Untangling False Positives, Statistical Power, Populations Structure, and Kinship in GWAS

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

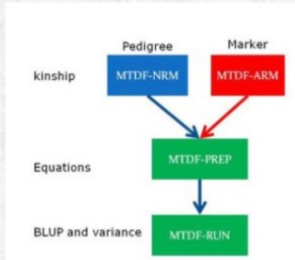
iPat

FarmCPU

Rooster

LADDER

mMAP



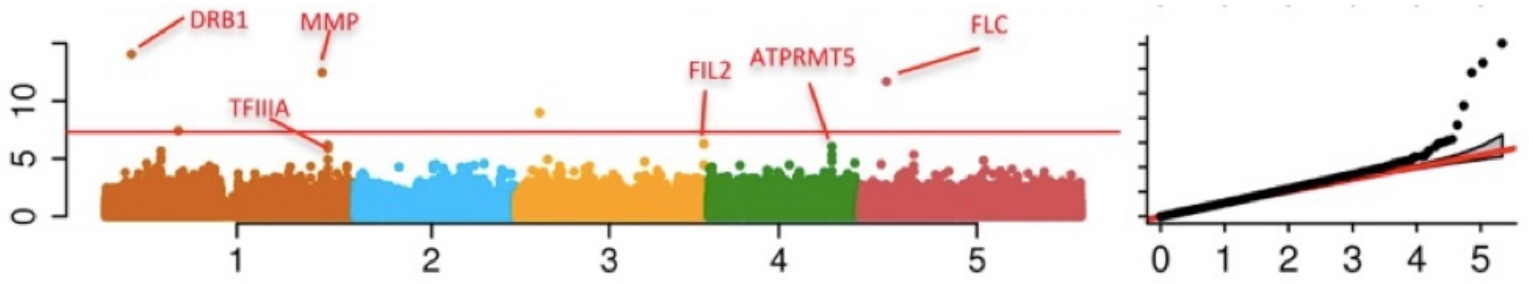
MTDFREML

Audio4EDU

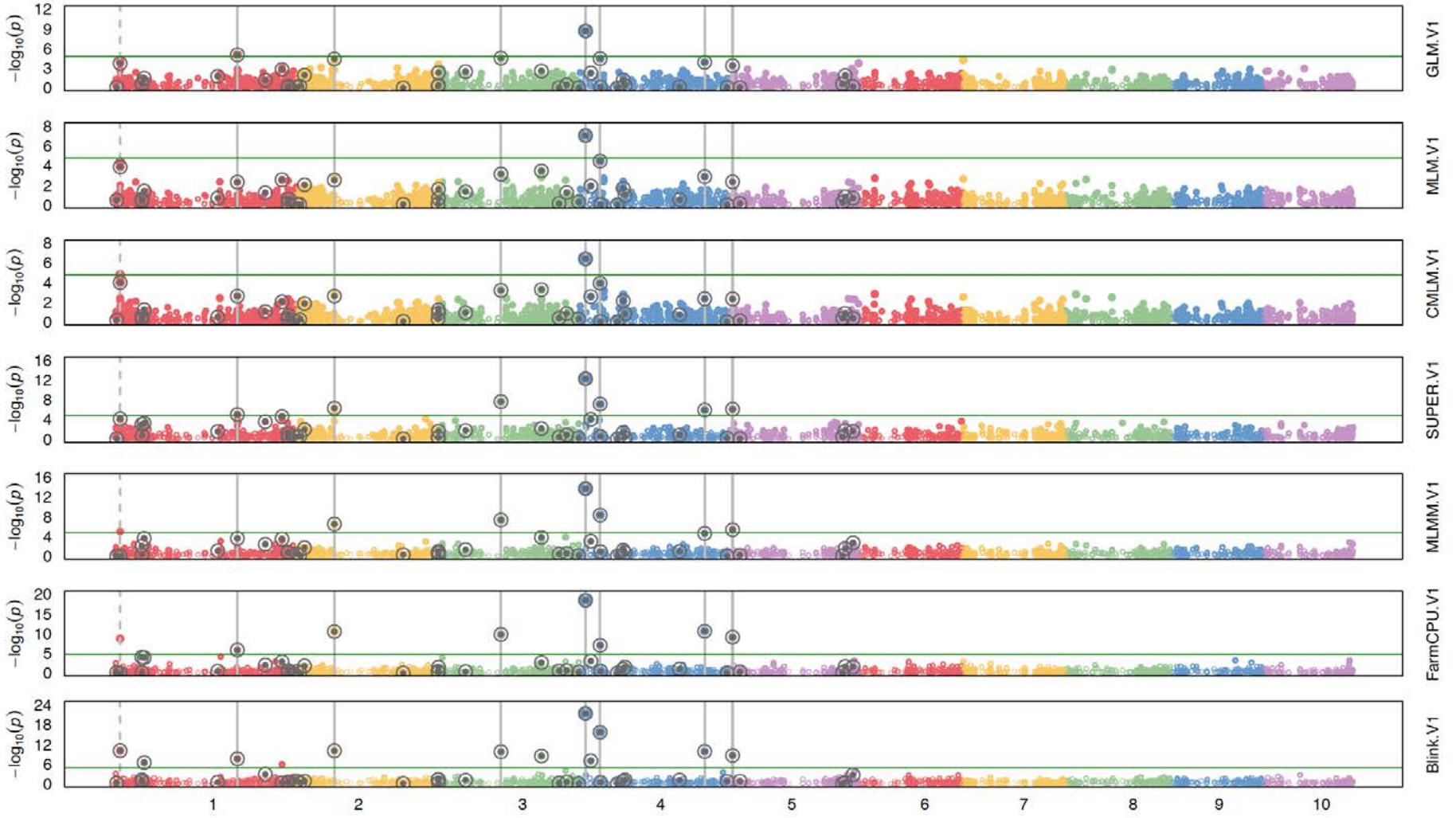
AI4EVER

Problems in GWAS

- Computing difficulties: millions of markers, individuals, and traits
- False positives, ex: “Amgen scientists tried to replicate 53 high-profile cancer research findings, but could only replicate 6”, Nature, 2012, 483: 531
- False negatives

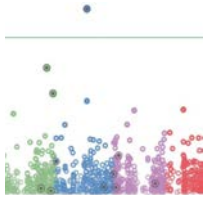


```
source("http://zzlab.net/GAPIT/gapit_functions.txt") #Import demo data
myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#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"))
```

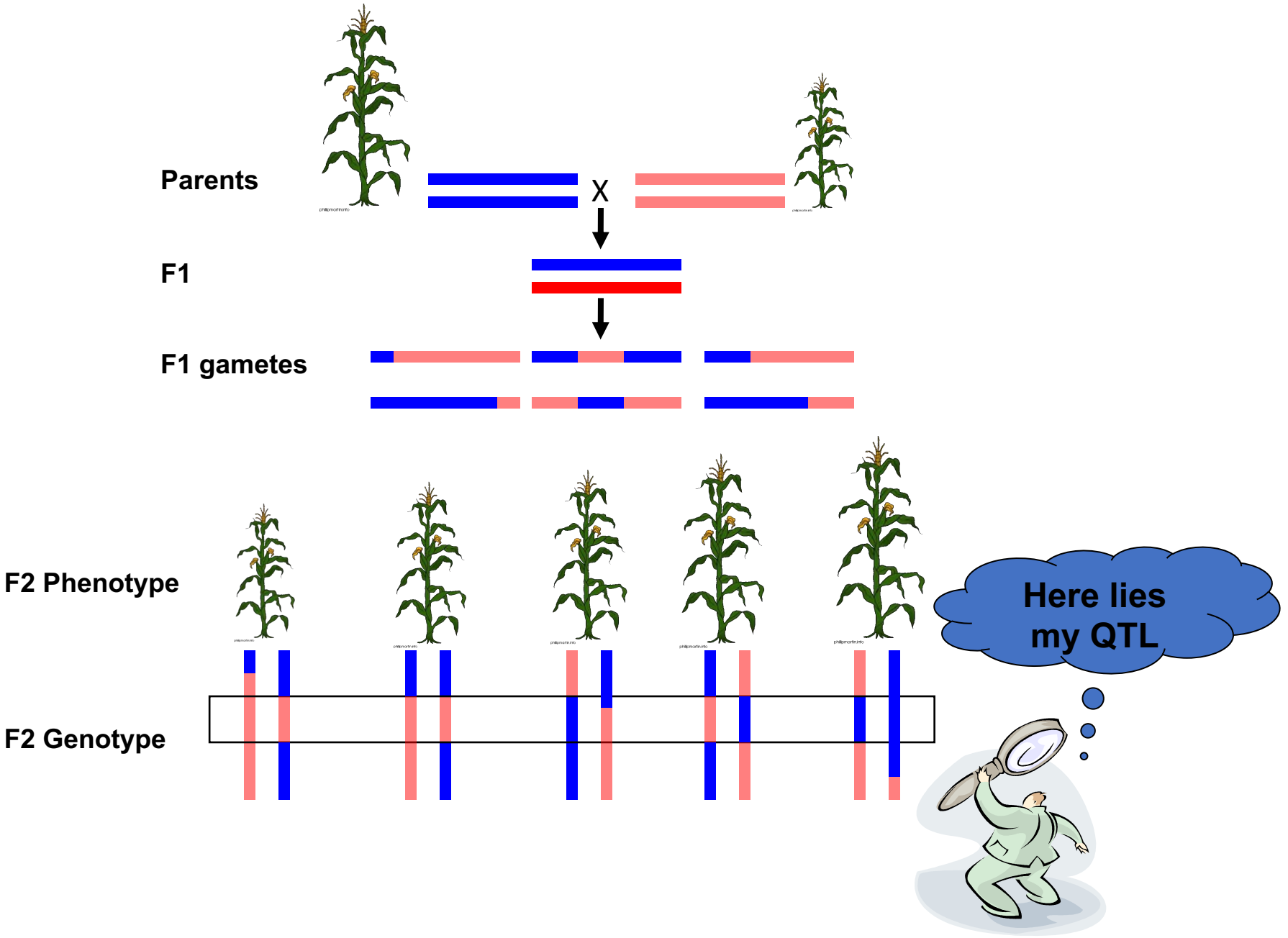


Outline

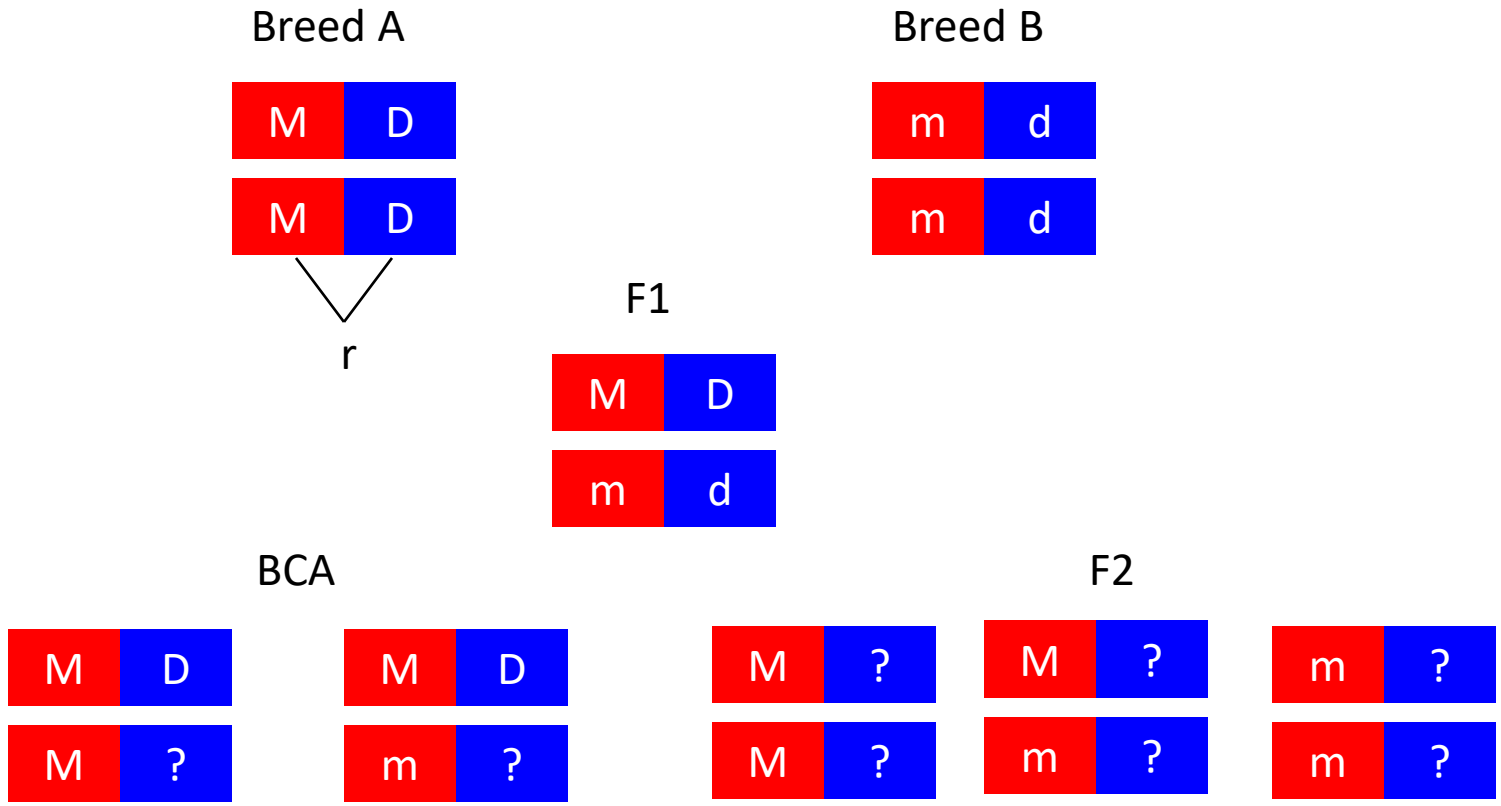
- Horrified community
- Saver Q+K
- 借名 (MLMM and FarmCPU)
- BLINK: Blackbox of GWAS



Linkage analysis

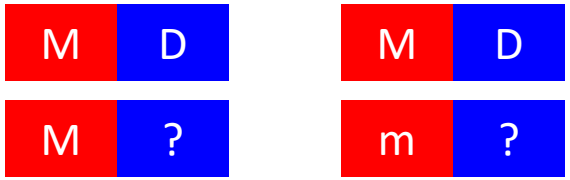


Crosses



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

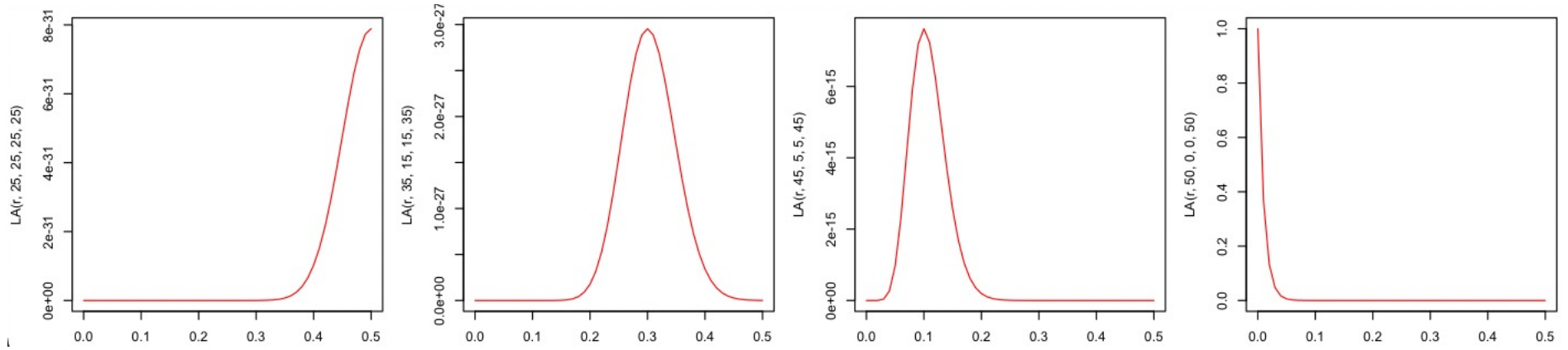
$$P = r^{(n2+n3)} (1-r)^{(n1+n4)}$$

	D	d
MM	25	25
Mm	25	25

	D	d
MM	35	15
Mm	15	35

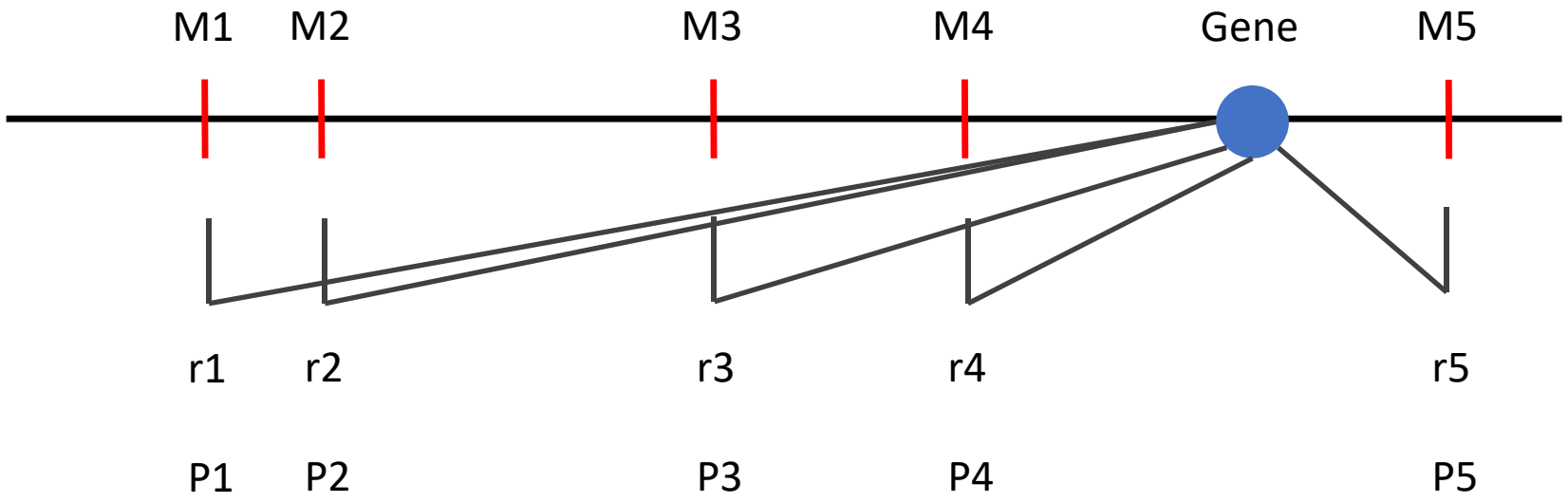
	D	d
MM	45	5
Mm	5	45

	D	d
MM	50	0
Mm	0	50



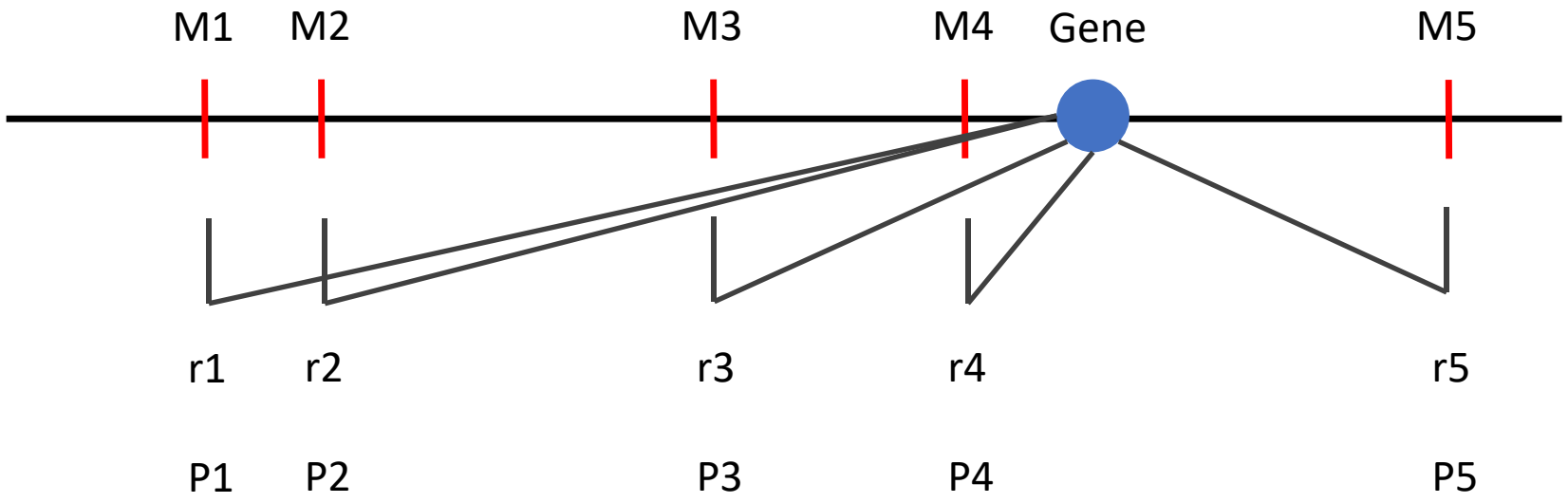

```
r=seq(0,.5,.01)
LA=function(r,n1,n2,n3,n4){return(r^(n2+n3)*(1-r)^(n1+n4))}
par(mfrow=c(1,4),mar = c(3,4,1,1))
plot(r,LA(r,25,25,25,25),type="l",col="red")
plot(r,LA(r,35,15,15,35),type="l",col="red")
plot(r,LA(r,45,5,5,45),type="l",col="red")
plot(r,LA(r,50,0,0,50),type="l",col="red")
```

Multiple markers

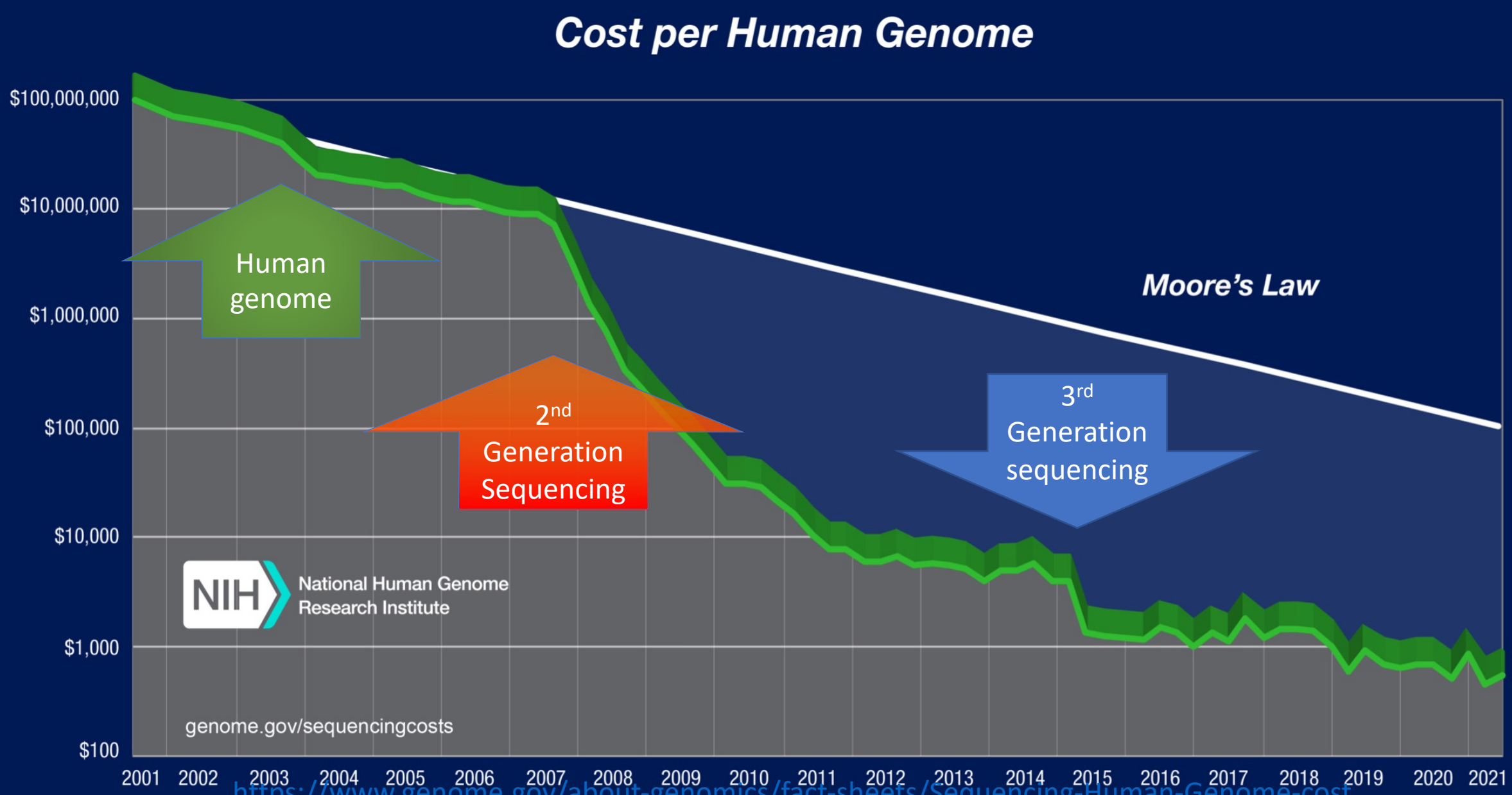


$P = P1*P2*P3*P4*P5$

Multiple markers



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

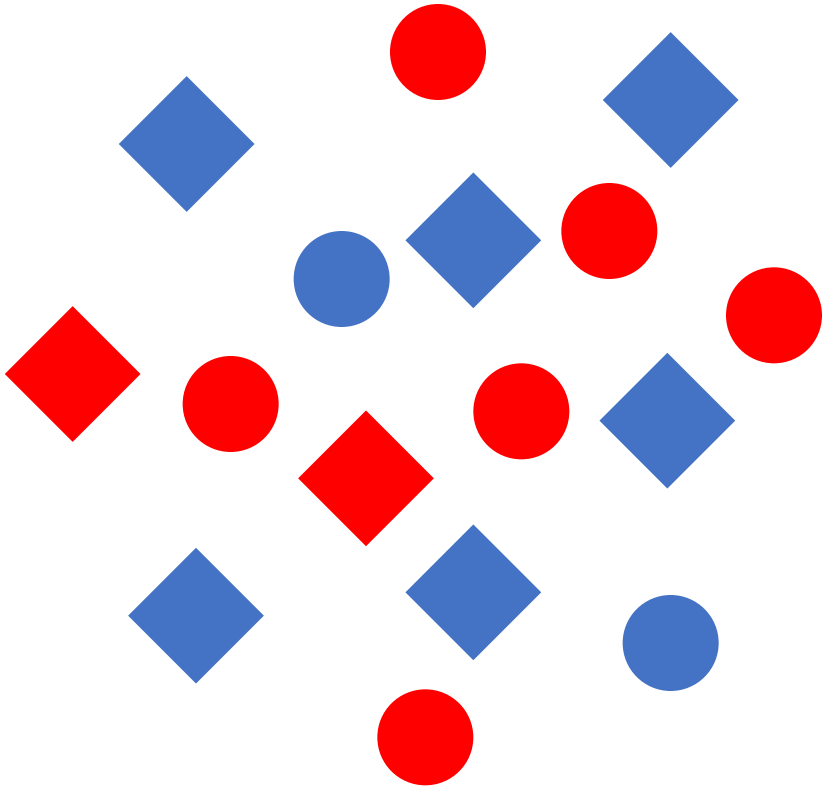


https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Comparison between linkage analysis and GWAS

Property	Linkage analysis	GWAS
Resolution		
Generation		
Genetic base		

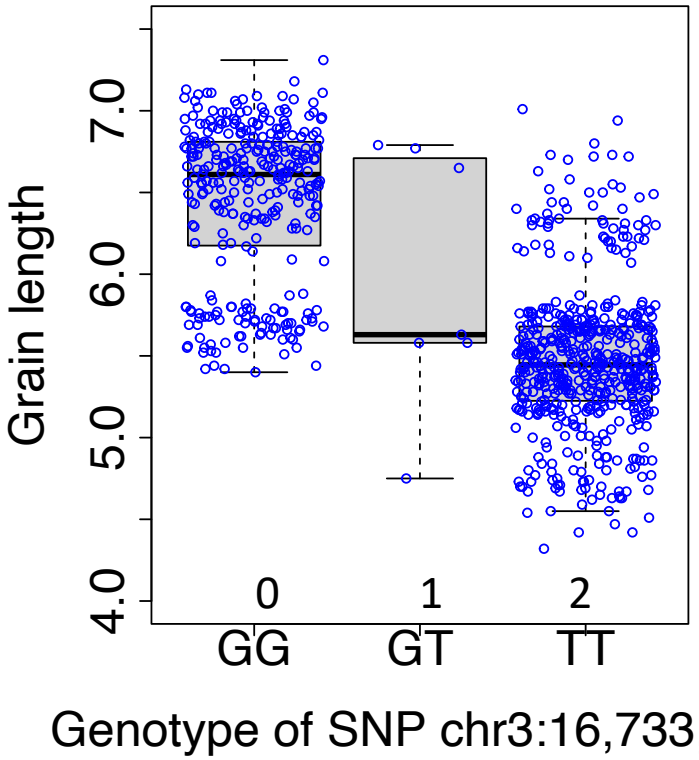
Association study



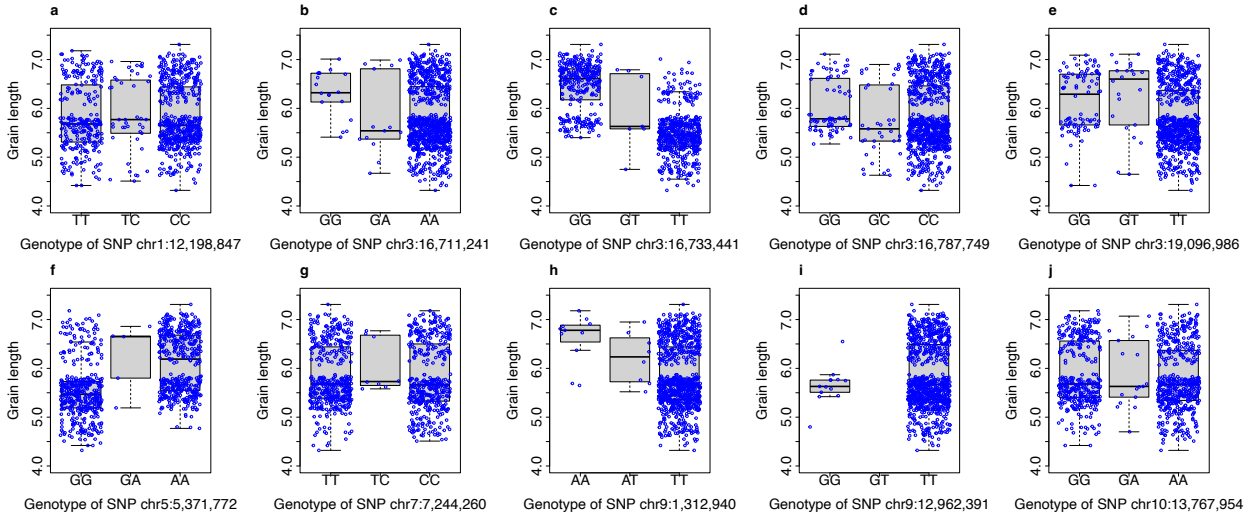
Marker	Control	Case
◇	6	2
○	2	6

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

Correlation



Genotype of SNP chr3:16,733,441



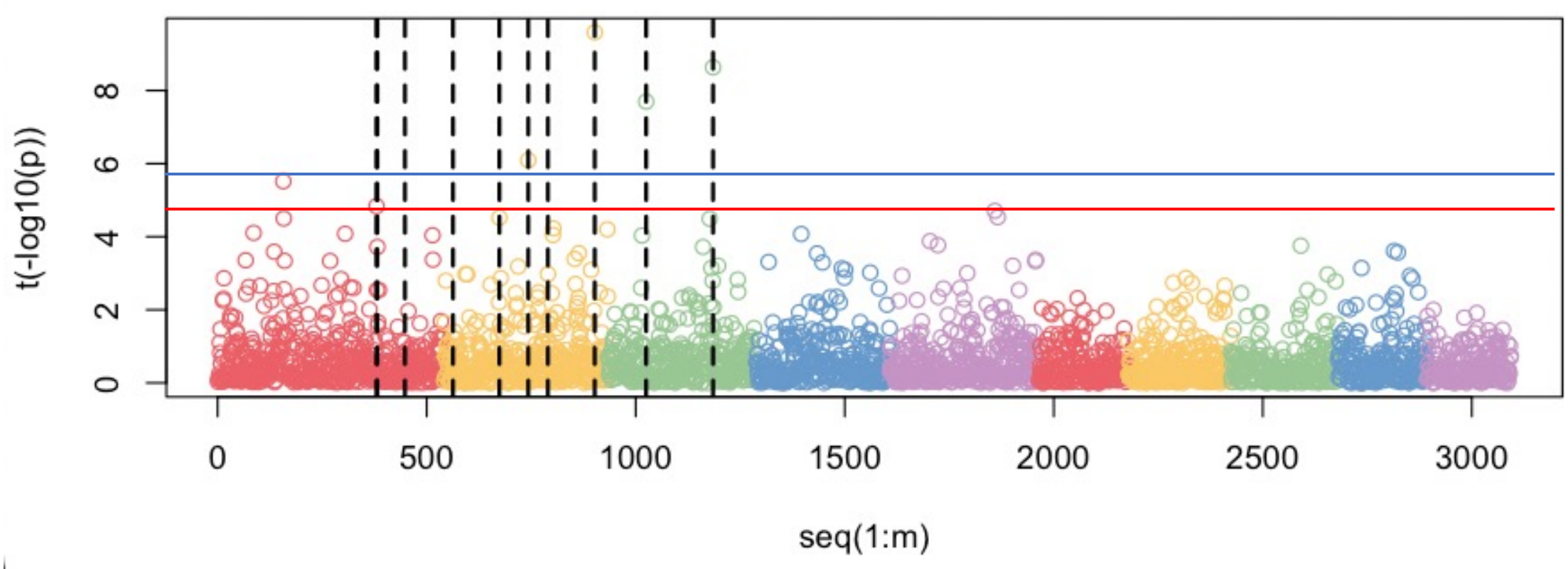
Genotype of SNP chr1:12,198,847
Genotype of SNP chr3:16,711,241
Genotype of SNP chr3:16,733,441
Genotype of SNP chr3:16,787,749
Genotype of SNP chr3:19,096,986
Genotype of SNP chr5:5,371,772
Genotype of SNP chr7:7,244,260
Genotype of SNP chr9:1,312,940
Genotype of SNP chr9:12,962,391
Genotype of SNP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

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)
```

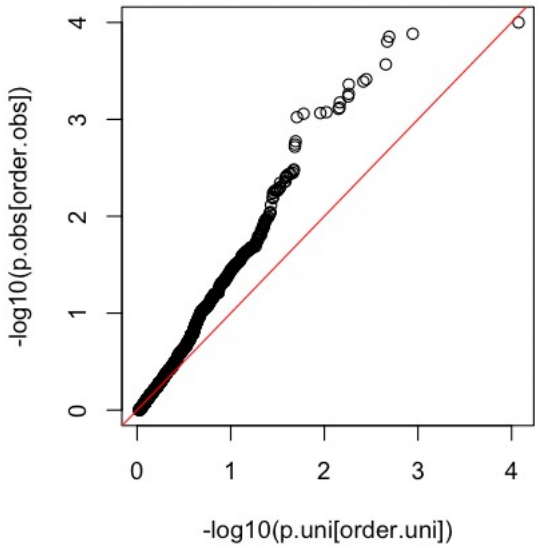
False positives

```
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")
```



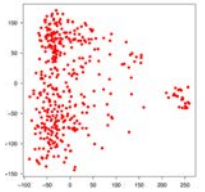
QQ plot

```
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")
```



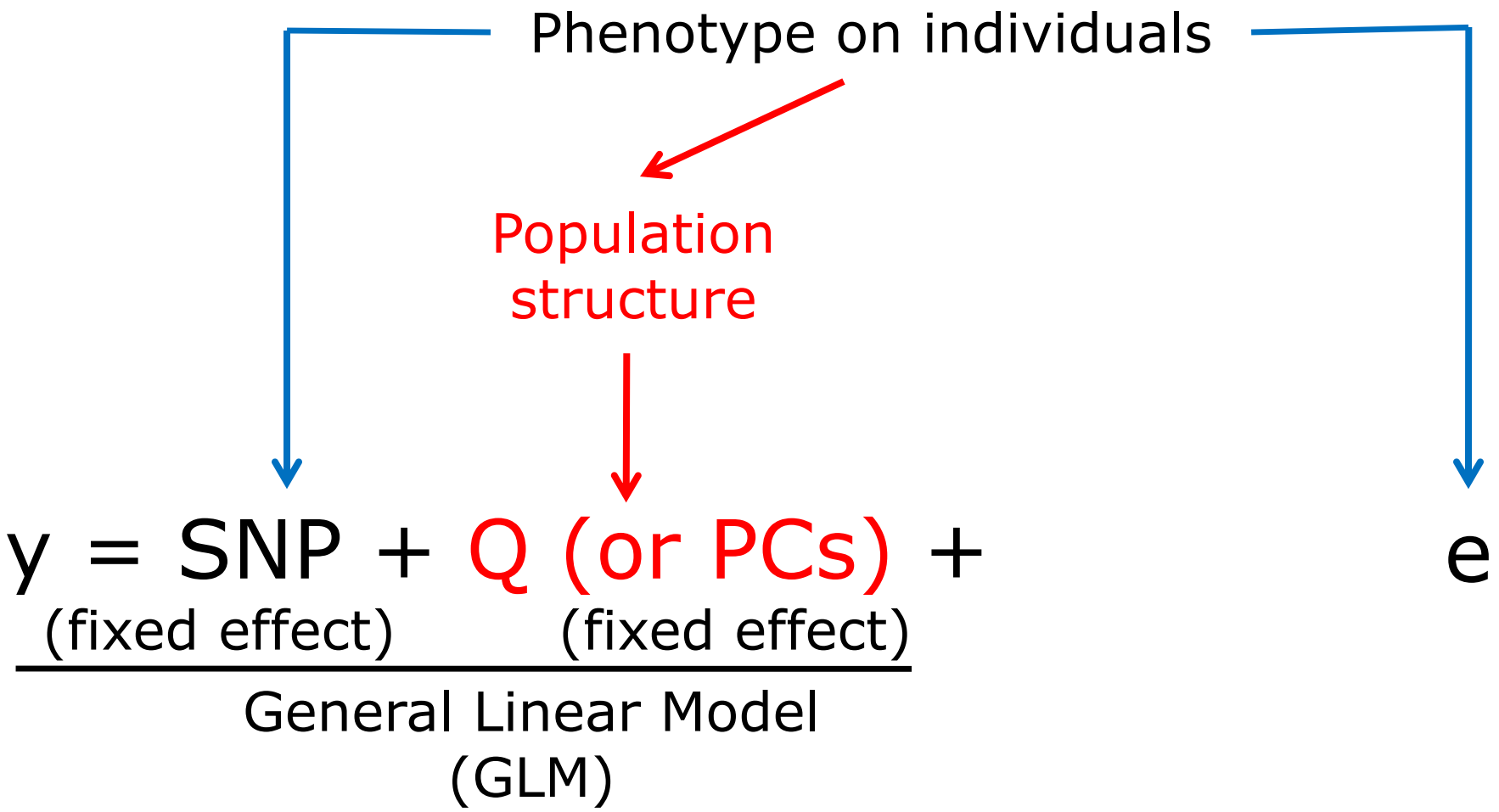
Outline

- Horrified community
- Saver Q+K
- 借名 (MLMM and FarmCPU)
- BLINK: Blackbox of GWAS




EMMAx
EMMA
PC+K
PC
CMLM
Q+K
Q
P3D
GCTA
SELECT
MLMM
FST-LMM
GEMMA
ECMLM
SUPER
FarmCPU
GenAbel
BLINK

GLM (Conceptual)



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

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

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

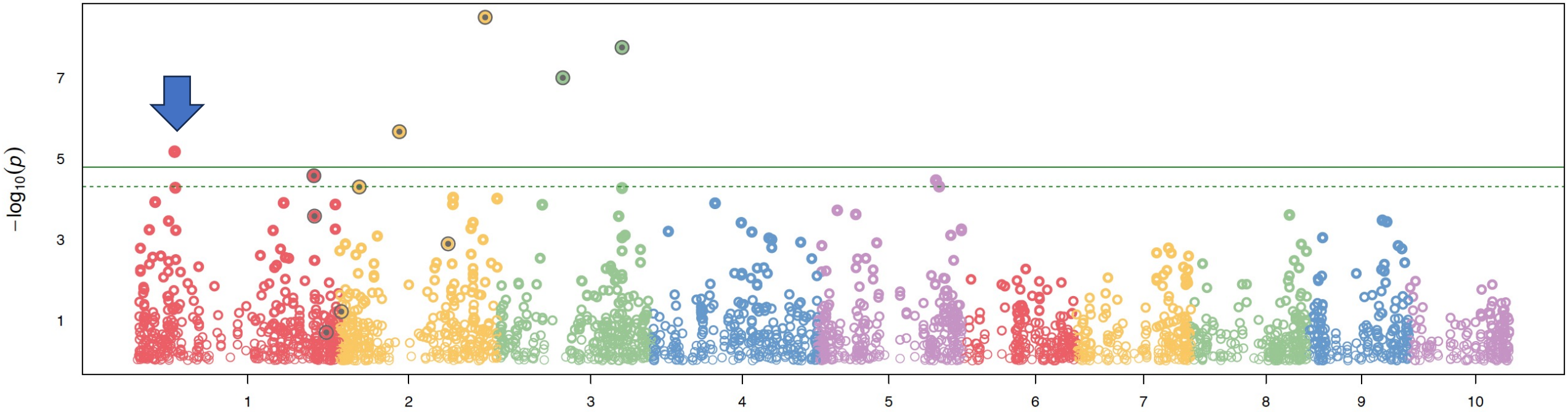
$$Var(\hat{b}) = [X'X]^{-1}\sigma_e^2$$

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

T Test

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

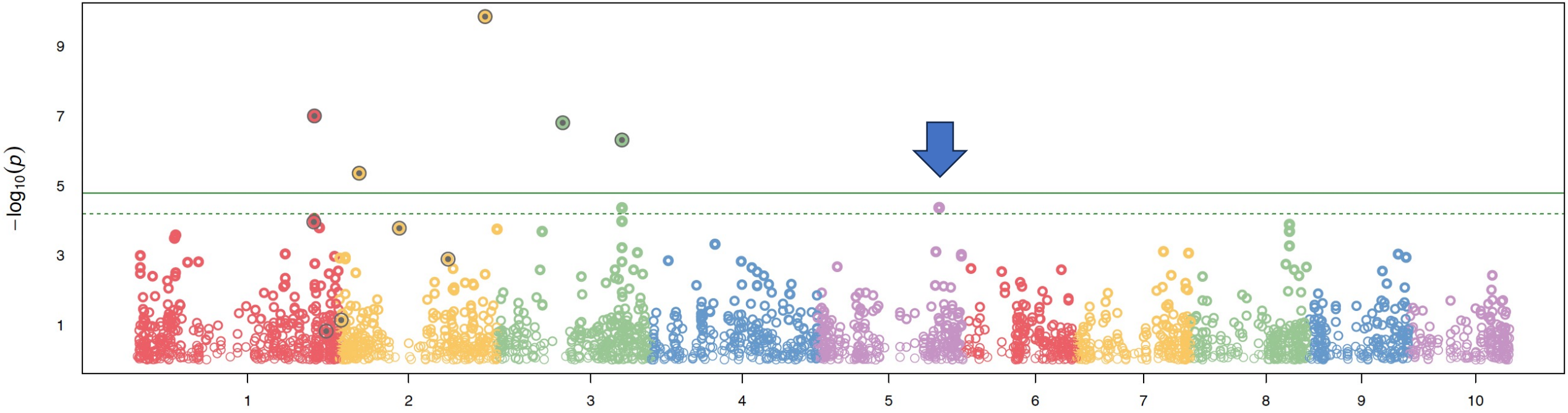
```
#GWAS by GAPIT
myGAPIT=GAPIT(
  Y=myY,
  GD=myGD,
  GM=myGM,
  QTN.position=mySim$QTN.position,
  PCA.total=0,
model="GLM",
memo="tTest")
```



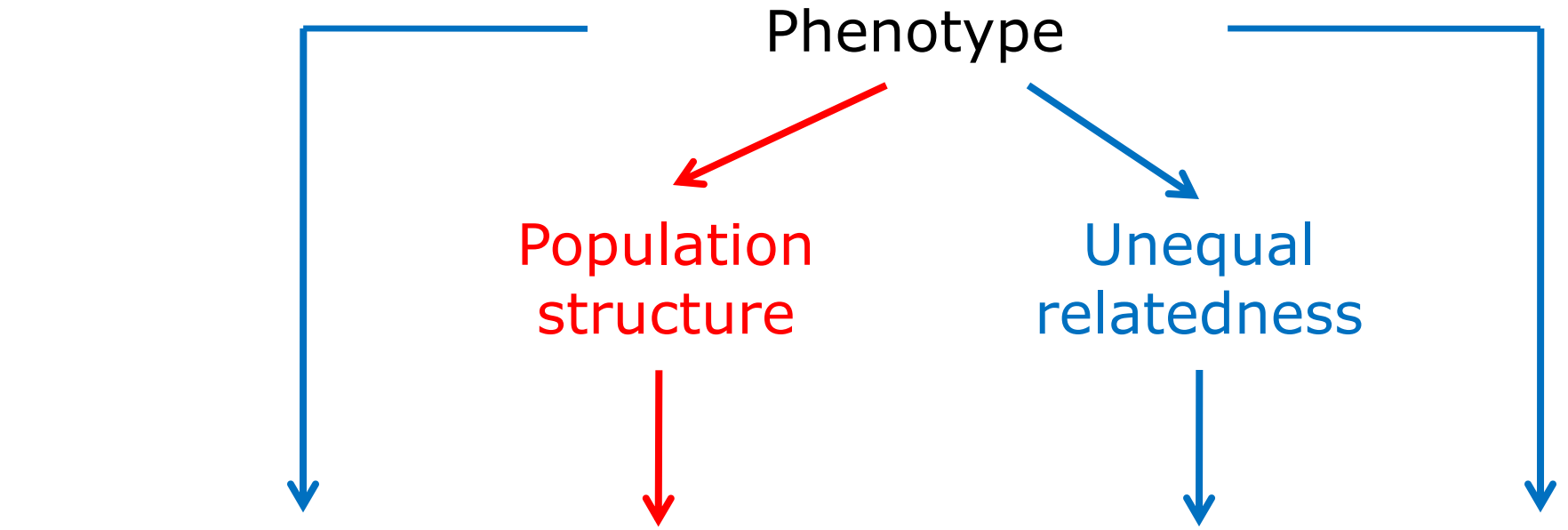
```
setwd("~/Desktop/temp")
myY=(cbind(myGD[,1], as.data.frame(mySim$y)))
source("http://zzlab.net/GAPIT/gapit_functions.txt")
```

GLM

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



MLM for GWAS



(Yu et al. 2005, Nature Genetics)

Mixed Linear Model (MLM)

$y = Xb + Zu + e$

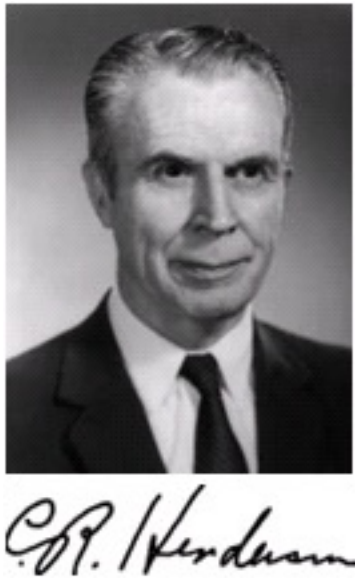
$Var(y)=V=Var(u)+Var(e)$

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

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

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

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

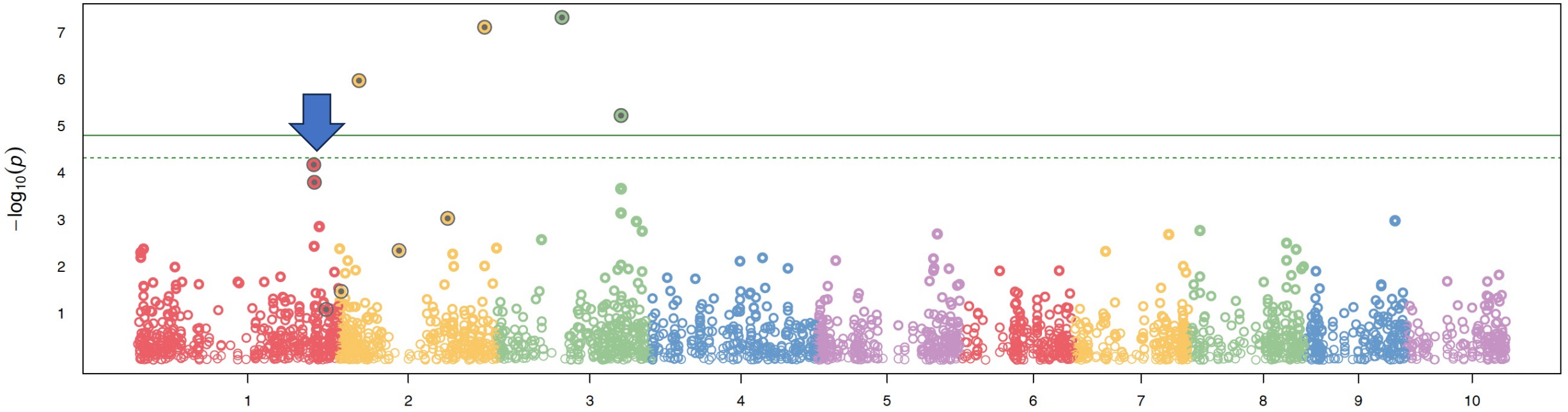
J. Dairy Sci. 2008. 91 (11) 4414-4423. Efficient Methods to Compute Genomic Predictions P. M. VanRaden

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=0,
model="MLM",
memo="MLM_0PC")
```



Outline

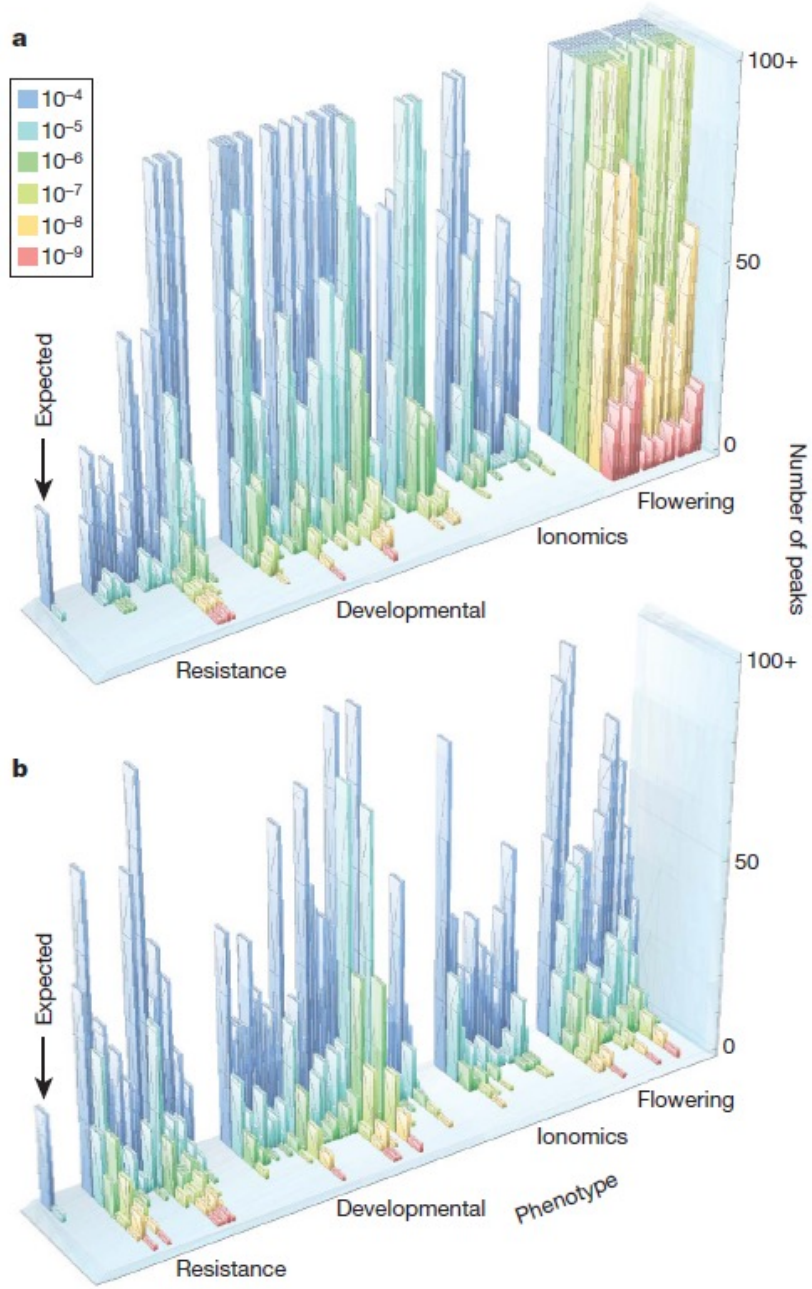
- Horrified community
- Saver Q+K
- 借名 (MLMM and FarmCPU)
- BLINK: Blackbox of GWAS

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

MLMM

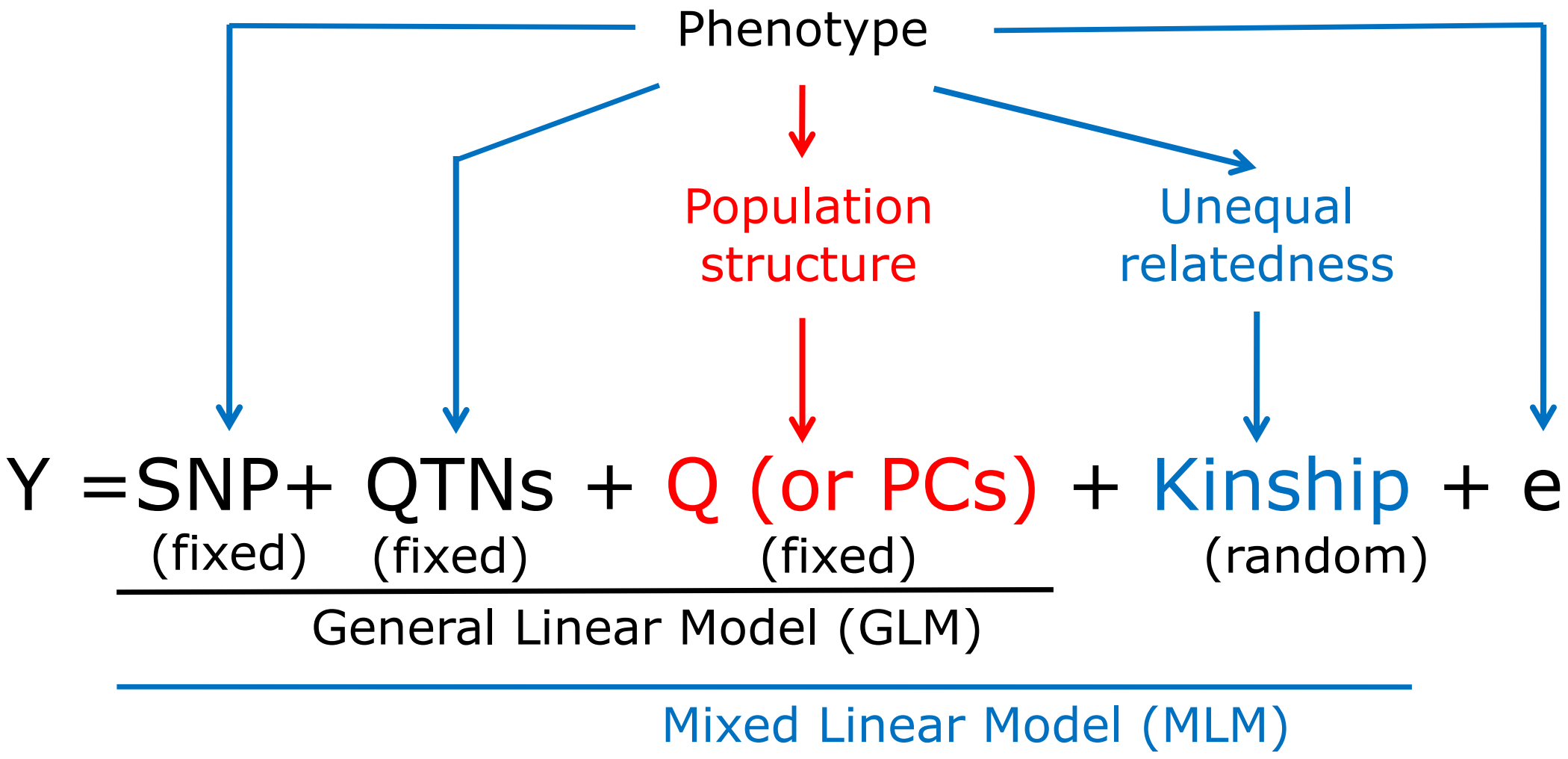
nature genetics

Published: 17 June 2012

An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations

Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long & Magnus Nordborg ✉

Multiple Loci Mixed Model (MLMM)



(Sagura et al Nature 2012)

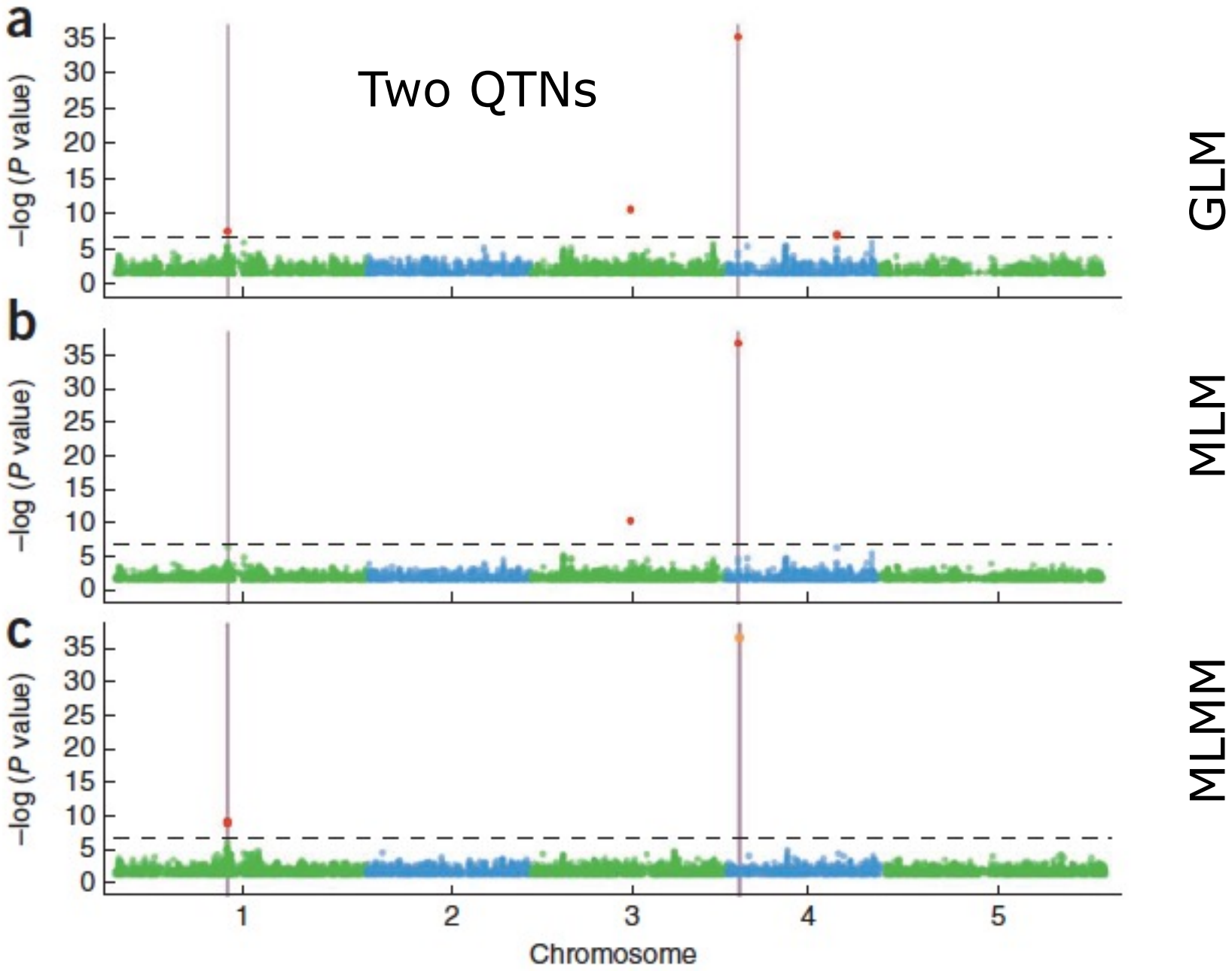
Forward regression

$$y = SNP + QTN1 + QTN2 + \ldots + Q + K + e$$

$$\frac{Var(u)}{Var(y)}$$

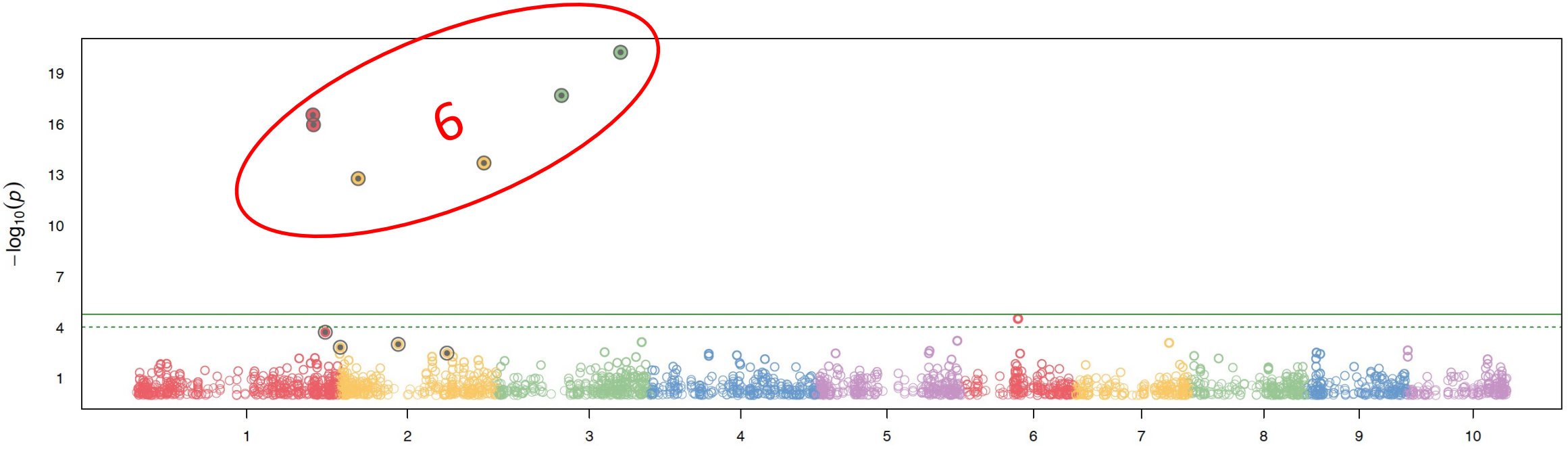
Stop when the ratio close to zero

Nature Genetics, 2012, 44, 825-830

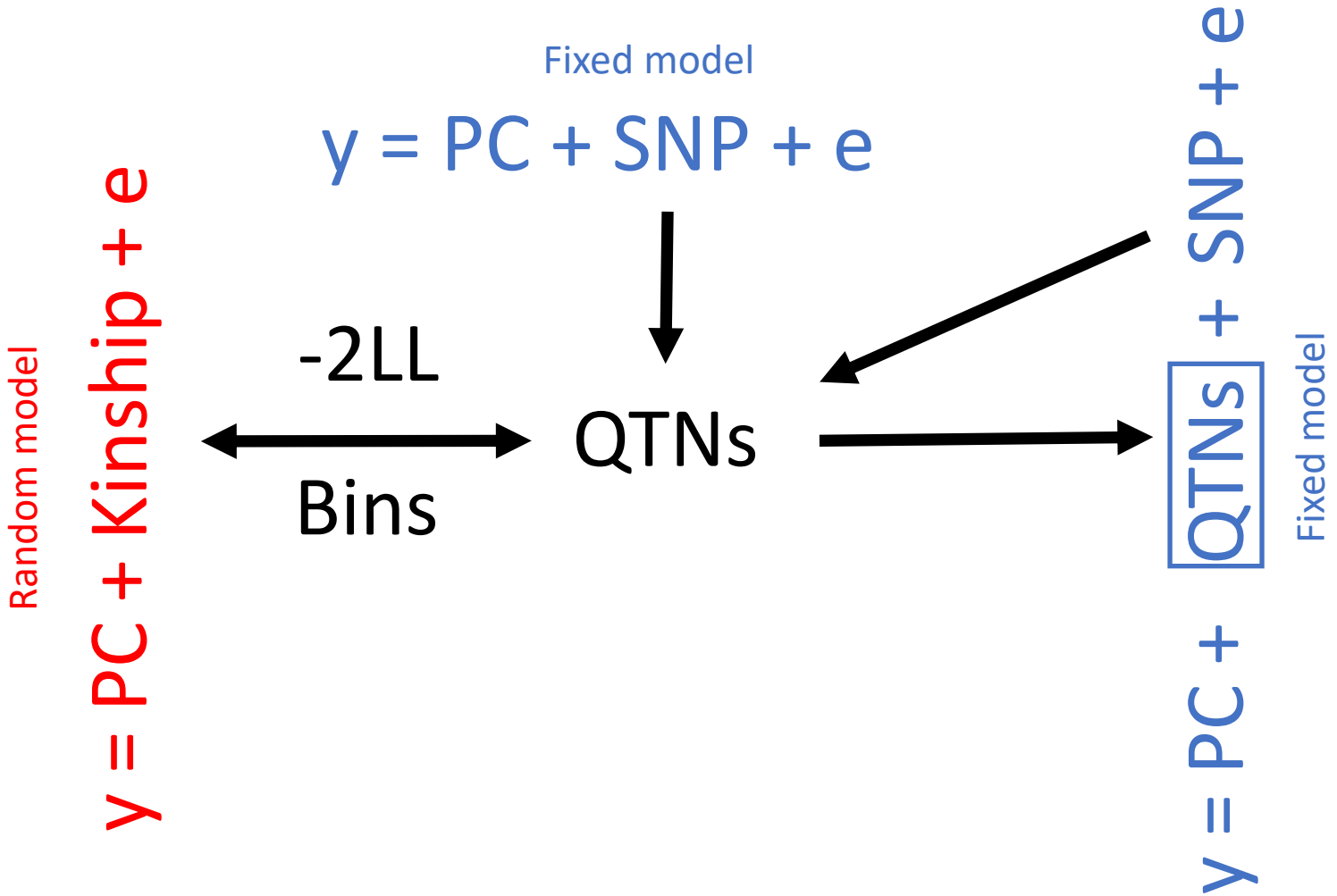


MLMM within GAPIT

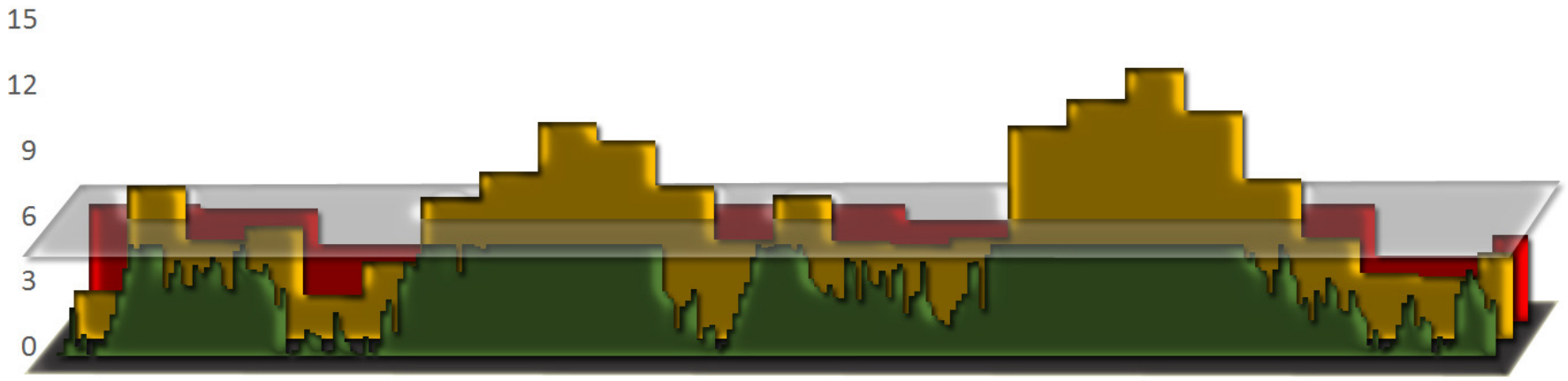
```
myGAPIT_MLM <- GAPIT(
Y=myY,
GD=myGD,#Genotype
GM=myGM,#Map information
PCA.total=0,
QTN.position=mySim$QTN.position,
model="MLMM",# Can choose MLM CMLM GLM SUPER MLMM FarmCPU Blink
memo="MLMM_0PC")
```



FarmCPU algorithm

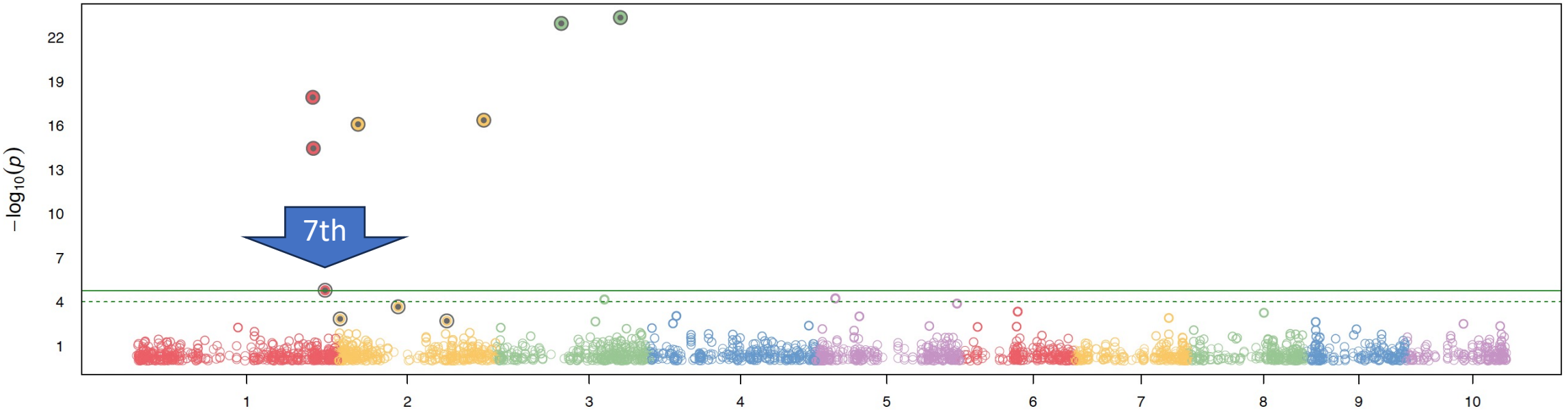


QTN selection (bin approach)



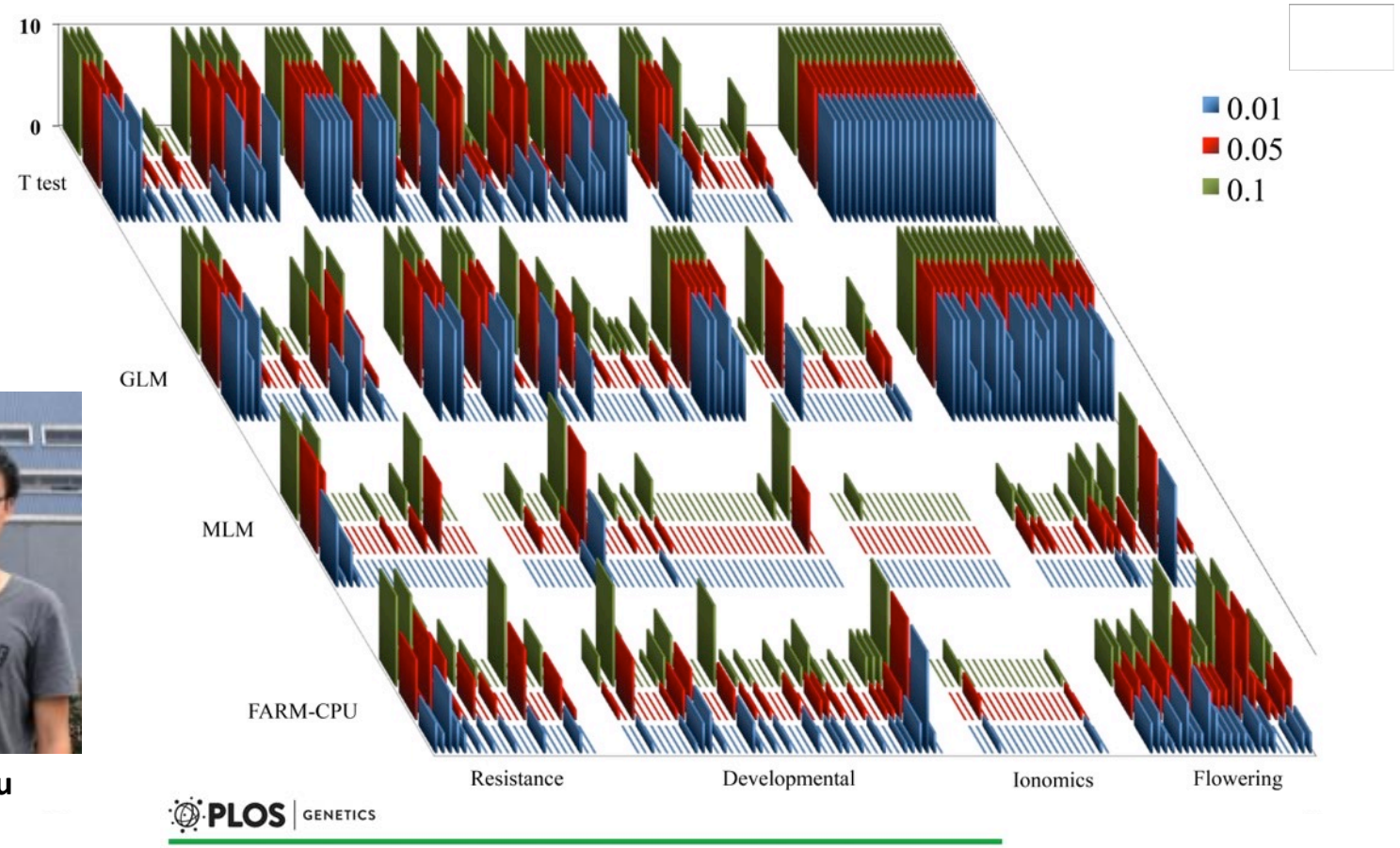
FarmCPU via GAPIT

```
myGAPIT_MLM <- GAPIT(
Y=myY,
GD=myGD,#Genotype
GM=myGM,#Map information
PCA.total=0,
QTN.position=mySim$QTN.position,
model="FarmCPU",
memo="FarmCPU_0PC ")
```



Re-analysis of *Arabidopsis* data

Xiaolei Liu

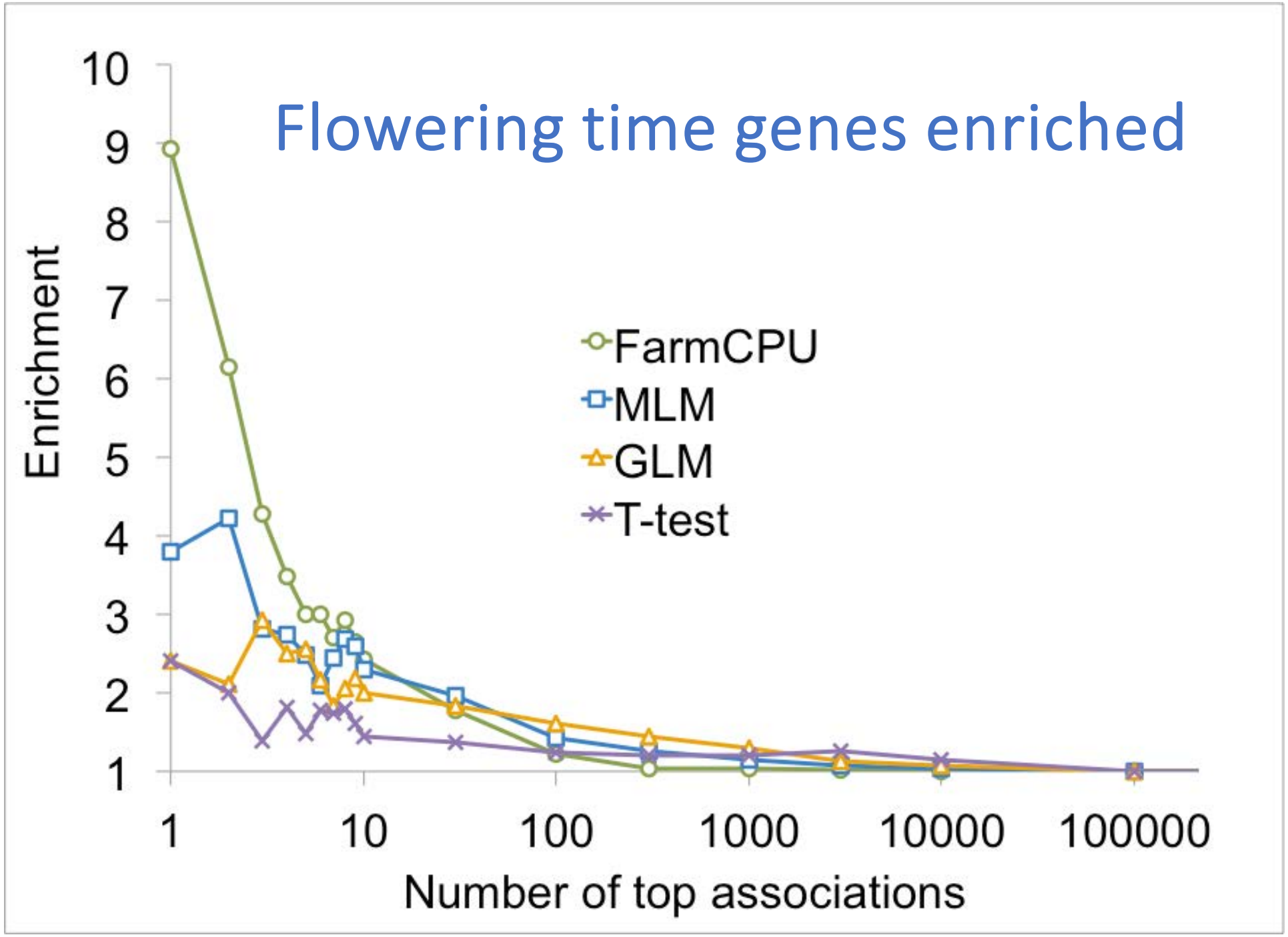


PLOS | GENETICS

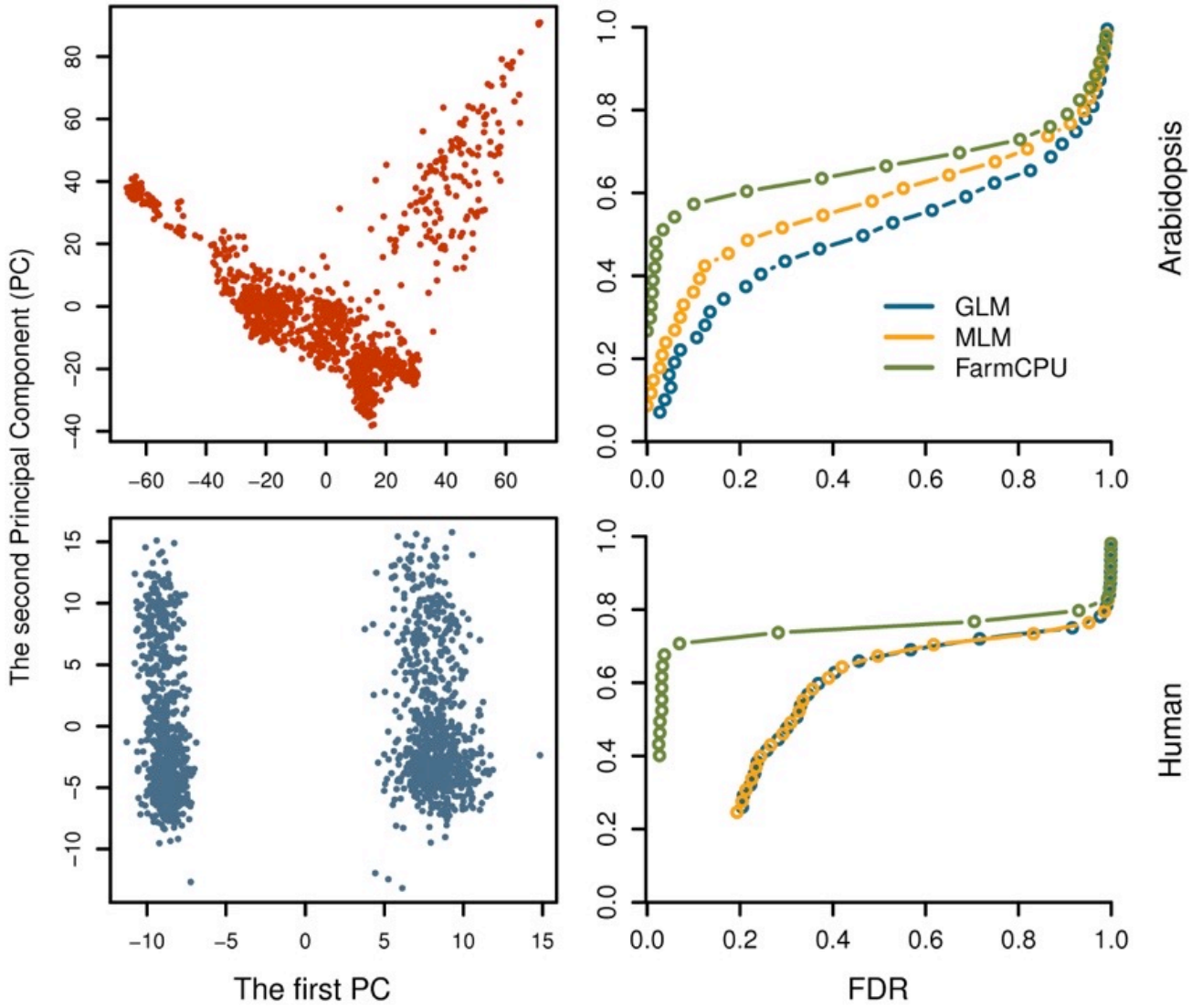
RESEARCH ARTICLE

Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies

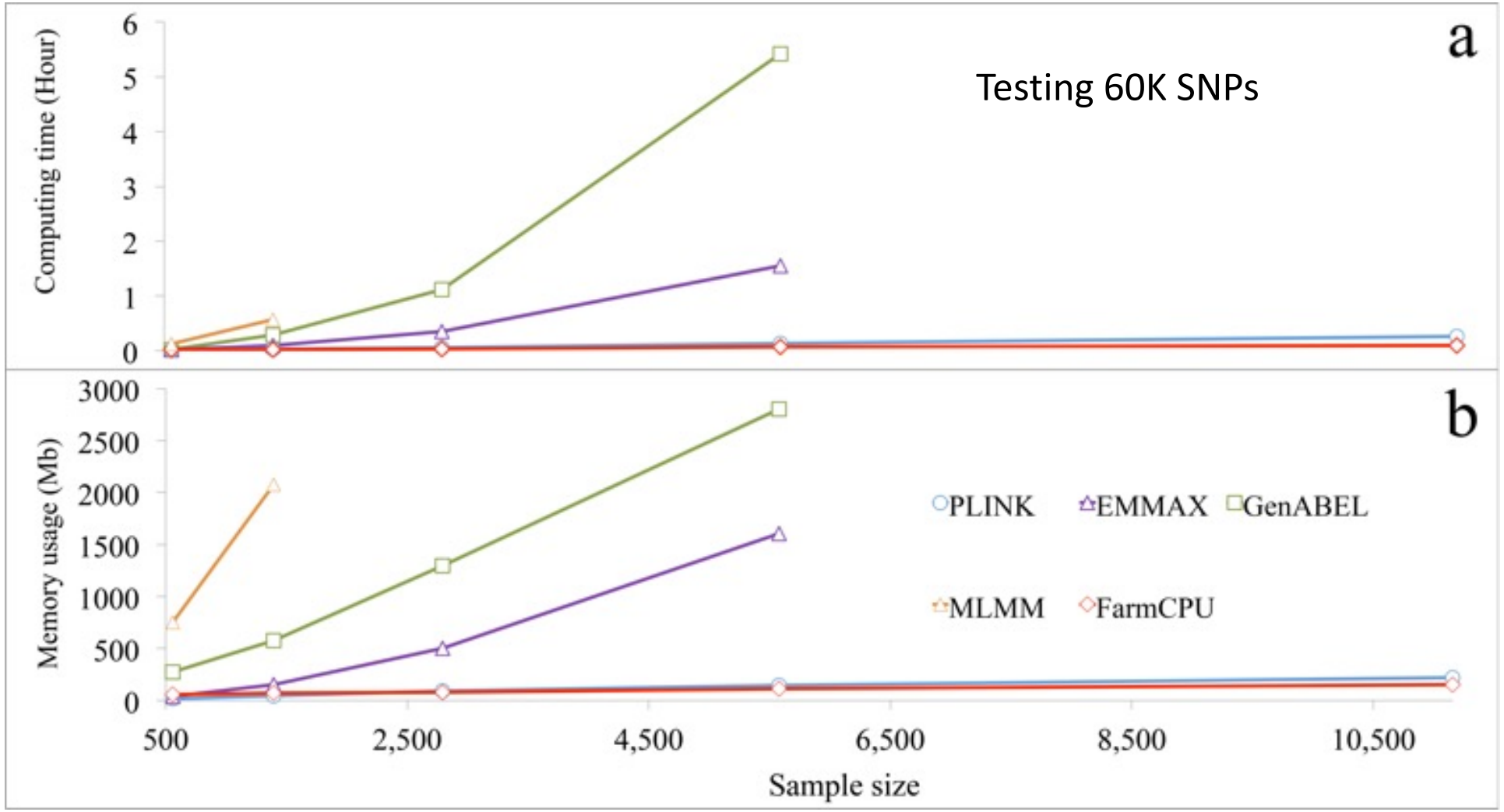
Xiaolei Liu[1,2], Meng Huang[3], Bin Fan[1], Edward S. Buckler[2,4], Zhiwu Zhang[3,5]*


Flowering time genes enriched
Enrichment
10
9
8
7
6
5
4
3
2
1
FarmCPU
MLM
GLM
T-test
1
10
100
1000
10000
100000
Number of top associations

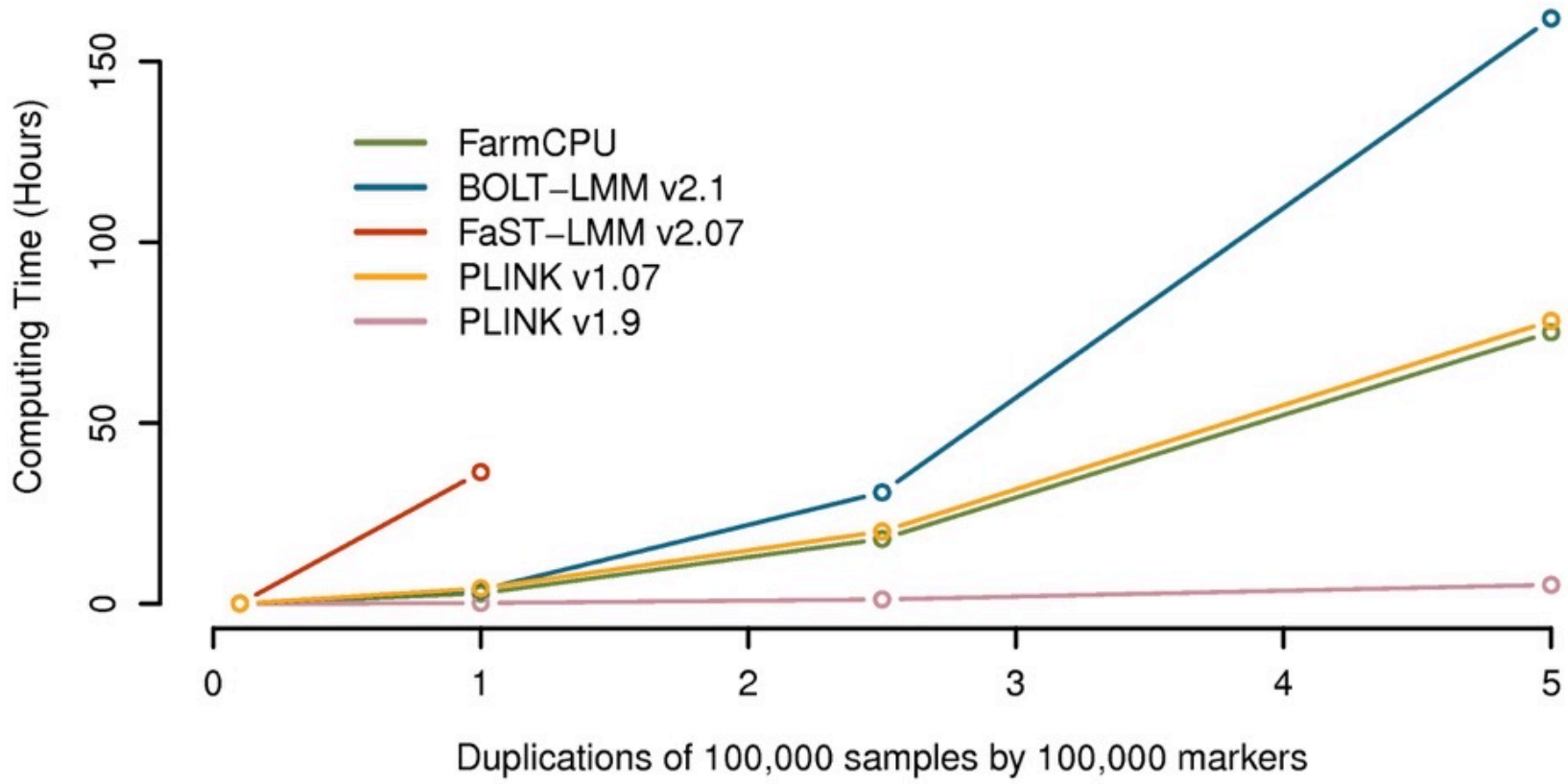
It is time for human geneticists to move forward



FarmCPU is computing efficient



Half million individuals, half million SNPs: three days

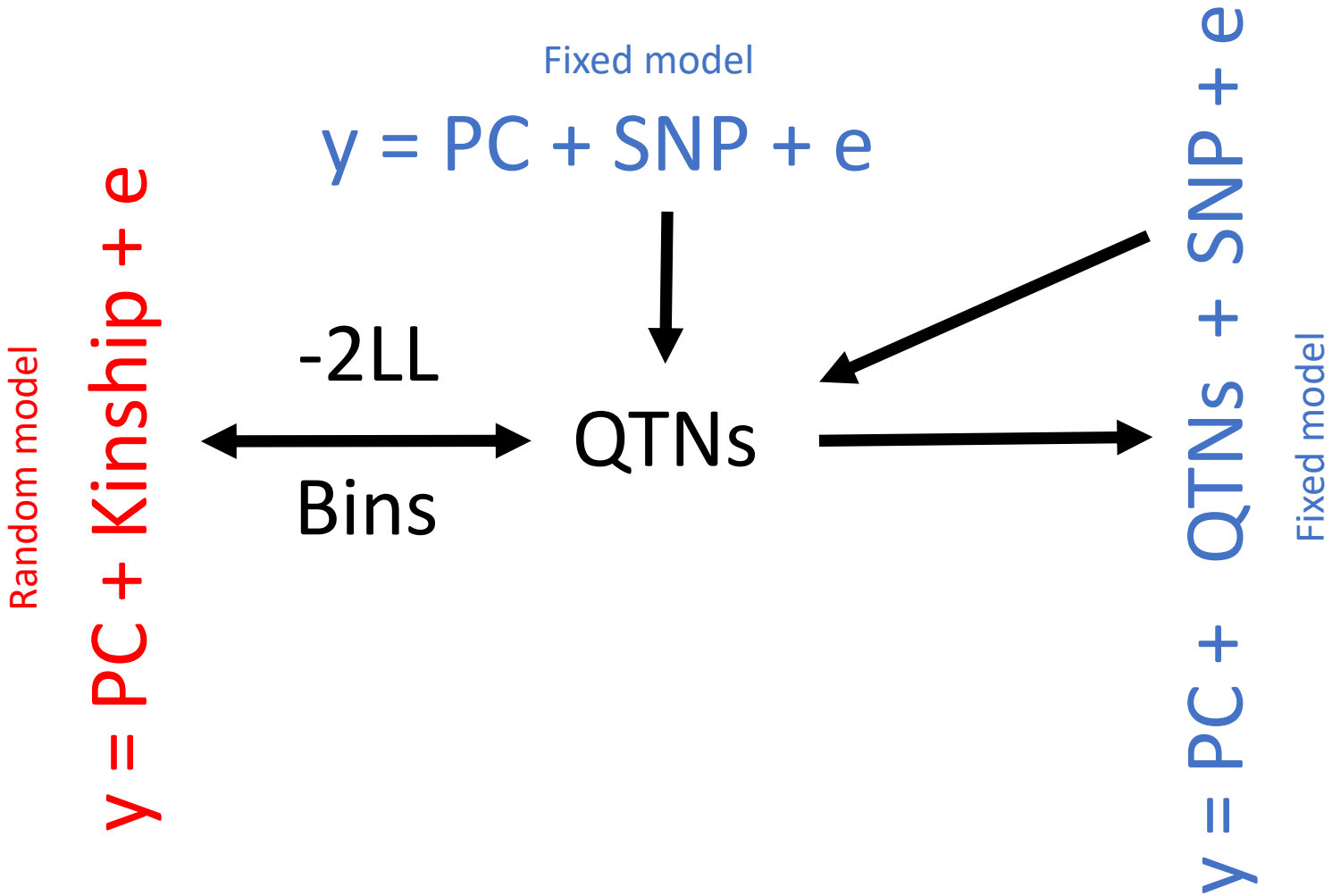


But, PINK new version is faster

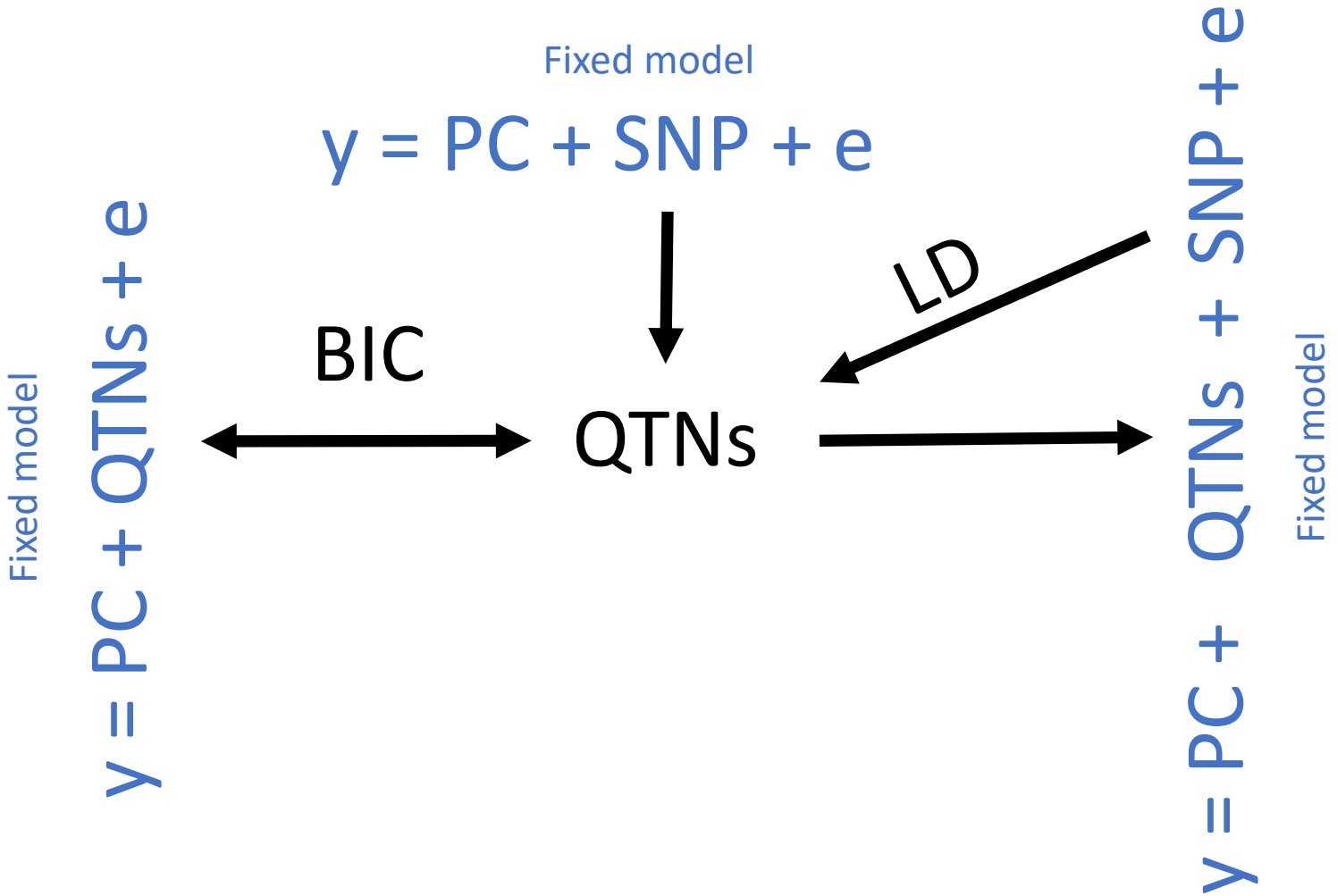
Outline

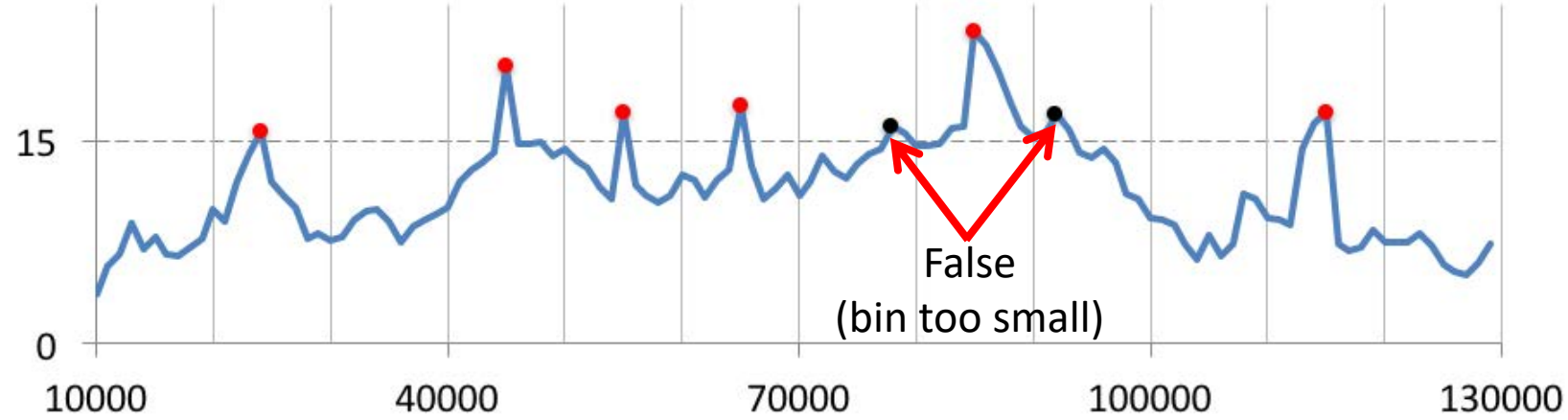
- Horrified community
- Saver Q+K
- 借名 (MLMM and FarmCPU)
- BLINK: Blackbox of GWAS

FarmCPU algorithm

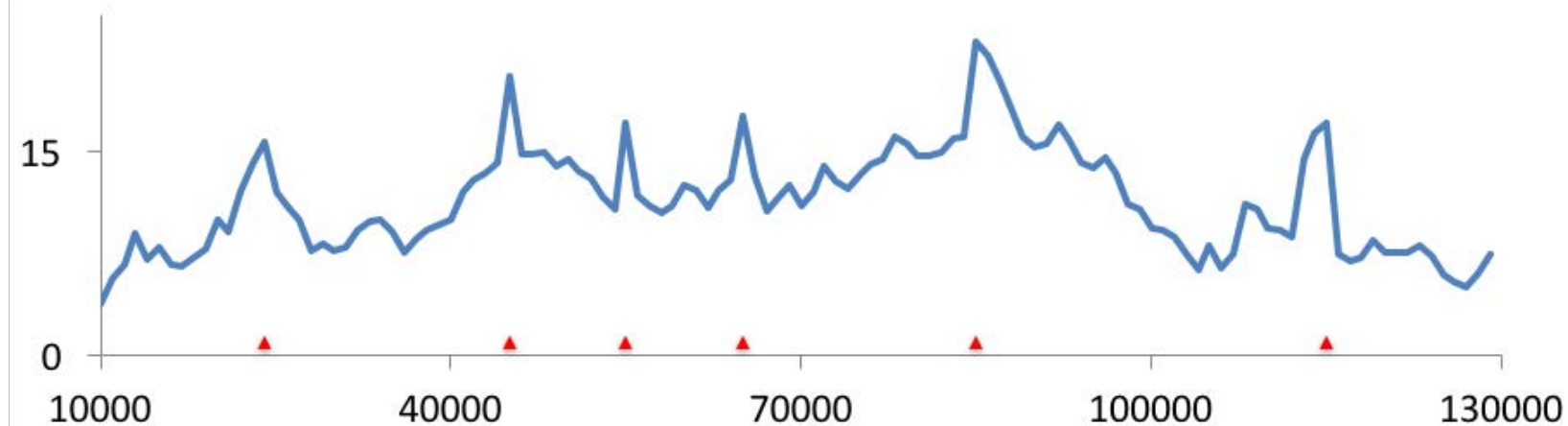


BLINK algorithm

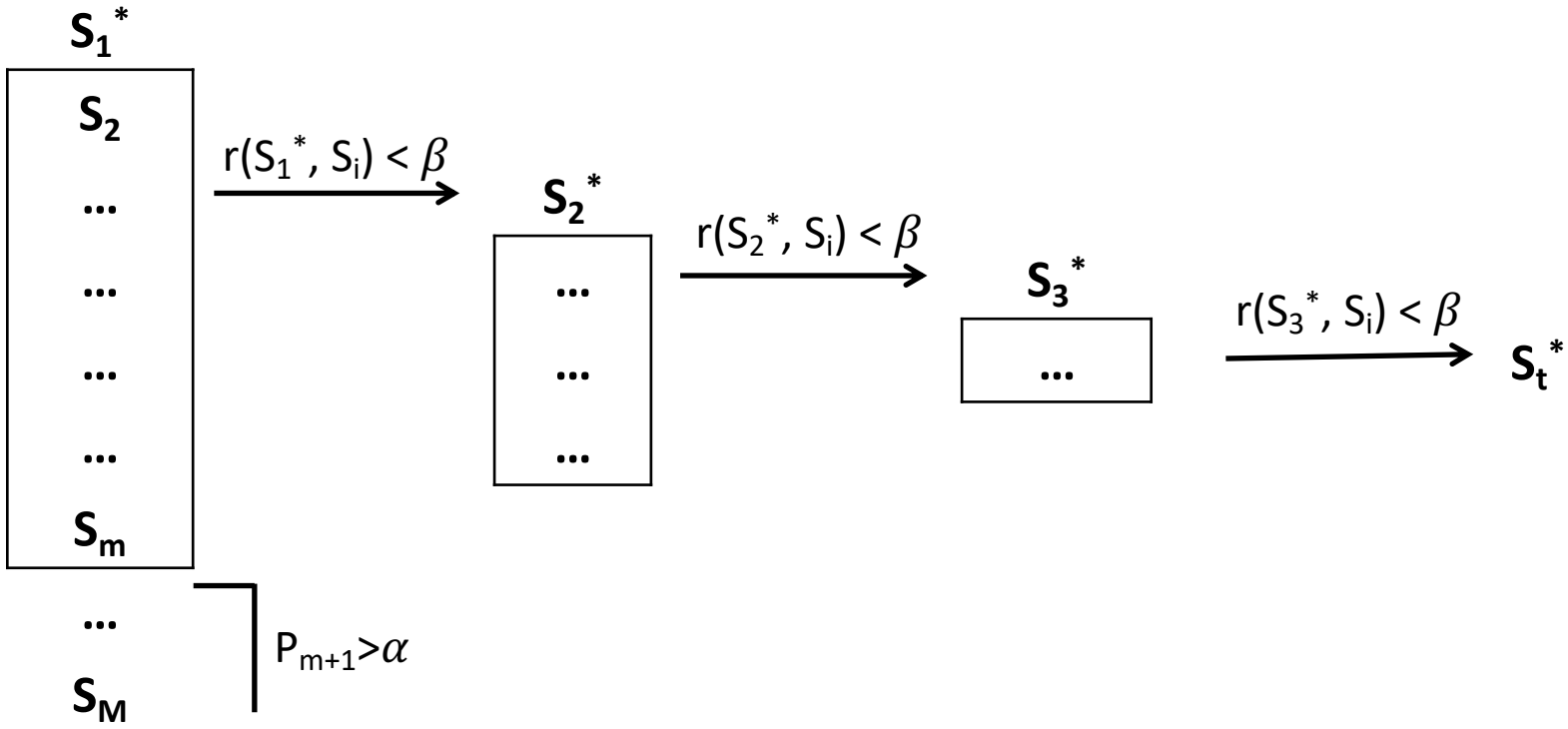



15
0
10000
40000
70000
100000
130000
False
(bin too small)


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


15
0
10000
40000
70000
100000
130000

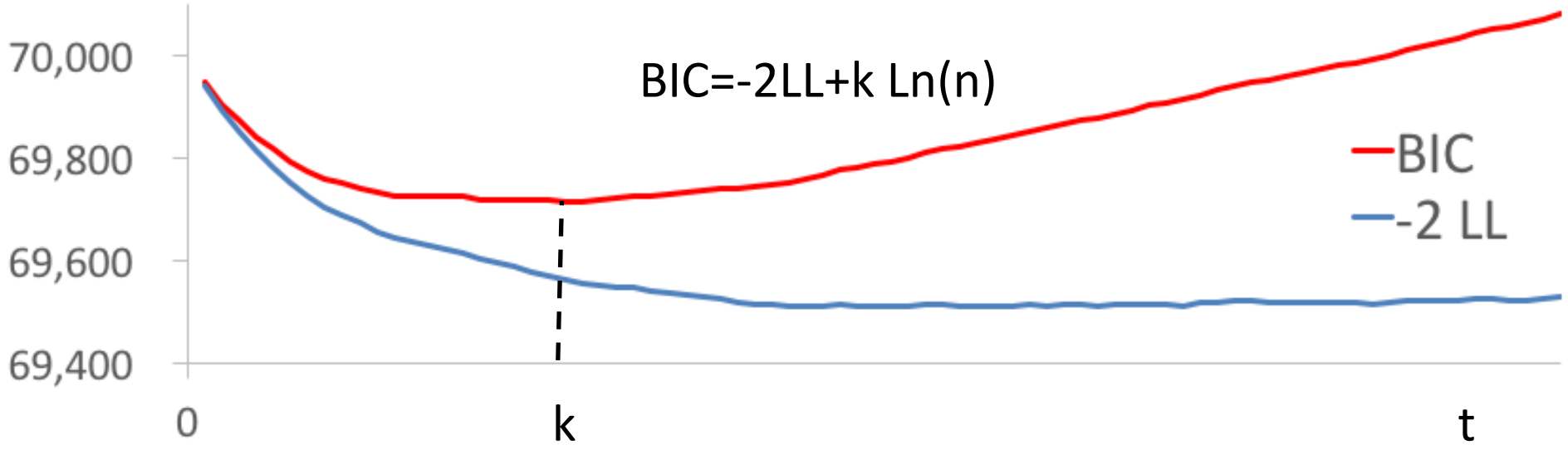
Elimination of markers with LD

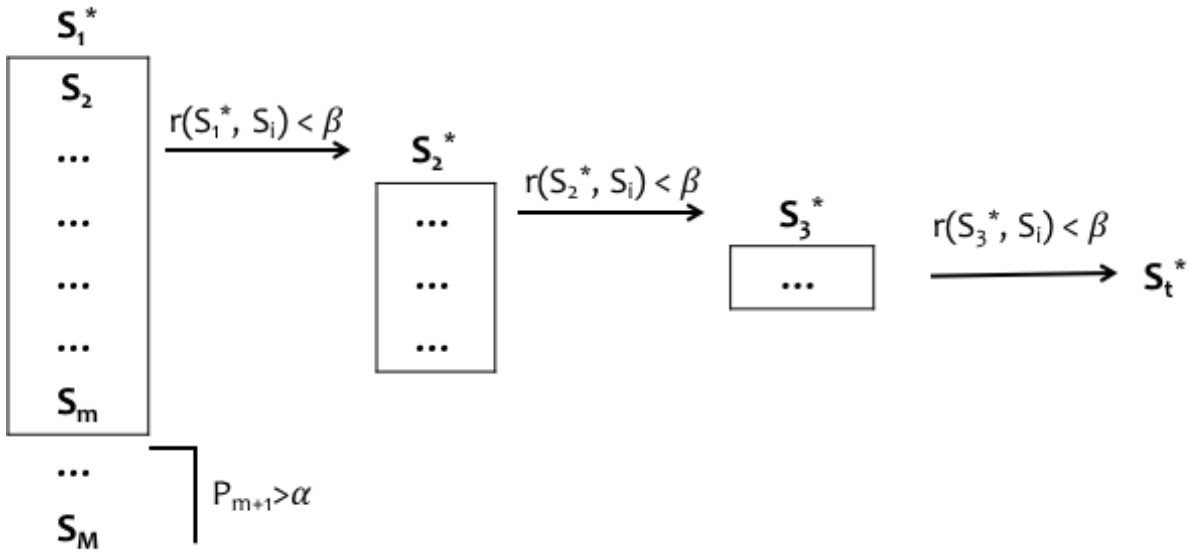


$$y = s_i + S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e, \text{ where } i = 1 \text{ to } M$$

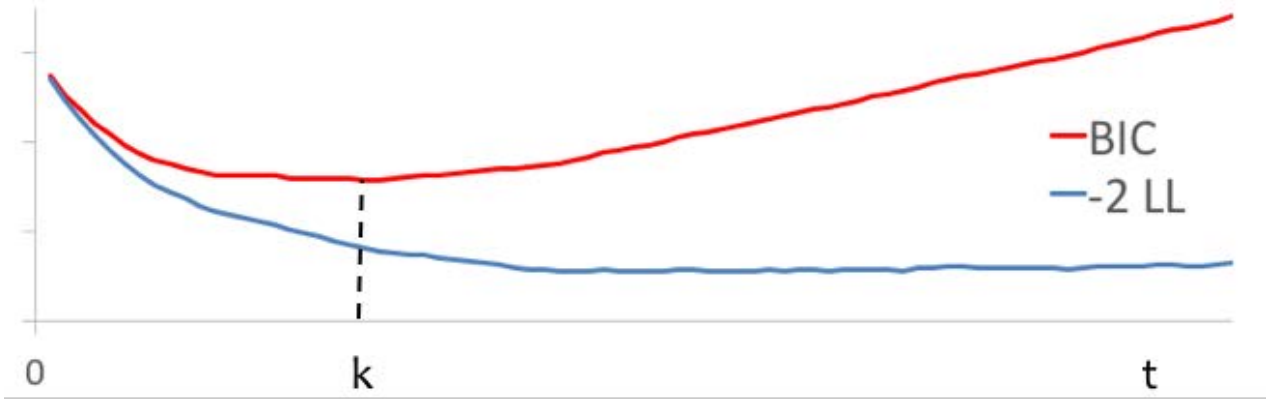
Bayesian information criterion

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





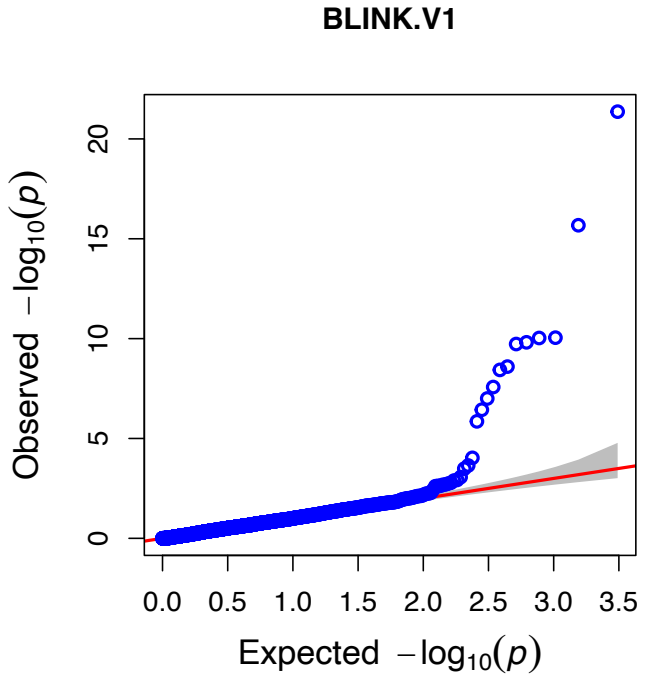
$y = S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e$, where $k \leq t$ maximizes BIC

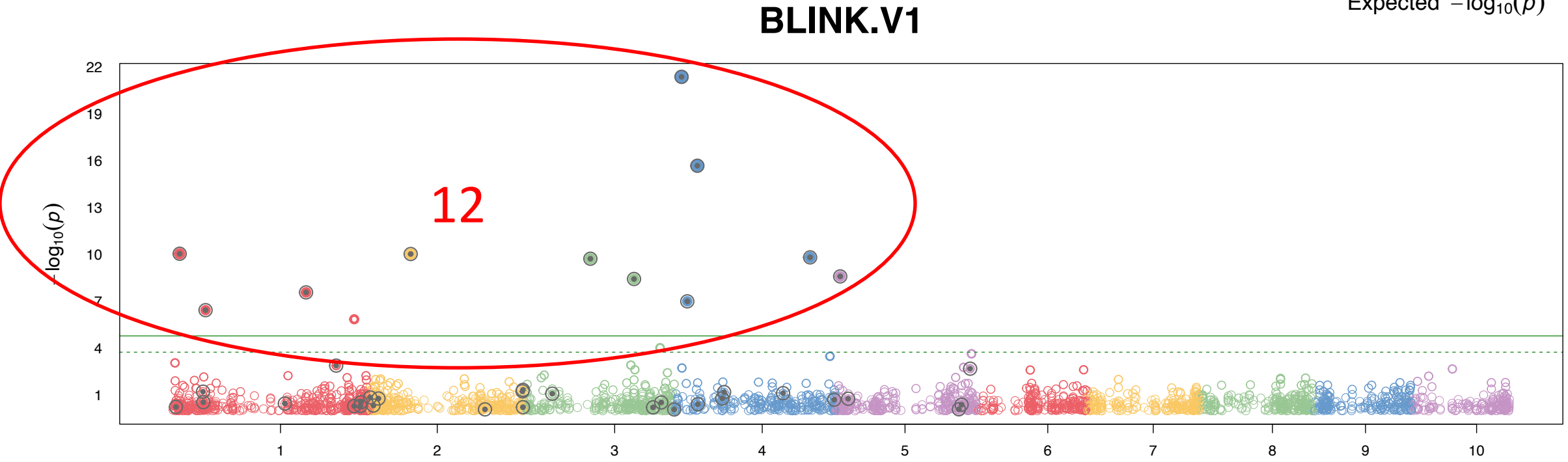


$y = s_i + S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e$, where i = 1 to M

BLINK (R in GAPIT)

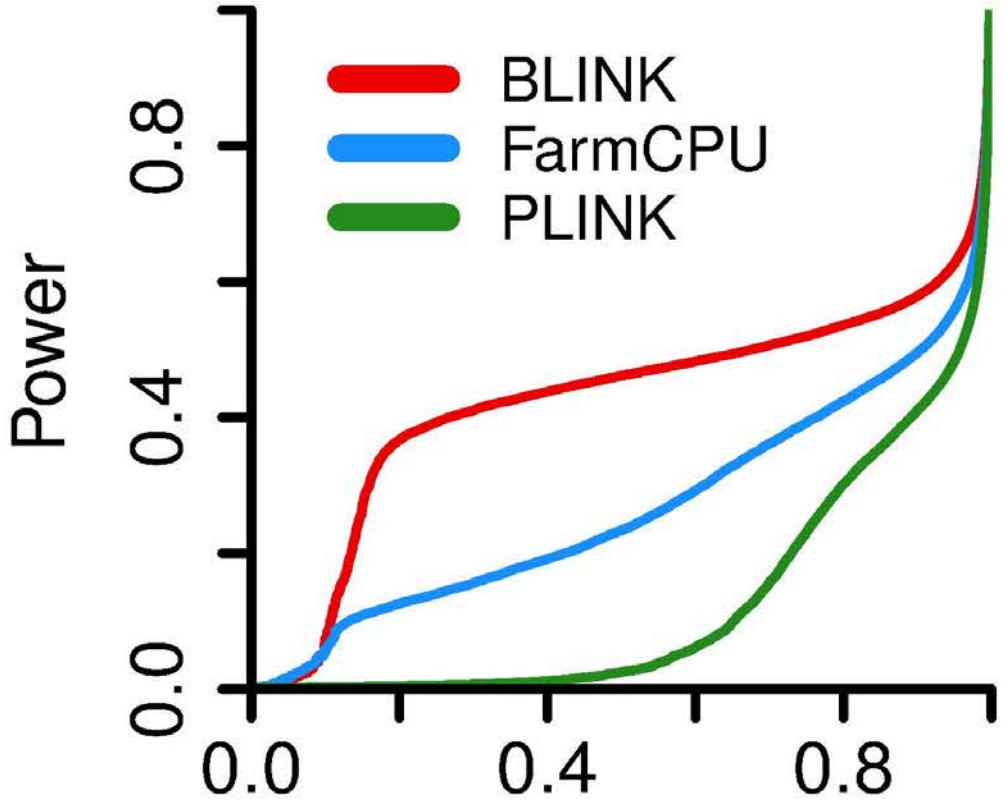
```
myGAPIT=GAPIT(
  Y=mySim$Y,
  GD=myGD,
  GM=myGM,
  QTN.position=mySim$QTN.position,
  PCA.total=3,
  model="BLINK")
```



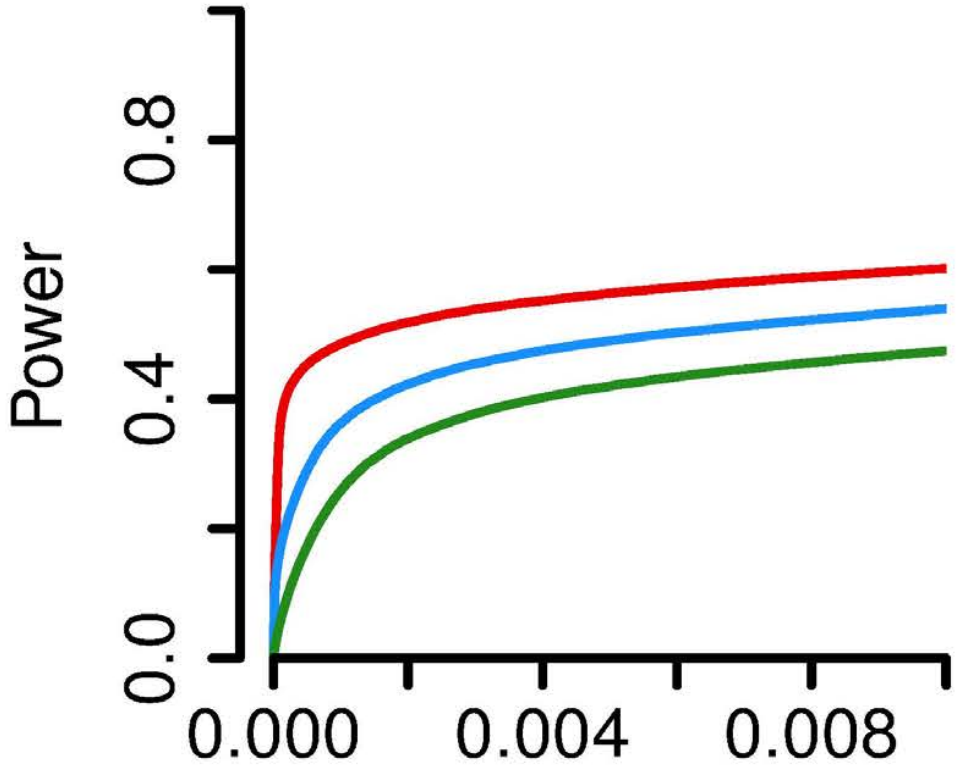


Simulation study with human data

Meng Huang

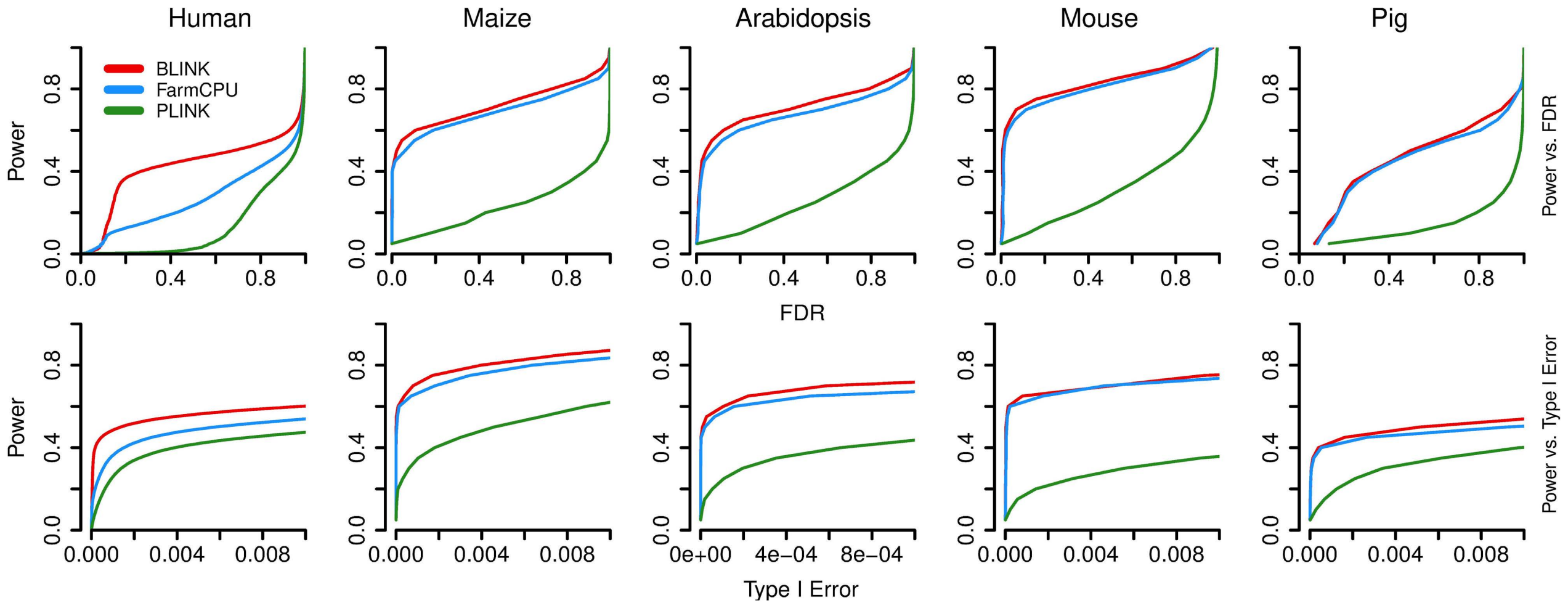


FDR

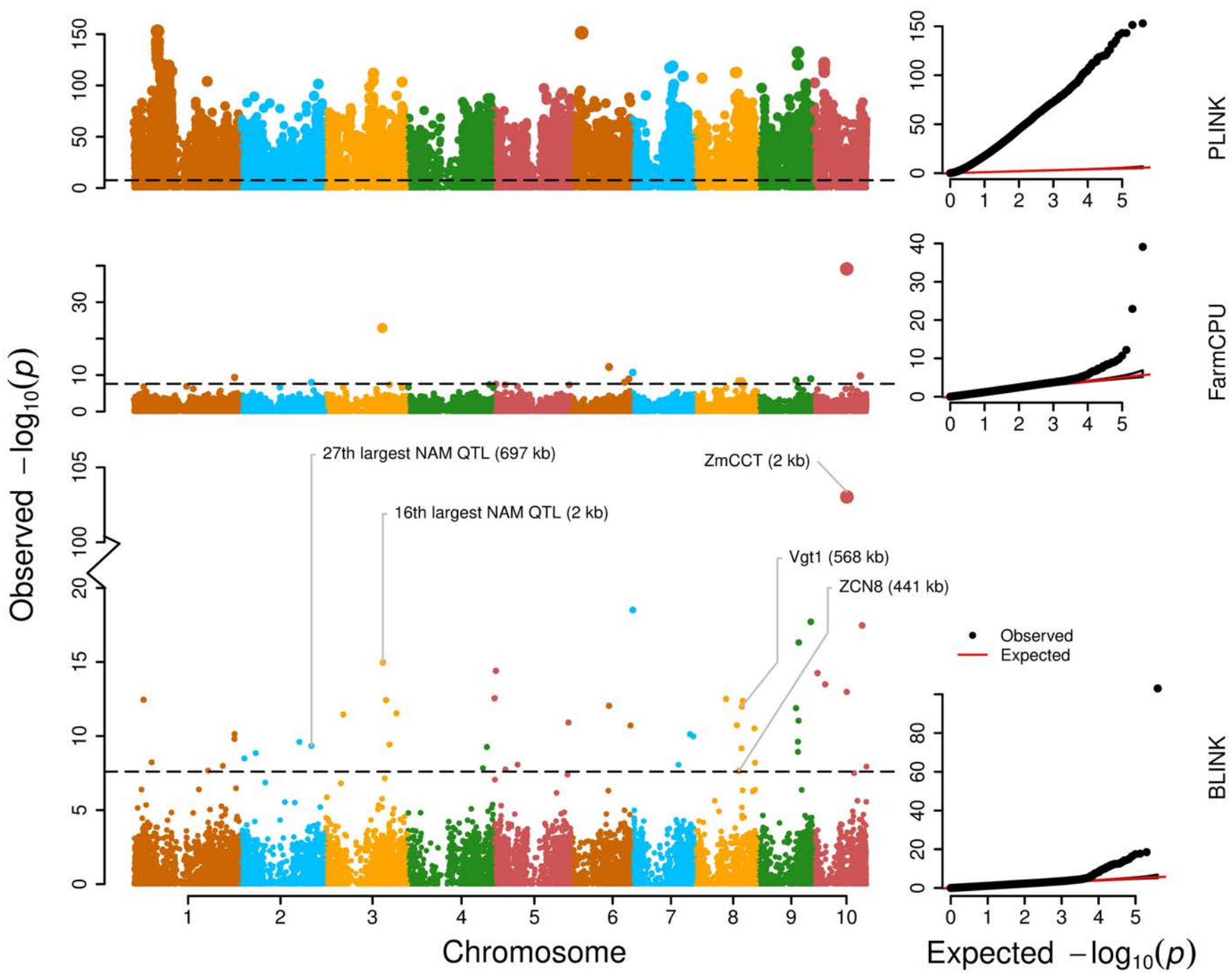


Type I error

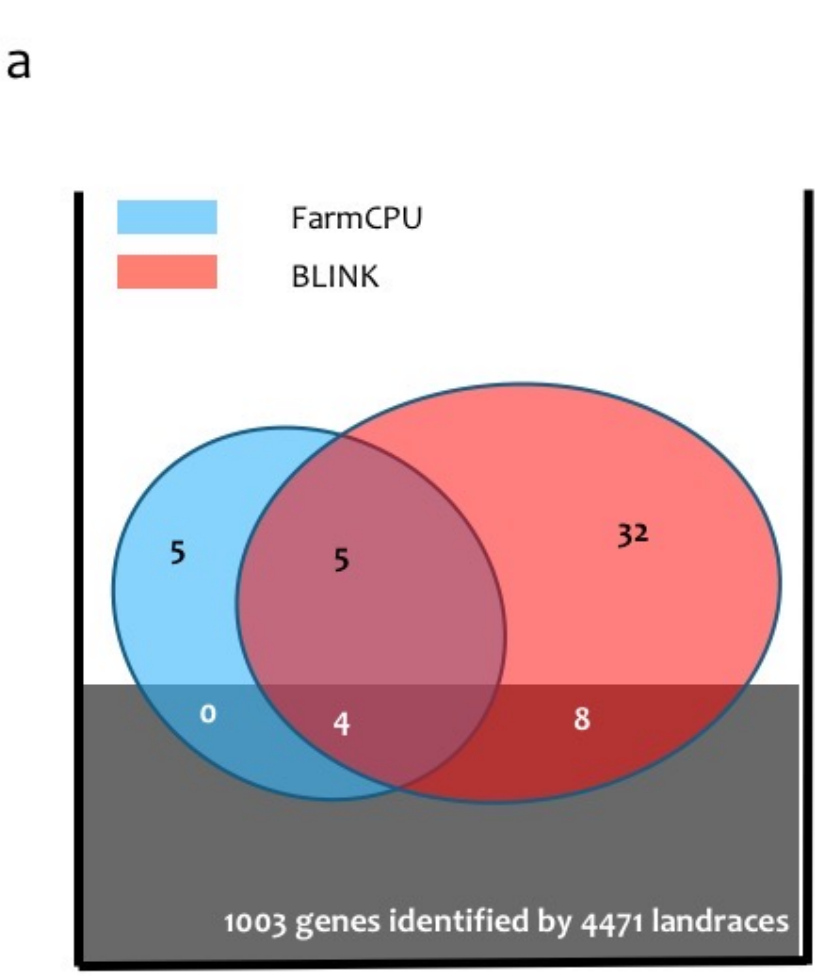
Same trend across species

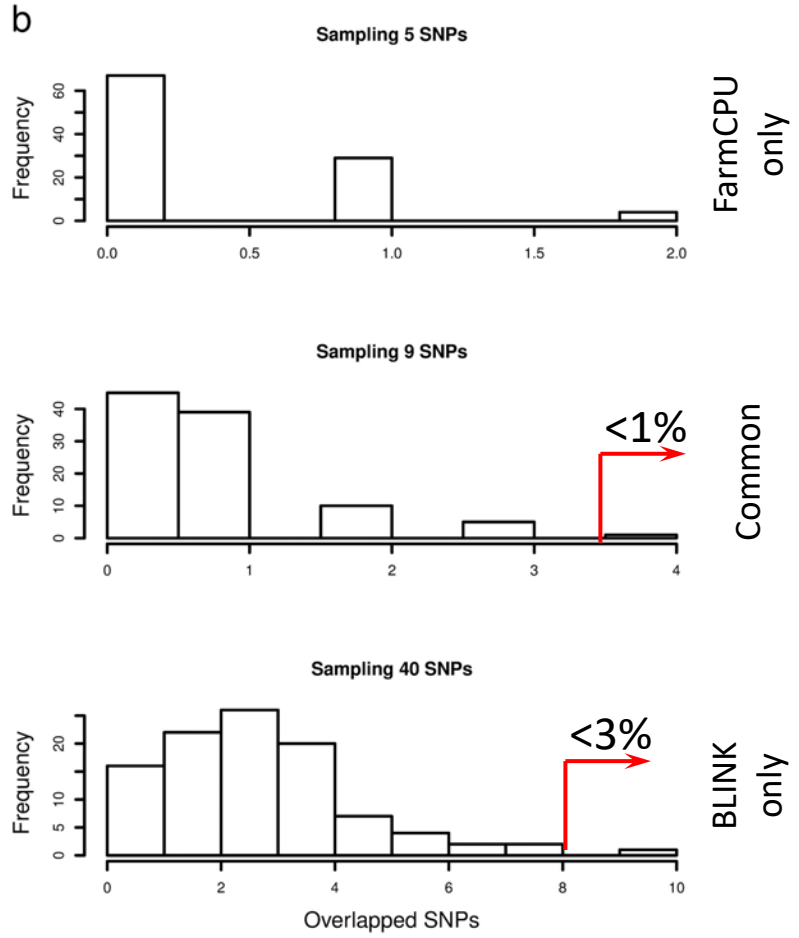


Application in Maize

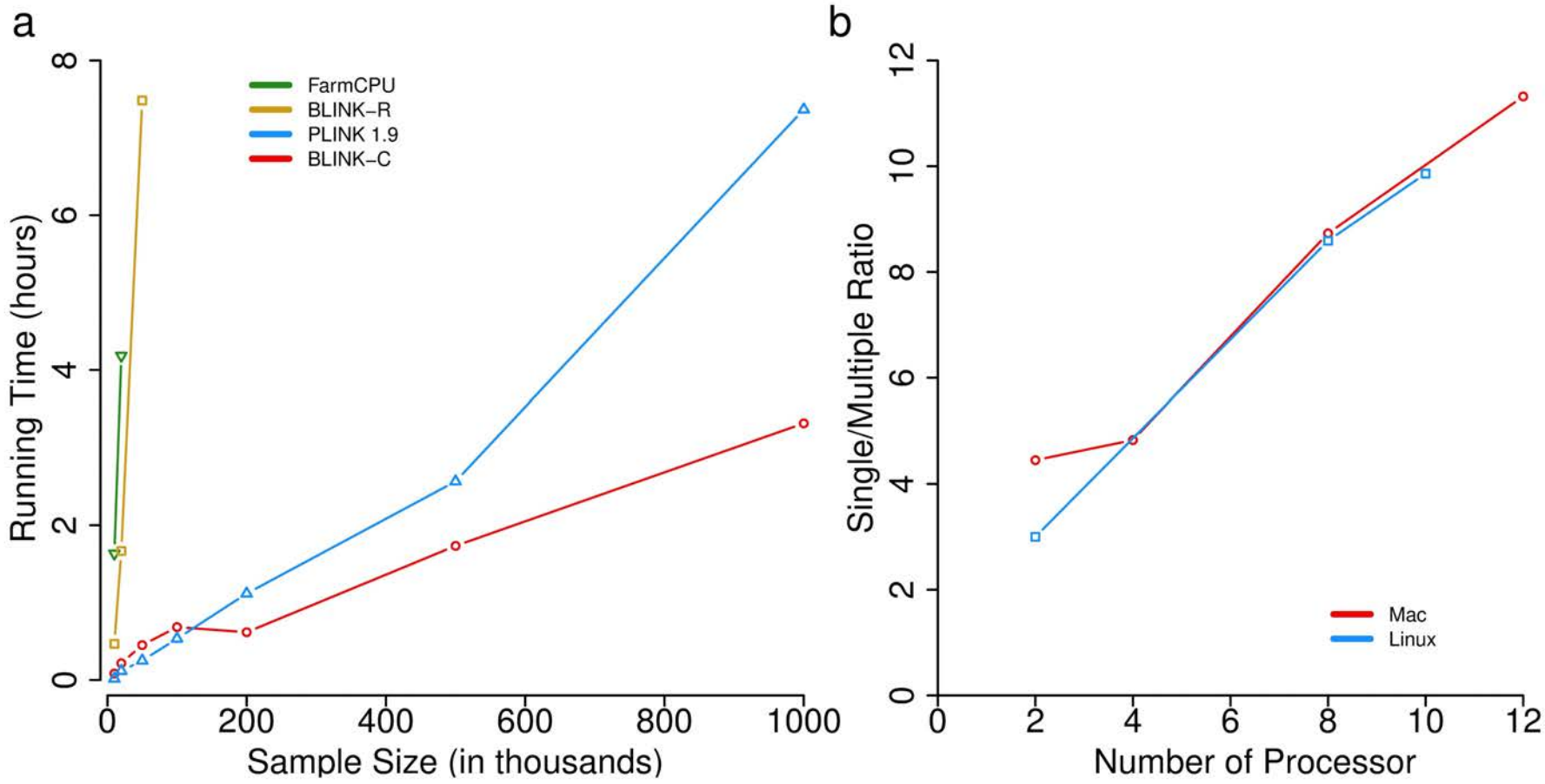


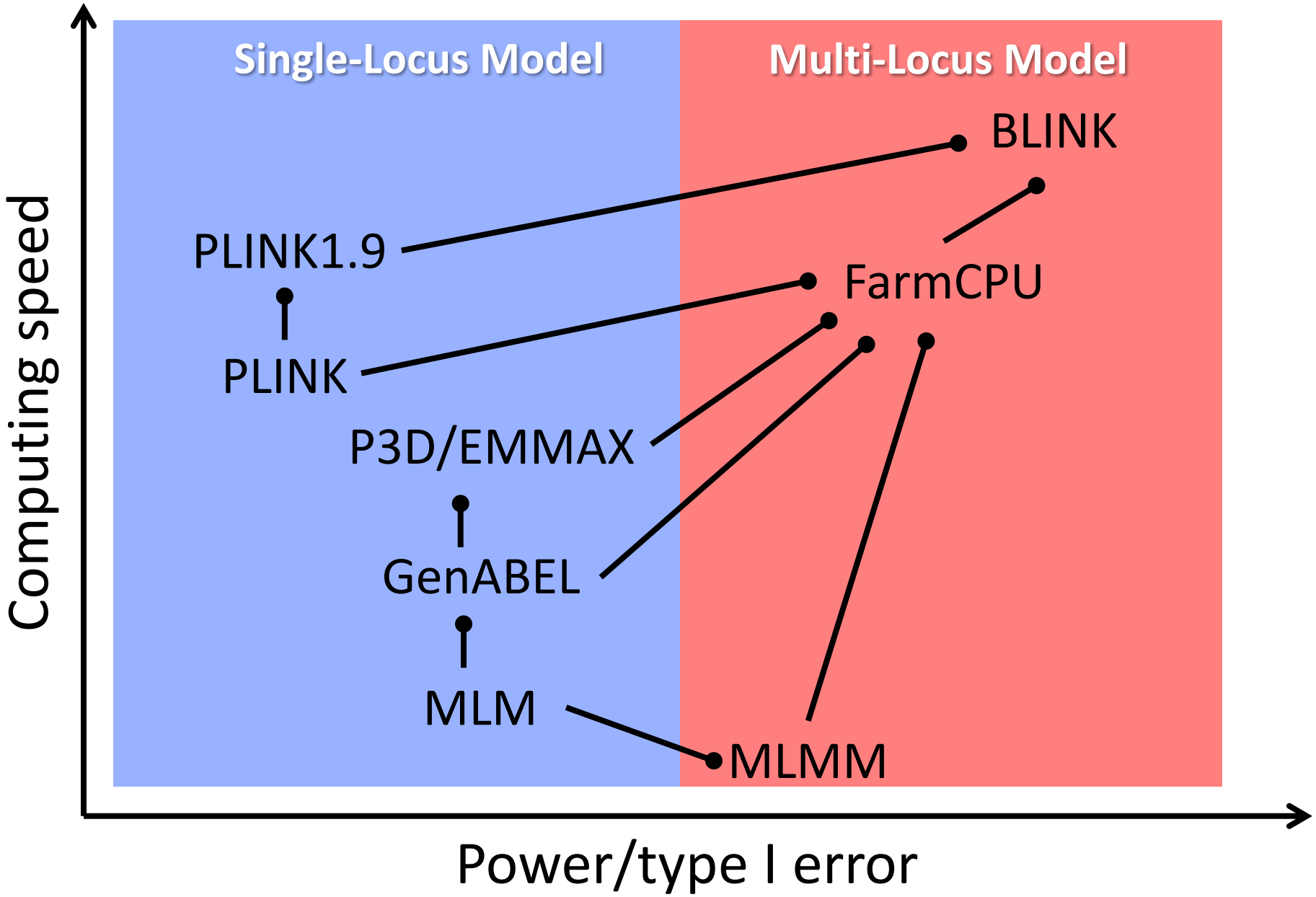
Enrichment


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

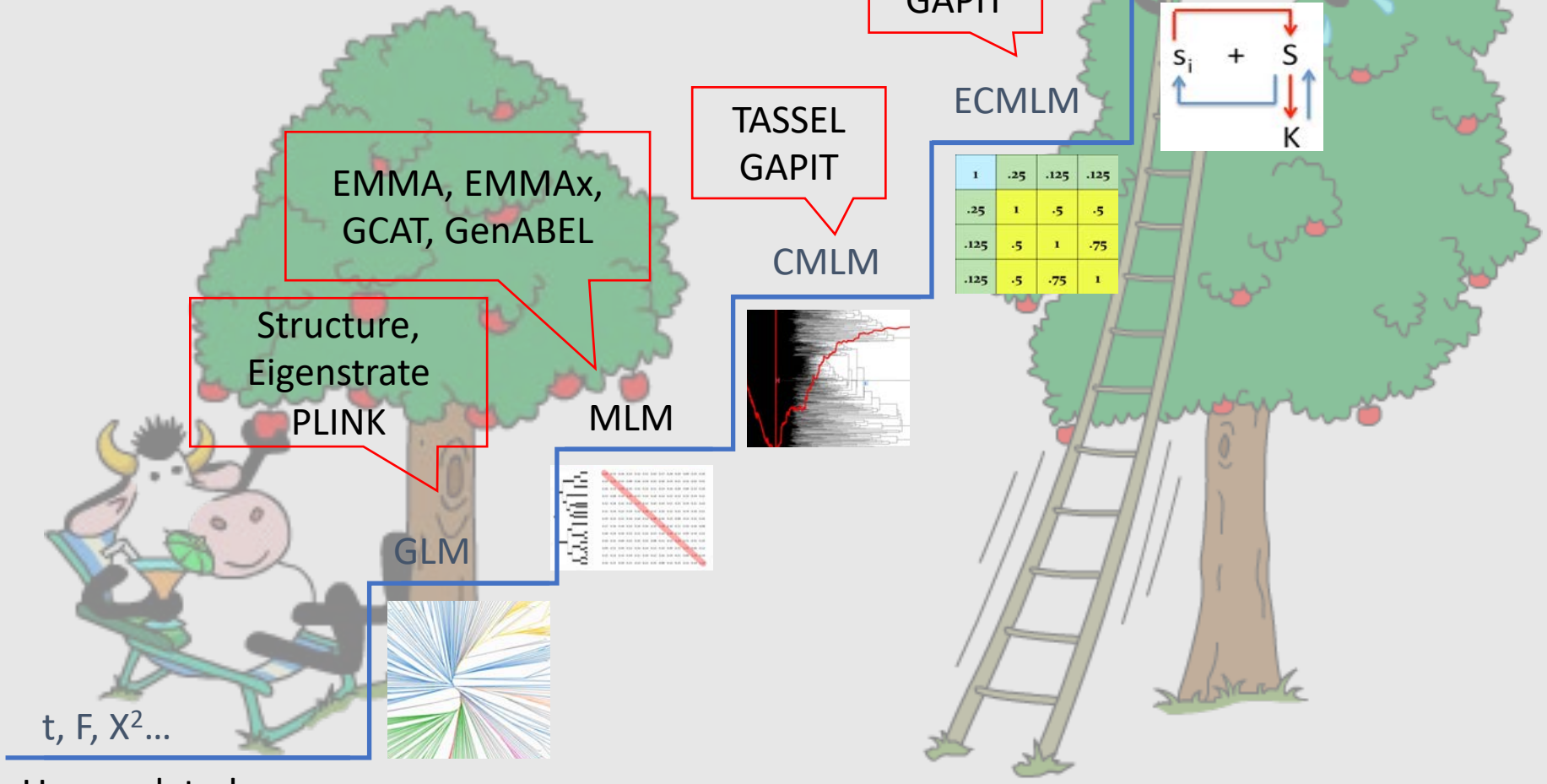

b
Sampling 5 SNPs
Frequency
FarmCPU only
Sampling 9 SNPs
<1%
Common
Sampling 40 SNPs
<3%
BLINK only
Overlapped SNPs

Computation efficiency

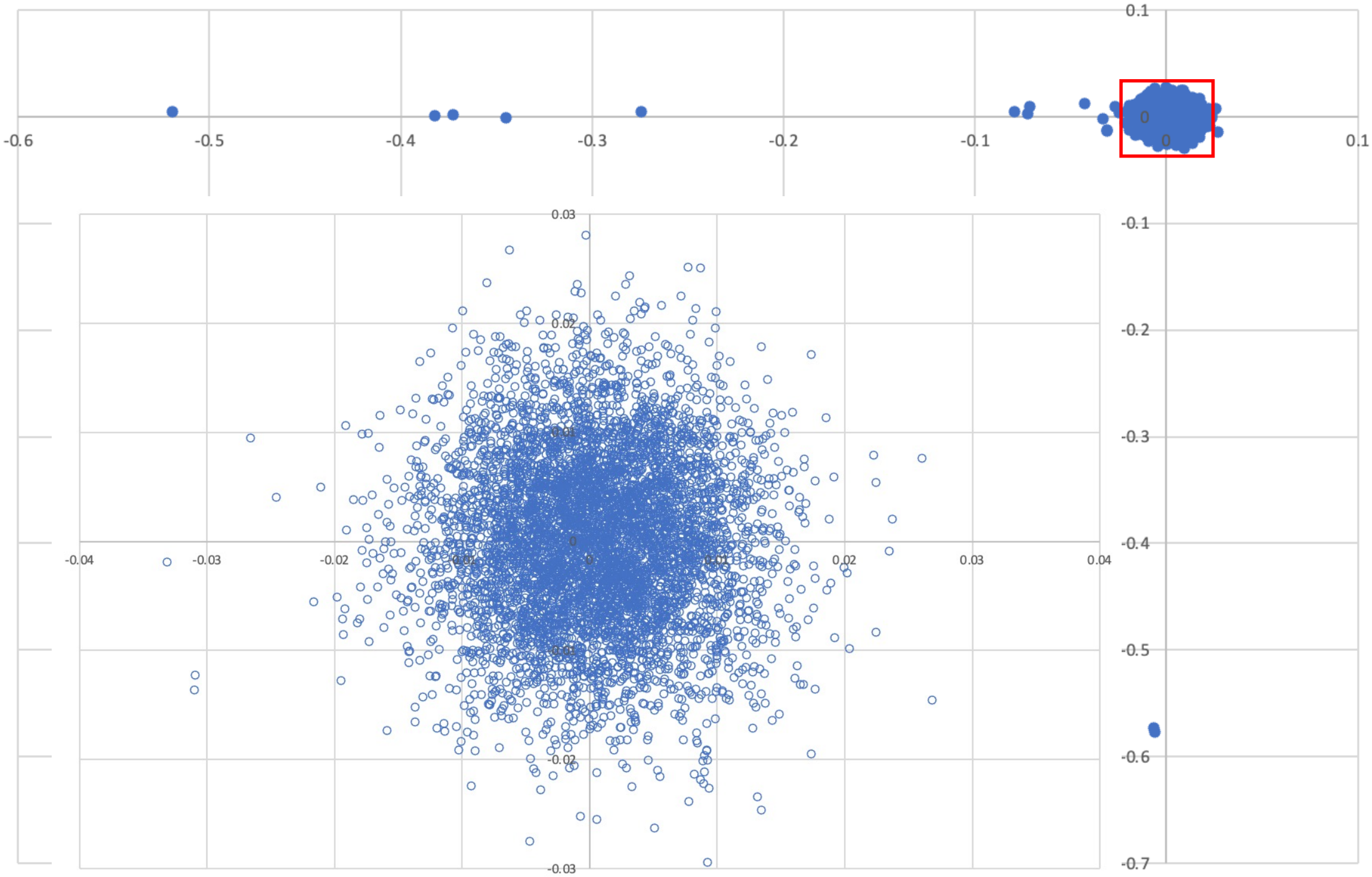



Single-Locus Model
Multi-Locus Model
BLINK
PLINK1.9
FarmCPU
PLINK
P3D/EMMAX
GenABEL
MLM
MLMM
Computing speed
Power/type I error

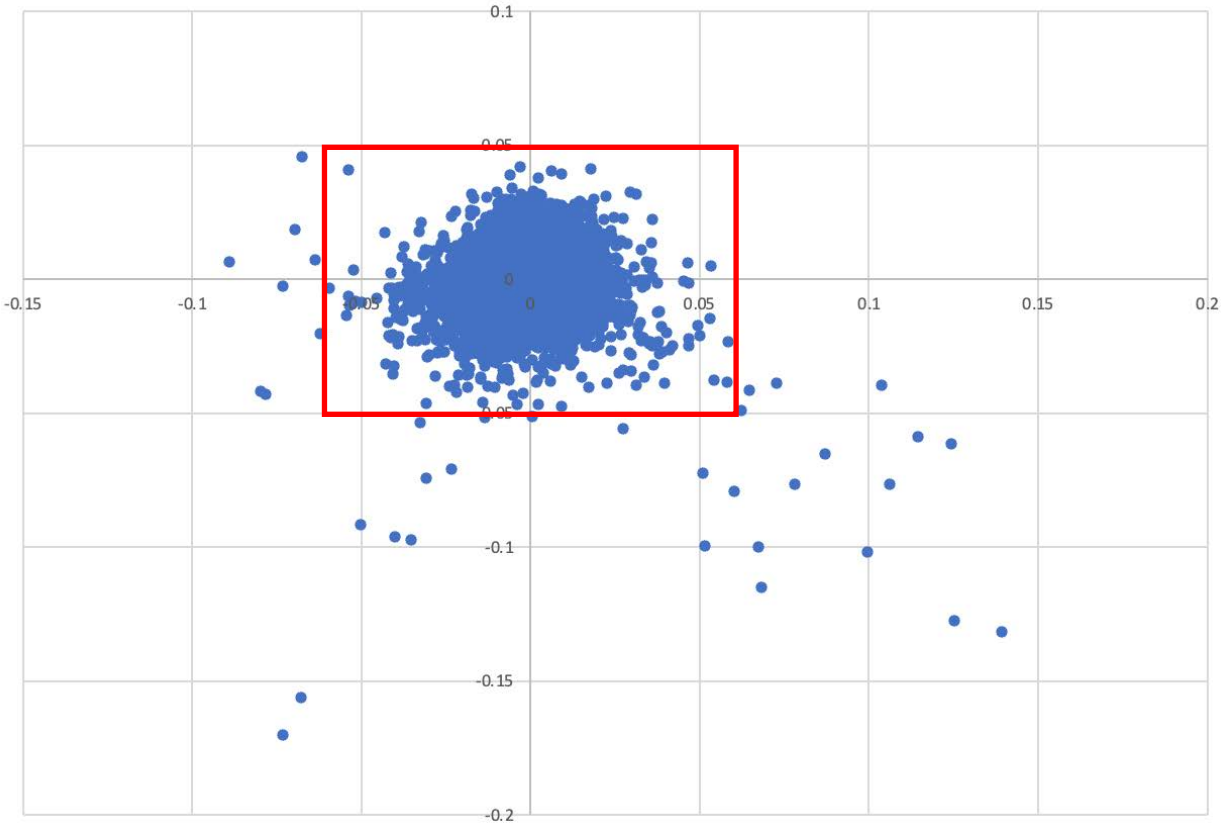
GAPIT

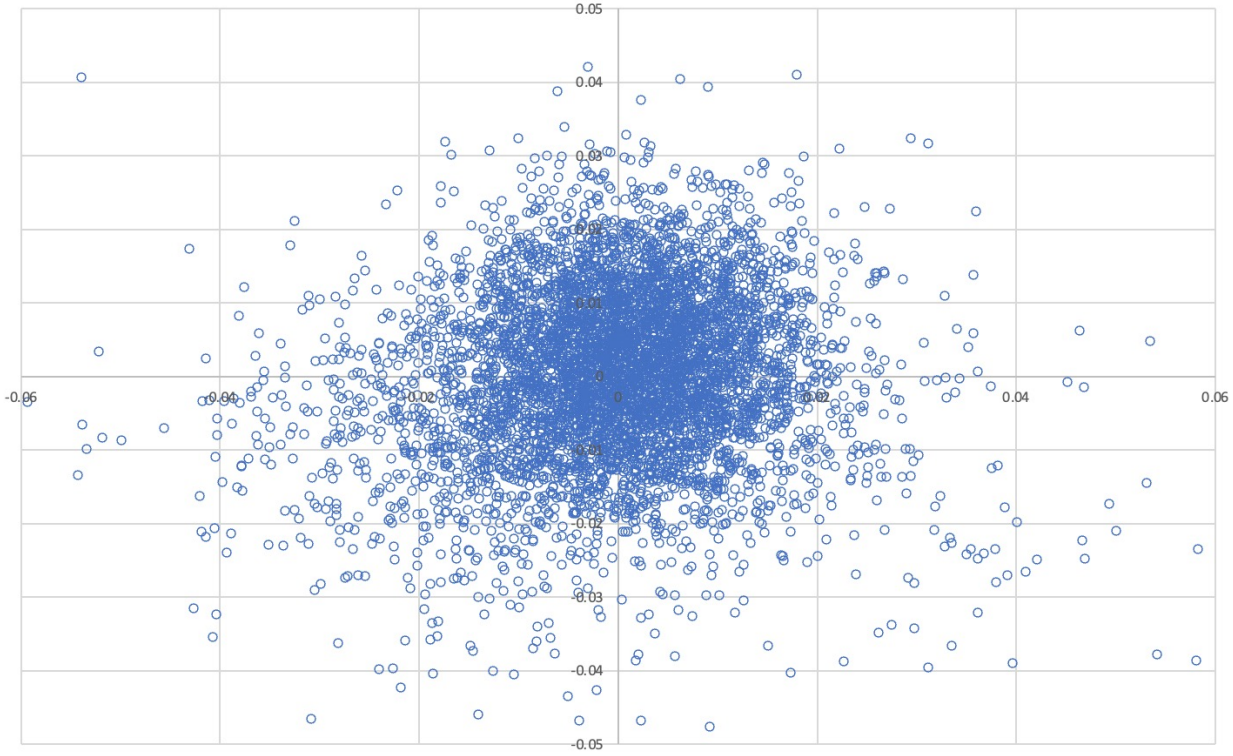

BLINK
FARM CPU
GAPIT
ECMLM
TASSEL
GAPIT
CMLM
EMMA, EMMAx,
GCAT, GenABEL
Structure,
Eigenstrate
PLINK
MLM
GLM
t, F, X2…
Uncorrelated or
equally correlated

PC3 vs 4


0.1
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7

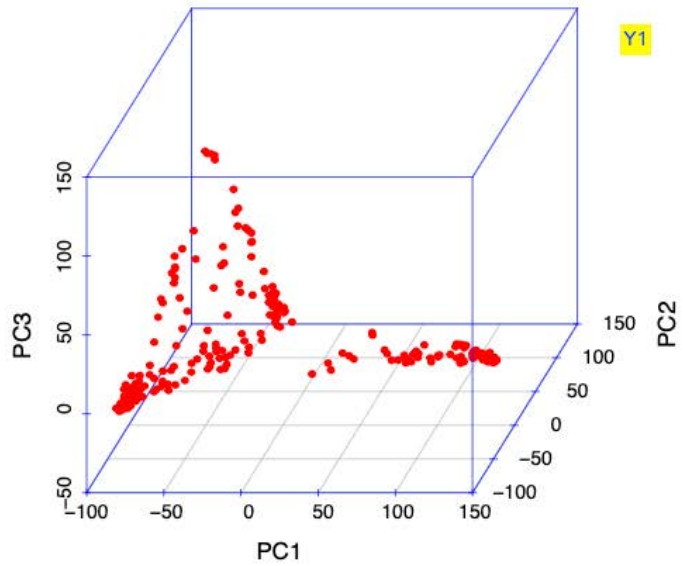
PC5 vs 6

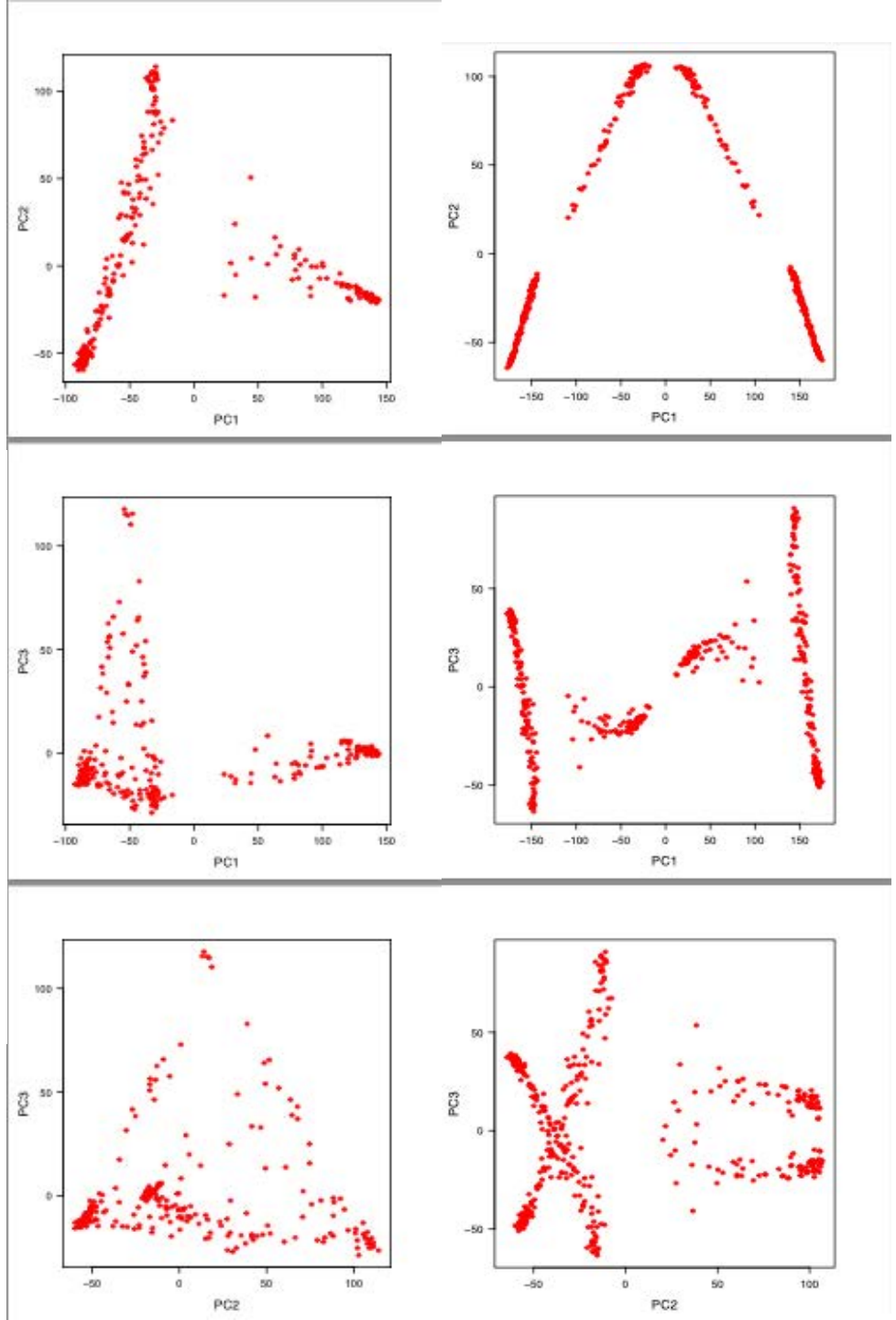

0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2


0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.04
-0.02
0
0.02
0.04
0.06

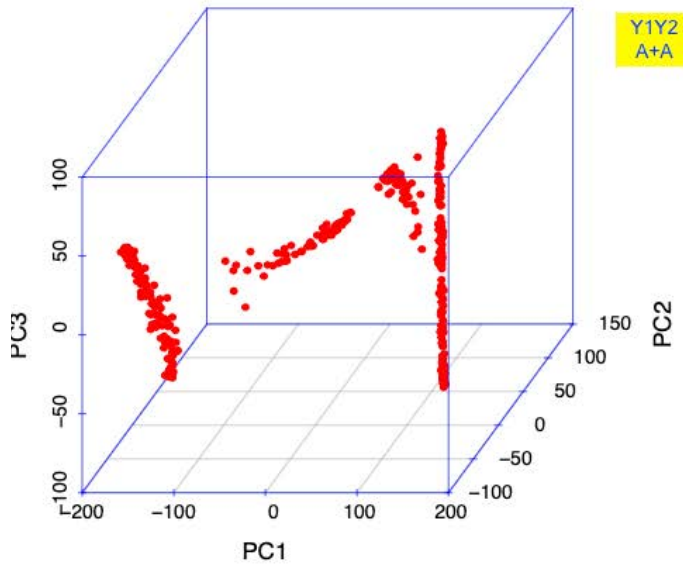
PC=f(G)
PCA.Total=3

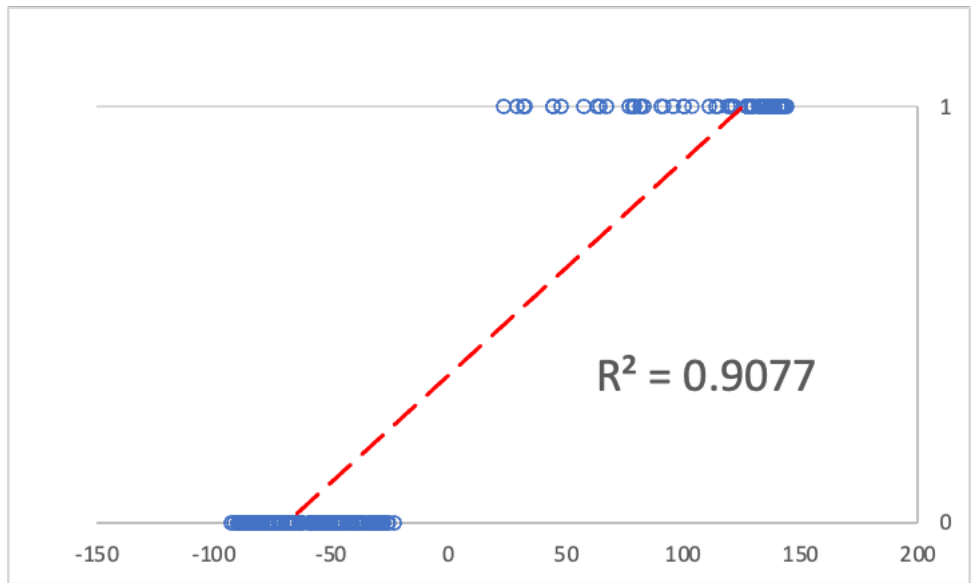
G=A



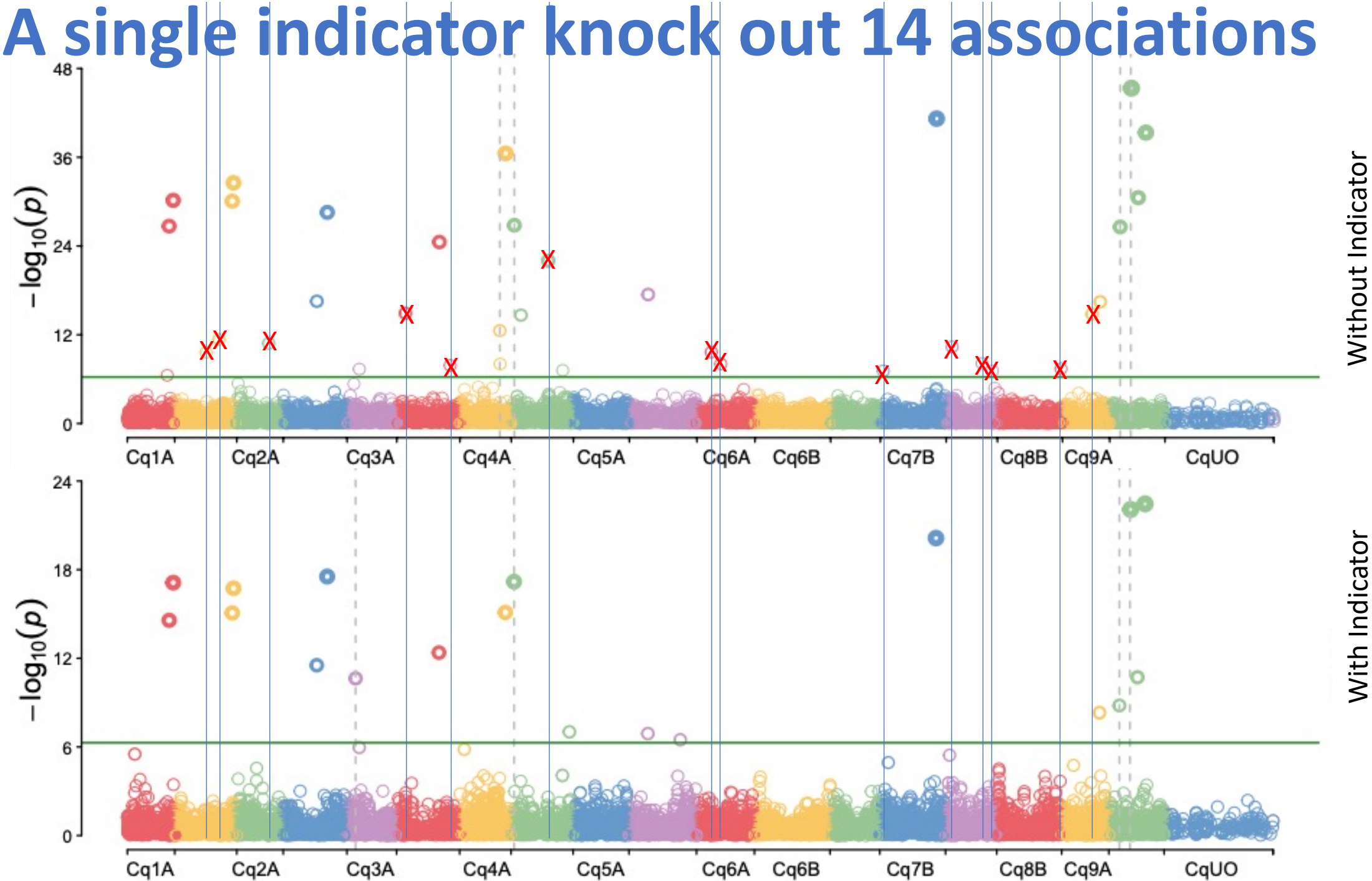


G=AO
OA



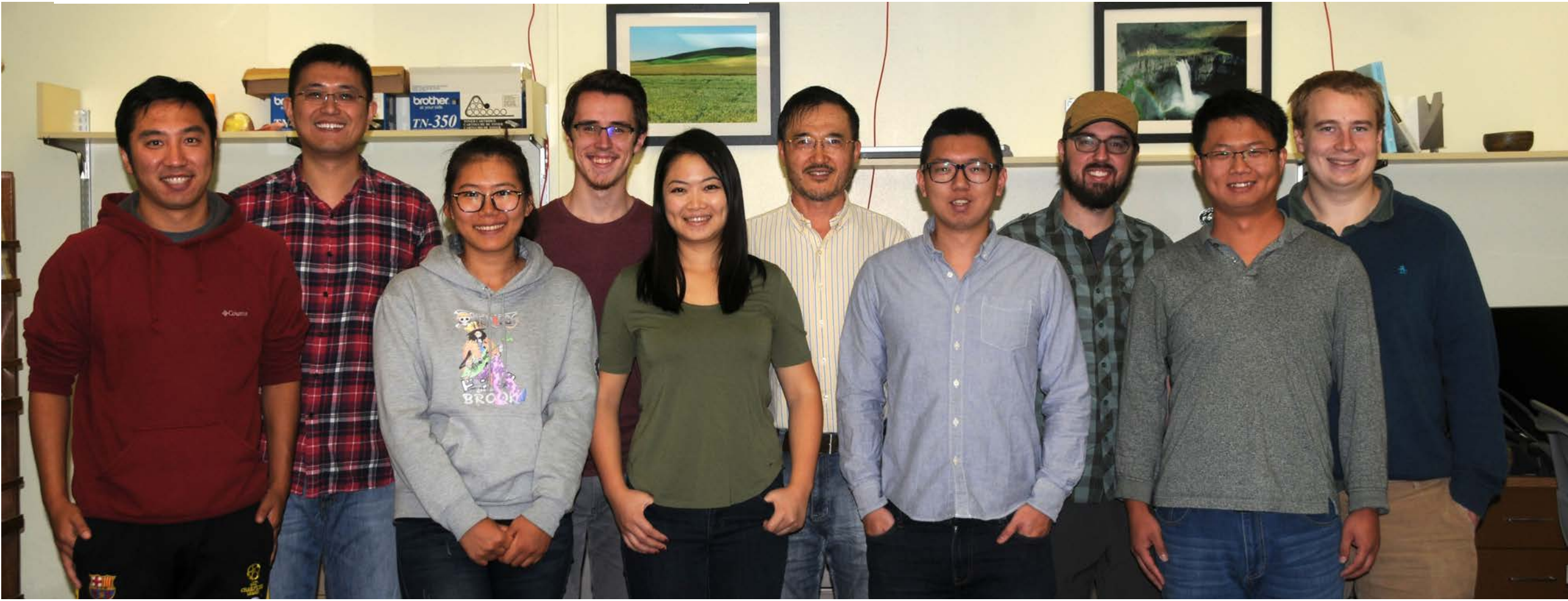

1
0
$R^2 = 0.9077$
-150
-100
-50
0
50
100
150
200

A single indicator knock out 14 associations



Zhiwu Zhang Laboratory
for Statistical Genomics

WASHINGTON STATE UNIVERSITY

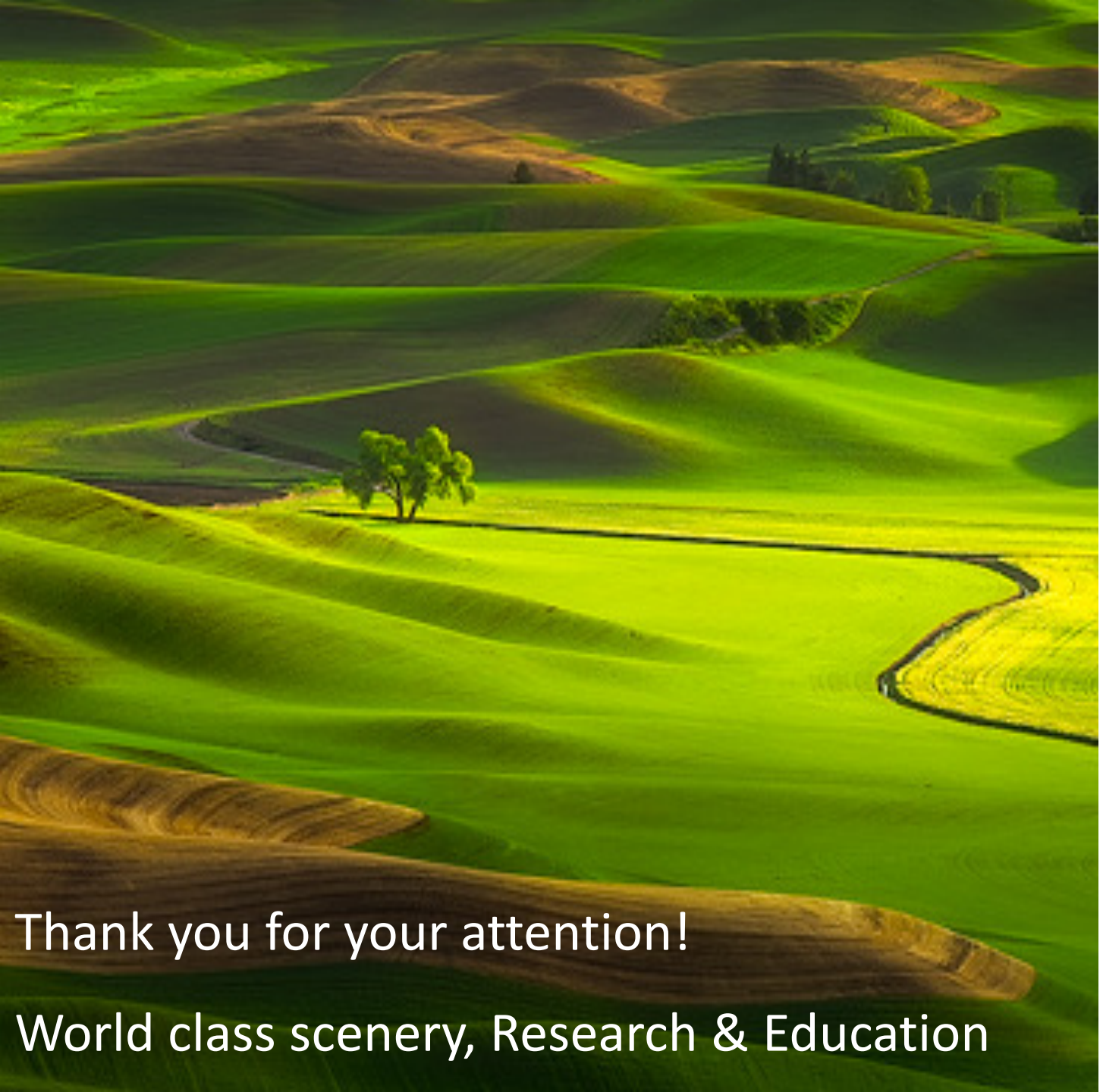


USDA

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

NSF

M.J. MURDOCK CHARITABLE TRUST



WASHINGTON STATE UNIVERSITY
Department of Crop and Soil Sciences

张志武教授在中国定向招收博士研究生数名

学习地点： 美国华盛顿州立大学，Pullman, WA, USA
招生对象： 国家留学基金委(CSC)奖学金(四年)获得者
资助内容： 提供学费、医疗保险，年度生活费补贴五千美金、GRE 豁免与协助 CSC 奖学金申请
资助来源： 美国农业部、华盛顿州谷物协会和华盛顿州立大学研究生院
研究方向： 图像处理或核酸数据分析
偏重技能： AI, GIS, Fixed and Random Effect Mixed Model, Bayesian Analysis 与计算编程
申报条件： 英语 TOEFL 80 分或 IELTS 7 分以上(英语成绩和申请推荐信可后补)
递交申请： http://css.wsu.edu/graduate-studies：Financial aid 填写"Pending CSC Application"，CSC 申请与咨询致信(Zhiwu.Zhang@WSU.EDU)或微信(zhiwu-zhang)
导师信息： Zhiwu Zhang Laboratory (http://zzlab.net)

PO Box 646420, Pullman, WA 99164-6420
509-335-3475 • Fax: 509-335-8674 • http://css.wsu.edu