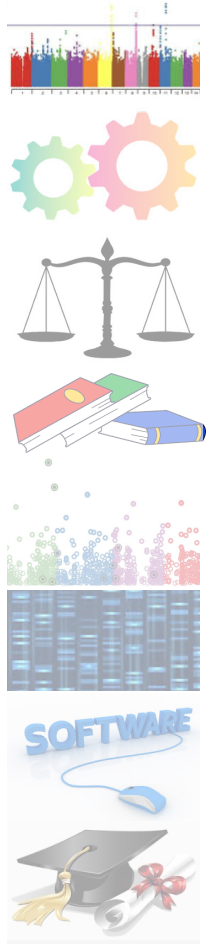
Genome wide Association Study

Zhiwu Zhang

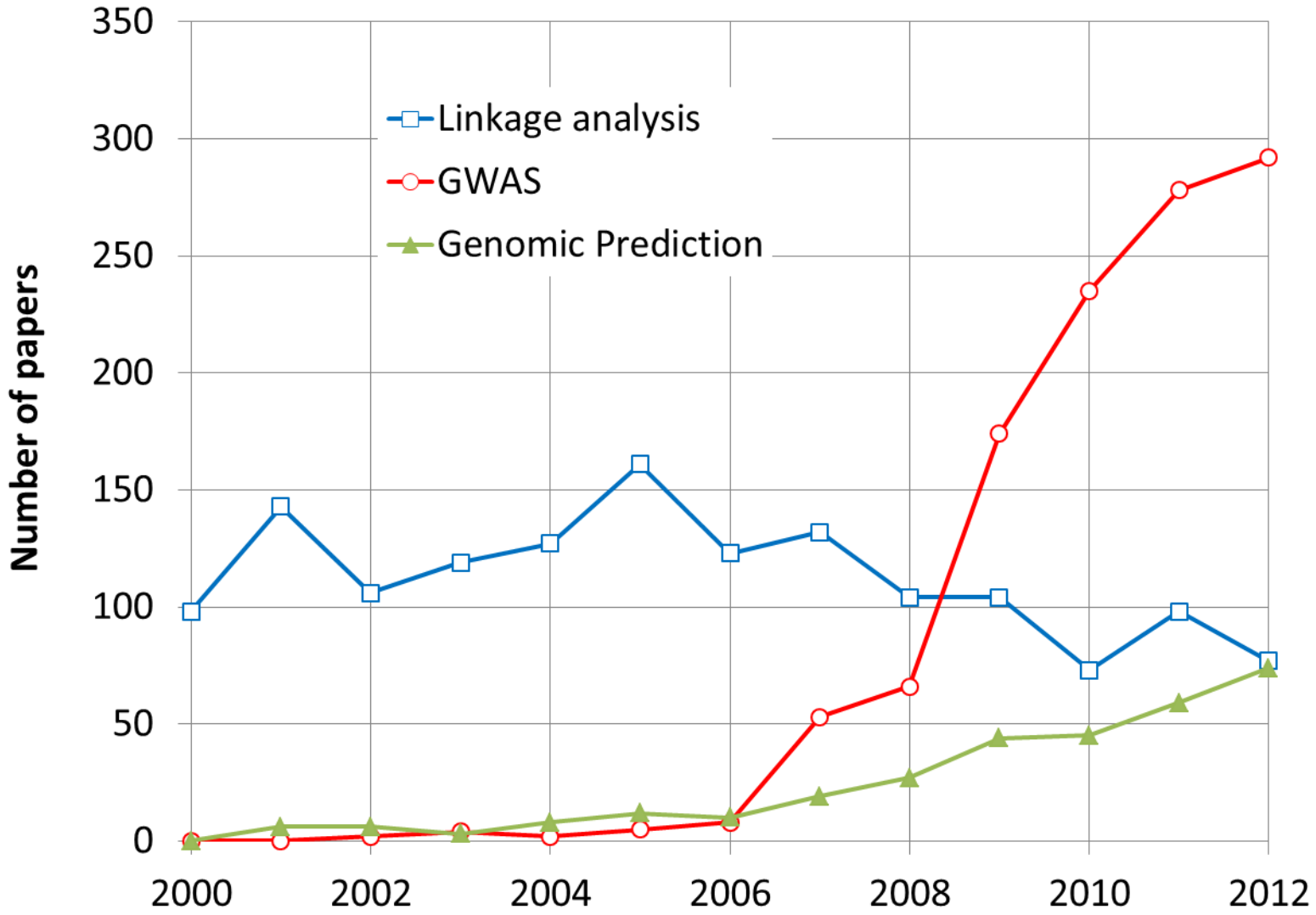
Washington State University

Outline

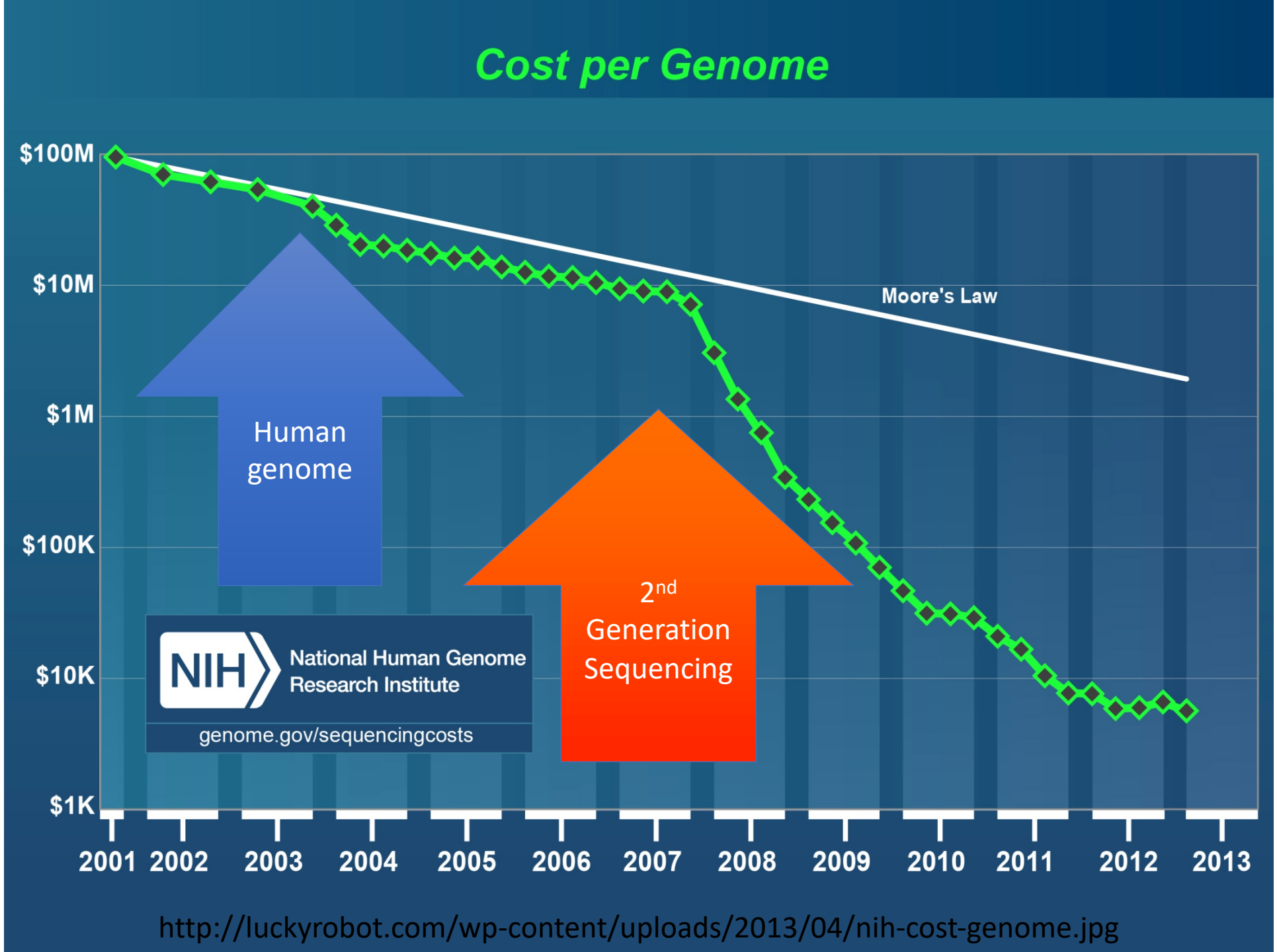
- **Why GWAS?**
- How does GWAS work?
- How to evaluate GWAS results?
  - Literature
  - Simulation
  - Enrichment analysis
- GWAS Software
- GWAS course at Washington State University



More Research on GWAS and GS



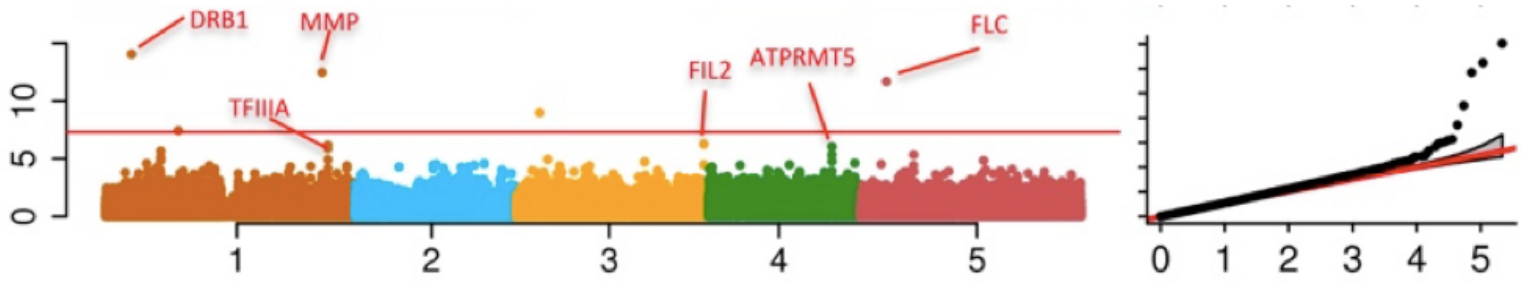
By May 31, 2013



http://luckyrobot.com/wp-content/uploads/2013/04/nih-cost-genome.jpg

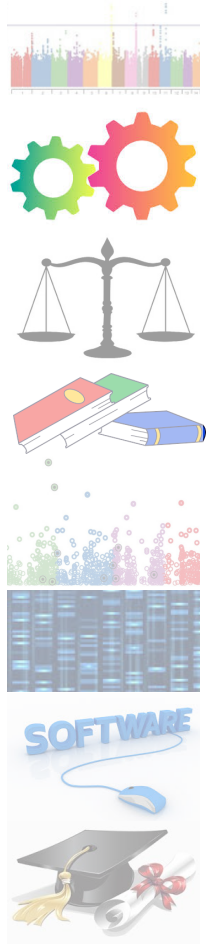
Problems in GWAS

- 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

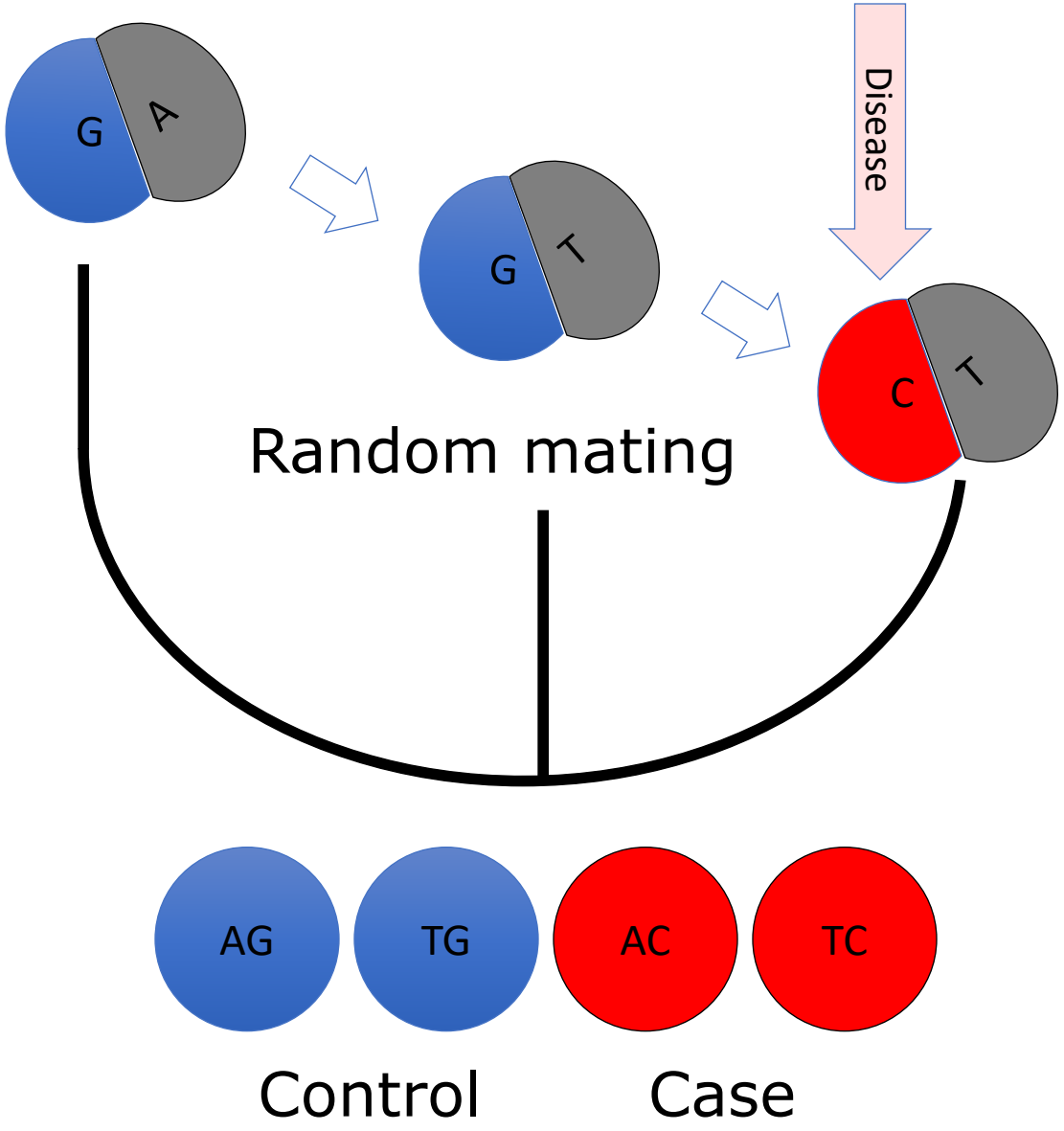


Outline

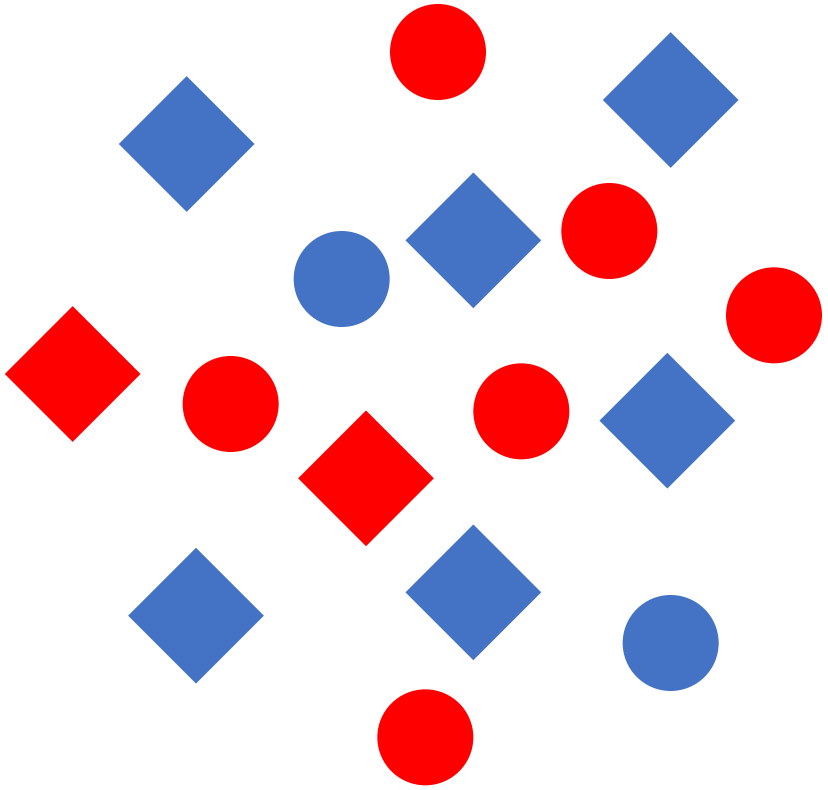
- Why GWAS?
- **How does GWAS work?**
- How to evaluate GWAS results?
  - Literature
  - Simulation
  - Enrichment analysis
- GWAS Software
- GWAS course at Washington State University



Linkage equilibrium



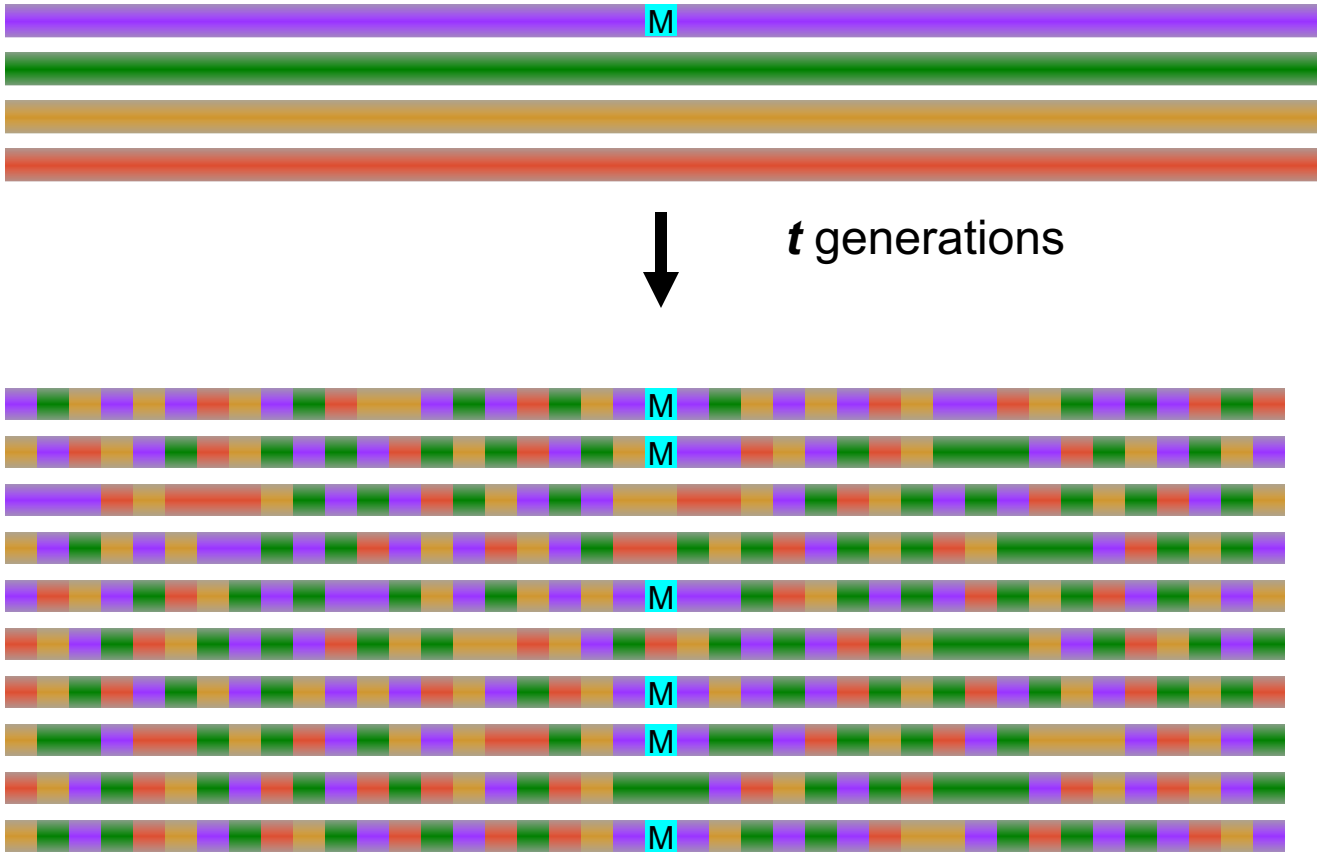
Association study



Marker	Control	Case
◇	6	2
○	2	6

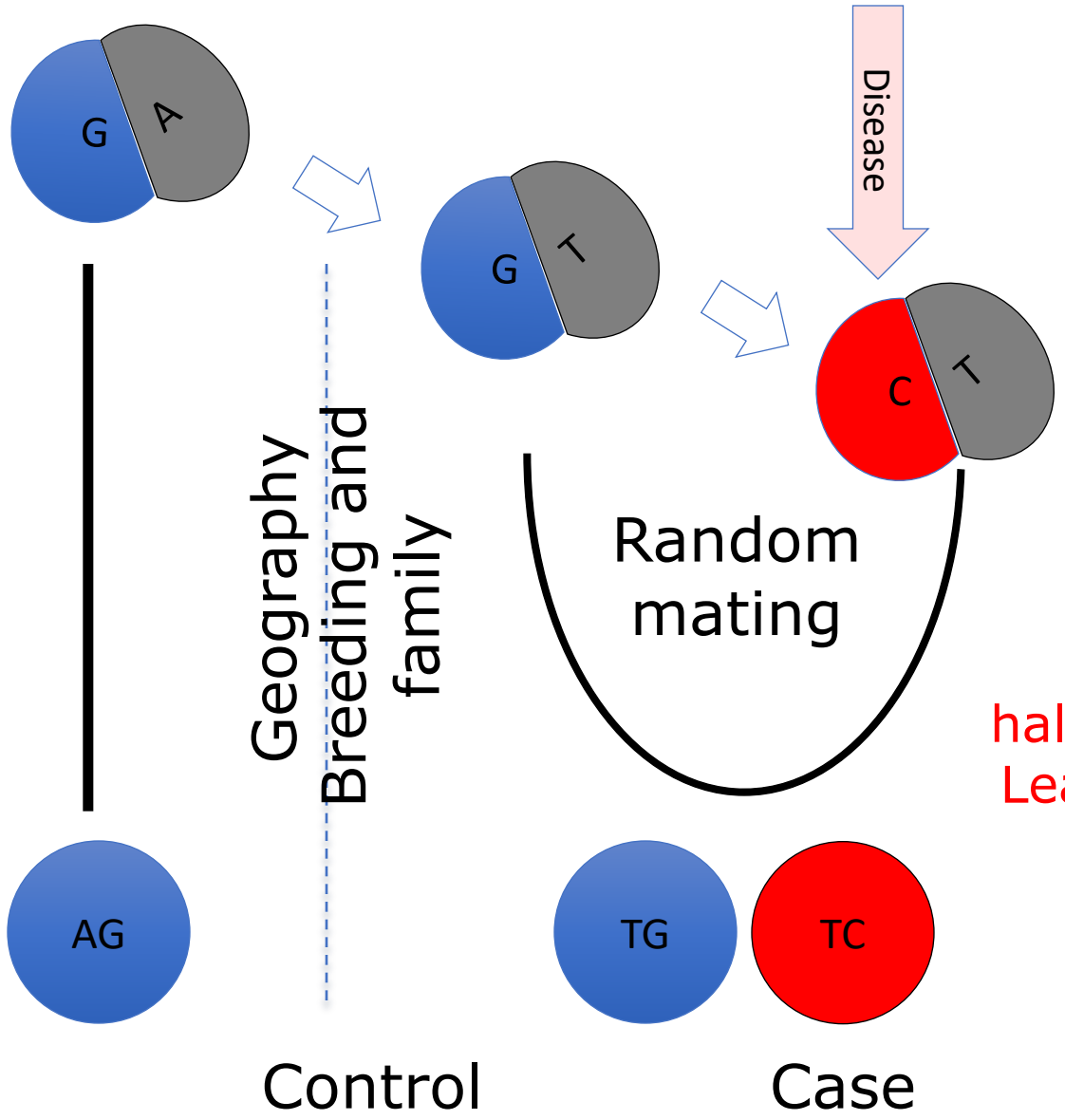
$X^2=4(2*2/4)=4$, df=1,
P=4.5%

Association study via linkage disequilibrium

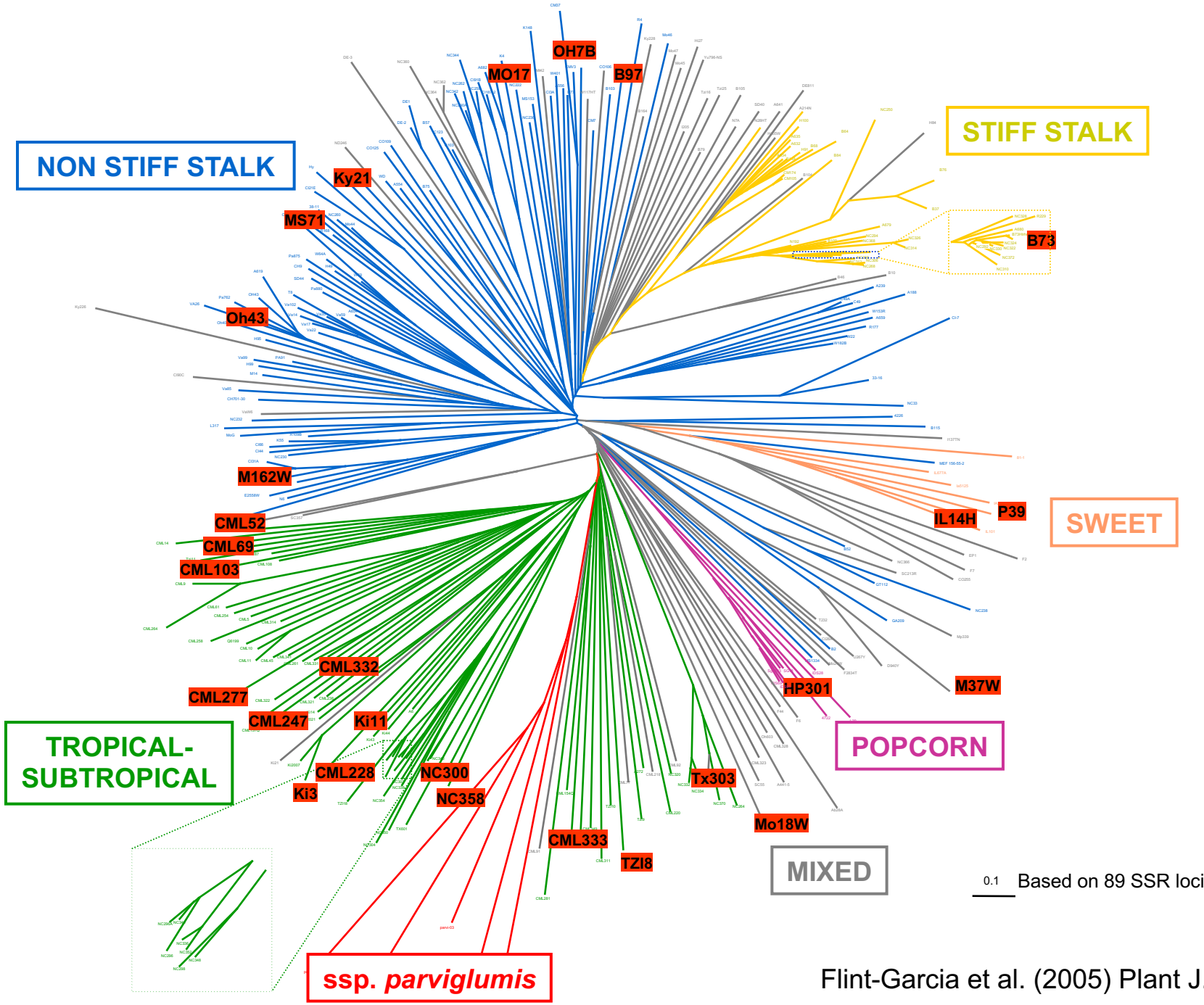


Jianming Yu, 2011

Linkage disequilibrium (LD)

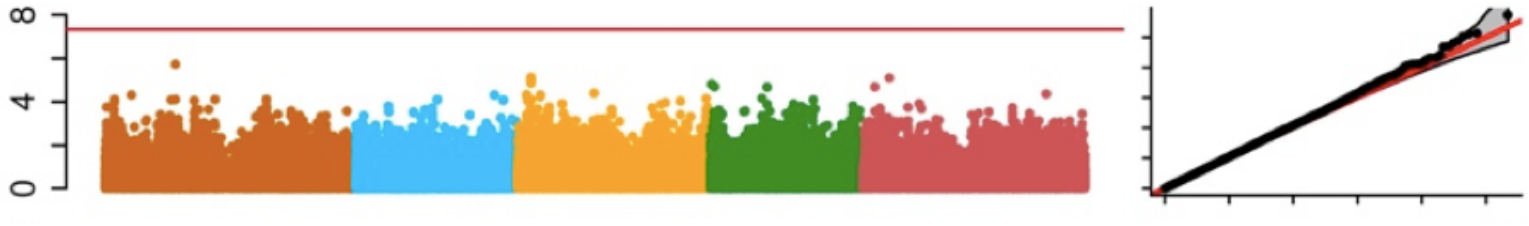


Association
half T as case, none for A
Lead to mistake if G/C is
not a marker

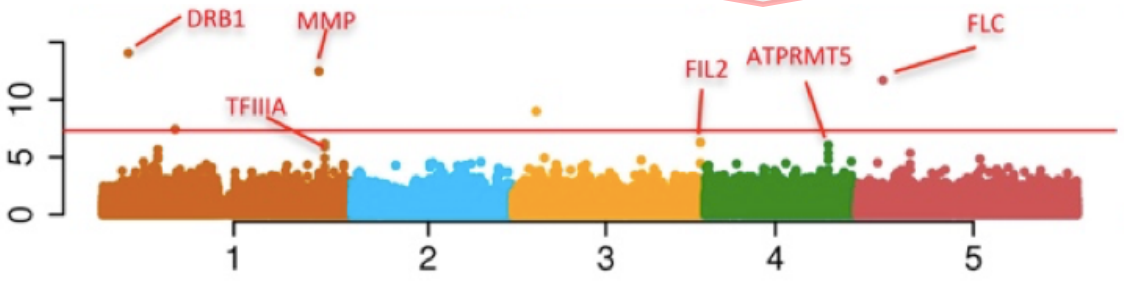


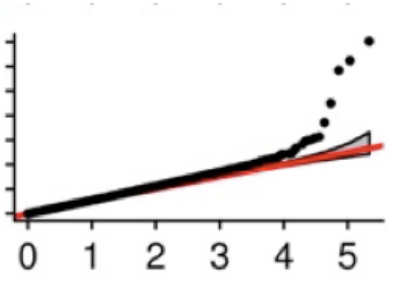
Flint-Garcia et al. (2005) Plant J. 44: 1054

Factors Affecting Statistical Power



- Number of genes
- Gene effect size
- Heritability
- Population size
- Marker density
- Population structure
- Resolution
- LD decade
- Multiple test correction
- Statistical methods





Outline

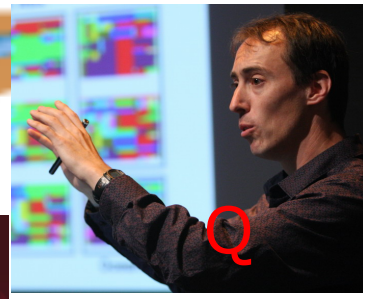
- Why GWAS?
- How does GWAS work?
- **How to evaluate GWAS results?**
  - **Literature**
  - Simulation
  - Enrichment analysis
- GWAS Software
- GWAS course at Washington State University


EMMAx


EMMA


PC+K


PC


Q


Q+K

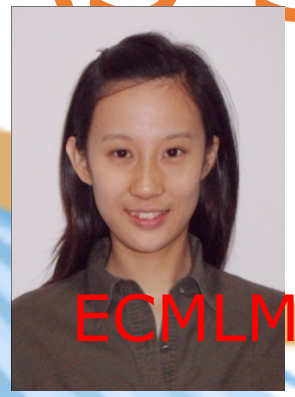

P3D


SELECT

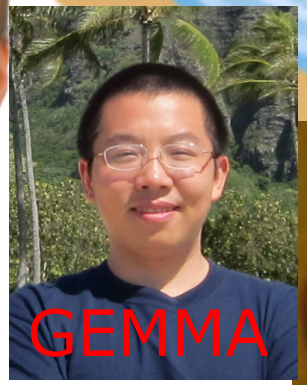

MLMM


CMLM


GCTA


ECMLM


FST-LMM

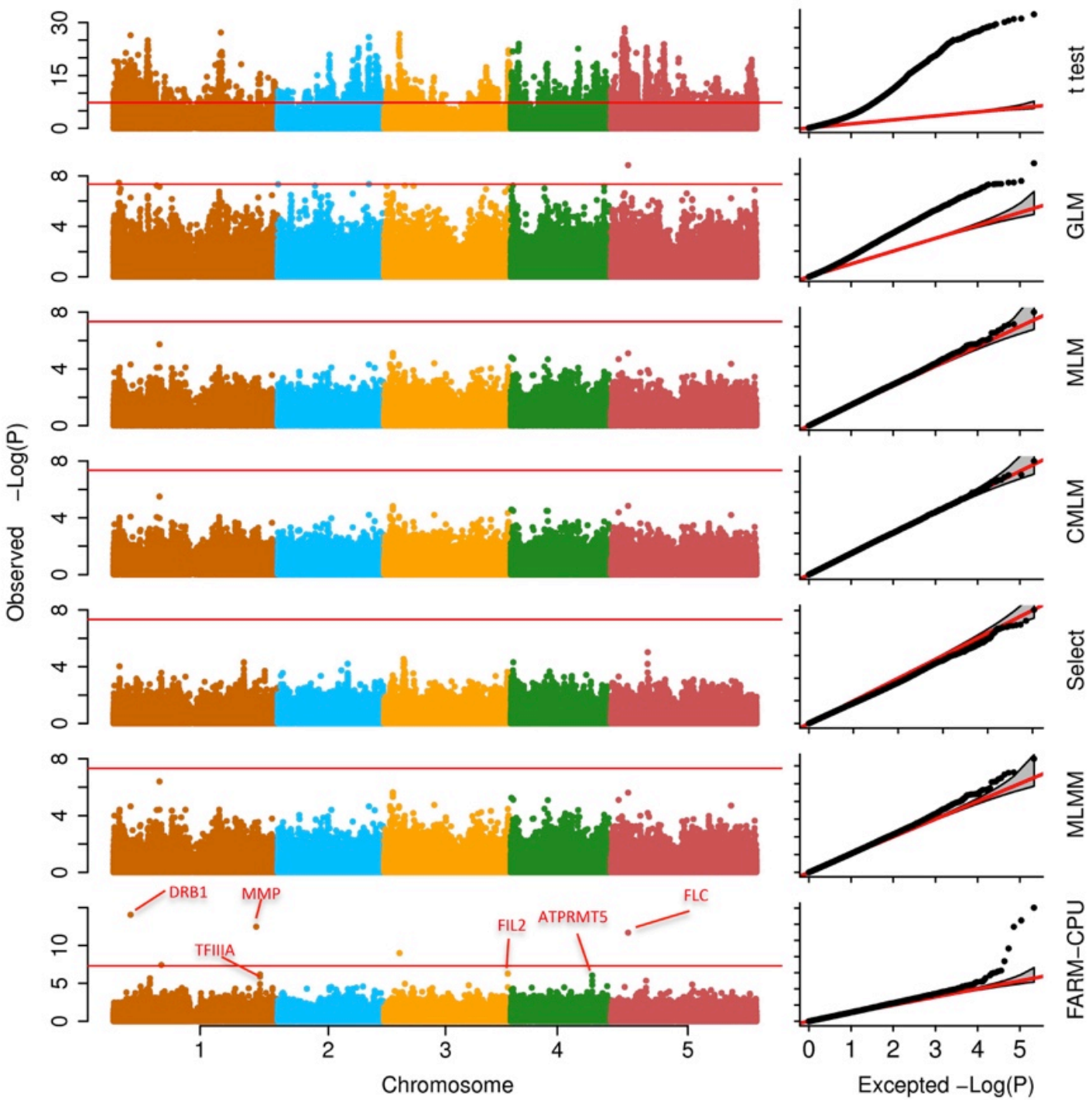

GEMMA

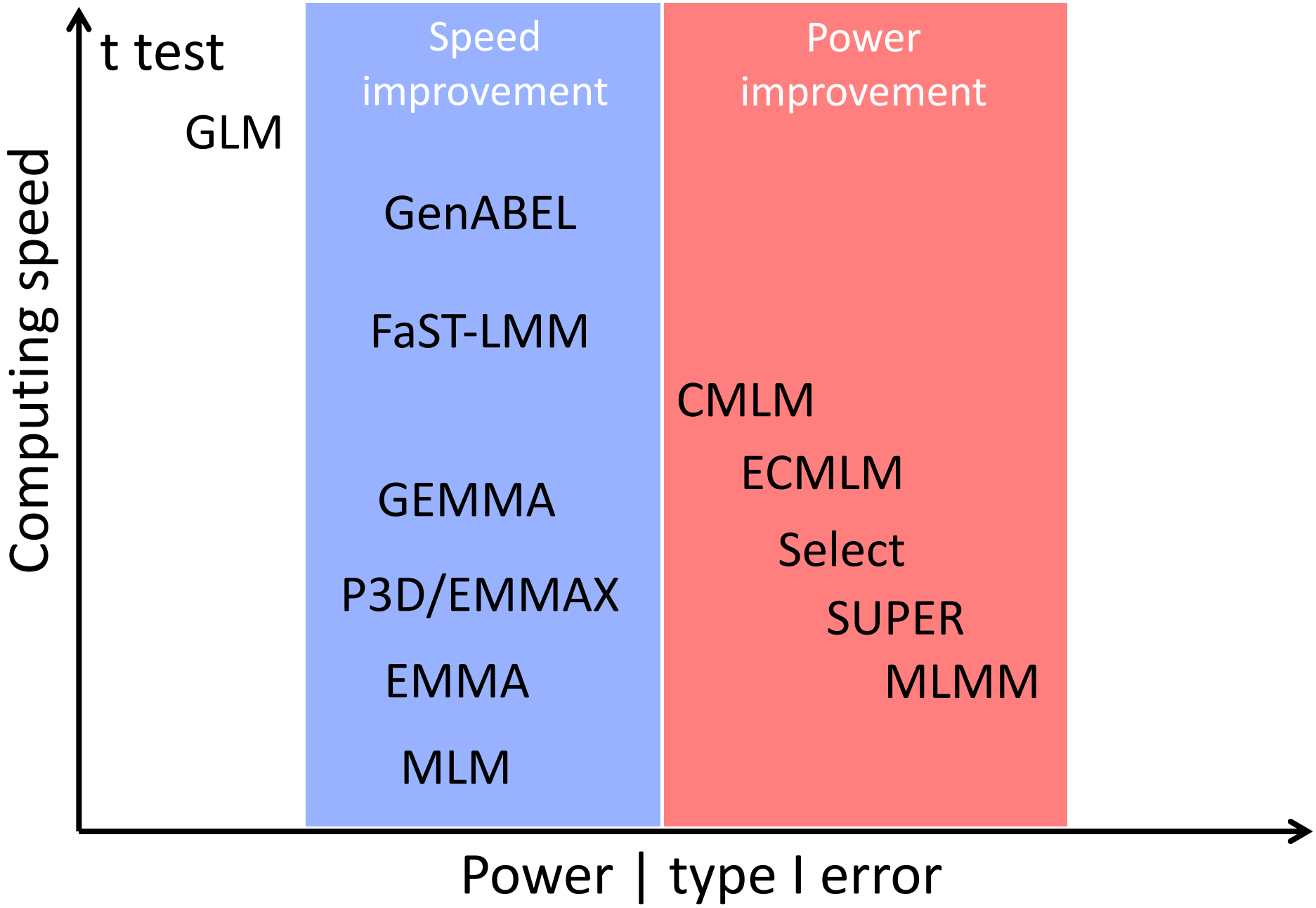

GenAbel


FarmCPU


BLINK

Associations on flowering time




t test
GLM
Speed improvement
Power improvement
GenABEL
FaST-LMM
CMLM
ECMLM
GEMMA
Select
P3D/EMMAX
SUPER
EMMA
MLMM
MLM
Computing speed
Power | type I error

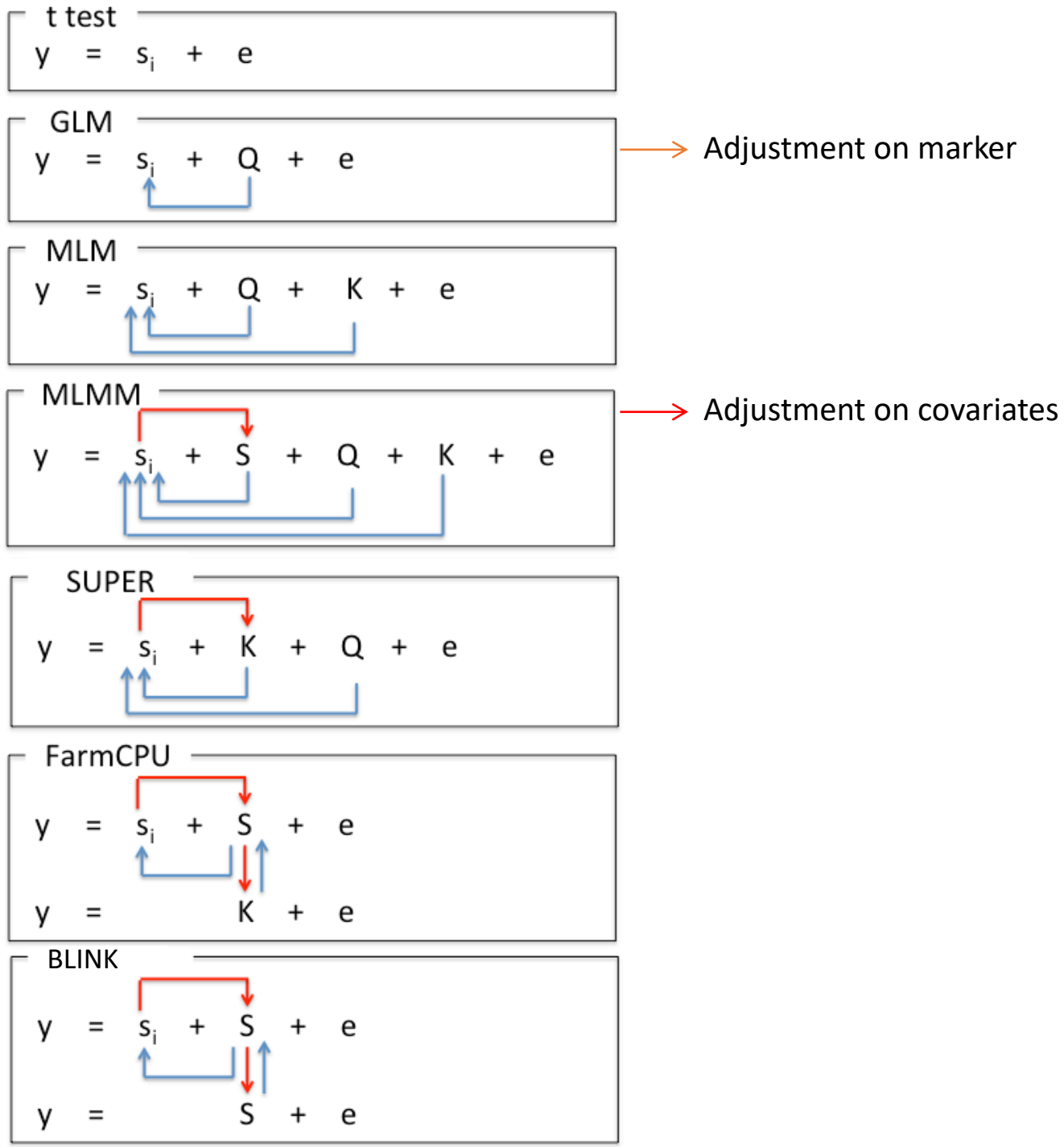
Model Development

S_i: Testing marker

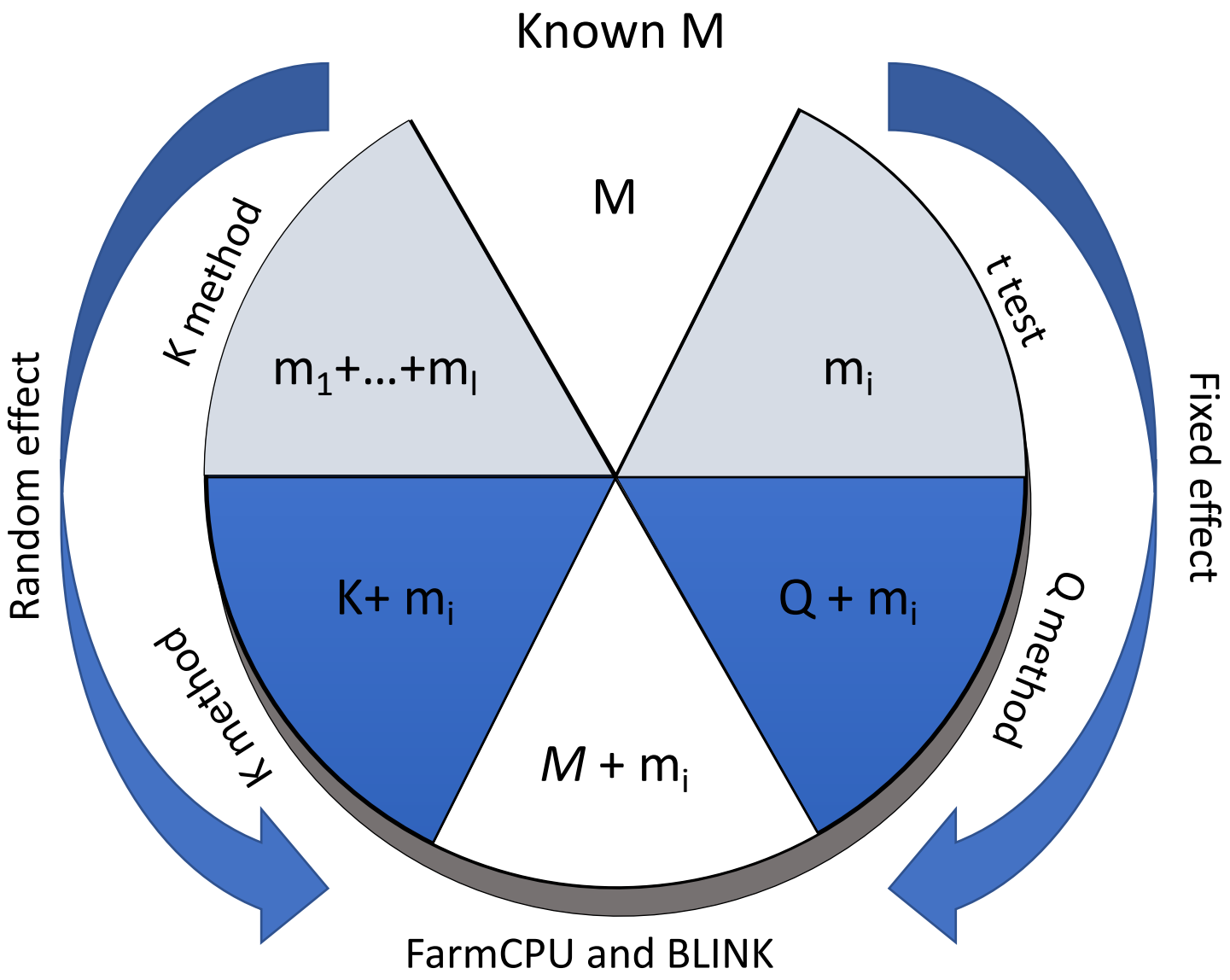
Q: Population structure

K: Kinship

S: Pseudo QTNs

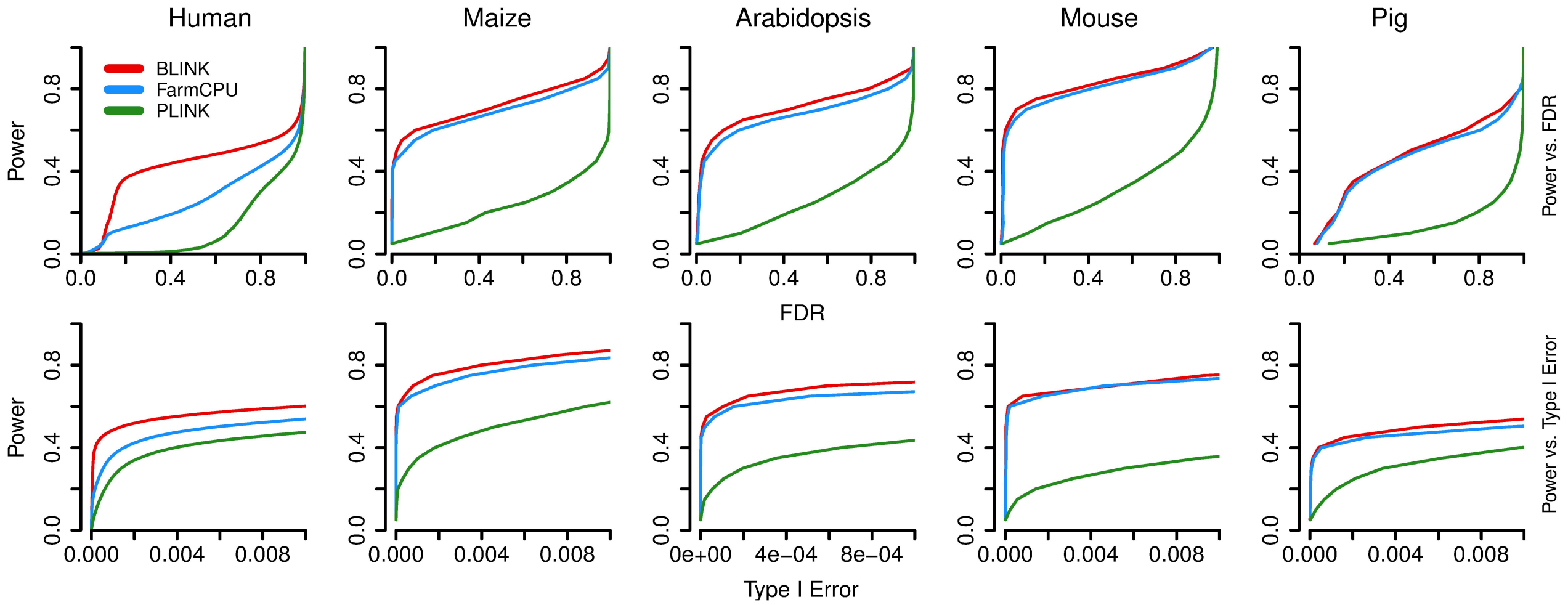


GWAS methods



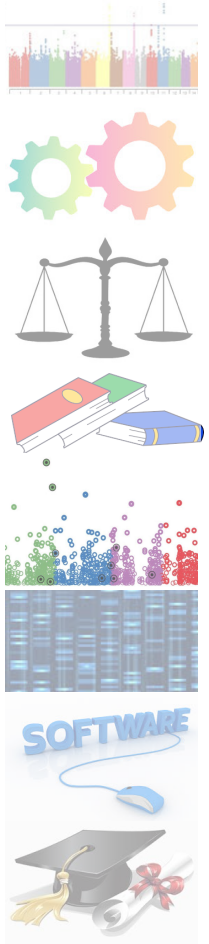
m: marker, M: Mutations, *M*: Estimated mutations

Same trend across species



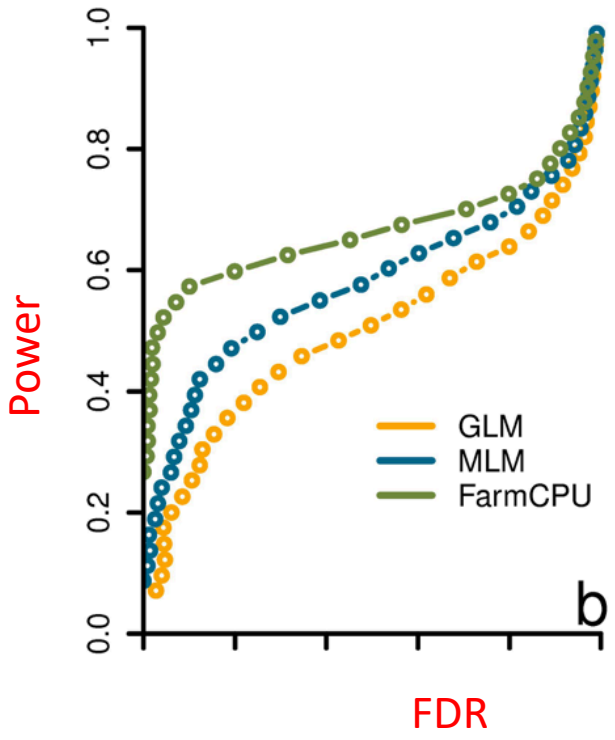
Outline

- Why GWAS?
- How does GWAS work?
- How to evaluate GWAS results?
  - Literature
  - **Simulation**
  - Enrichment analysis
- GWAS Software
- GWAS course at Washington State University



ROC curve

- Receiver Operating Characteristic
- "The curve is created by plotting the true positive rate against the false positive rate at various threshold settings." -Wikipedia



Liu et. al. PLoS Genetics, 2016

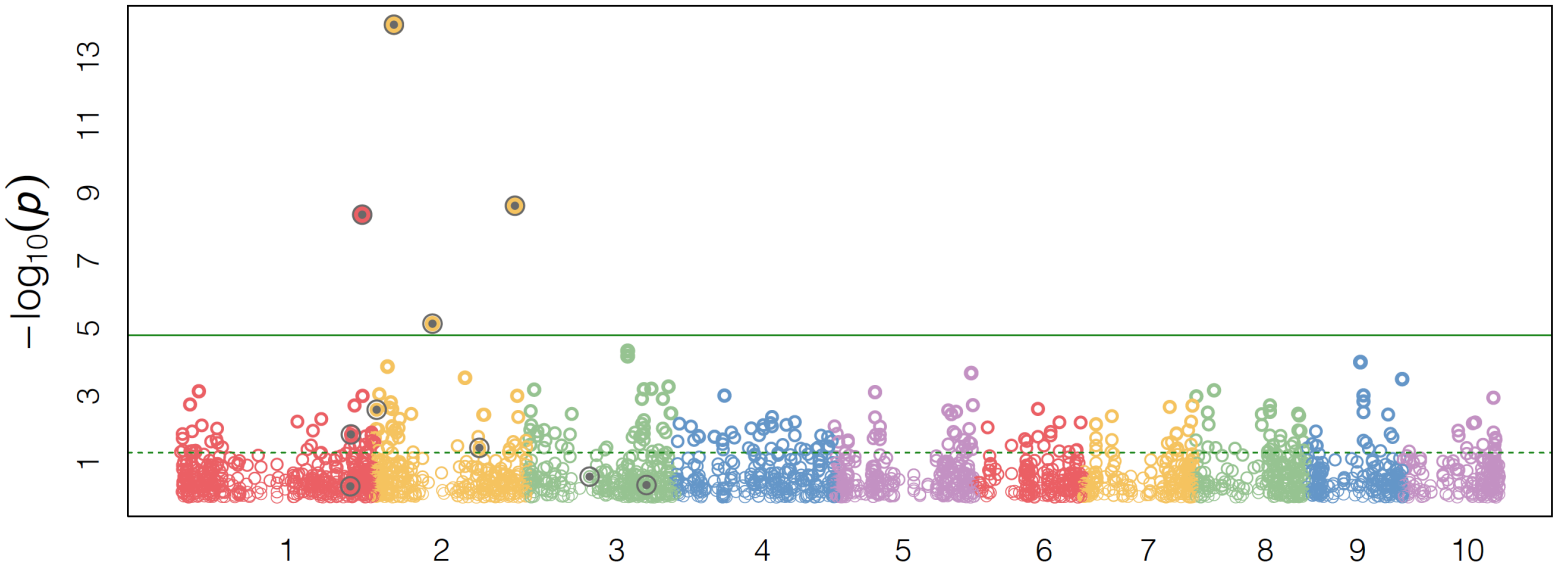
Genotypes

taxa	PZB00859.1	PZA01271.1	PZA03613.2	PZA03613.1	PZA03614.2	PZA03614.1	PZA00258.3
33-16	2	0	0	2	2	2	2
38-11	2	2	0	2	2	2	0
4226	2	0	0	2	2	2	0
4722	2	2	0	2	2	2	1
A188	0	0	0	2	2	2	0
A214N	2	0	2	0	2	0	0
A239	0	0	2	2	0	0	0
A272	0	0	2	2	0	0	2
A441-5	2	0	0	2	2	2	0
A554	2	2	2	2	0	2	0
A556	2	0	0	2	2	2	1
A6	0	0	2	2	0	0	0
A619	2	2	0	2	2	2	0
A632	2	0	2	0	2	0	0
A634	2	0	2	0	2	0	0
A635	2	0	2	0	2	0	0

SNP	Chromosome	Position
PZB00859.1	1	157104
PZA01271.1	1	1947984
PZA03613.2	1	2914066
PZA03613.1	1	2914171
PZA03614.2	1	2915078
PZA03614.1	1	2915242
PZA00258.3	1	2973508
PZA02962.13	1	3205252
PZA02962.14	1	3205262
PZA00599.25	1	3206090

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

GWAS by GLM in GAPIT



Import functions and data → Restrict genes on CHR1-5 → Simulate phenotype → GWAS with GLM using GAPIT

```
library(compiler) #required for cmpfun
source("http://www.zzlab.net/GAPIT/emma.txt")
source("http://www.zzlab.net/GAPIT/gapit_functions.txt")
source("http://www.zzlab.net/StaGen/2021/R/G2P.R")
myGD=read.table(file="http://www.zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

X=myGD[,-1]
index1to5=myGM[,2]<6
X1to5 = X[,index1to5]

set.seed(99164)
setwd("~/Desktop/temp")
mySim=G2P(X=X1to5,h2=.75,alpha=1,NQTN=10,distribution="normal")
myY=cbind(as.data.frame(myGD[,1]), mySim$y)

myGAPIT=GAPIT(
Y=myY,
GD=myGD,
GM=myGM,
QTN.position=mySim$QTN.position,
PCA.total=3,
group.from=1,
group.to=1,
group.by=10,
memo="GLM",
file.output=TRUE,)
```

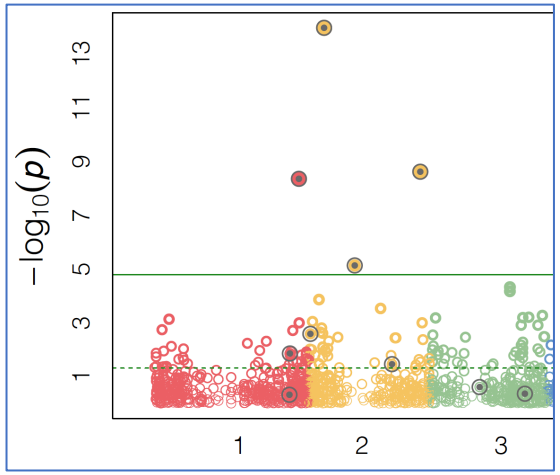
Bins (e.g. 100Kb)

```
bigNum=1e9
resolution=100000
bin=round((myGM[,2]*bigNum+myGM[,3])/resolution)
myGWAS=cbind(myGM,myGAPIT$mp,bin)
head(myGWAS)
```

N	SNP	Chromosome	Position	myGAPIT$mp	bin
1	PZB00859.1	1	157104	0.04532318	10002
2	PZA01271.1	1	1947984	0.68535236	10019
3	PZA03613.2	1	2914066	0.01337254	10029
4	PZA03613.1	1	2914171	0.11339977	10029
5	PZA03614.2	1	2915078	0.01822921	10029
6	PZA03614.1	1	2915242	0.42560772	10029

Minimum p value within bin

Bins of QTNs



QTN.bin=result[mySim$QTN.position,]
QTN.bin

N	SNP	Chromosome	Position	myGAPIT$mp	bin
382	PZA03457.1	1	262715518	1.380424e-02	12627
563	PZB01233.6	2	3376157	2.564230e-03	20034
448	PZA00610.15	1	280215046	4.368796e-09	12802
902	PZA00527.10	2	216833071	2.379767e-09	22168
380	PZA00709.19	1	261859081	4.722836e-01	12619
790	PZA00367.6	2	161879515	3.440120e-02	21619
1185	PZA03647.1	3	185318330	4.331027e-01	31853
1025	PZA02742.1	3	97441783	2.445469e-01	30974
674	PZA03121.2	2	29977973	1.053286e-14	20300
743	PZA01902.1	2	89520665	7.319398e-06	20895

Sorted bins of QTNs

index.qtn.p=order(QTN.bin[,4])

QTN.bin[index.qtn.p,]

N	SNP	Chromosome	Position	myGAPIT$mp	bin
674	PZA03121.2	2	29977973	1.053286e-14	20300
902	PZA00527.10	2	216833071	2.379767e-09	22168
448	PZA00610.15	1	280215046	4.368796e-09	12802
743	PZA01902.1	2	89520665	7.319398e-06	20895
563	PZB01233.6	2	3376157	2.564230e-03	20034
382	PZA03457.1	1	262715518	1.380424e-02	12627
790	PZA00367.6	2	161879515	3.440120e-02	21619
1025	PZA02742.1	3	97441783	2.445469e-01	30974
1185	PZA03647.1	3	185318330	4.331027e-01	31853
380	PZA00709.19	1	261859081	4.722836e-01	12619

FDR and type I error

Total number of bins: 1365 (bin size of 100kb)

N	SNP	CHR	BP	P	Bin
674	PZA03121.2	2	29977973	1.05E-14	20300
902	PZA00527.10	2	216833071	2.38E-09	22168
448	PZA00610.15	1	280215046	4.37E-09	12802
743	PZA01902.1	2	89520665	7.32E-06	20895
563	PZB01233.6	2	3376157	2.56E-03	20034
382	PZA03457.1	1	262715518	1.38E-02	12627
790	PZA00367.6	2	161879515	3.44E-02	21619
1025	PZA02742.1	3	97441783	2.45E-01	30974
1185	PZA03647.1	3	185318330	4.33E-01	31853
380	PZA00709.19	1	261859081	4.72E-01	12619

Power	#False bins	FDR	Typel Error
0.1	0	0	0
0.2	0	0	0
0.3	0	0	0
0.4	0	0	0
0.5	32	0.8648649	0.02344322
0.6	105	0.9459459	0.07692308
0.7	181	0.962766	0.13260073
0.8	608	0.987013	0.44542125
0.9	611	0.9854839	0.44761905
1	930	0.9893617	0.68131868

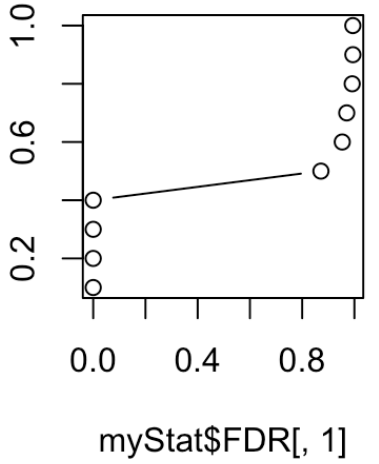
0.8648649=32/(322+4)

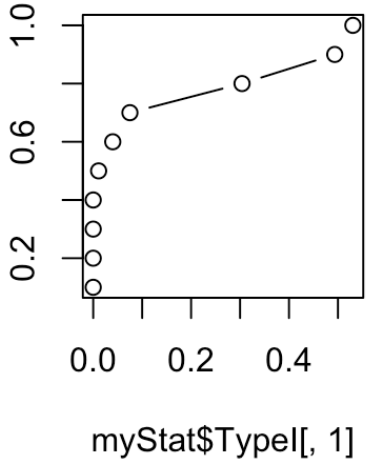
0.02344322=32/1365

GAPIT.FDR.TypeI Function for Area Under Curve

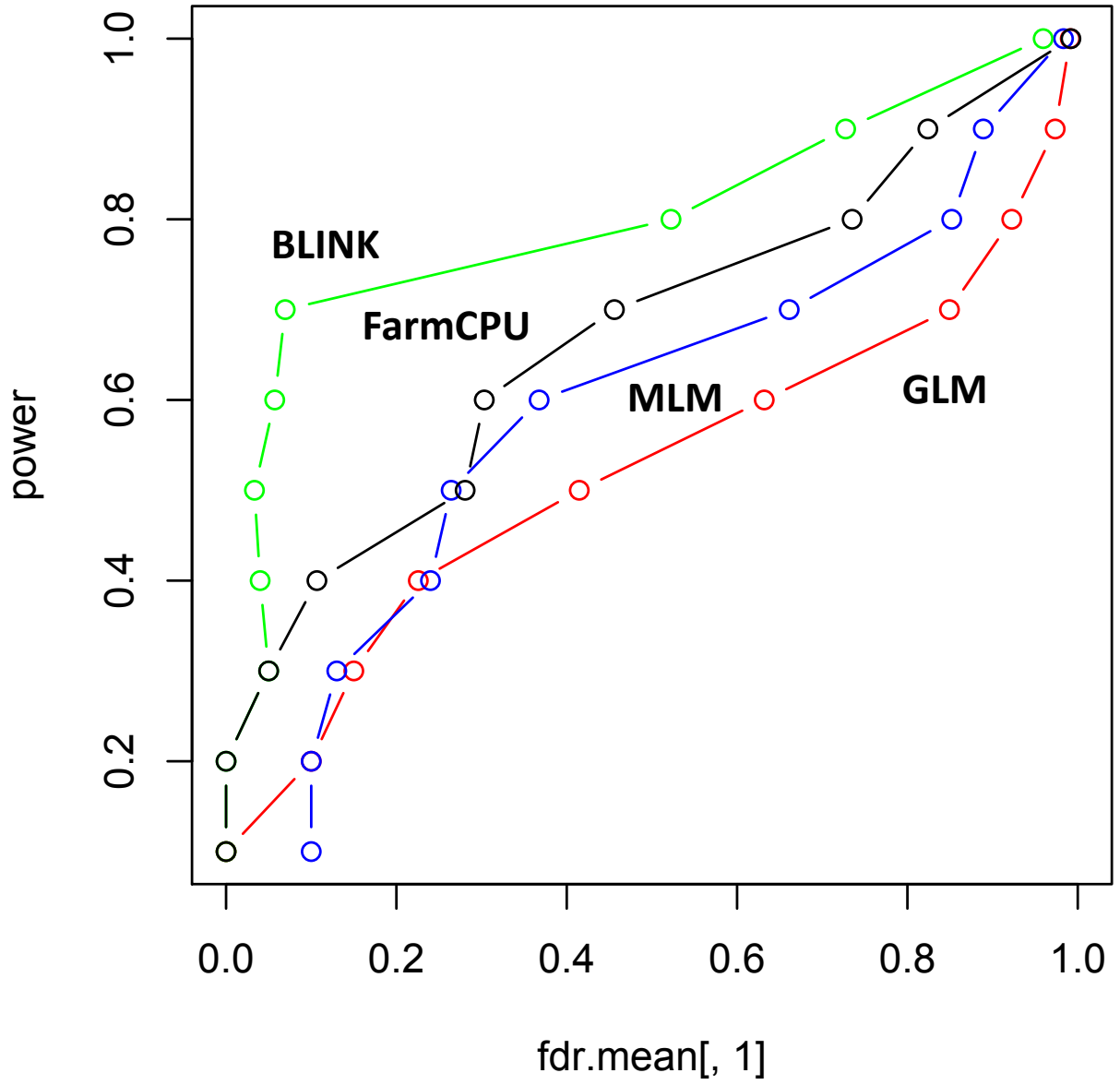
```
library(compiler) #required for cmpfun
source("http://www.zzlab.net/GAPIT/gapit_functions.
txt")
myStat=GAPIT.FDR.TypeI(
WS=c(1e0,1e3,1e4,1e5), GM=myGM,
seqQTN=mySim$QTN.position,
GWAS=myGWAS)

str(myStat)
par(mfrow=c(2,1),mar = c(5,2,5,2))
plot(myStat$FDR[,1],myStat$Power,type="b")
plot(myStat$TypeI[,1],myStat$Power,type="b")
```



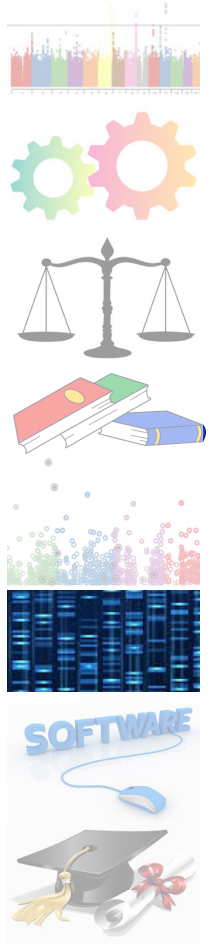


http://zzlab.net/GAPIT/data/Workshop_Iowa.R

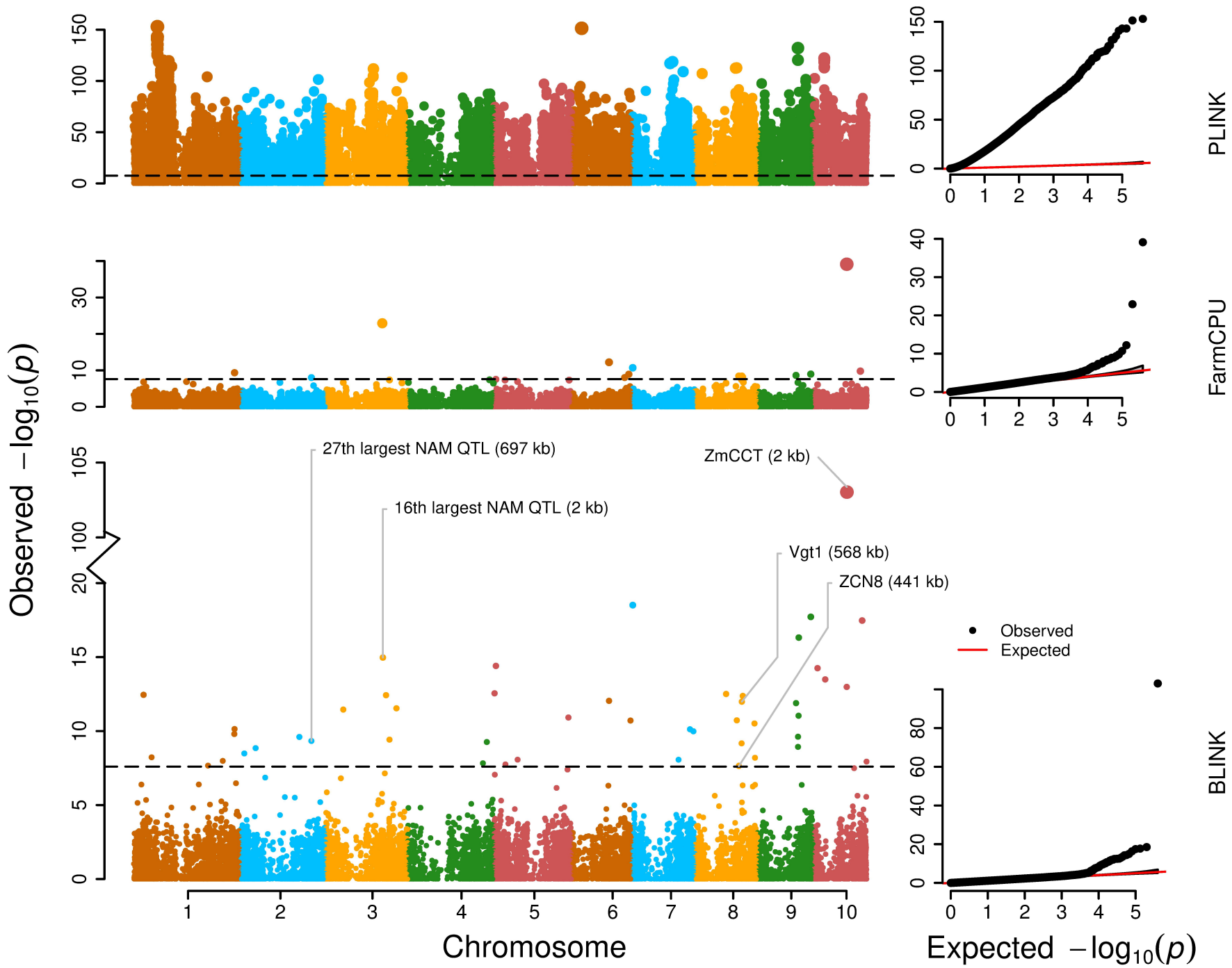


Outline

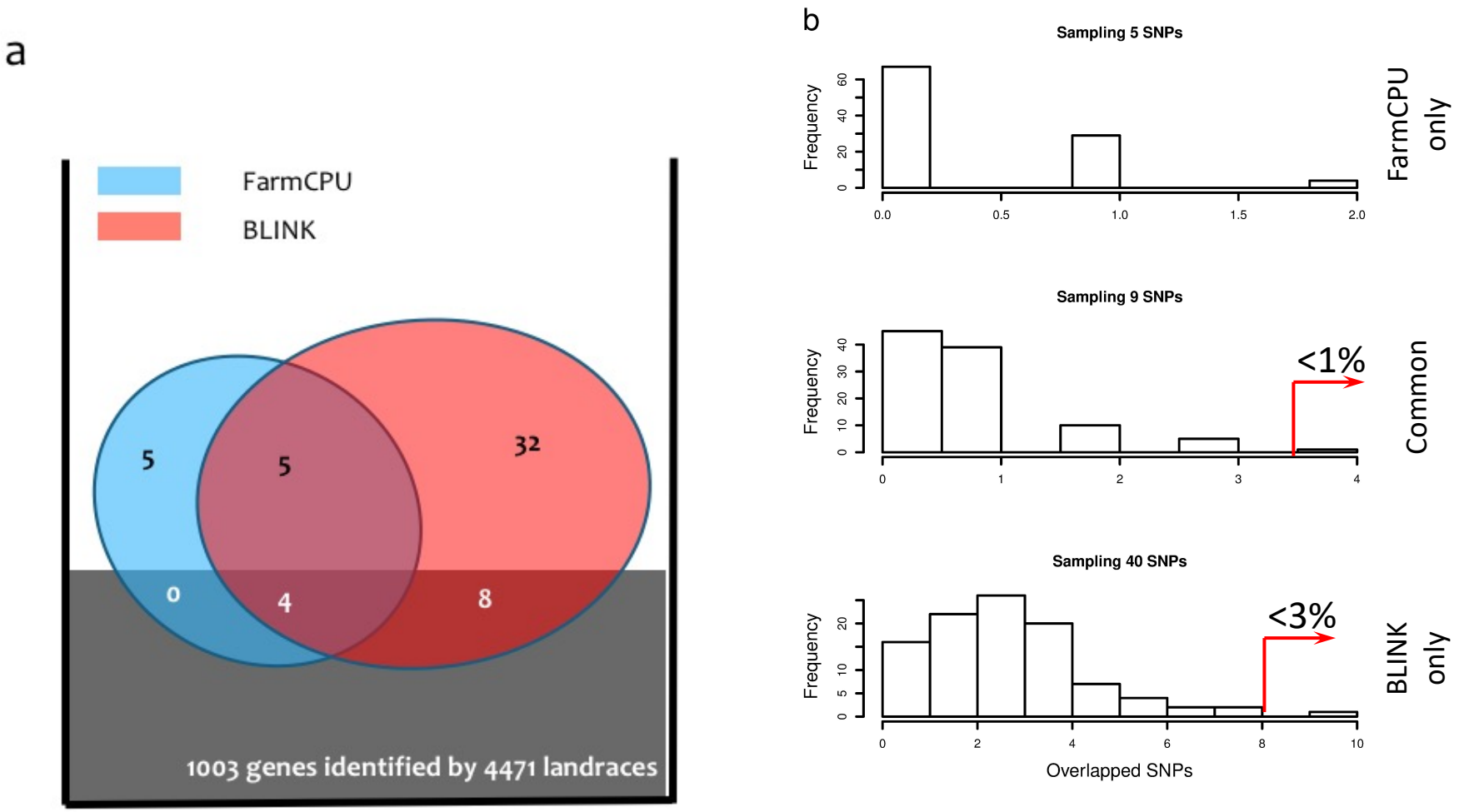
- Why GWAS?
- How does GWAS work?
- How to evaluate GWAS results?
  - Literature
  - Simulation
  - **Enrichment analysis**
- GWAS Software
- GWAS course at Washington State University



Application in Maize

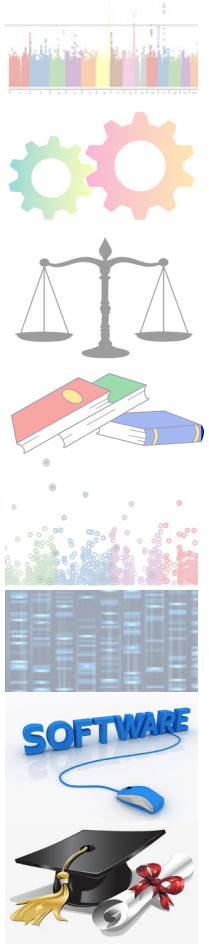


Enrichment

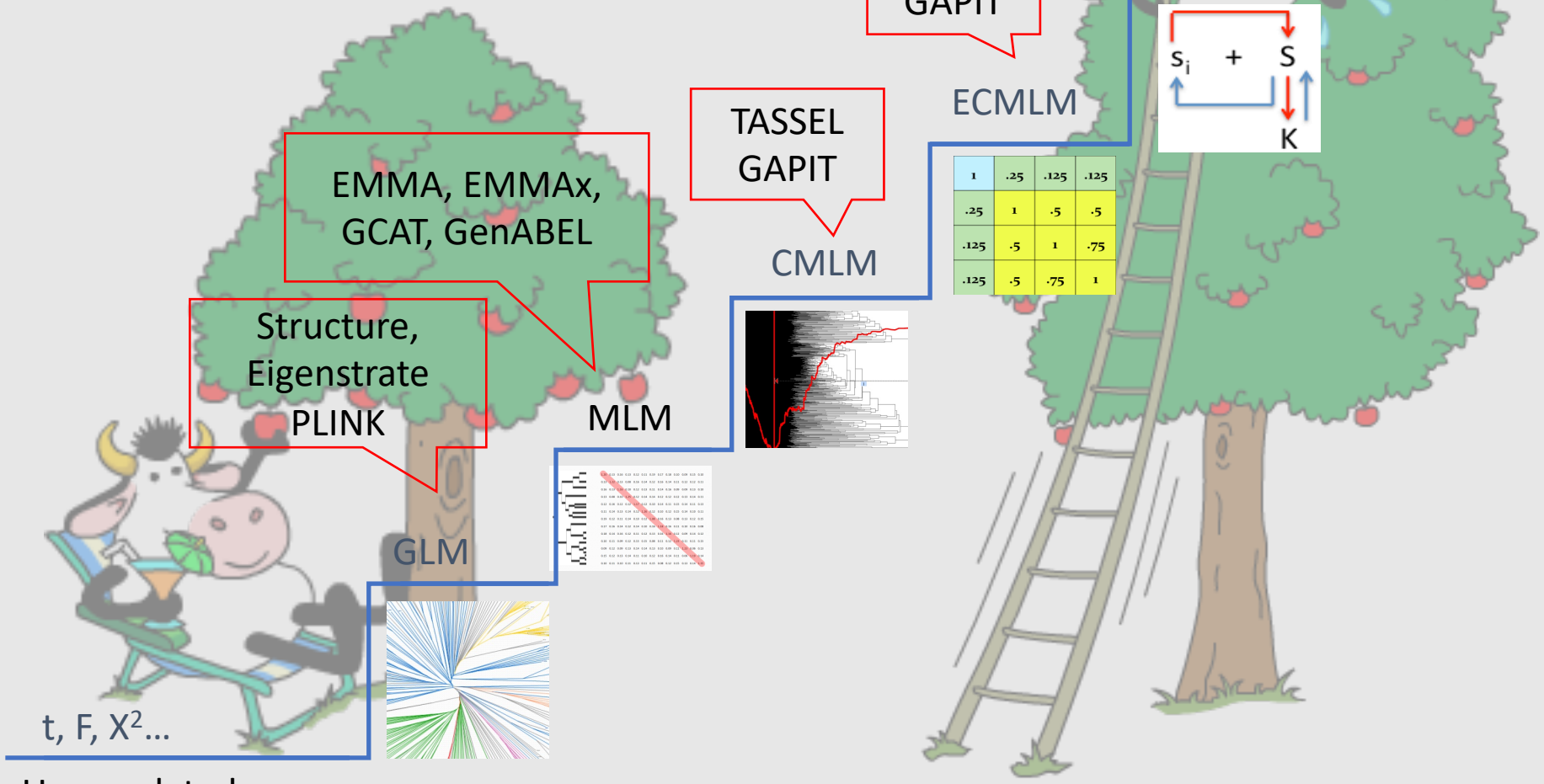

a
FarmCPU
BLINK
5
5
32
0
4
8
1003 genes identified by 4471 landraces
b
Sampling 5 SNPs
Frequency
FarmCPU only
Sampling 9 SNPs
<1%
Common
Sampling 40 SNPs
<3%
BLINK only
Overlapped SNPs

Outline

- Why GWAS?
- How does GWAS work?
- How to evaluate GWAS results?
  - Literature
  - Simulation
  - Enrichment analysis
- **GWAS Software**
- **GWAS course at Washington State University**



GAPIT

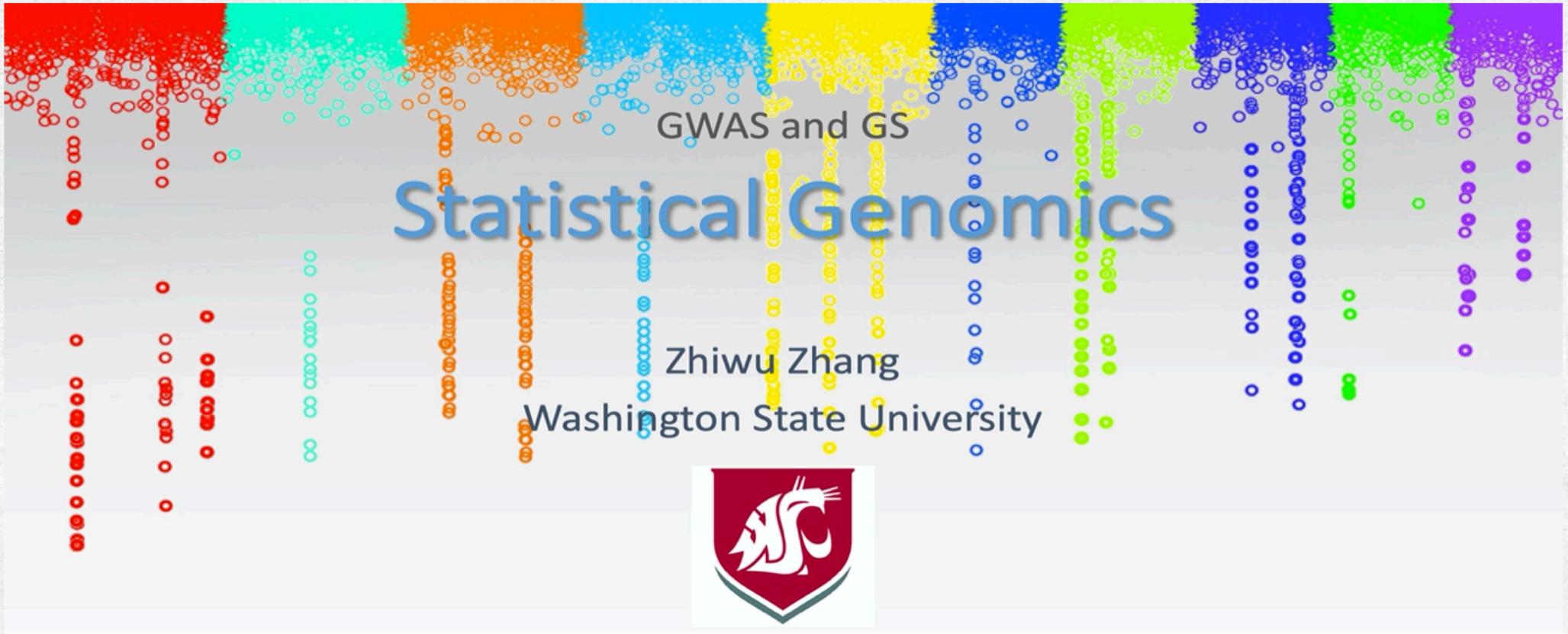

BLINK
FARM CPU
GAPIT
ECMLM
TASSEL
GAPIT
CMLM
EMMA, EMMAx,
GCAT, GenABEL
Structure,
Eigenstrate
PLINK
MLM
GLM
t, F, X2...
Uncorrelated or
equally correlated

iPat

Zhiwu Zhang Laboratory for Statistical Genomics

http://zzlab.net

Home People Publication Research Teaching Software Outreach Jobs



- Flyer
- Syllabus
- Student evaluation
- Public Genomic Data Resources
- Lecture slides (PPT)
- Source code (R)
- Lab
- Quizzes
- Homework

Offered in spring semesters (2015, 2016, 2017, 2018, 2020, and 2021), this graduate course primarily covers Genome Wide Association Study and Genomic Prediction (Selection). The objective is to develop concepts in quantitative genetics and analytical skills in statistics and computation through critical thinking. The course is cross listed by five departments in CAHNRS and College of Art and Science (ANIM_SCI 545, BIOLOGY 545, CROP_SCI 545, HORT 545, and PLP 545). There is no strict prerequisite, however, experience in R programming is strongly recommended. The flyer, syllabus, student evaluation, lecture slides (PPT), and source code (R) are available for all teaching cycles.