM",
memo="MLM_3PC")
```

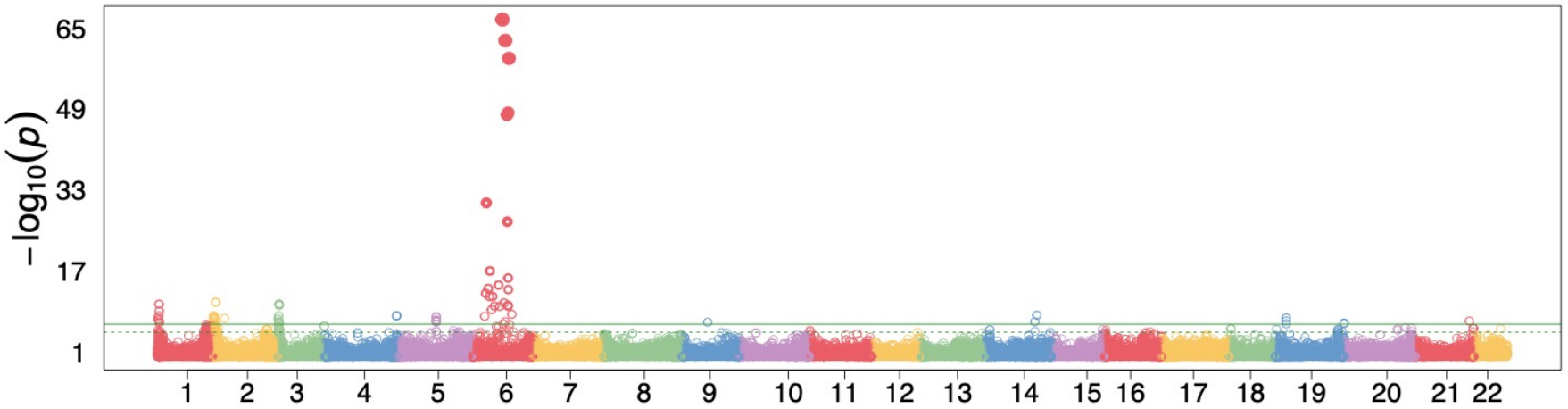


Physical mapping of the wheat genes in low-recombination regions: radiation hybrid mapping of the C-locus, Kajla and et al., TAG, 2023

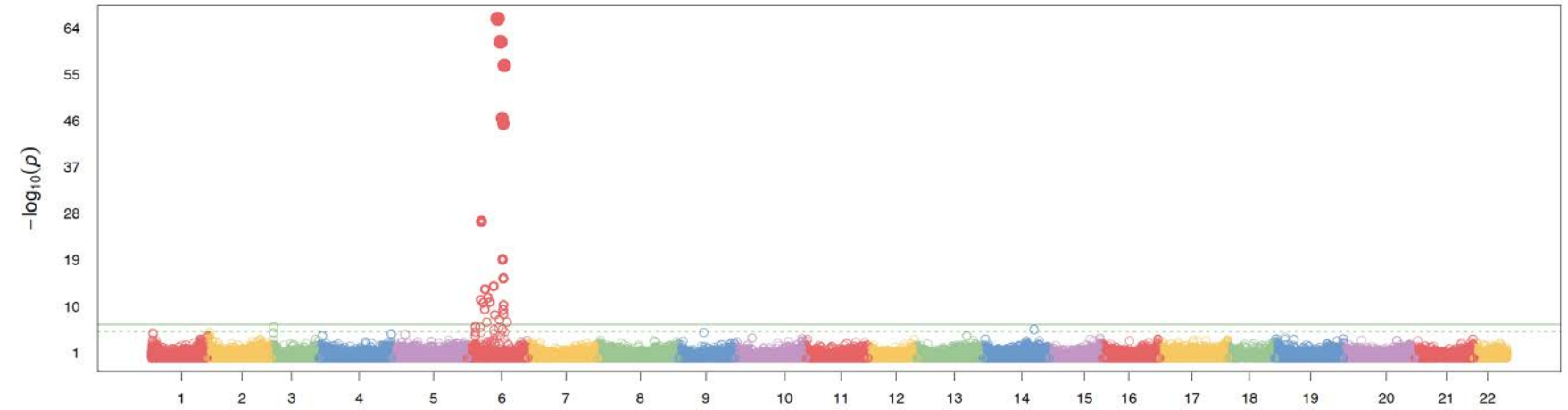




Y=S+E



Y=S+PC+E



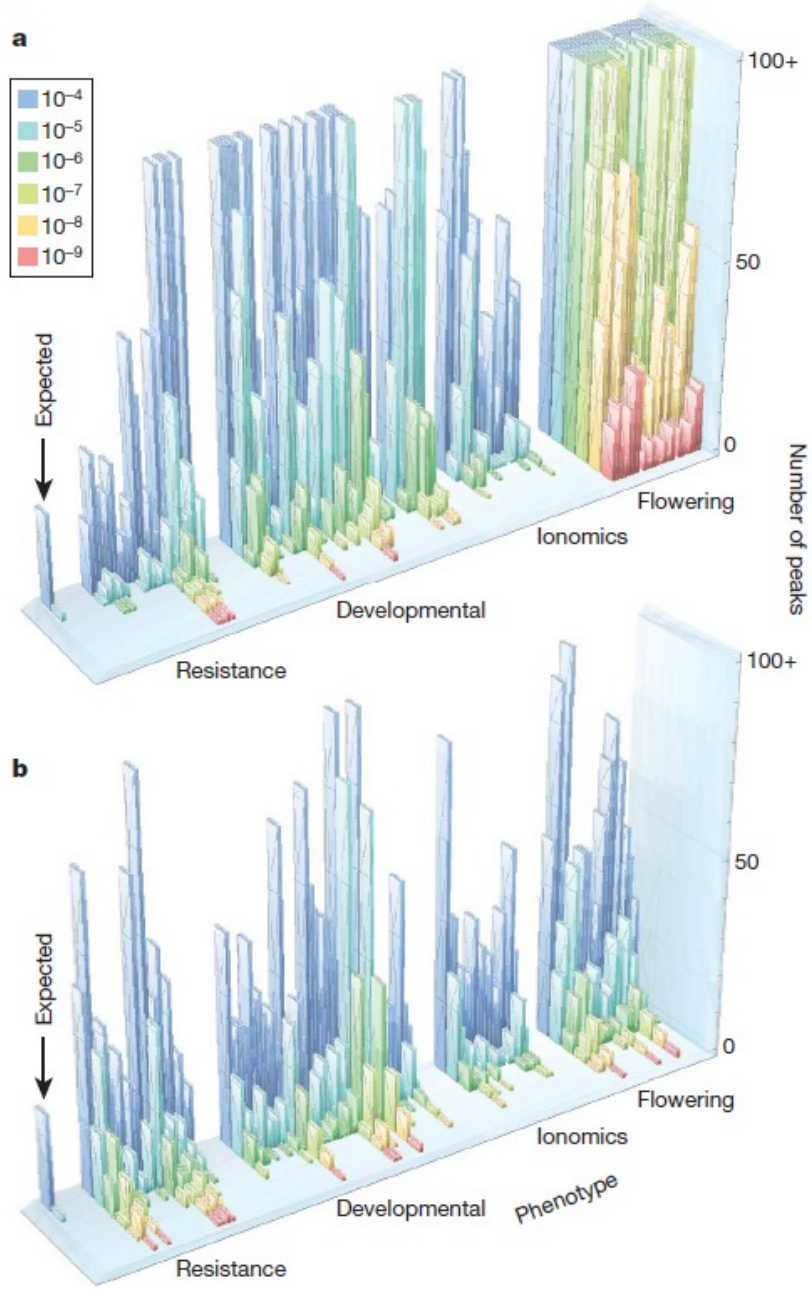
Y=S+PC+K+E

Atwell et al Nature 2010

a, No correction test

b, Correction with MLM

Magnus Norborg



GWAS does not work for traits associated with structure

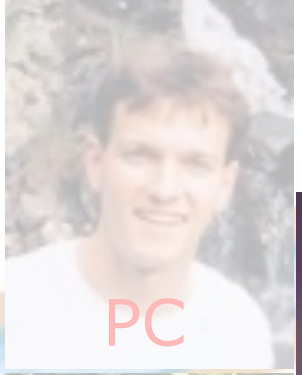
Queen + King

Flour>Water>Yeast>Salt


EMMAx


EMMA


PC+K

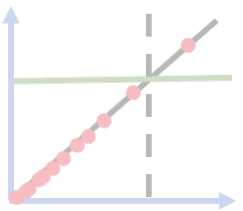

PC

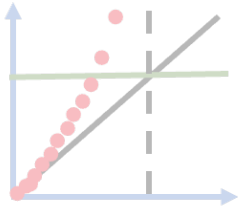

CMLM

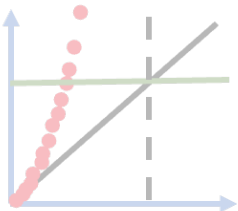

Q+K


Q


GWAS

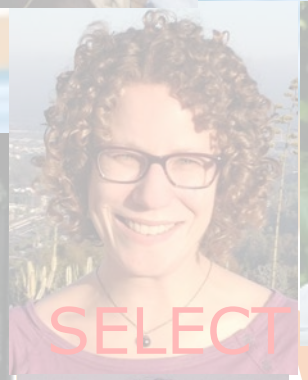


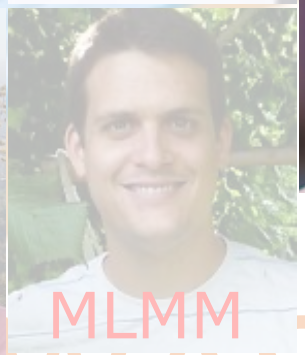





P3D


GCTA


SELECT


MLMM

WAS Stream


ECMLM


SUPER


FST-LMM


GEMMA


GenAbel

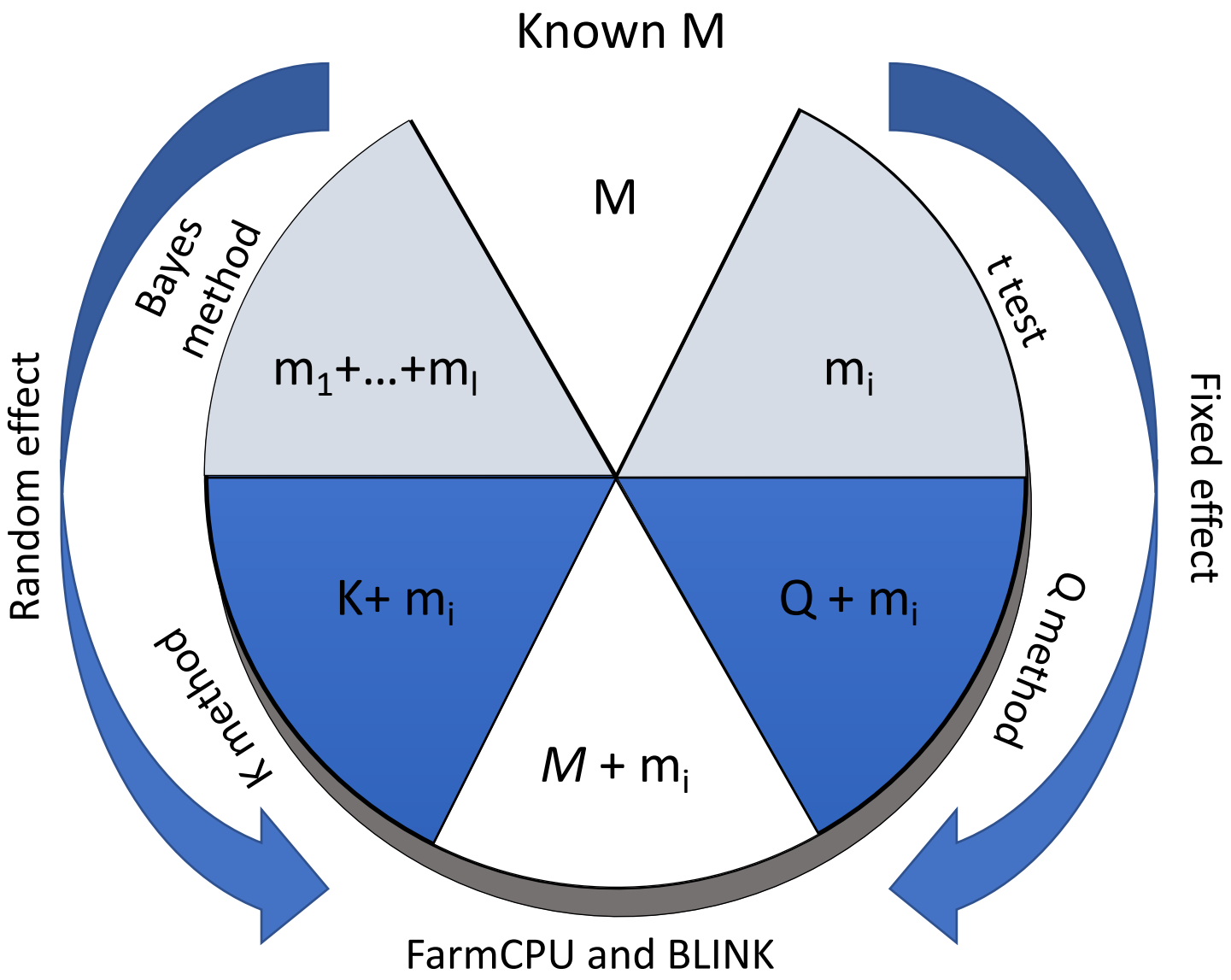

FarmCPU


BLINK

Outline

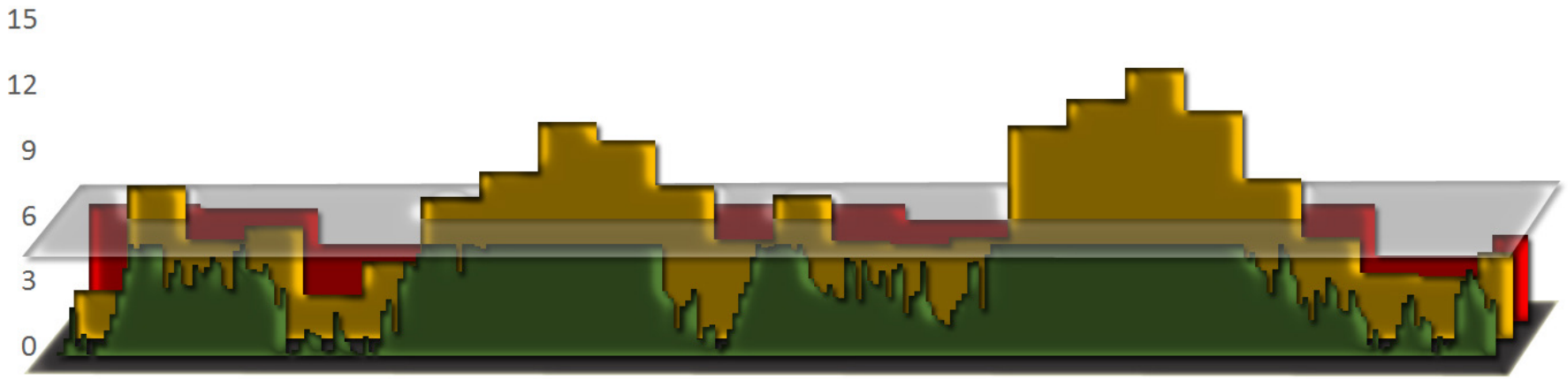
- Linkage analysis
- Association study
- Population structure and GLM
- Kinship and MLM
- BLINK

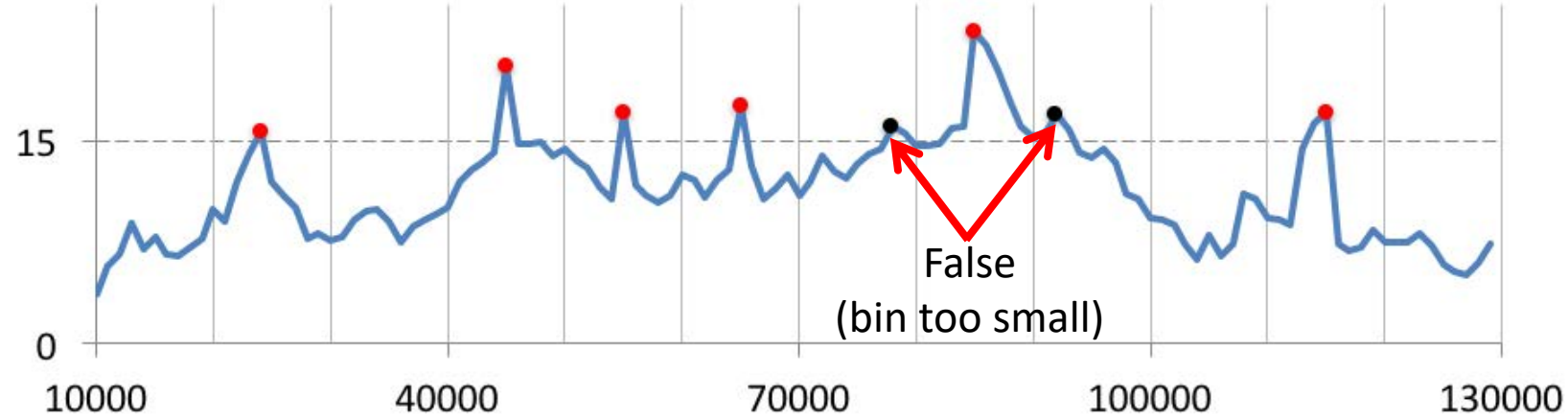
GWAS methods



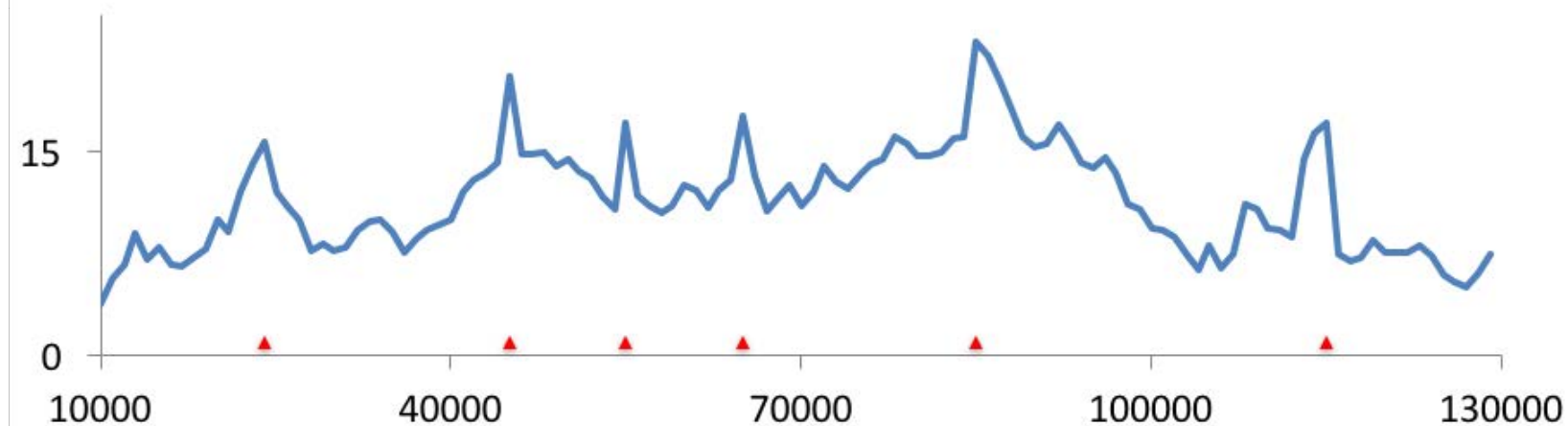
m: marker, M: Mutations, *M*: Estimated mutations

Bin approach

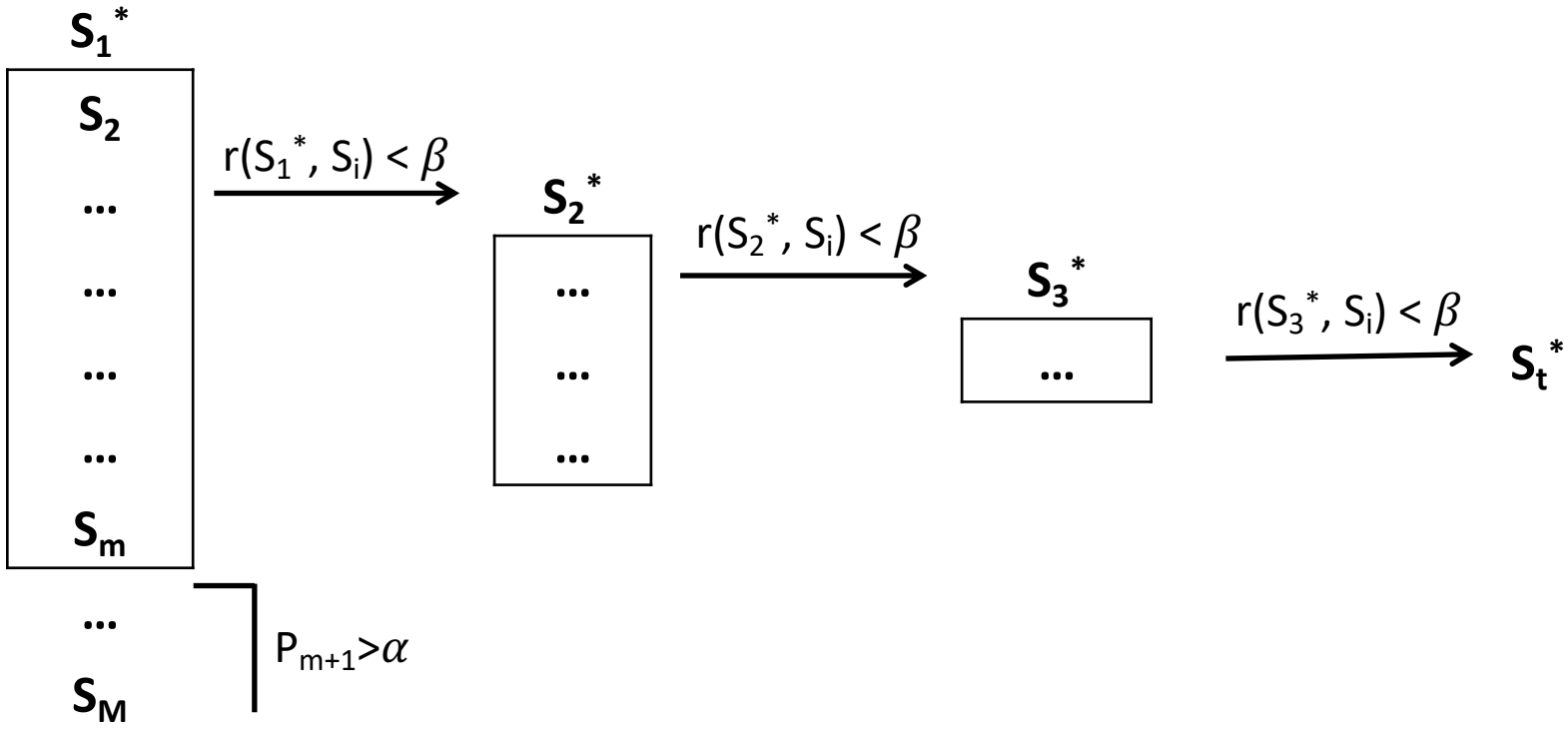



15
0
10000
40000
70000
100000
130000
False
(bin too small)


15
0
10000
40000
70000
100000
130000
Missing
(bin too big)


15
0
10000
40000
70000
100000
130000

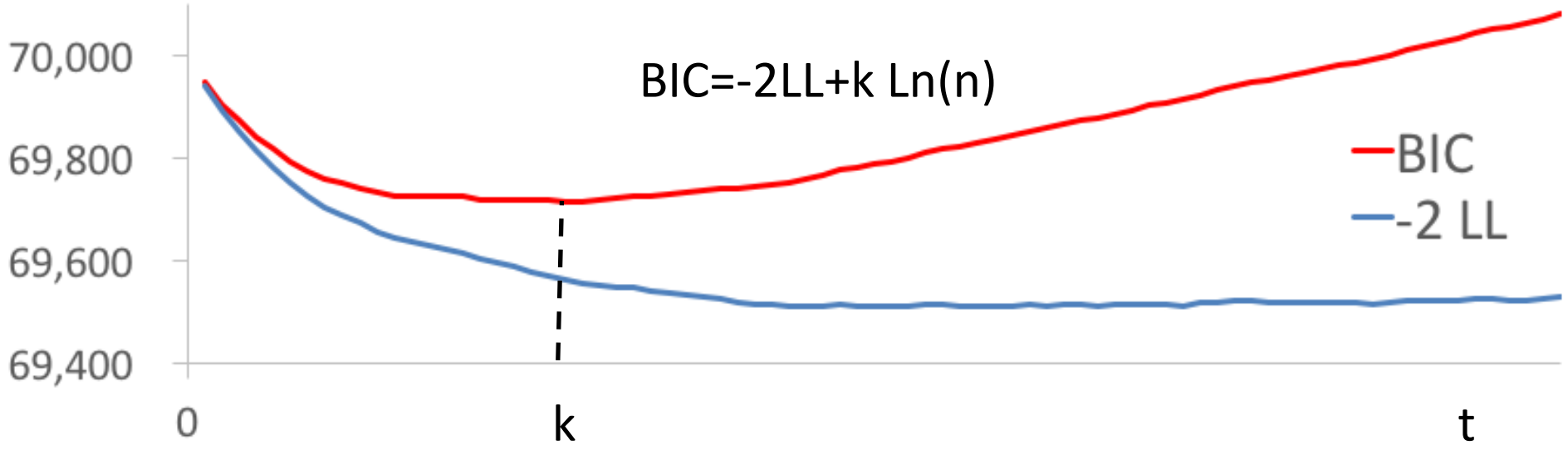
Elimination of markers with LD

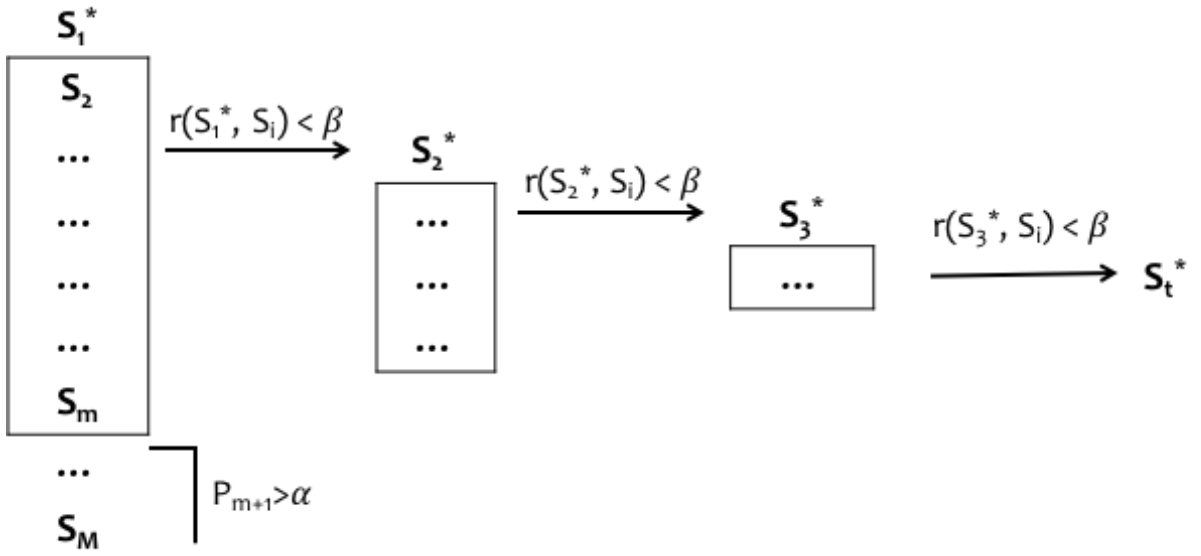


$$y = s_i + S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e, \text{ where } i = 1 \text{ to } M$$

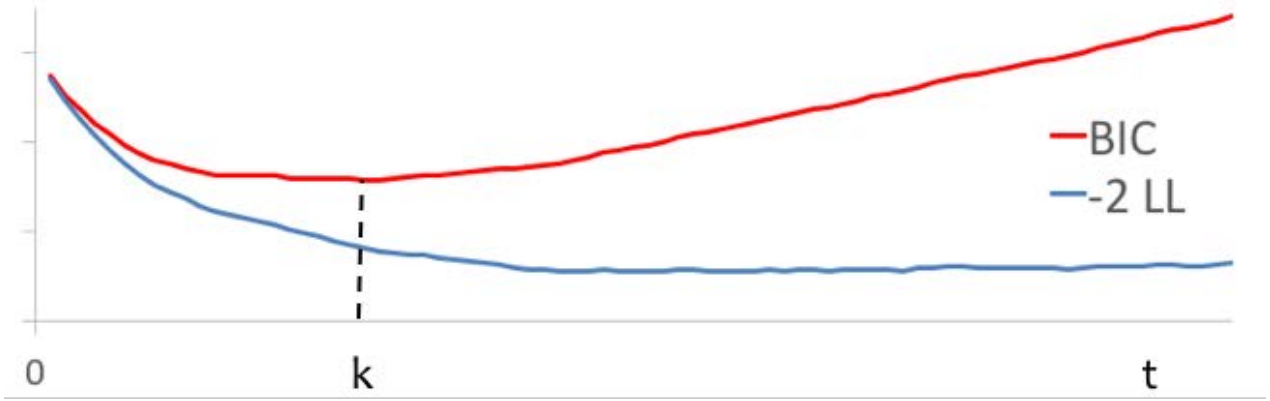
Bayesian information criterion

$y = S_1^* + S_2^* + S_3^* + \ldots + S_t^* + e$, where $k \leq t$ maximizes BIC





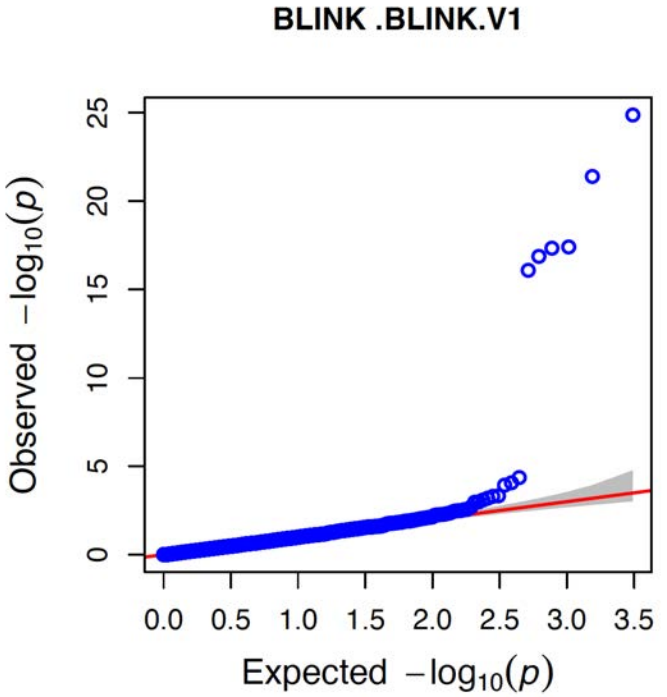
$y = S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e$, where $k \leq t$ maximizes BIC

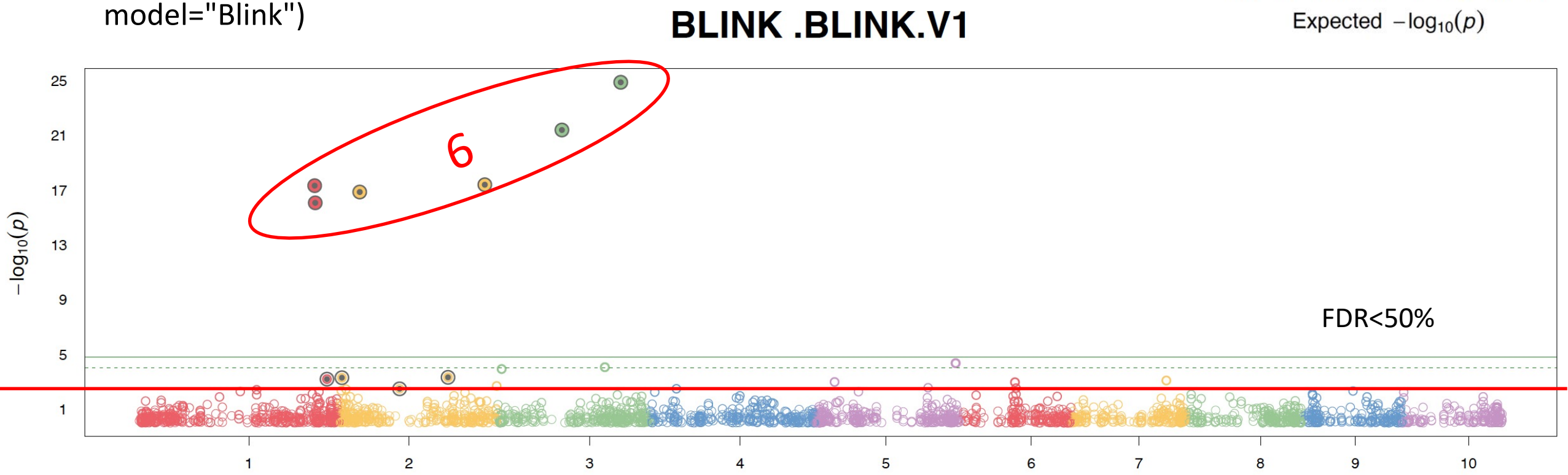


$y = s_i + S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e$, where i = 1 to M

BLINK (R in GAPIT)

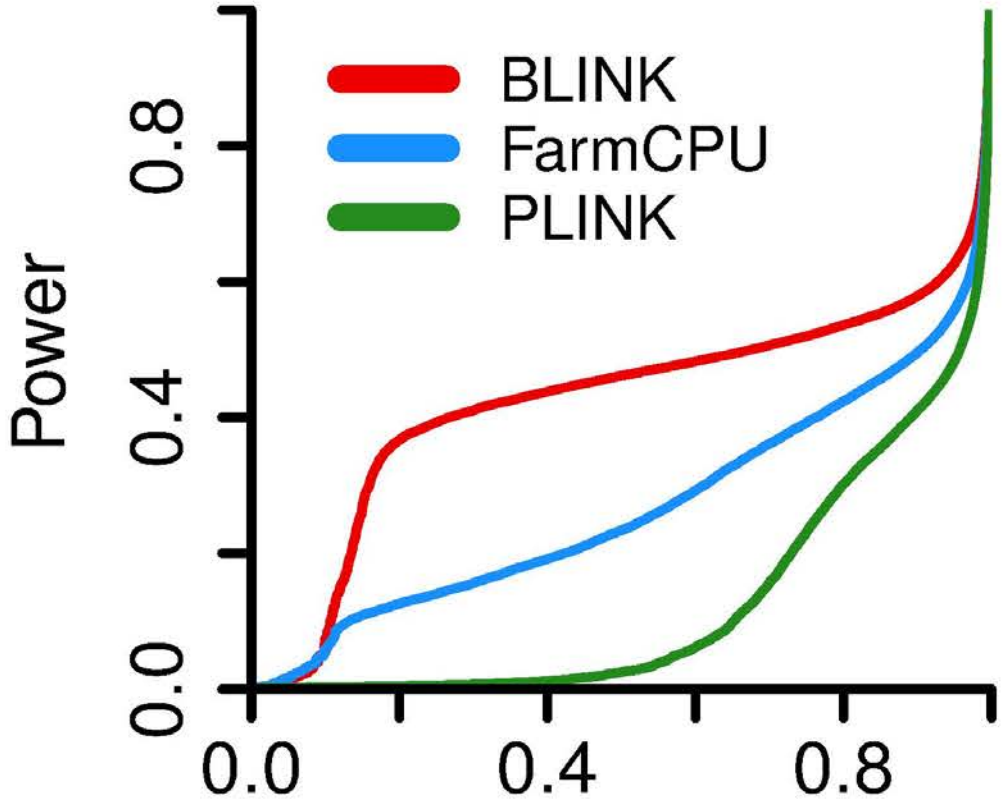
```
myGAPIT=GAPIT(
  Y=mySim$Y,
  GD=myGD,
  GM=myGM,
  QTN.position=mySim$QTN.position,
  PCA.total=3,
  model="Blink")
```



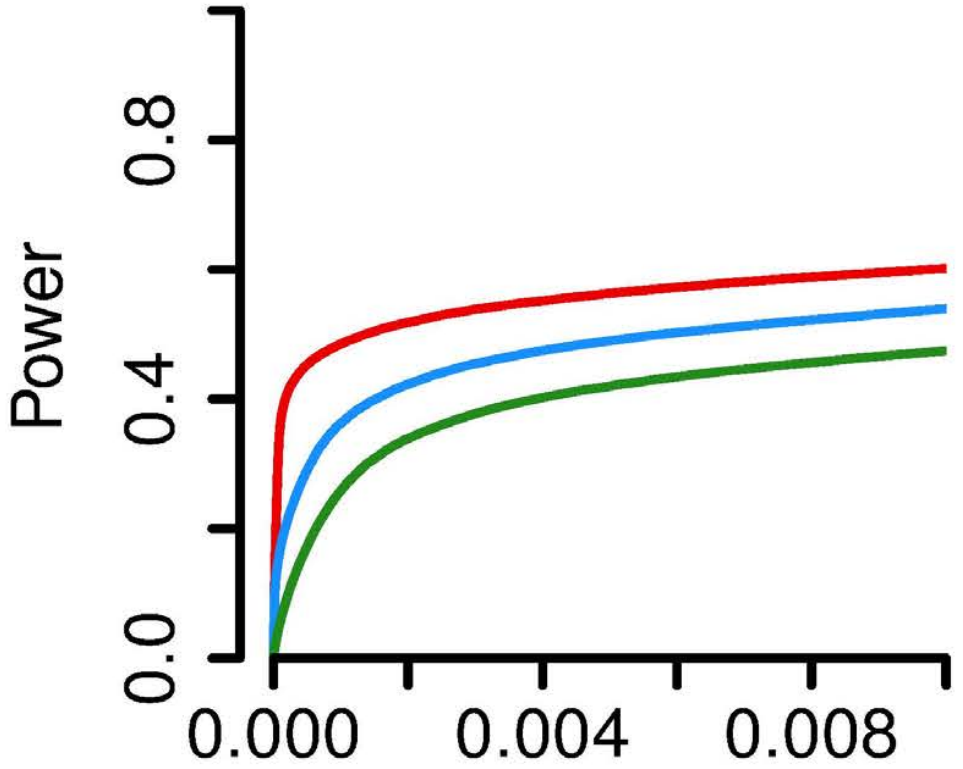


Simulation study with human data

Meng Huang

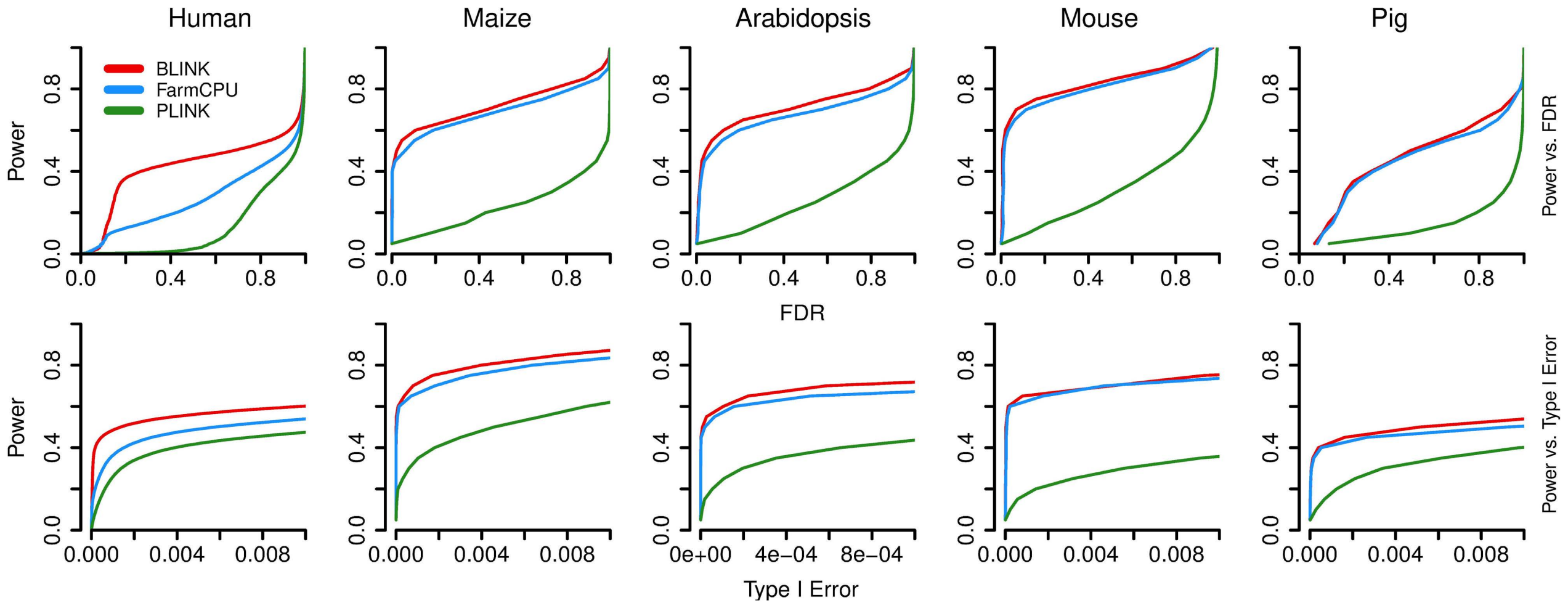


FDR

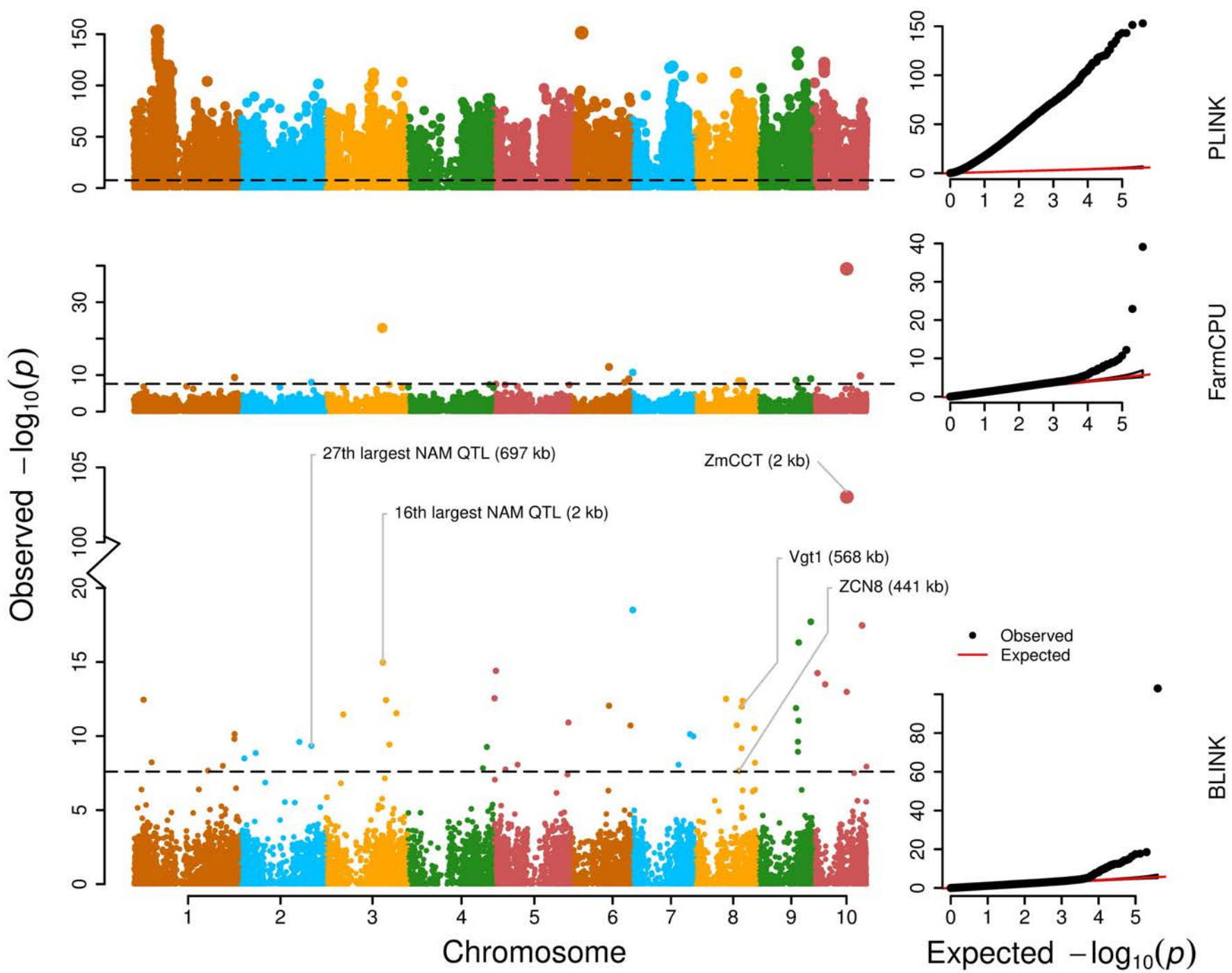


Type I error

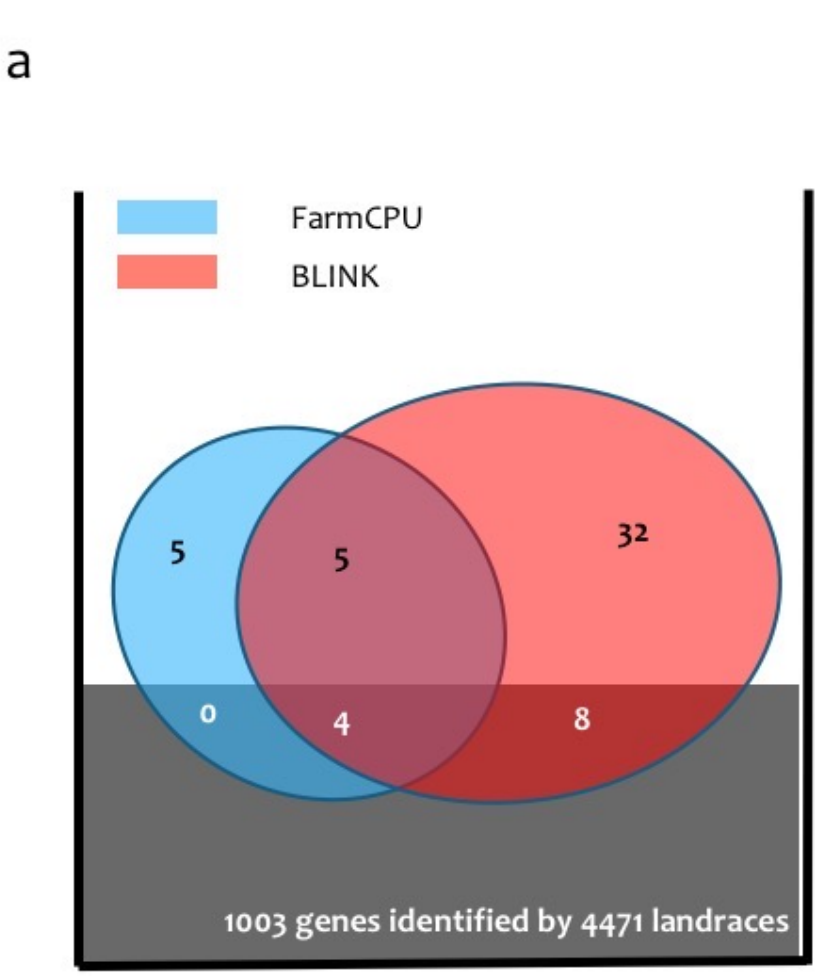
Same trend across species

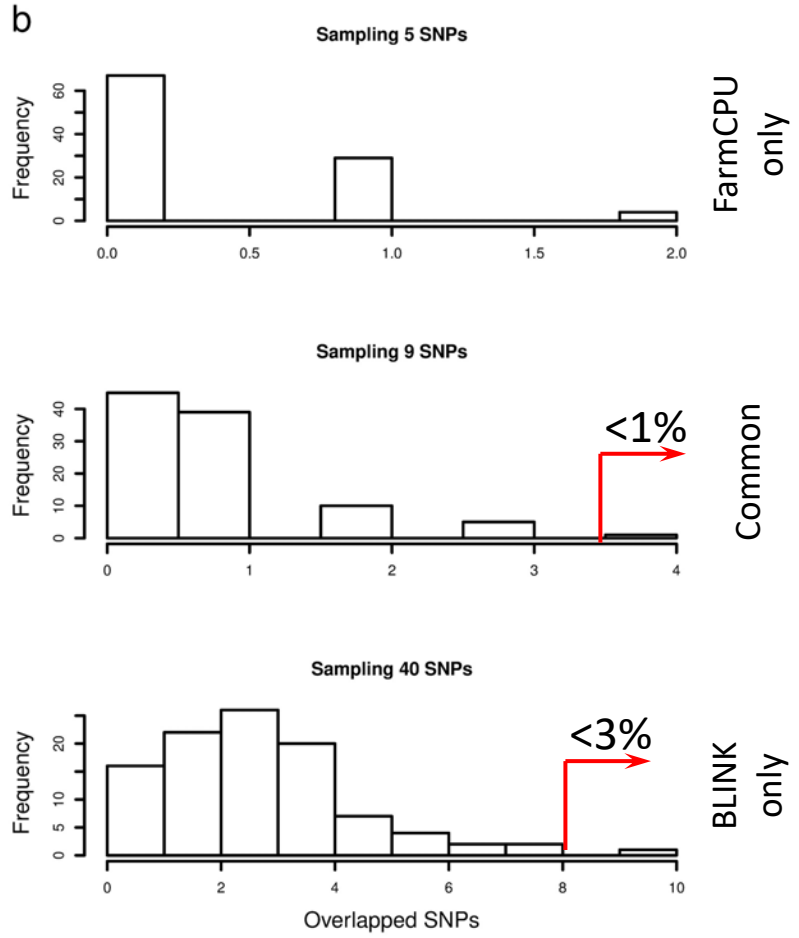


Application in Maize

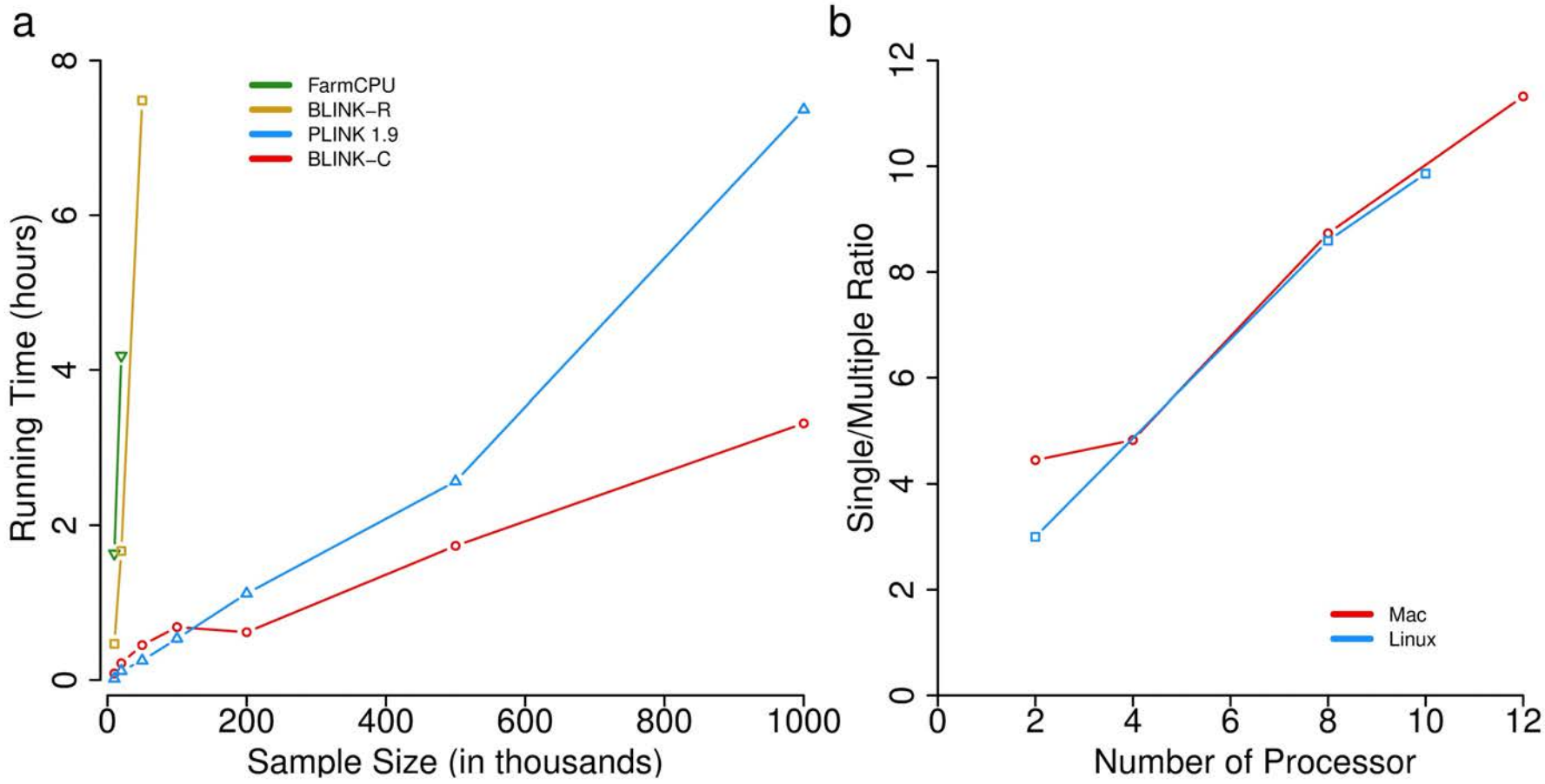


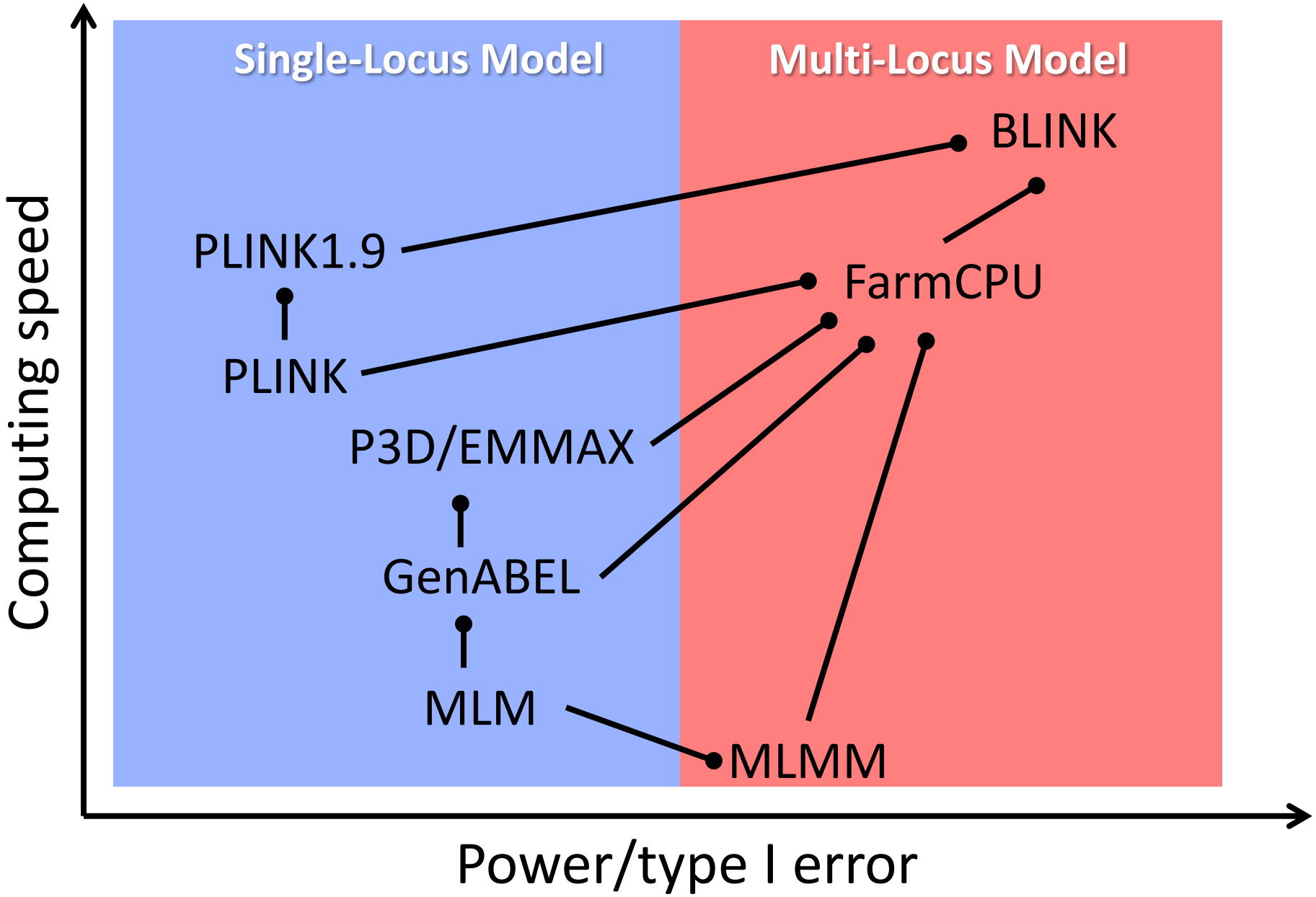
Enrichment


a
FarmCPU
BLINK
5
5
32
0
4
8
1003 genes identified by 4471 landraces

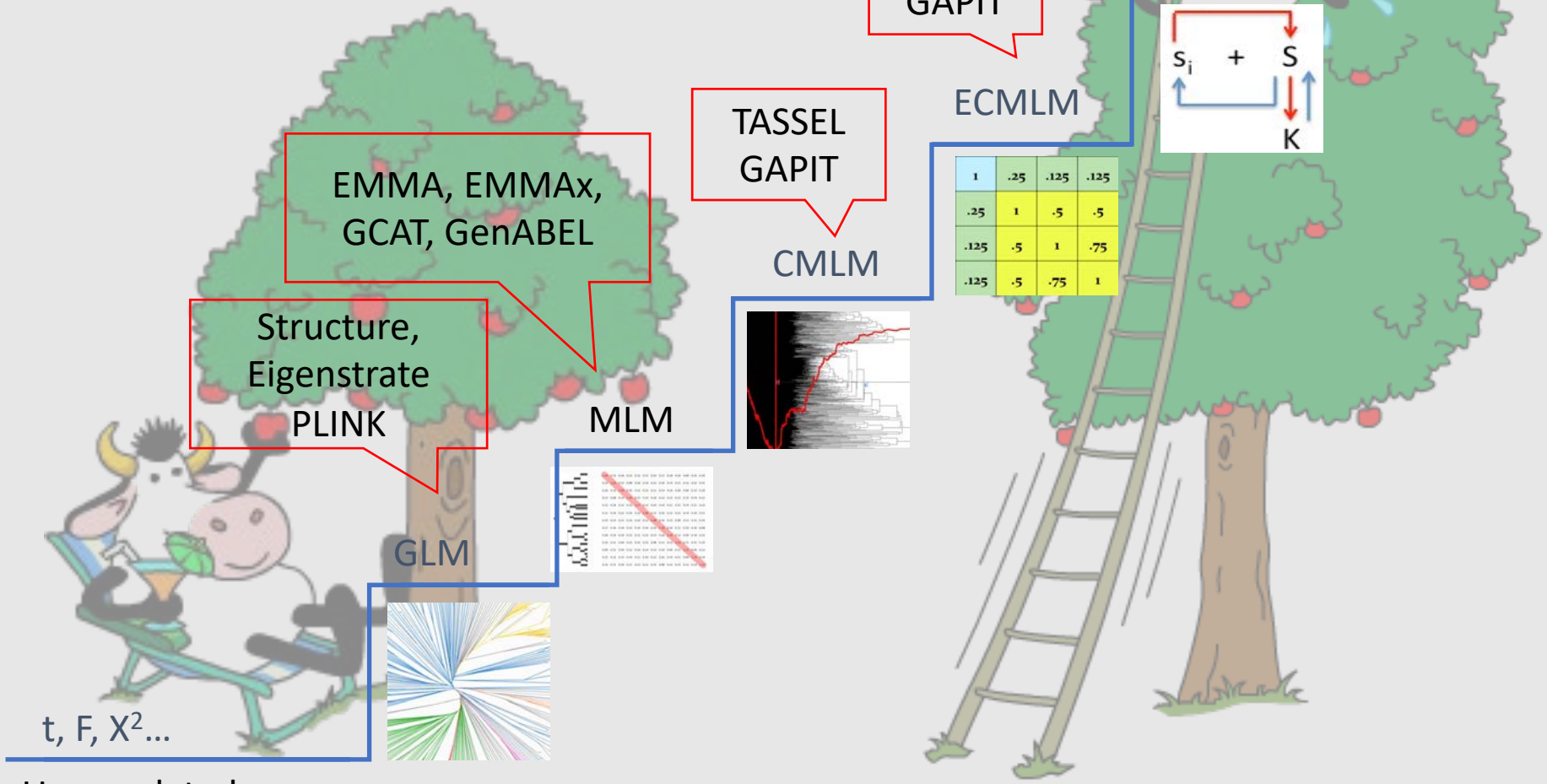

b
Sampling 5 SNPs
Frequency
FarmCPU only
Sampling 9 SNPs
<1%
Common
Sampling 40 SNPs
<3%
BLINK only
Overlapped SNPs

Computation efficiency




Single-Locus Model
Multi-Locus Model
BLINK
PLINK1.9
FarmCPU
PLINK
P3D/EMMAX
GenABEL
MLM
MLMM
Computing speed
Power/type I error

GAPIT

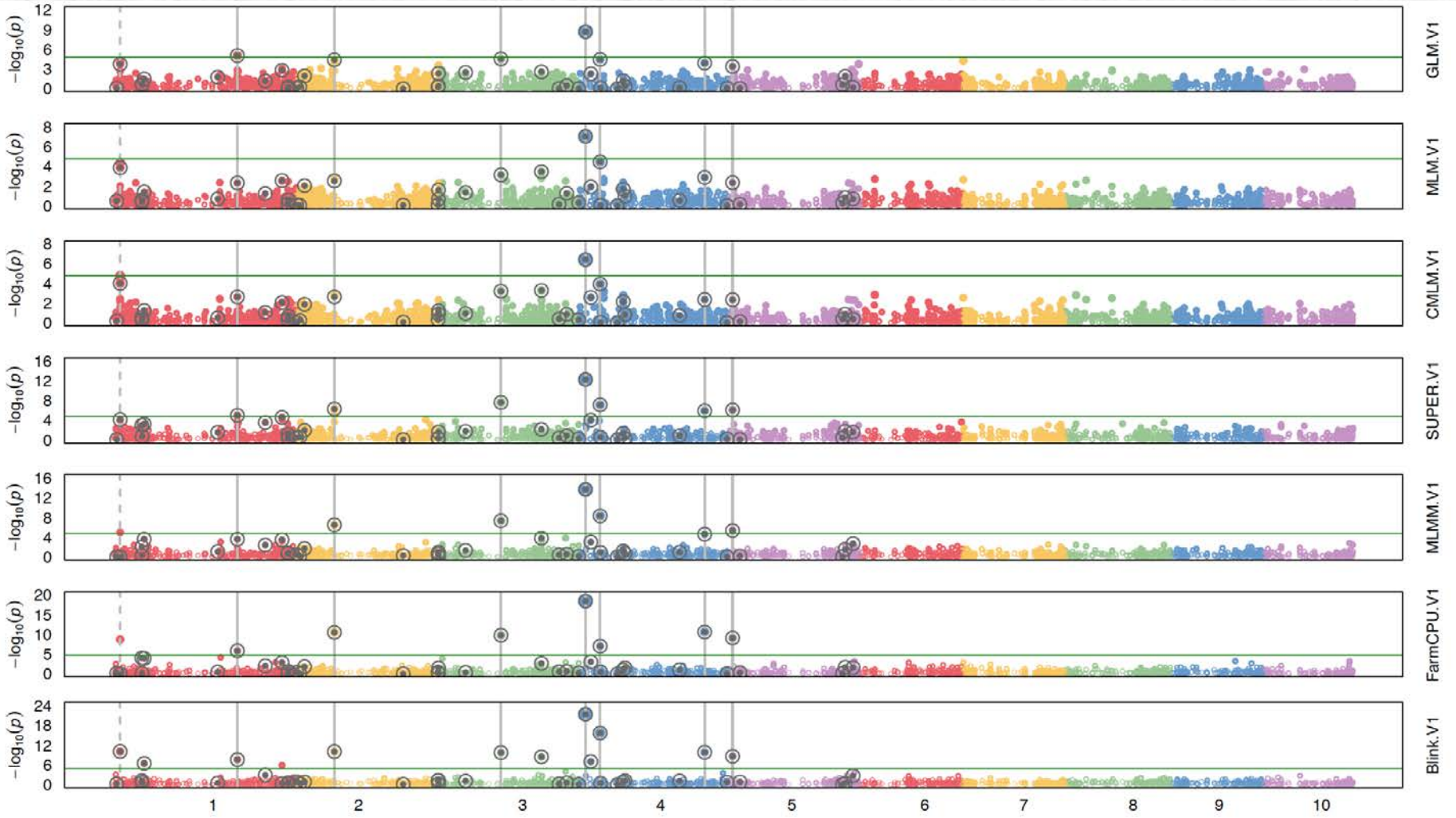

BLINK
FARM CPU
GAPIT
ECMLM
TASSEL
GAPIT
CMLM
EMMA, EMMAx,
GCAT, GenABEL
Structure,
Eigenstrate
PLINK
MLM
GLM
t, F, X2…
Uncorrelated or
equally correlated

iPat

Zhiwu Zhang Laboratory
for Statistical Genomics

Home People Publication Research Teaching Software Outreach Jobs

```
source("http://zzlab.net/GAPIT/gapit_functions.txt") #Import demo data
myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#Simultate 10 QTN on the first half chromosomes
index1to5=myGM[,2]<6 set.seed(99164)
mySim=GAPIT.Phenotype.Simulation(GD=myGD[,c(TRUE,index1to5)],GM=myGM[index1to5,],h2=.7,NQTN=40, effectunit=.95,QTNDist="normal")

#GWAS with GAPIT
myGAPIT=GAPIT(Y=mySim$Y,GD=myGD,GM=myGM,PCA.total=3,
QTN.position=mySim$QTN.position,
model=c("GLM", "MLM", "CMLM", "SUPER", "MLMM", "FarmCPU", "Blink"))
```



Collaborators and funding


Arron Carter


Mike Pumphrey


Karen Sanguinet


Kawamu Tanaka


Sindhuja Sankaran


Longxi Yu


Jack Brown


Ananth Kalyanaraman


Kim Campbell


Deven See


Camille Steber


Mike Peel

USDA

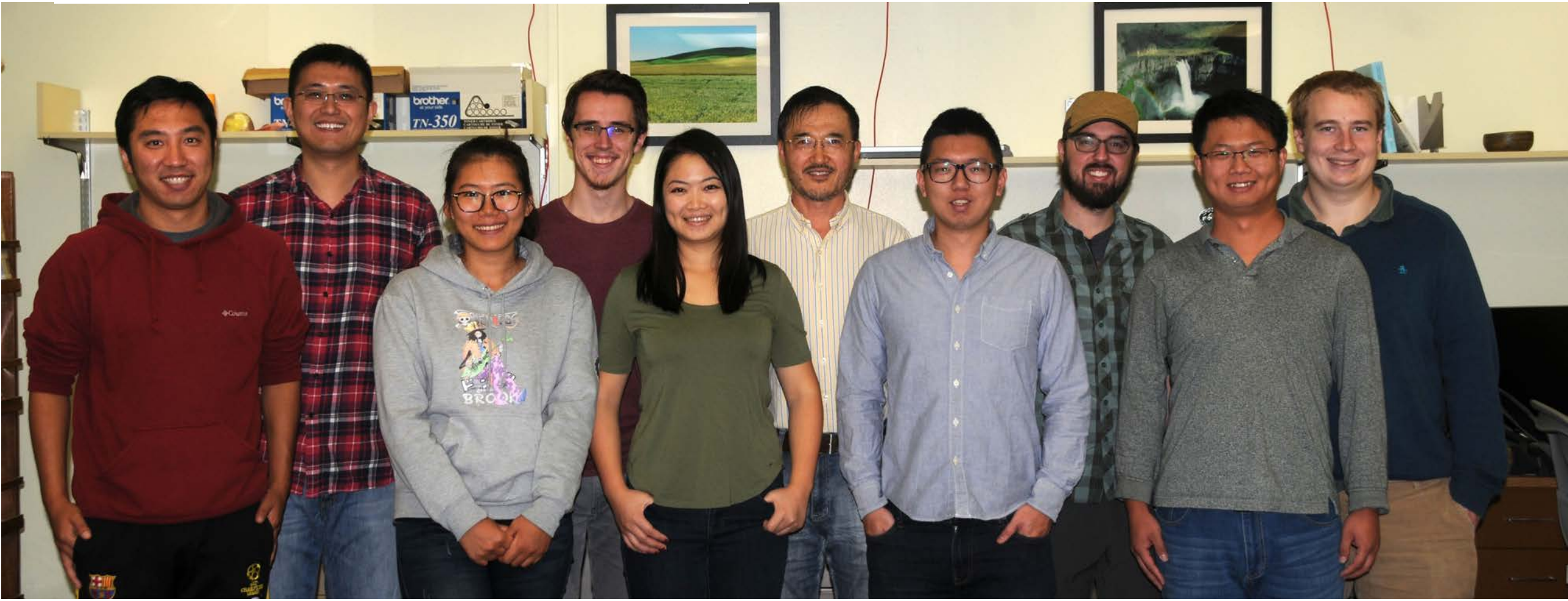
DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

NSF

M.J. MURDOCK CHARITABLE TRUST

Zhiwu Zhang Laboratory
for Statistical Genomics

WASHINGTON STATE UNIVERSITY



USDA

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

NSF

M.J. MURDOCK CHARITABLE TRUST


Thank you for your attention!