Emerging Marker Assisted Selection and Genomic Selection

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

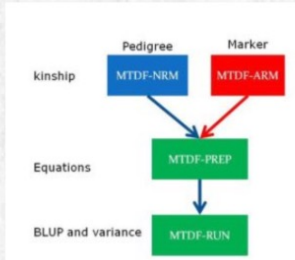
iPat

FarmCPU

Rooster

LADDER

mMAP



MTDFREML

Audio4EDU

AI4EVER

Genomic study

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

} Backward

} Forward

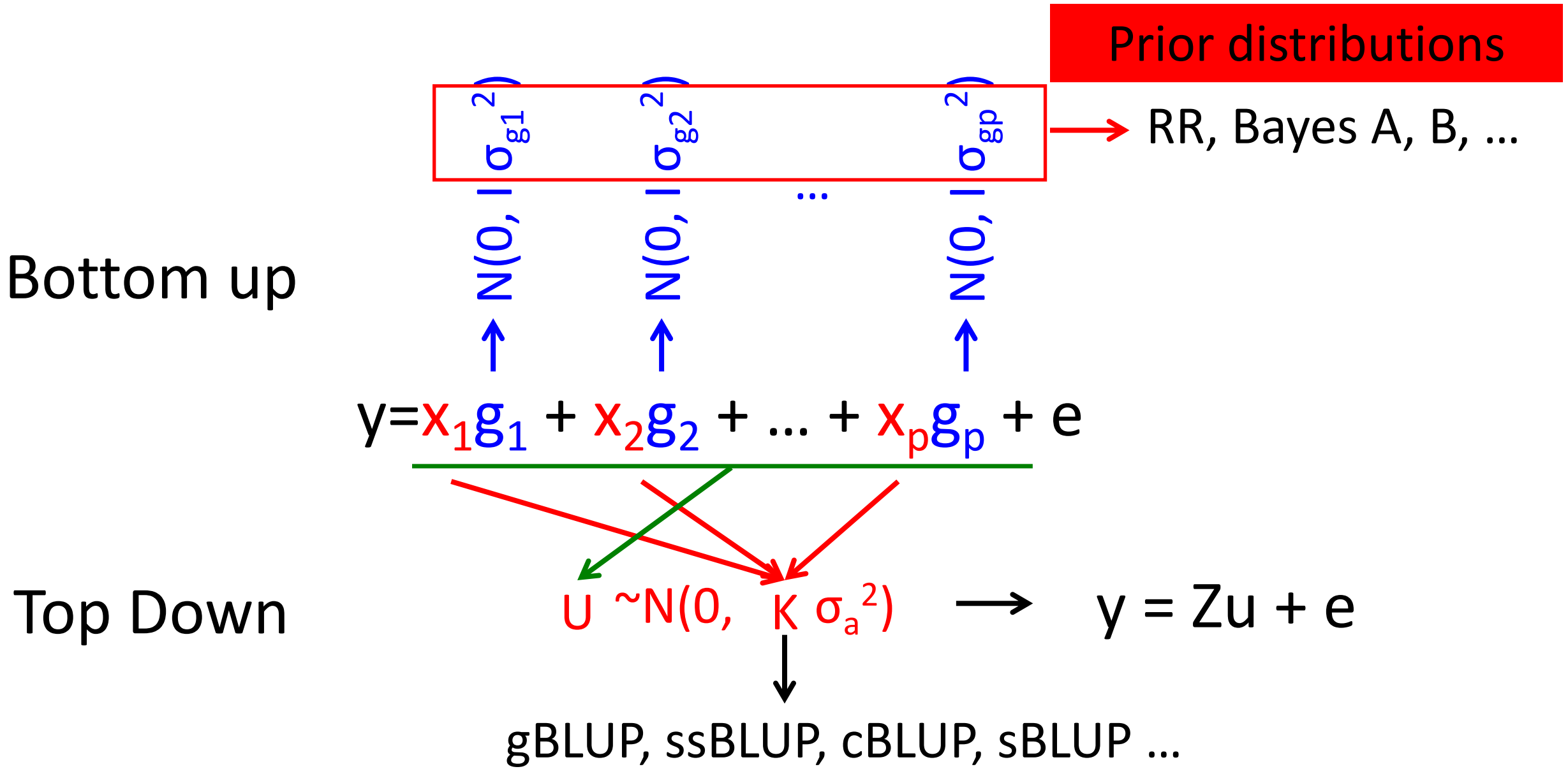
Validation

❖Prediction
- MAS
- GS
- GWAS+GS
- AI

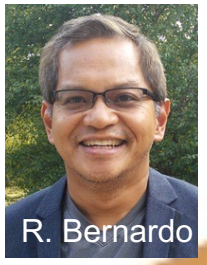
Outline

- Overview
- Two types of problems: too bad and too good
- Right in the middle

Genomic Prediction

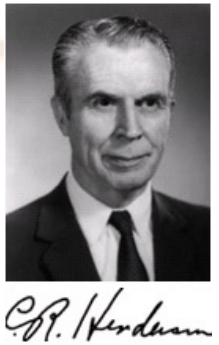


Use of Marker Based Relationships with MTDFREML
J. Animal Sci., 2007



RFLP kinship in maize
Crop Science
1994

MAS using BLUP
GSE, 1989



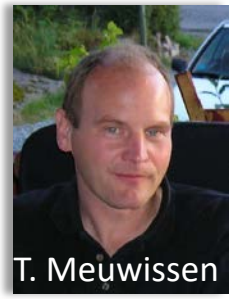
BLUP
Biometrics,1975

MAS

Efficient kinship
J. Dairy Science
2008

gBLUP

Prediction of total genetic value using genome wide dense marker maps
Genetics, 2001



Ridge Regression, Bayes
A, B, Cpi, …

ssBLUP

Pedigree & Marker kinship
single step
GES, 2011

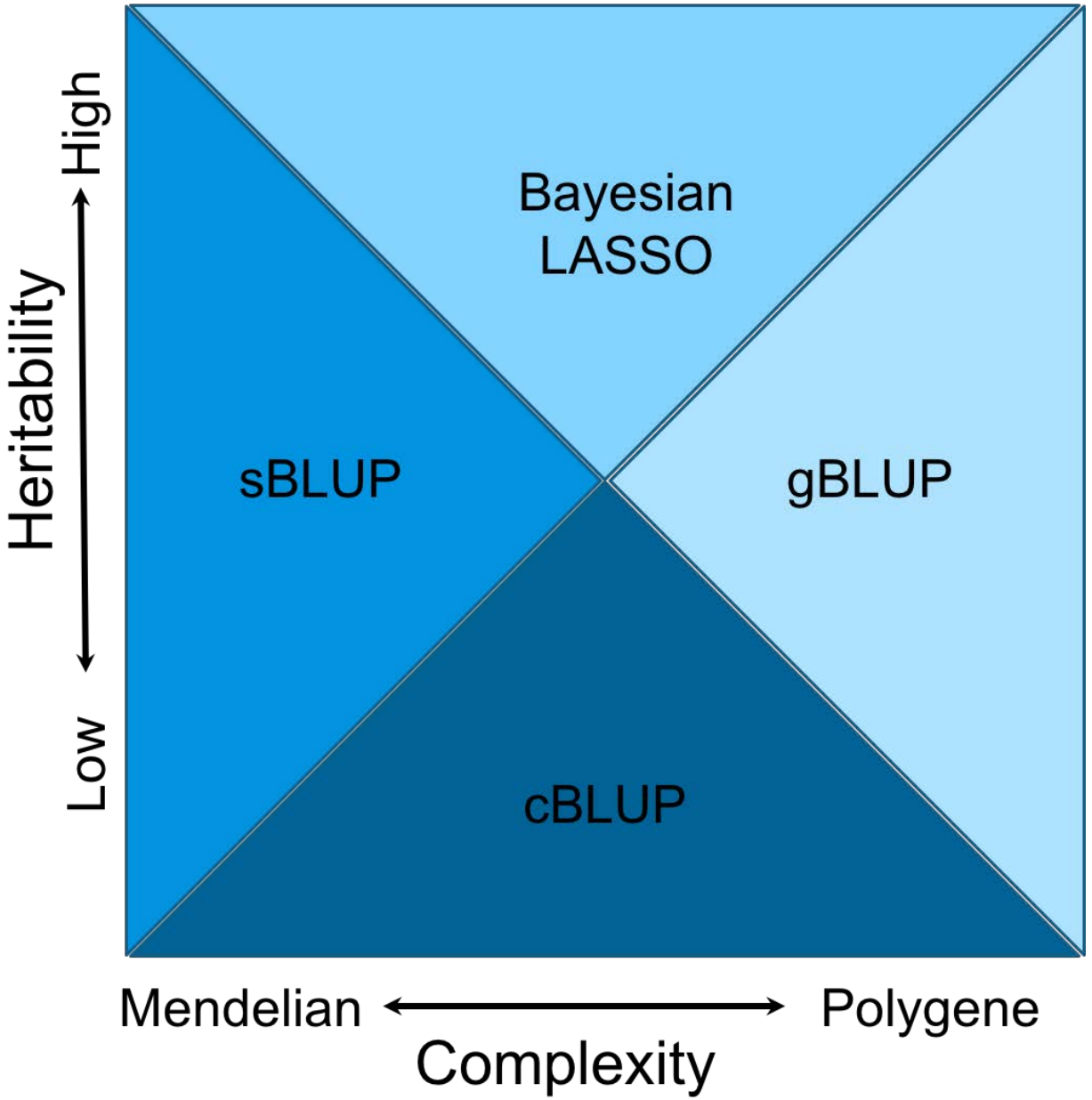
cBLUP

sBLUP

Super and compression
Heredity
2018

Genomic prediction

Interaction between data and methods

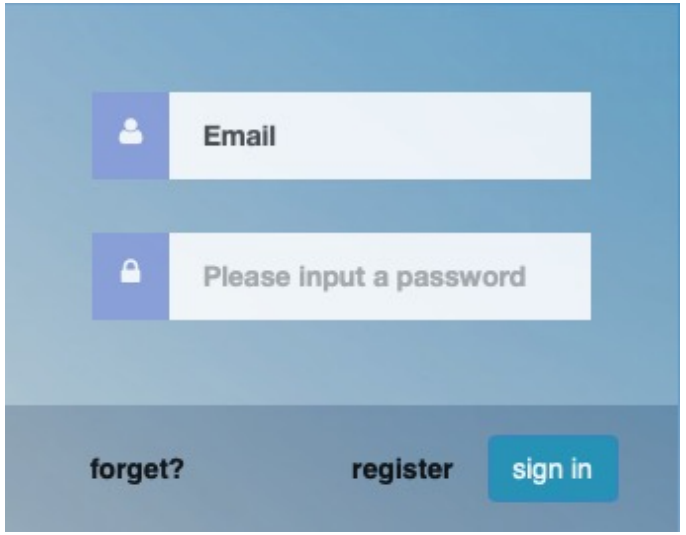


Jiabo Wang

Heredity **volume 121**, pages 648–662 (2018)

mMap: An Online Computing Platform to Transform Genotypes to Phenotypes by Mining the Maximum Accuracy of Prediction

You Tang



Select the best method using machine learning

mMAP website: http://zzlab.net/mMAP

Bioinformatics, Volume 37, Issue 9, May 2021, Pages 1324–1326, https://doi.org/10.1093/bioinformatics/btaa824


We are not luckier than this

Type I problem: too bad than can be

Negative prediction accuracy

Table 2 Testcross population parameters, parental contributions, testcross heritabilities, and RR-BLUP prediction accuracies within 14 maize biparental populations

Pedigree	Tester	N^a	$N^b_{Markers}$	Parental contribution[c]		Heritability				Prediction accuracy (r_{MG})	
				Mean	Range	Grain yield	Grain moisture	Root lodging	Stalk lodging	Grain yield	Grain moisture
S1 × S2	N1	214	115	0.47	(0.30, 0.63)	0.19*[d]	0.67*	–0.06	0.11	0.21*	0.23*
S1 × S3	N1	177	214	0.49	(0.19, 0.80)	0.48*	0.7*	0.01	0.19	0.32*	0.38*
S4 × S5	N1	177	231	0.35	(0.12, 0.88)	0.45*	0.75*	–0.03	0.15	–0.30*	0.33*
S4 × S6	N2	185	239	0.48	(0.26, 0.74)	0.59*	0.32*	0.21*	–0.04	0.16*	0.00
S7 × S1	N2	151	203	0.48	(0.17, 0.78)	0.54*	0.86*	0.22*	0.31*	0.14	0.25*
S8 × S9	N2	292	197	0.33	(0.10, 0.90)	0.44*	0.73*	–0.01	–0.02	0.36*	0.39*
S10 × S11	N3	184	232	0.47	(0.14, 0.86)	0.73*	0.63*	–0.08	0.04	0.30*	0.26*
N4 × N5	S10	141	249	0.34	(0.13, 0.52)	0.25*	0.83*	0.12	0.07	0.18	–0.14
N4 × N5	S12	77	249	0.34	(0.16, 0.52)	0.5*	0.56*	0.35*	0.06	0.33*	–0.33*
N6 × N4	S7	171	203	0.43	(0.15, 0.85)	0.39*	0.77*	0.02	–0.05	–0.08	–0.14
N6 × N4	S12	71	203	0.44	(0.14, 0.86)	0.35*	0.28*	0.07	–0.11	–0.12	–0.11
N7 × N3	S4	109	249	0.44	(0.38, 0.62)	0.32*	0.48*	0.11	0.09	0.07	–0.42*
N8 × N6	S13	211	338	0.33	(0.17, 0.48)	0.43*	0.83*	0.002	0.11	0.21*	–0.08
N7 × N9	S4	114	243	0.33	(0.22, 0.78)	0.37*	0.72*	0.11	–0.04	–0.10	–0.24*

* Significant at $P = 0.05$

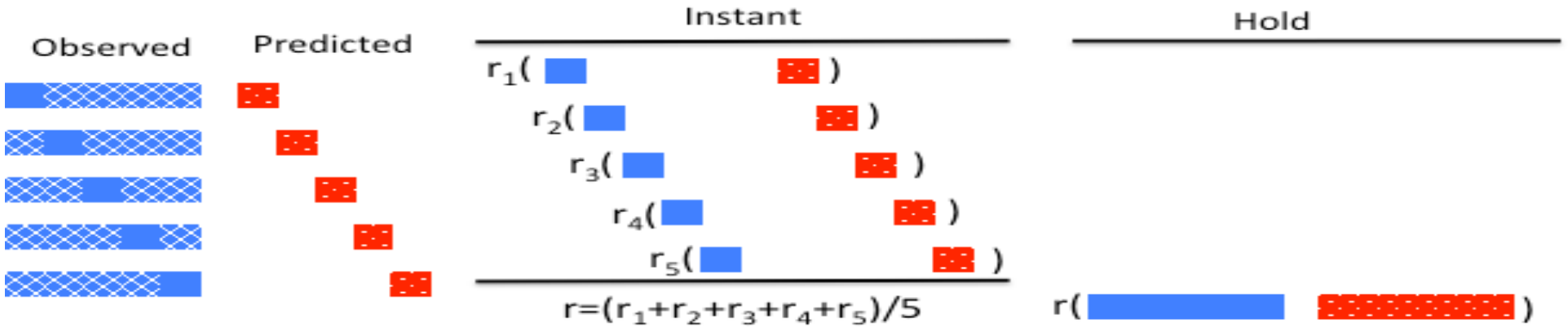
[a] Number of individuals in the biparental population

[b] Number of polymorphic SNPs in the biparental population

[c] Parental contribution of the less-represented parent

Theor Appl Genet. 2013 Jan;126(1):13-22
Genomewide predictions from maize single-cross data.
Massman JM1, Gordillo A, Lorenzana RE, Bernardo R.

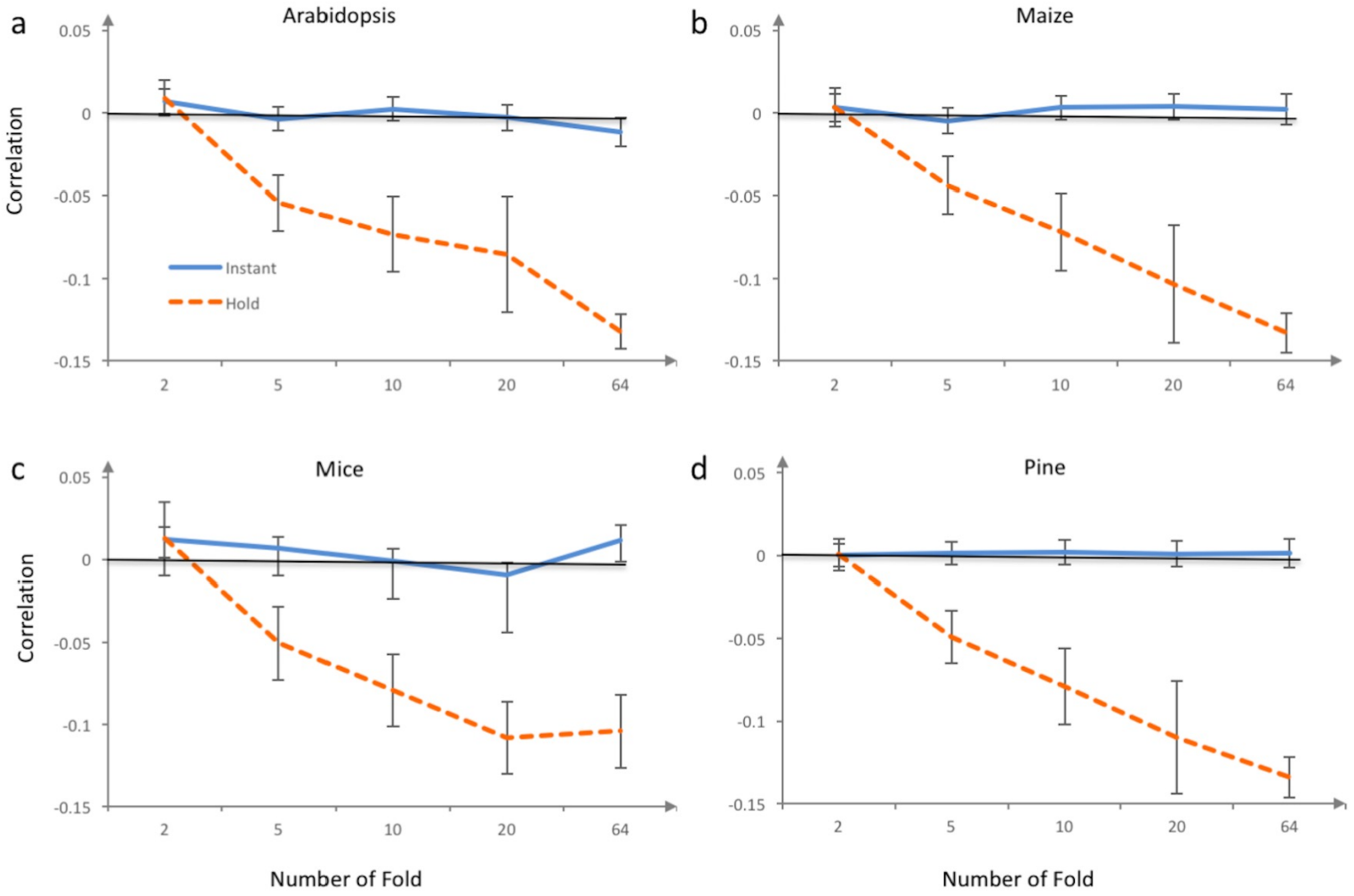
Two ways of calculating correlation



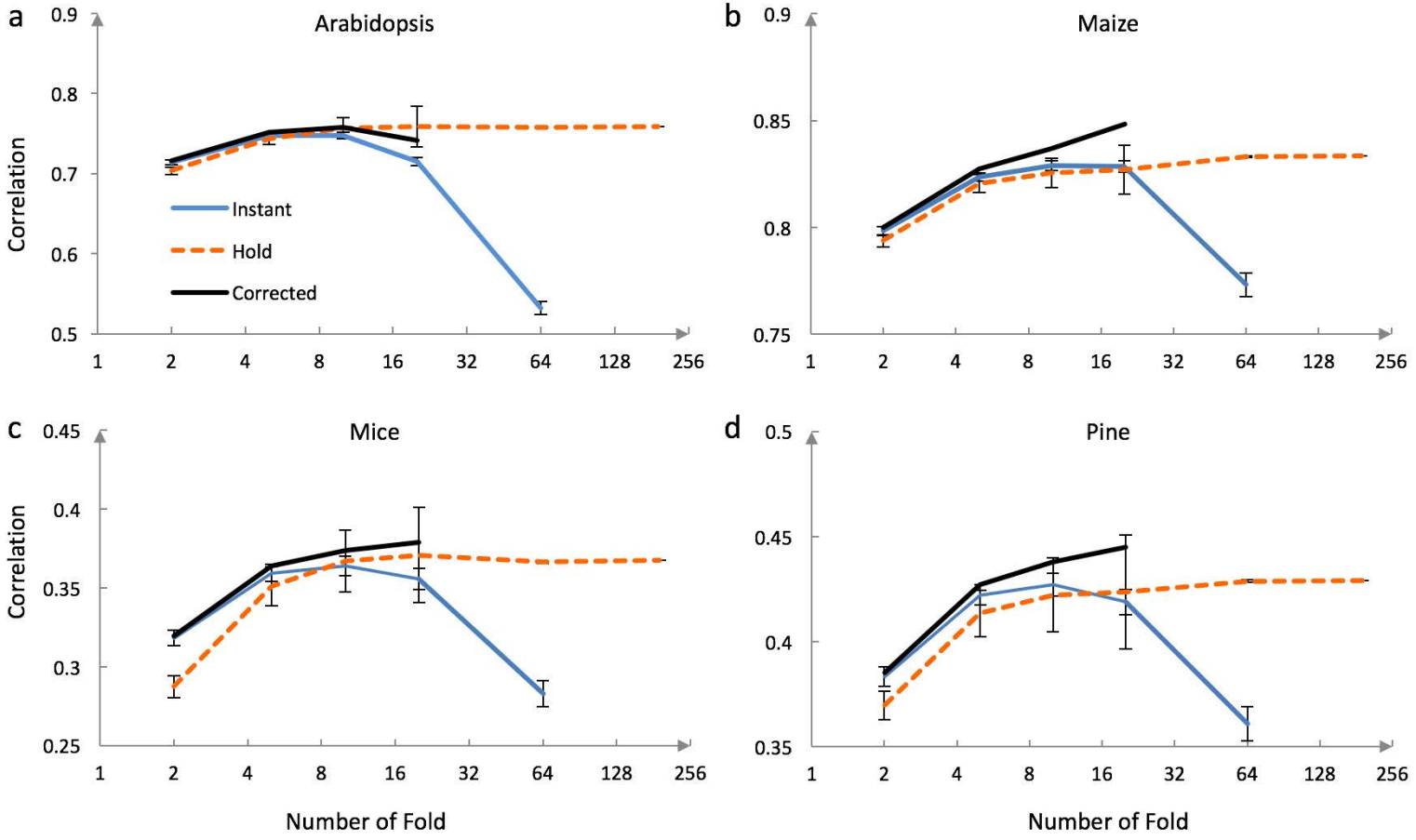
Yao Zhou and et al., Briefings in Bioinformatics, Volume 18, Issue 5, September 2017, Pages 744–753,
https://doi.org/10.1093/bib/bbw064

Underestimates of hold accuracy

Yao Zhou



Underestimates and correction of instant accuracy



$$\hat{\rho} = r\left[1 + \frac{(1-r^2)}{2(n-4)}\right]$$

Y Zhou, MI Vales, A Wang, Z Zhang Briefings in bioinformatics 18 (5), 744-753

Crop Science

Crop Breeding & Genetics

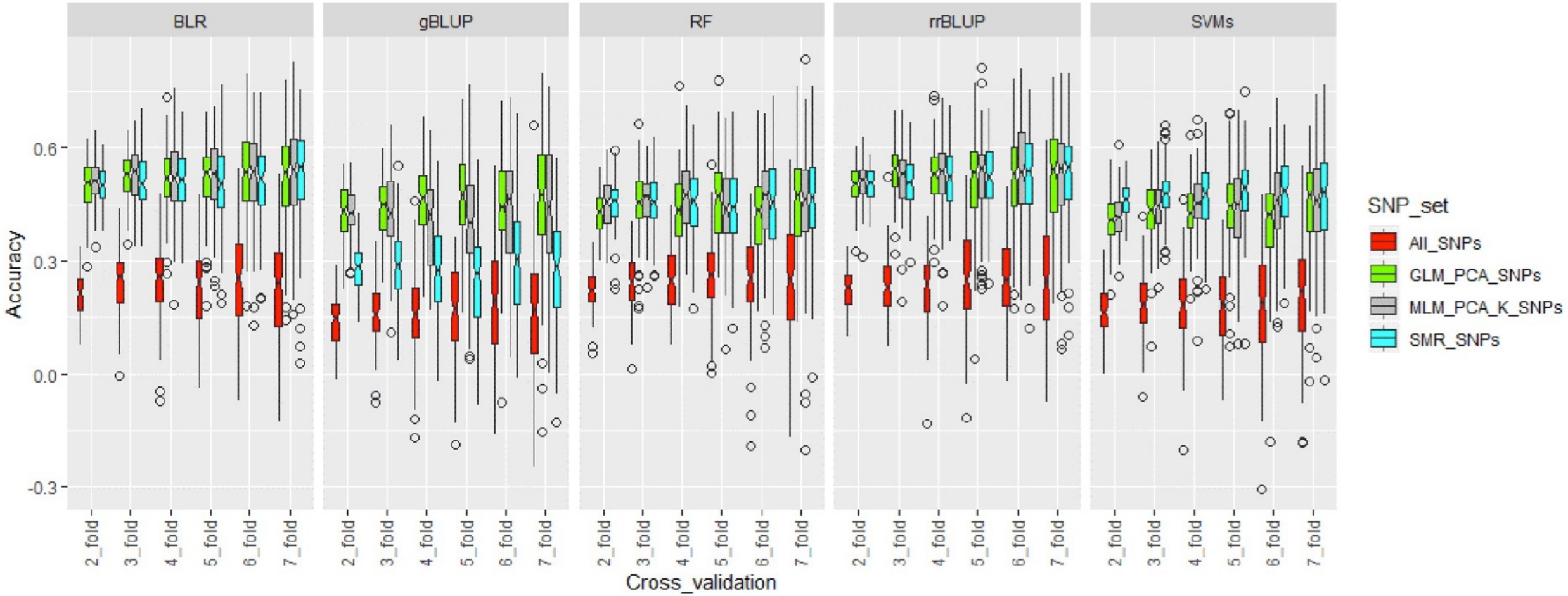
Genomewide Selection when Major Genes Are Known

Rex Bernardo

First published: 01 January 2014 | https://doi.org/10.2135/cropsci2013.05.0315 |
Citations: 161

Type II problem: too good to be true

Ravelombola, W. S. *et al.* Genome-wide association study and genomic selection for tolerance of soybean biomass to soybean cyst nematode infestation. *PLoS One* **15**, (2020).



Genomic selection accuracy increased to almost 2-fold at each level of cross validation when the GWAS-derived SNPs were incorporated into the genomic selection model.

Table 5. Comparison of genomic selection (GS) models in 13 phenotypic traits collected in the SolCAP potato diversity panel. Mean and standard deviation of Pearson's correlation obtained by 10-fold cross validation in 10 replicates. SNP weights for WGBLUP were obtained from $-\log_{10}$ *p*-values of different GWASpoly models.

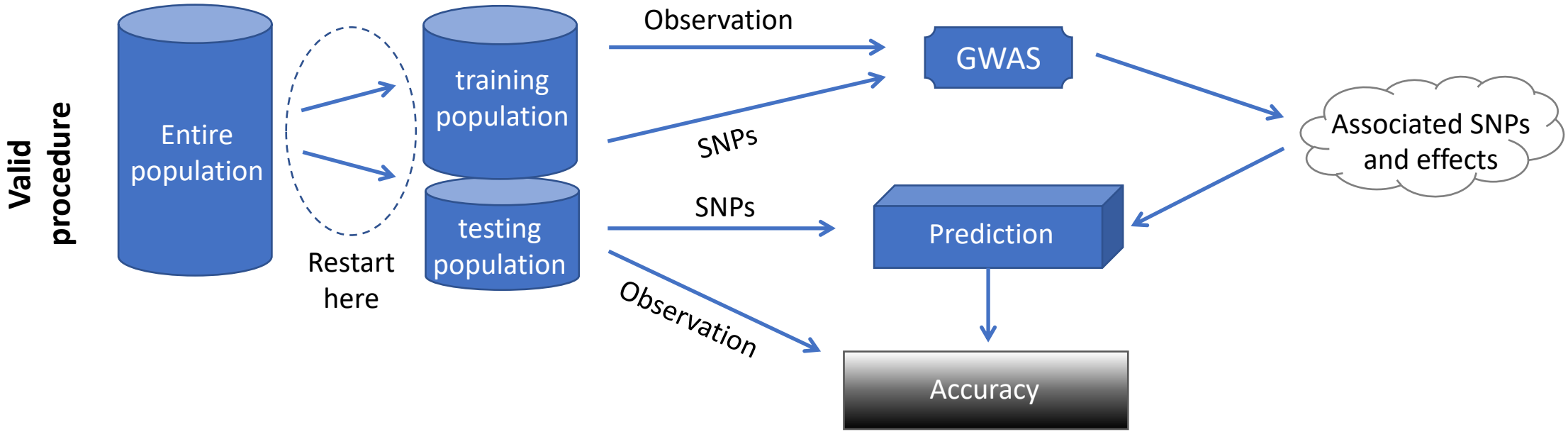
Trait	RRBLUP	GBLUP	WGBLUP							
			1-d-a	1-d-r	2-d-a	2-d-r	General	d-Gen	d-Add	Additive
Chip color	0.723 (±0.014)	0.721 (±0.015)	0.826 (±0.009)	0.798 (±0.011)	0.859 (±0.007)	0.850 (±0.013)	0.867 (±0.008)	0.849 (±0.009)	0.855 (±0.007)	0.896 (±0.007)
$\log_{10}$ fructose	0.682 (±0.024)	0.676 (±0.025)	0.819 (±0.014)	0.785 (±0.017)	0.845 (±0.007)	0.833 (±0.011)	0.868 (±0.011)	0.839 (±0.015)	0.855 (±0.003)	0.895 (±0.008)
$\log_{10}$ glucose	0.678 (±0.017)	0.668 (±0.030)	0.796 (±0.009)	0.809 (±0.016)	0.855 (±0.009)	0.849 (±0.009)	0.875 (±0.009)	0.844 (±0.011)	0.848 (±0.013)	0.91 (±0.007)
Malic acid	0.602 (±0.016)	0.598 (±0.027)	0.751 (±0.021)	0.745 (±0.022)	0.802 (±0.021)	0.801 (±0.016)	0.838 (±0.011)	0.808 (±0.016)	0.826 (±0.009)	0.876 (±0.007)
Sucrose	0.539 (±0.024)	0.519 (±0.034)	0.676 (±0.011)	0.675 (±0.022)	0.702 (±0.019)	0.716 (±0.015)	0.725 (±0.023)	0.722 (±0.011)	0.739 (±0.019)	0.806 (±0.011)
Total yield	0.132 (±0.023)	0.117 (±0.041)	0.401 (±0.026)	0.413 (±0.030)	0.418 (±0.031)	0.428 (±0.017)	0.470 (±0.029)	0.492 (±0.030)	0.504 (±0.030)	0.584 (±0.028)
Tuber eye depth	0.495 (±0.026)	0.478 (±0.019)	0.605 (±0.029)	0.655 (±0.016)	0.693 (±0.025)	0.717 (±0.014)	0.740 (±0.020)	0.693 (±0.020)	0.736 (±0.018)	0.812 (±0.007)
Tuber length	0.826 (±0.012)	0.821 (±0.014)	0.891 (±0.006)	0.884 (±0.009)	0.899 (±0.006)	0.889 (±0.012)	0.904 (±0.008)	0.908 (±0.008)	0.912 (±0.005)	0.928 (±0.009)
Tuber shape	0.775 (±0.018)	0.780 (±0.017)	0.865 (±0.010)	0.853 (±0.013)	0.886 (±0.008)	0.863 (±0.005)	0.896 (±0.010)	0.89 (±0.008)	0.891 (±0.009)	0.922 (±0.006)
Tuber size	0.501 (±0.024)	0.499 (±0.027)	0.641 (±0.019)	0.650 (±0.020)	0.679 (±0.020)	0.663 (±0.022)	0.666 (±0.024)	0.661 (±0.022)	0.679 (±0.019)	0.742 (±0.021)
Tuber width	0.635 (±0.023)	0.638 (±0.021)	0.752 (±0.020)	0.749 (±0.021)	0.782 (±0.016)	0.772 (±0.018)	0.805 (±0.012)	0.789 (±0.015)	0.803 (±0.013)	0.847 (±0.017)
Vine maturity 95 days	0.288 (±0.035)	0.286 (±0.042)	0.550 (±0.028)	0.538 (±0.020)	0.603 (±0.022)	0.589 (±0.028)	0.668 (±0.022)	0.632 (±0.019)	0.65 (±0.025)	0.746 (±0.017)
Vine maturity 120 days	0.321 (±0.047)	0.323 (±0.024)	0.495 (±0.026)	0.569 (±0.021)	0.636 (±0.021)	0.633 (±0.013)	0.669 (±0.025)	0.616 (±0.023)	0.666 (±0.026)	0.755 (±0.019)

RRBLUP, best linear unbiased prediction using ridge-regression; GBLUP, genomic best linear unbiased prediction using VanRaden G matrix; WGBLUP, weighted GBLUP; 1-d-a and 1-d-r, simplex dominant models; 2-d-a and 2-d-r, duplex dominant models; d-gen, diploidized general; d-add, diploidized additive.

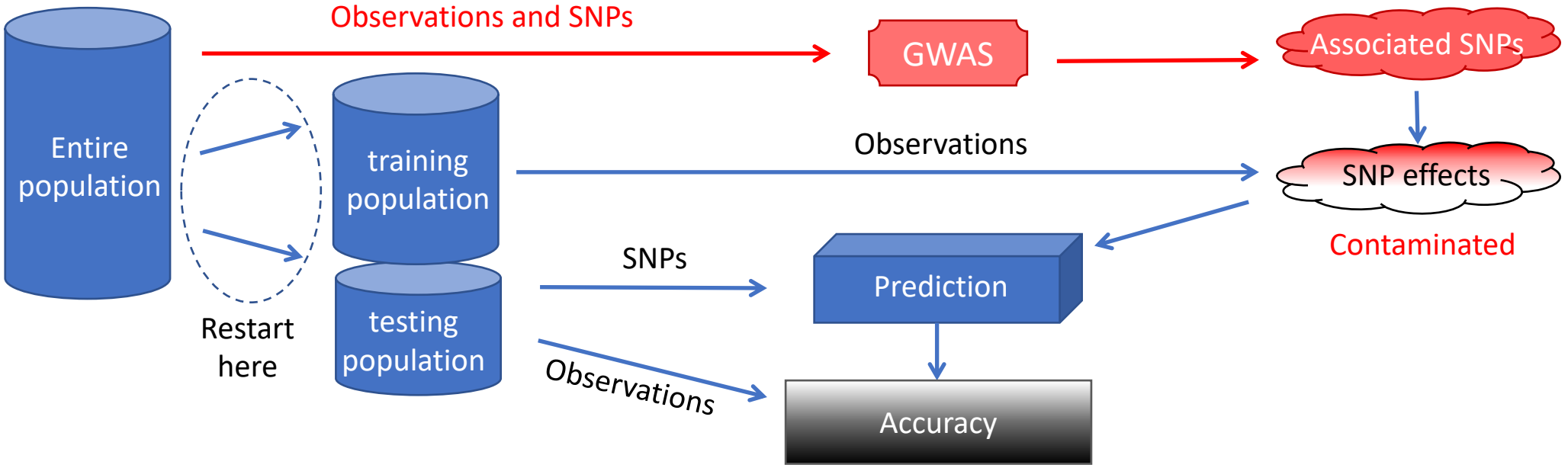
Medina, C. A., Kaur, H., Ray, I. & Yu, L. X. Strategies to increase prediction accuracy in genomic selection of complex traits in alfalfa (Medicago sativa l.). Cells vol. 10 Preprint at https://doi.org/10.3390/cells10123372 (2021).

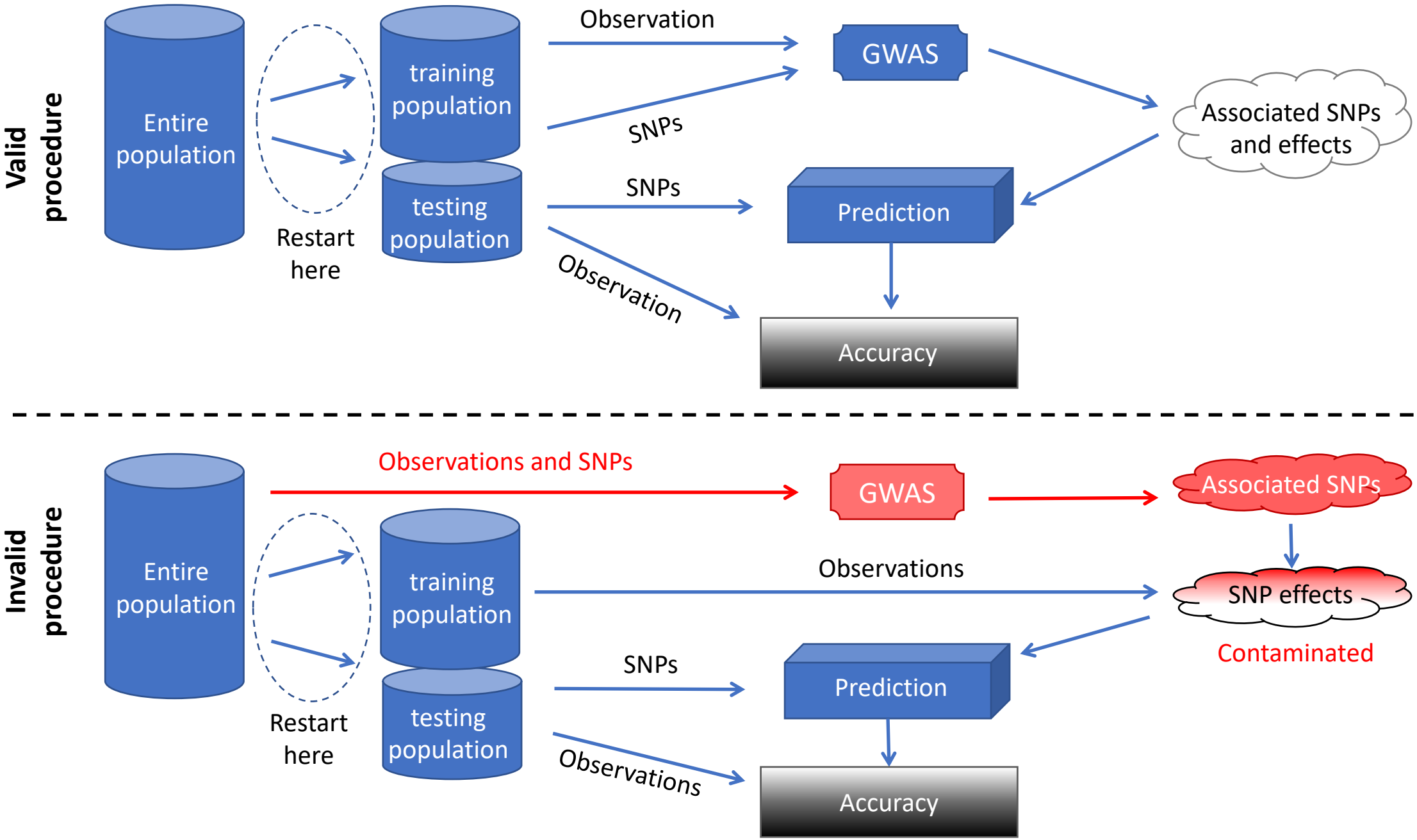
Higher prediction accuracies in all 13 agronomic traits of potato were obtained using the WGBLUP model with SNP $-\log_{10}$ *p*-values derived from the additive GWASpoly model. Traits of glucose, tuber length, or tuber shape showed accuracies higher than 0.9. It is important to point out that traits of tuber length or tuber shape had high accuracies (0.82 and 0.78 respectively using RRBLUP and GBLUP models) and the use of the WGBLUP model increased the prediction accuracy up to 0.93. Total yield had low prediction accuracies with RRBLUP or GBLUP models (0.132 and 0.117, respectively), and the use of the WGBLUP model increased prediction accuracy by almost five times (Table 5). These results agree with our previous results in alfalfa (Figure 2).

Valid procedure

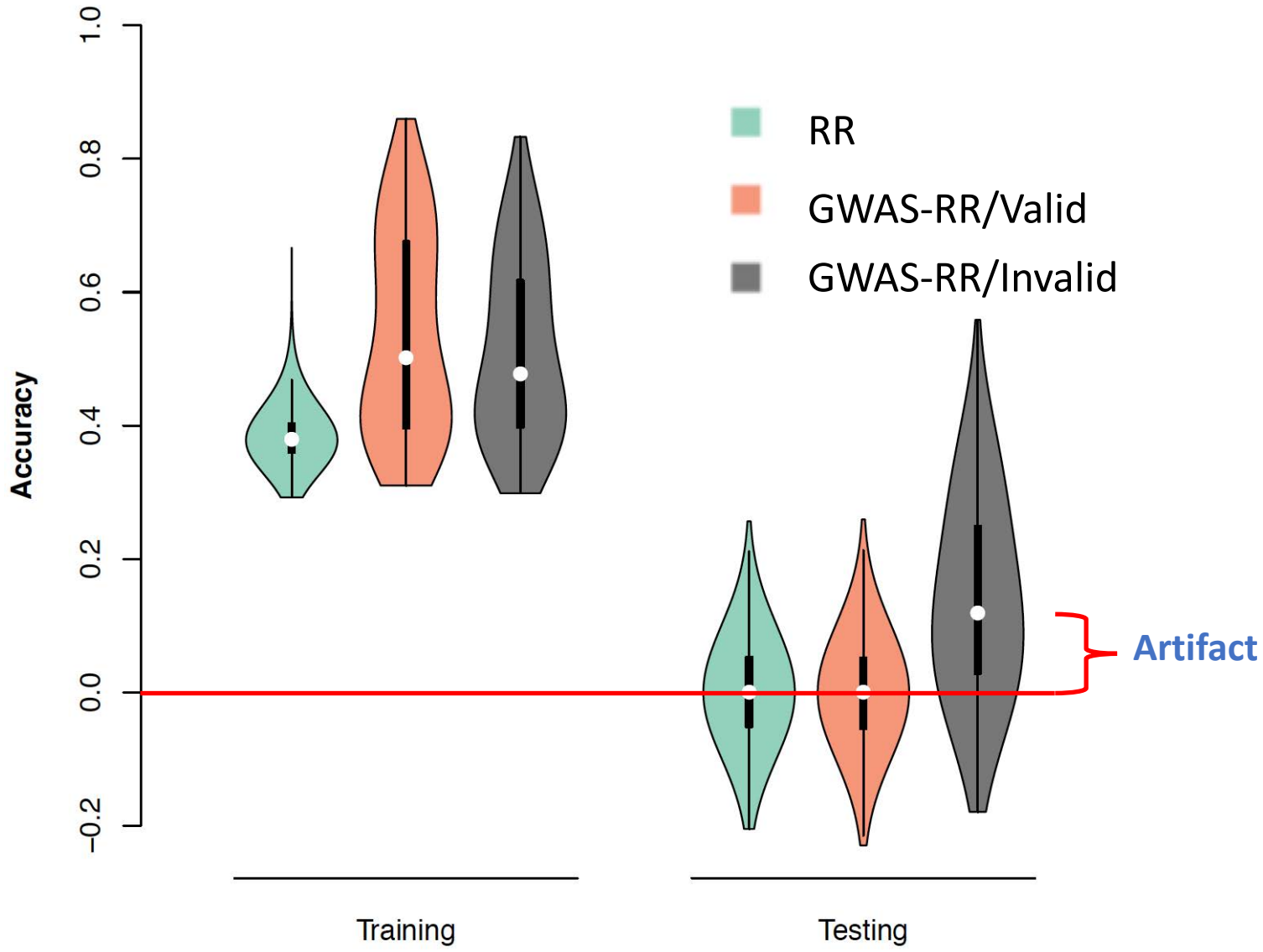


Invalid procedure




Valid procedure
Entire population
Restart here
training population
testing population
Observation
SNPs
GWAS
Associated SNPs and effects
SNPs
Prediction
Observation
Accuracy
Invalid procedure
Entire population
Restart here
Observations and SNPs
training population
testing population
GWAS
Associated SNPs
Observations
SNP effects
Contaminated
SNPs
Prediction
Observations
Accuracy

Invalid procedure create artifact



Matthew McGowan, Jiabo Wang, Haixiao Dong, Xiaolei Liu, Yi Jia, Xianfeng Wang, Hiroyoshi Iwata, Yutao Li, Alexander E Lipka, and **Zhiwu Zhang***. **Ideas in Genomic Selection with the Potential to Transform Plant Molecular Breeding: A Review**. *Plant Breeding Review Vol 45* (eds. Irwin Goldman), John Wiley & Sons, Inc. 2021. pp. 273-320, available as Preprint, Publication, and Google Book.

Right in the middle

The Plant Genome | ORIGINAL RESEARCH

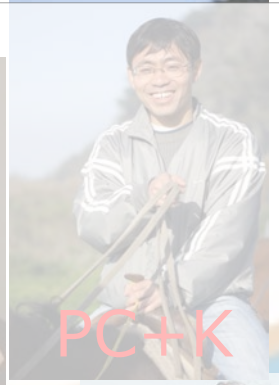
Evaluation of RR-BLUP Genomic Selection Models that Incorporate Peak Genome-Wide Association Study Signals in Maize and Sorghum

Brian Rice and Alexander E. Lipka*

was quantified. Of the 216 genetic architectures that we simulated, we identified 60 where the addition of fixed-effect covariates boosted prediction accuracy. However, for the majority of the simulated data, no increase or a decrease in prediction accuracy was observed. We also noted several instances where the


EMMAx


EMMA


PC+K

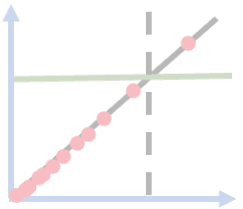

PC

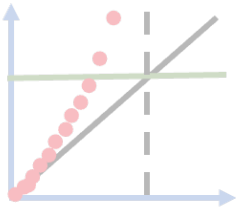

CMLM

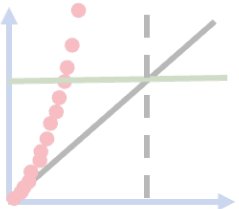

Q+K


Q


GWAS

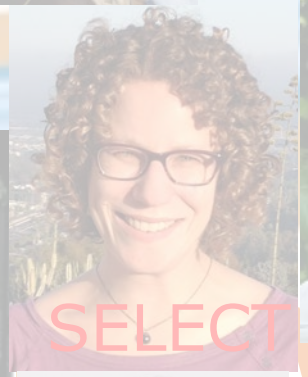







P3D


GCTA

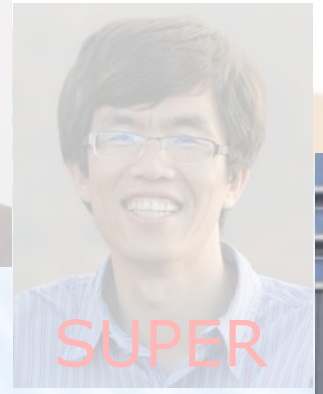

SELECT


MLMM

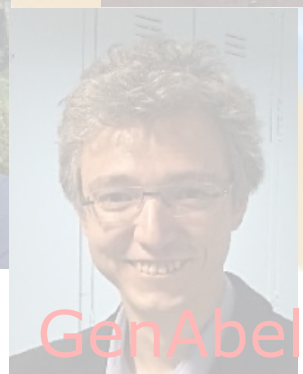
WAS Stream


FST-LMM


ECMLM


SUPER

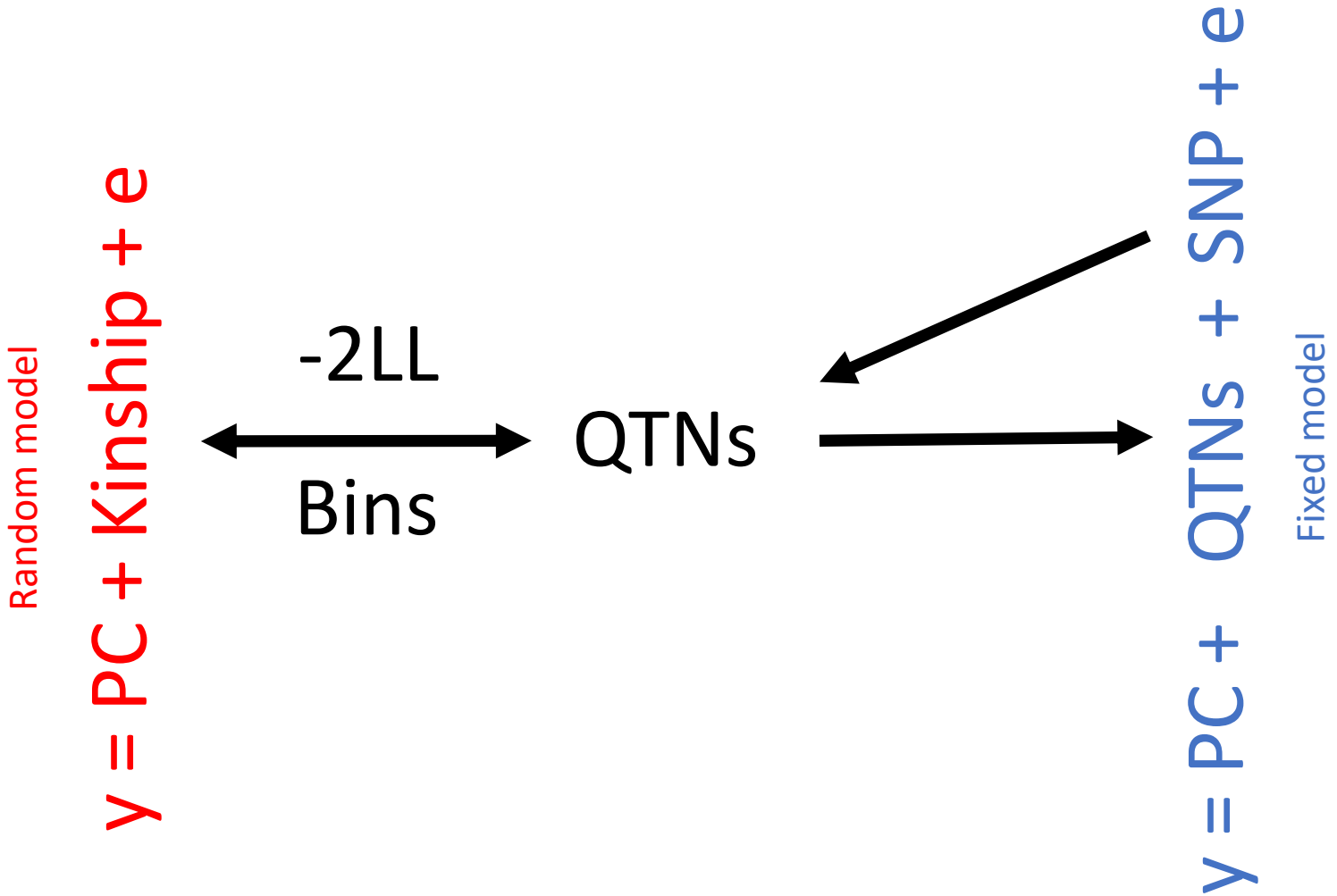

GEMMA


GenAbel

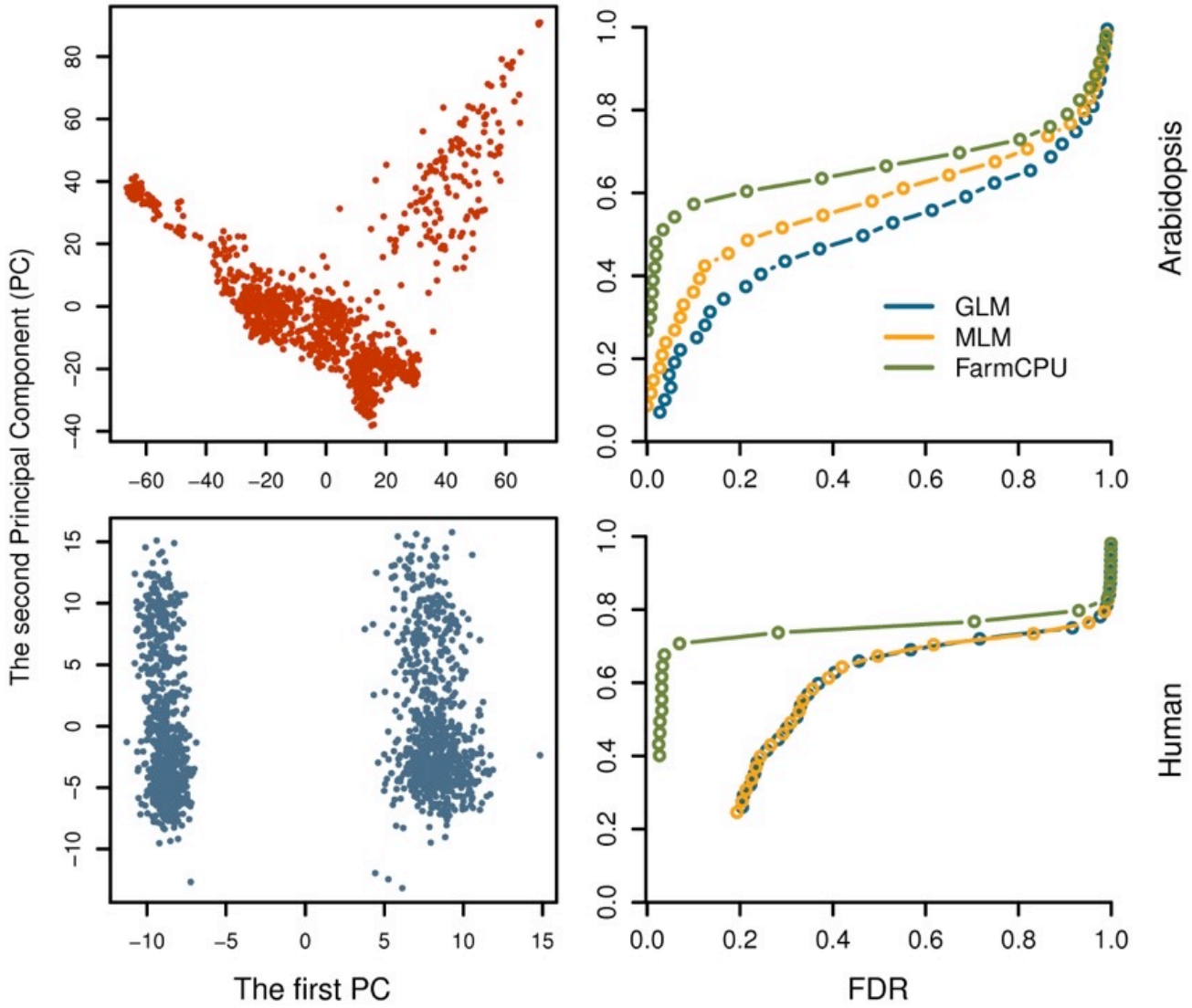

FarmCPU


BLINK

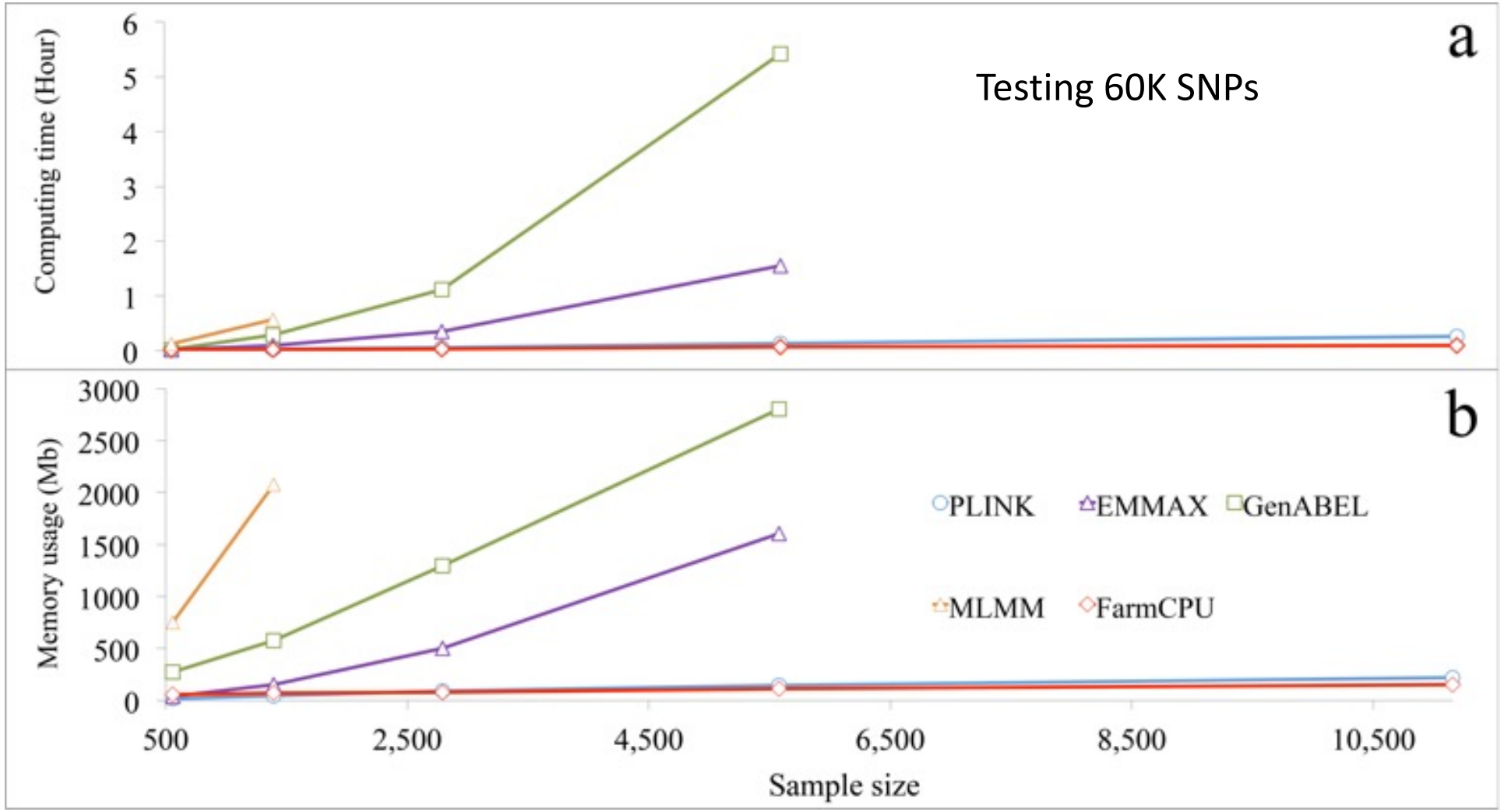
FarmCPU algorithm



It is time for human geneticists to move forward



FarmCPU is computing efficient



Plant Direct

Open Access

ORIGINAL RESEARCH | Open Access

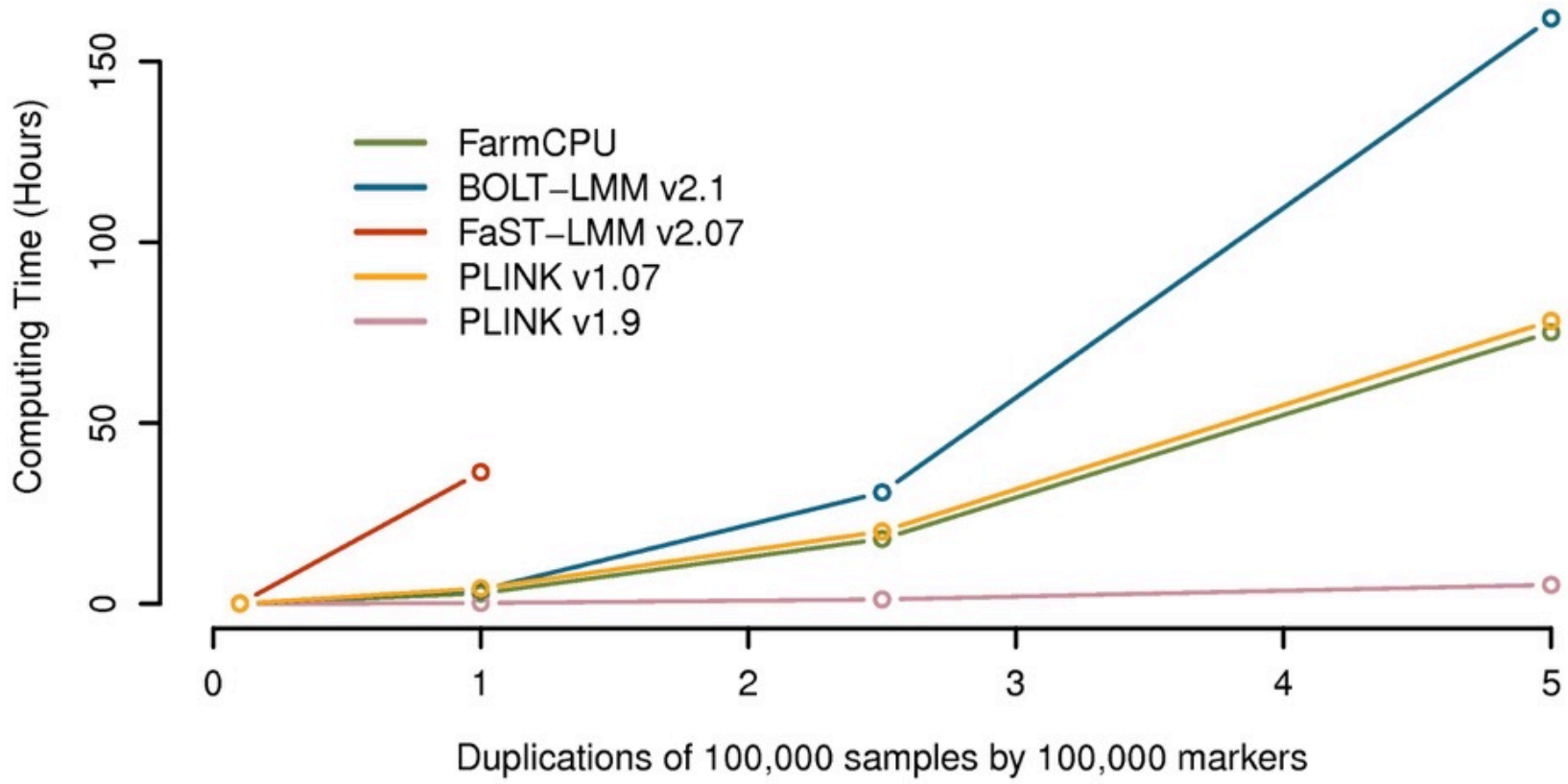
FarmCPUpp: Efficient large-scale genomewide association studies

Aaron Kusmec, Patrick S. Schnable

First published: 10 April 2018 | https://doi.org/10.1002/pld3.53 | Citations: 24

The preprint of this article can be found at
https://www.biorxiv.org/content/early/2017/12/24/238832.

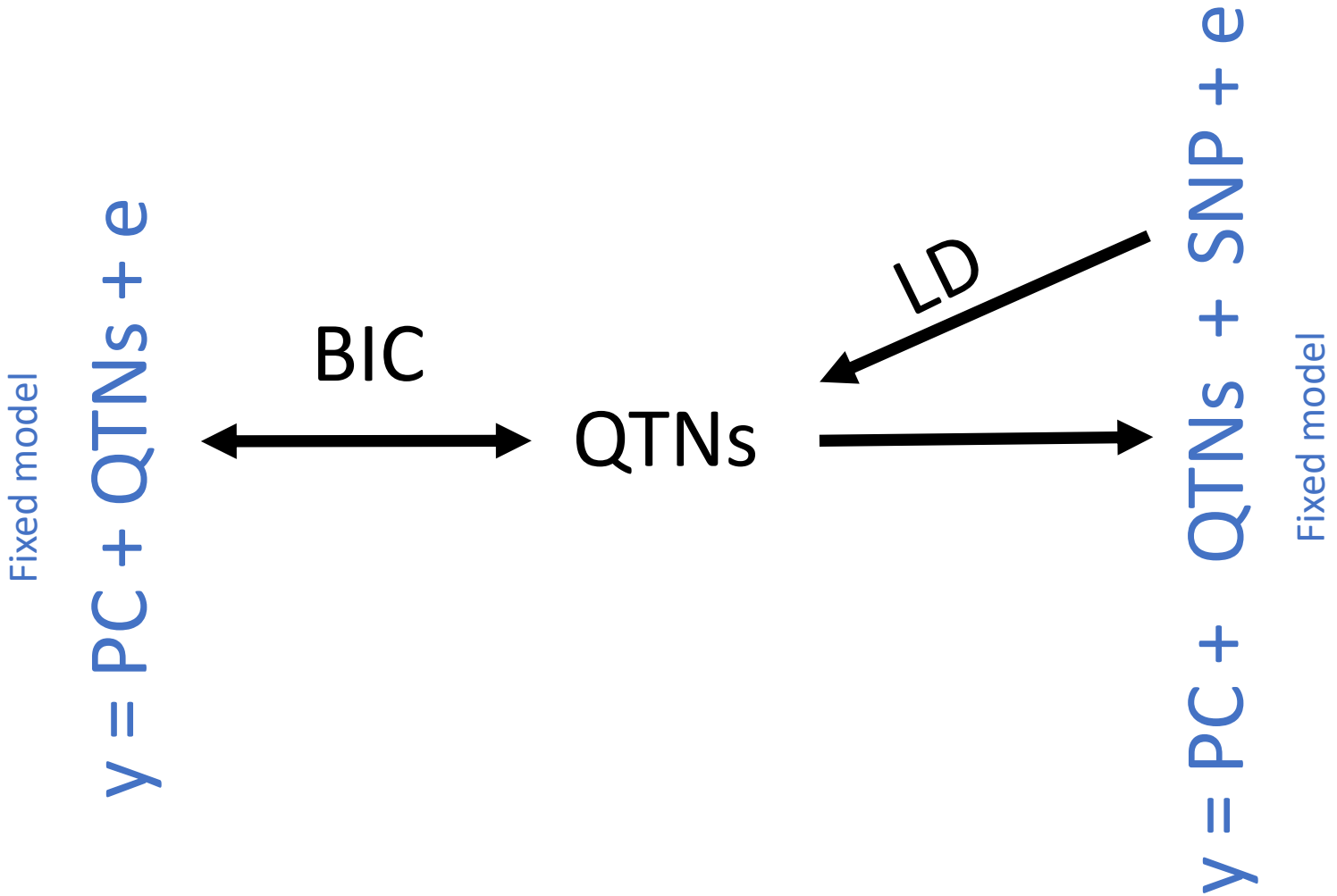
Half million individuals, half million SNPs: three days

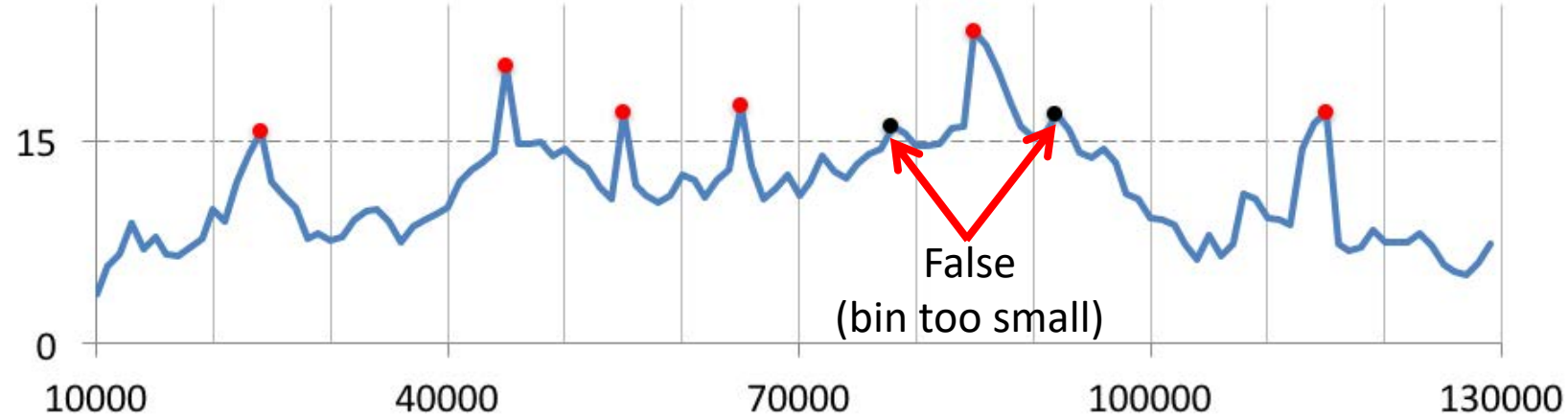


Crowdsource

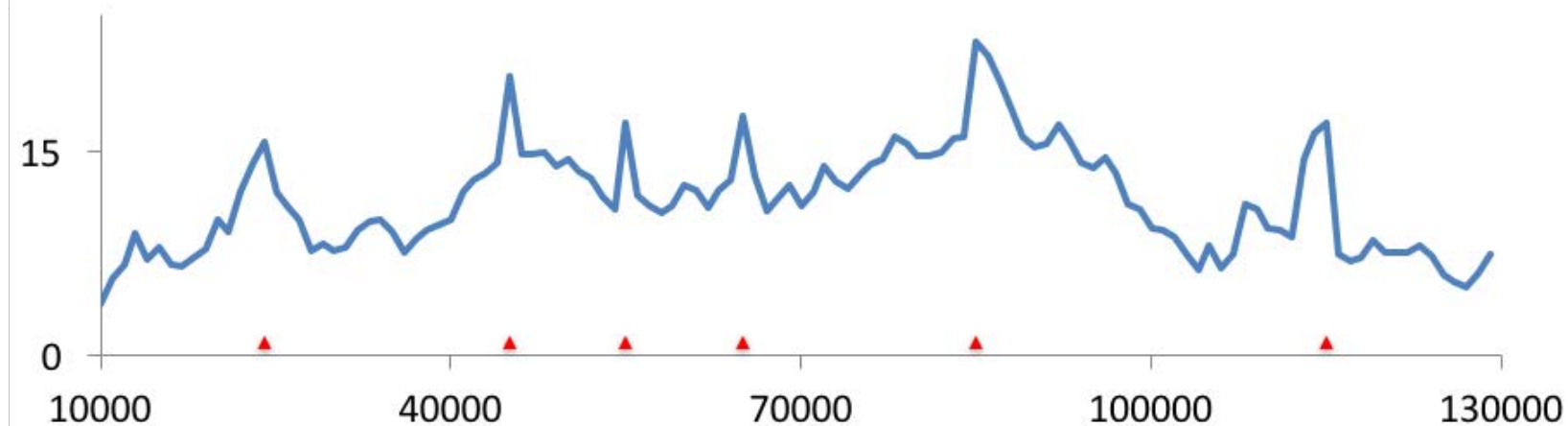
But, PINK new version is faster

BLINK algorithm

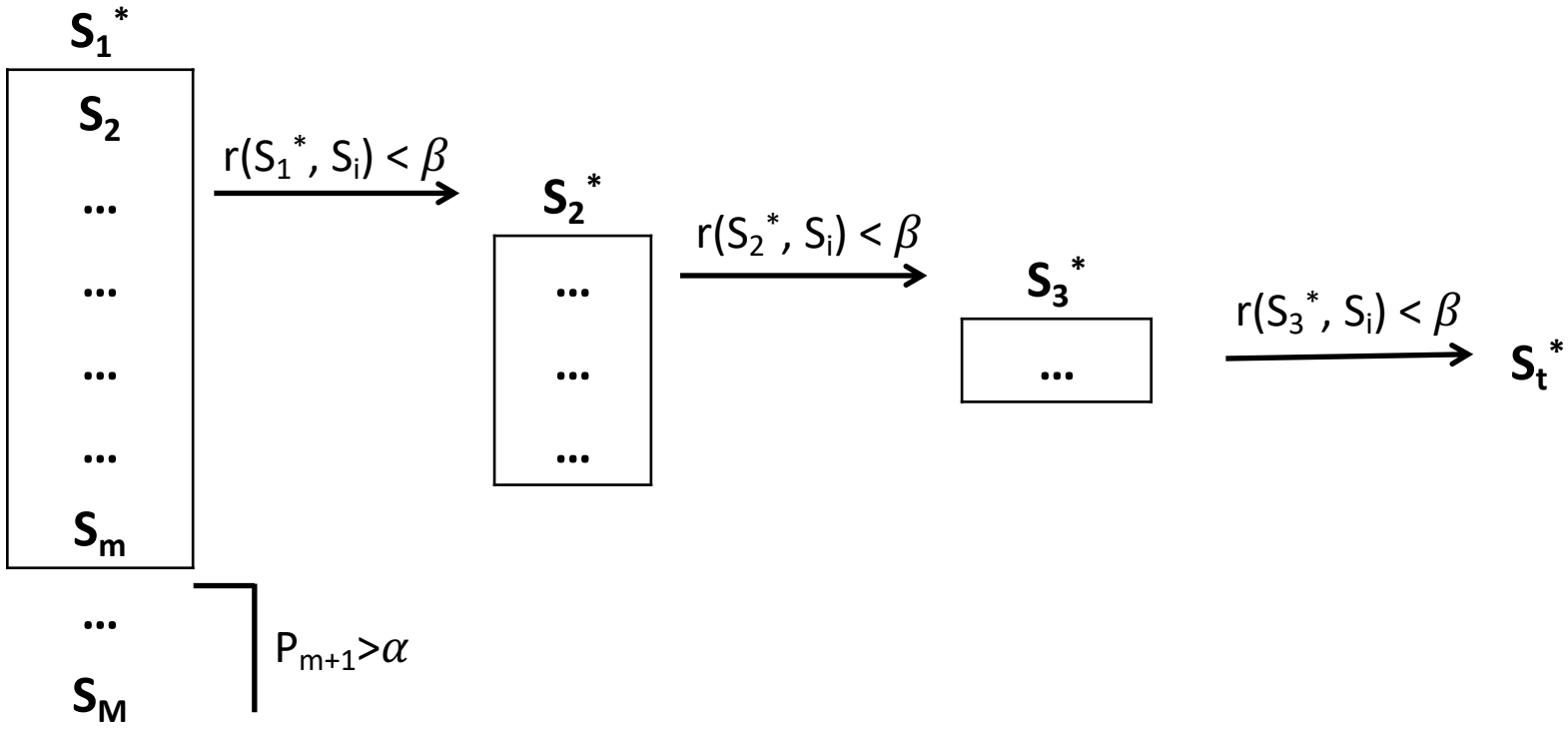



15
0
10000
40000
70000
100000
130000
False
(bin too small)


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


15
0
10000
40000
70000
100000
130000

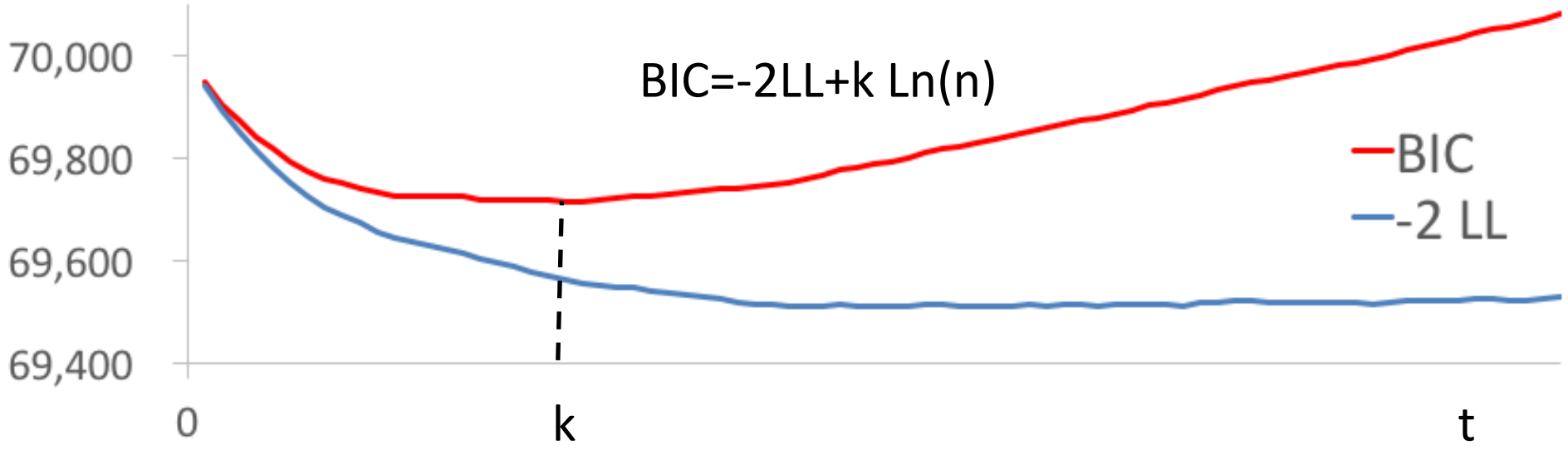
Elimination of markers with LD

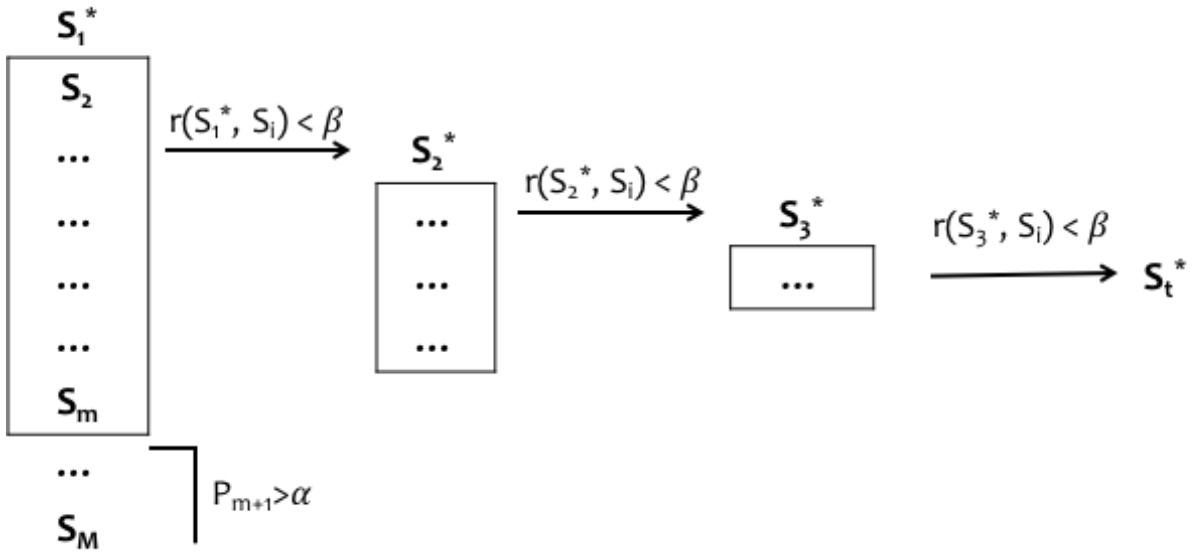


$$y = s_i + S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e, \text{ where } i = 1 \text{ to } M$$

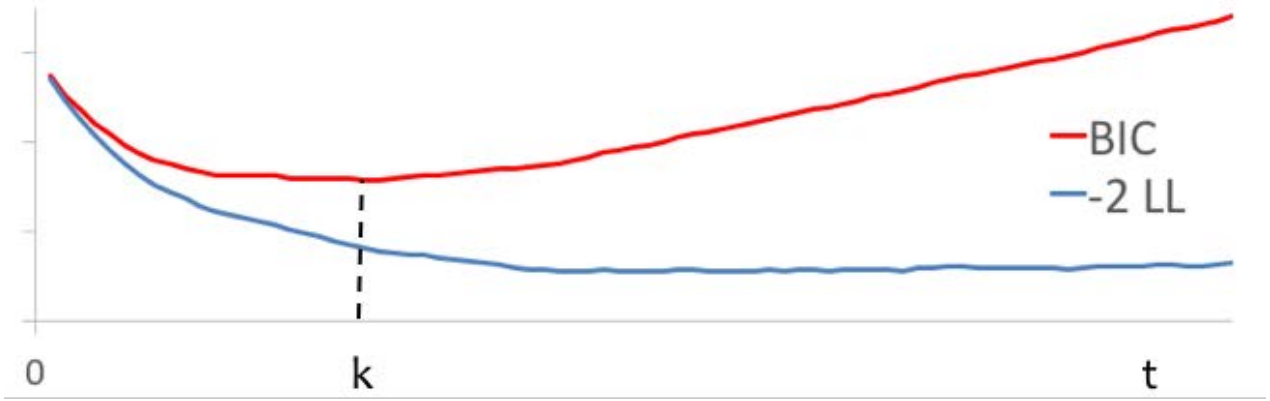
Bayesian information criterion

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





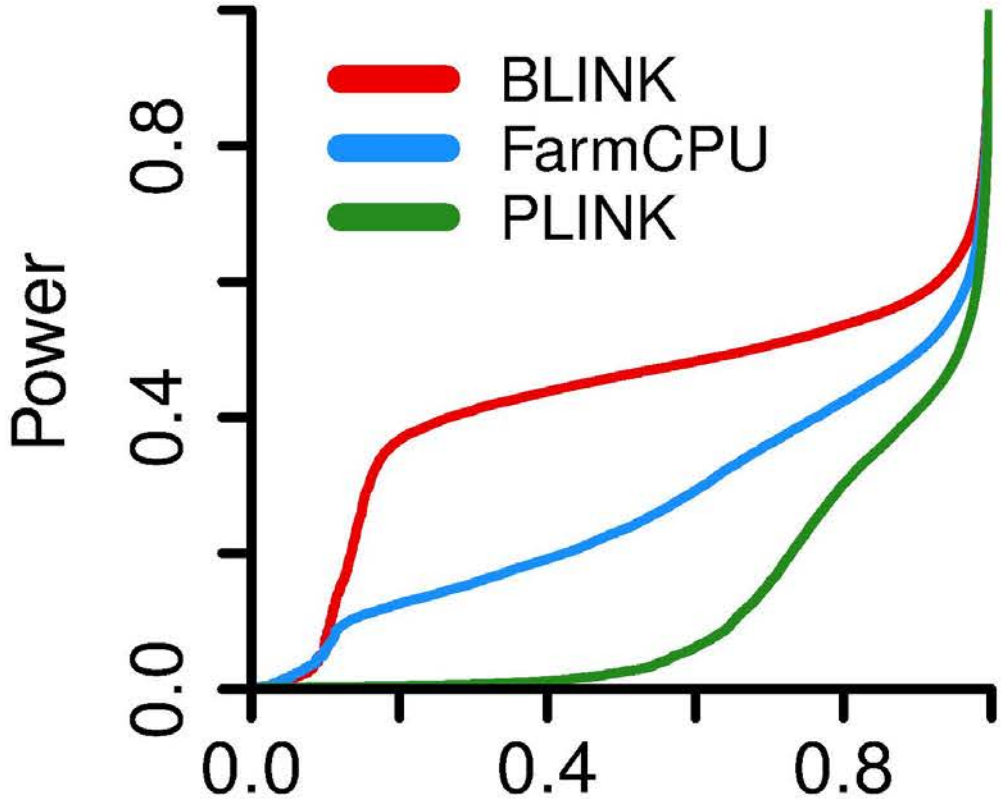
$y = S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e$, where $k \leq t$ maximizes BIC



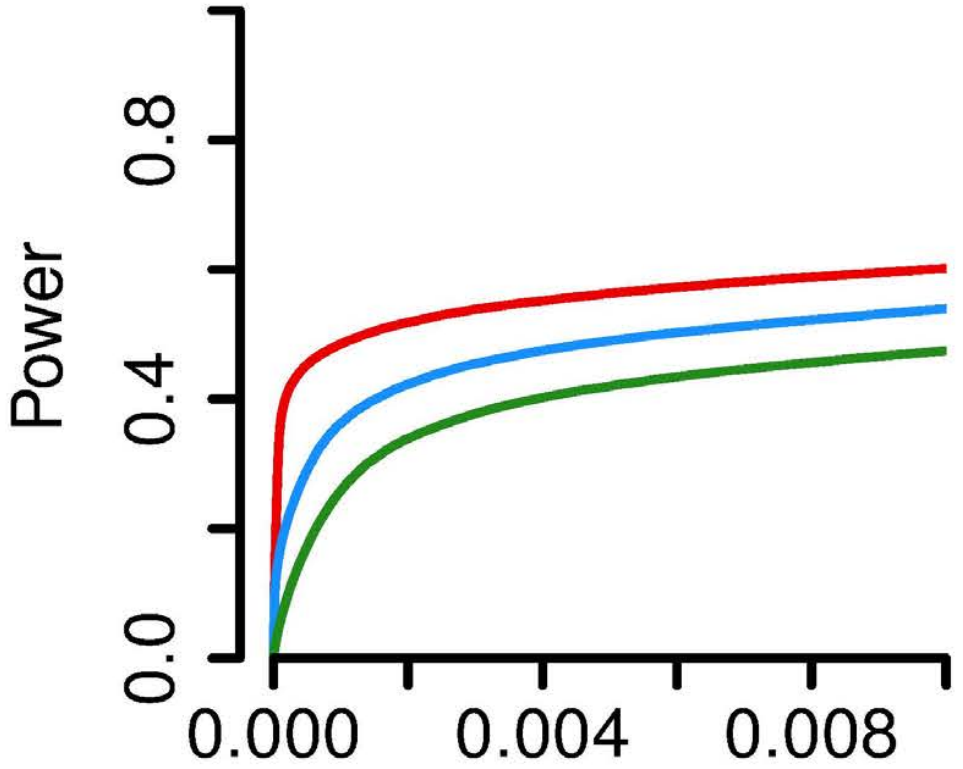
$y = s_i + S_1^* + S_2^* + S_3^* + \ldots + S_k^* + e$, where i = 1 to M

Simulation study with human data

Meng Huang



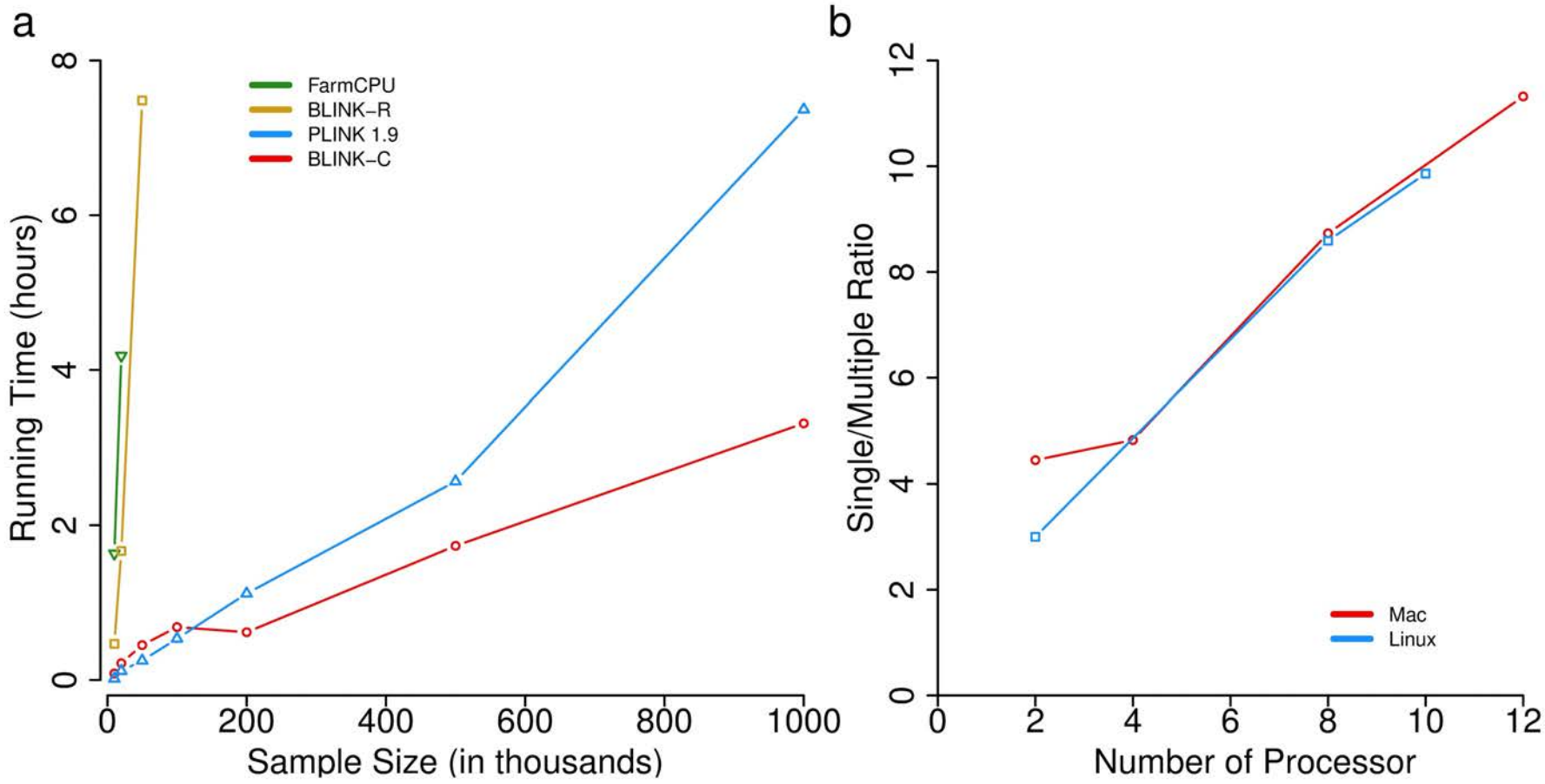
FDR



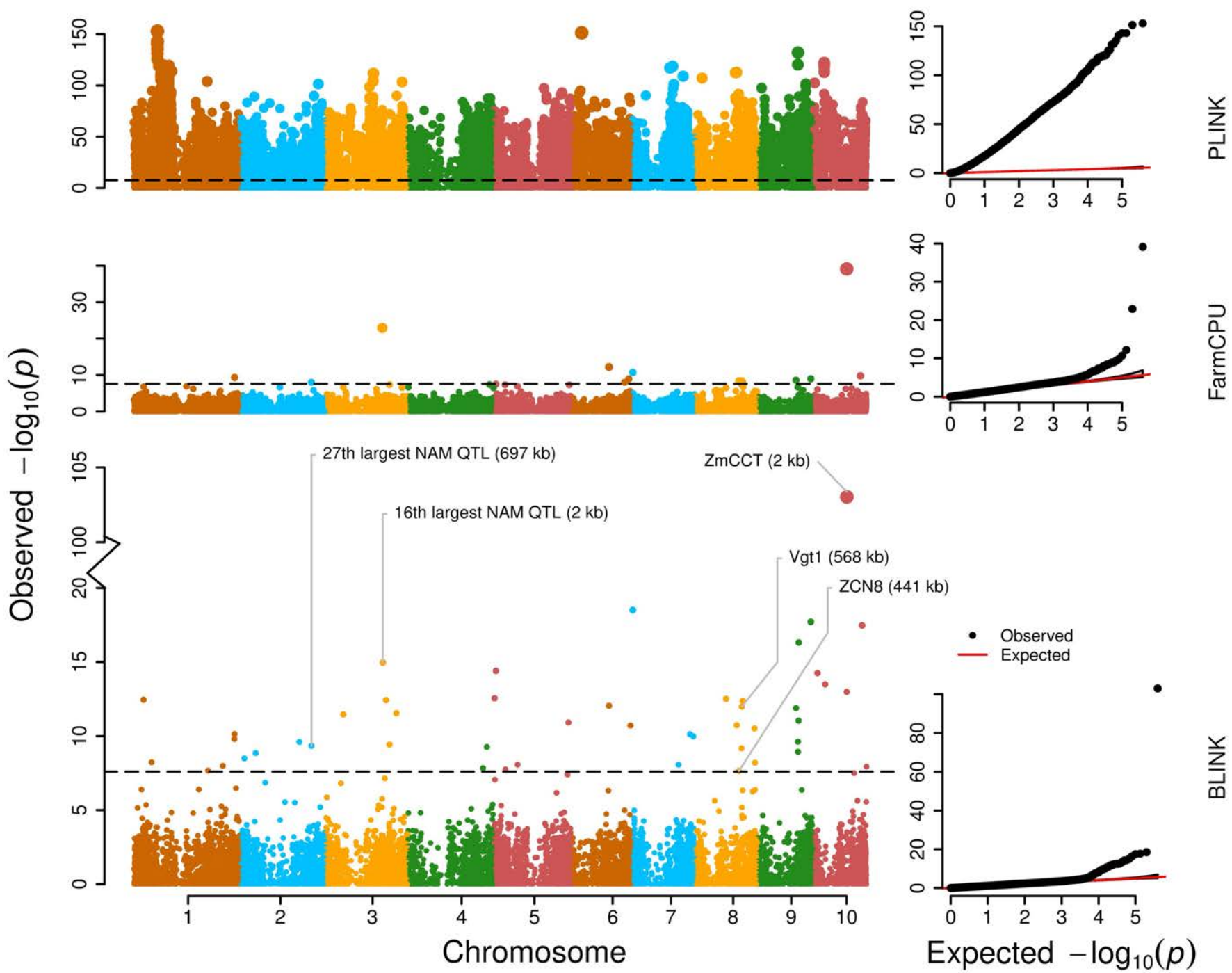
Type I error

1M SNPs + 1M Samples = 3 hours

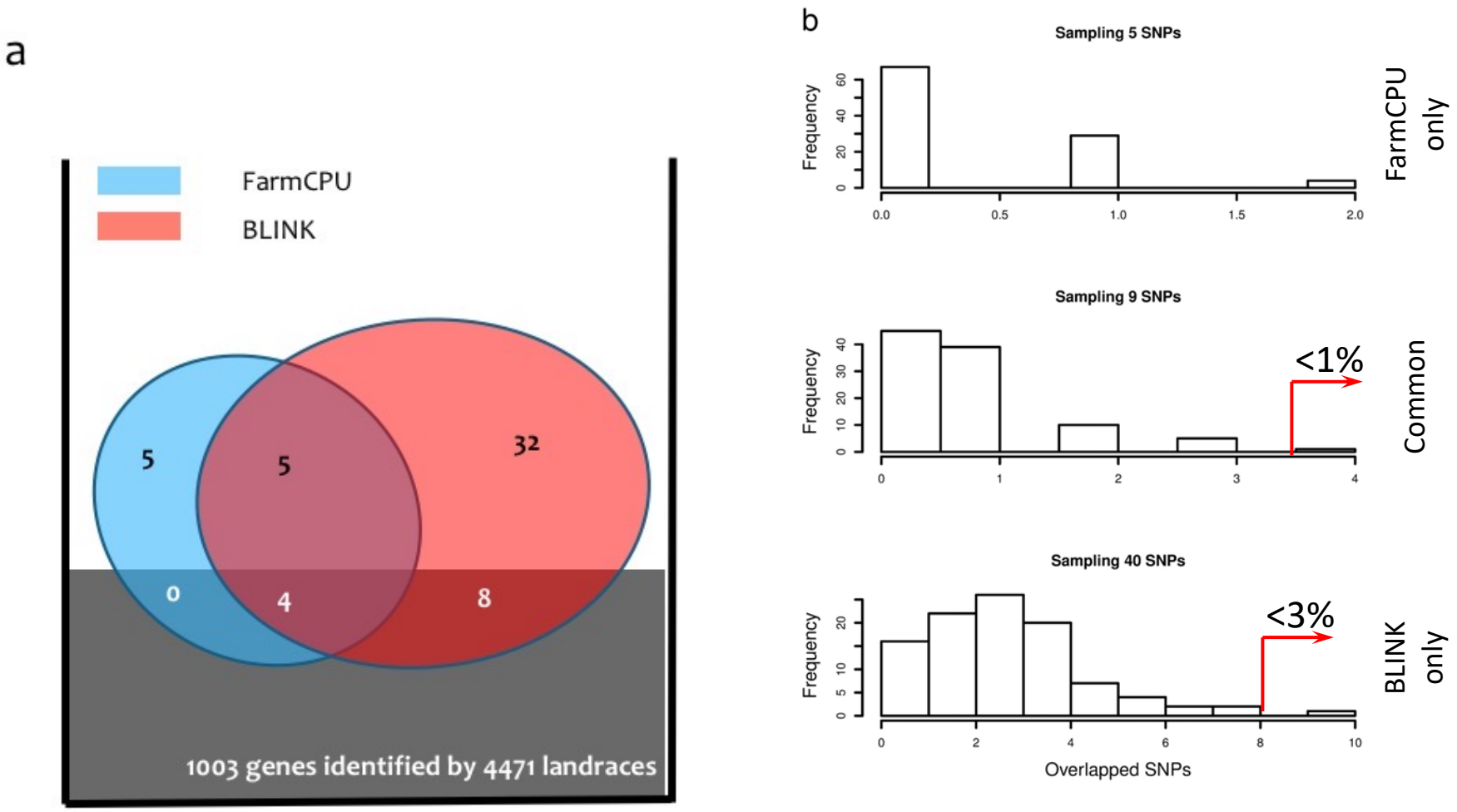
Meng Huang



Application in Maize



Enrichment



Meng Huang and et al, Giga Science, 227249

All the versions have been implemented to GAPIT with R

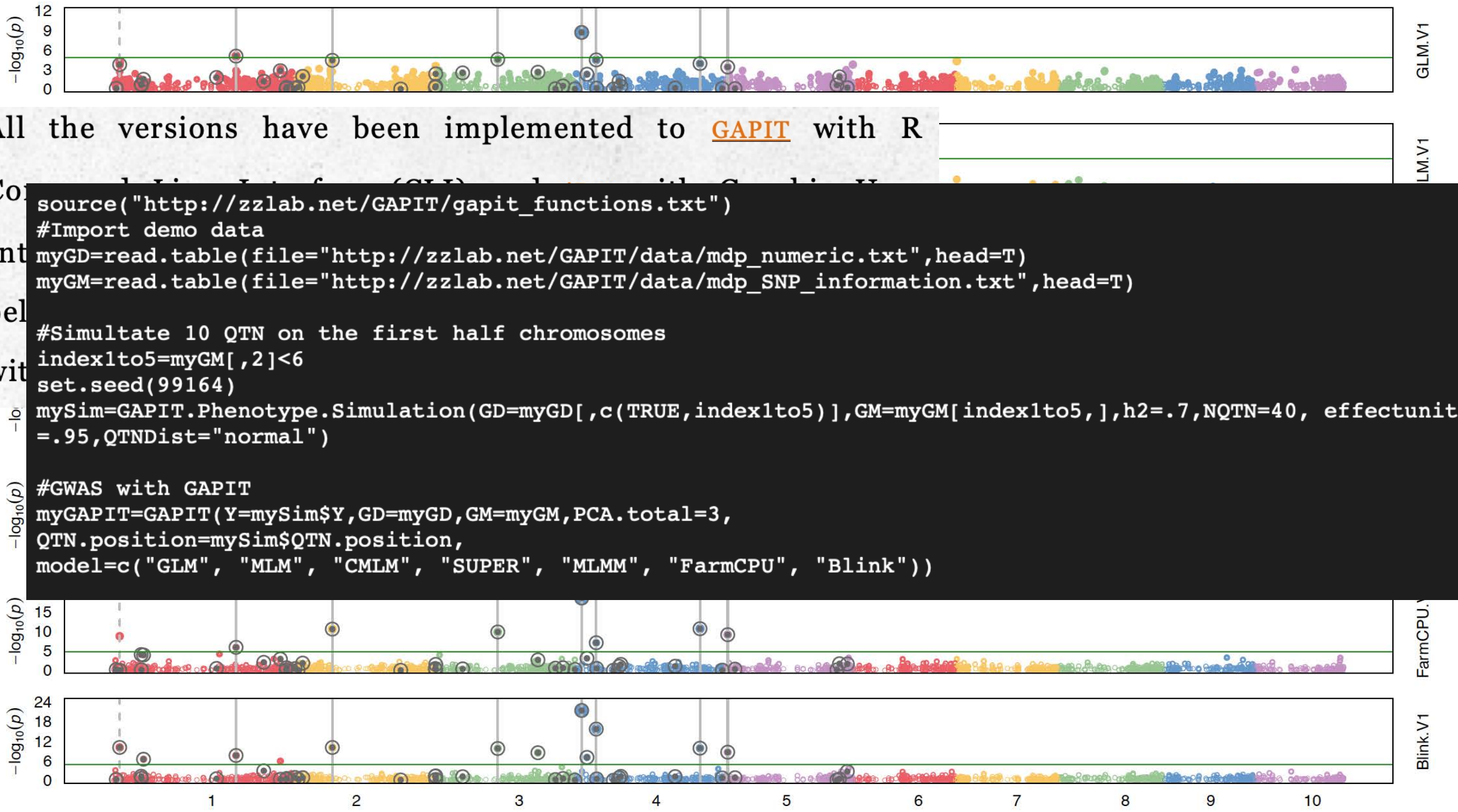
Co

Int

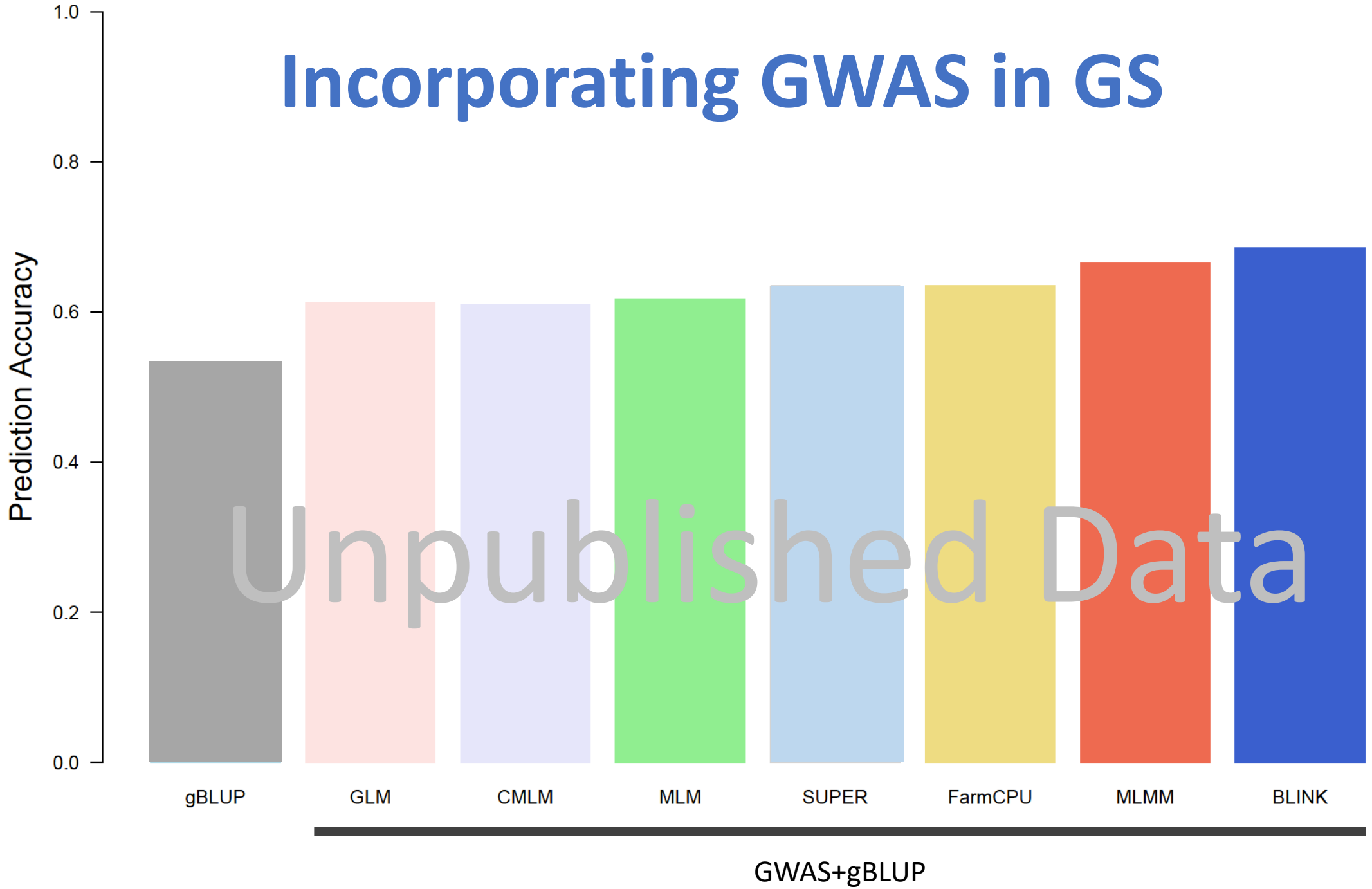
bel

wit

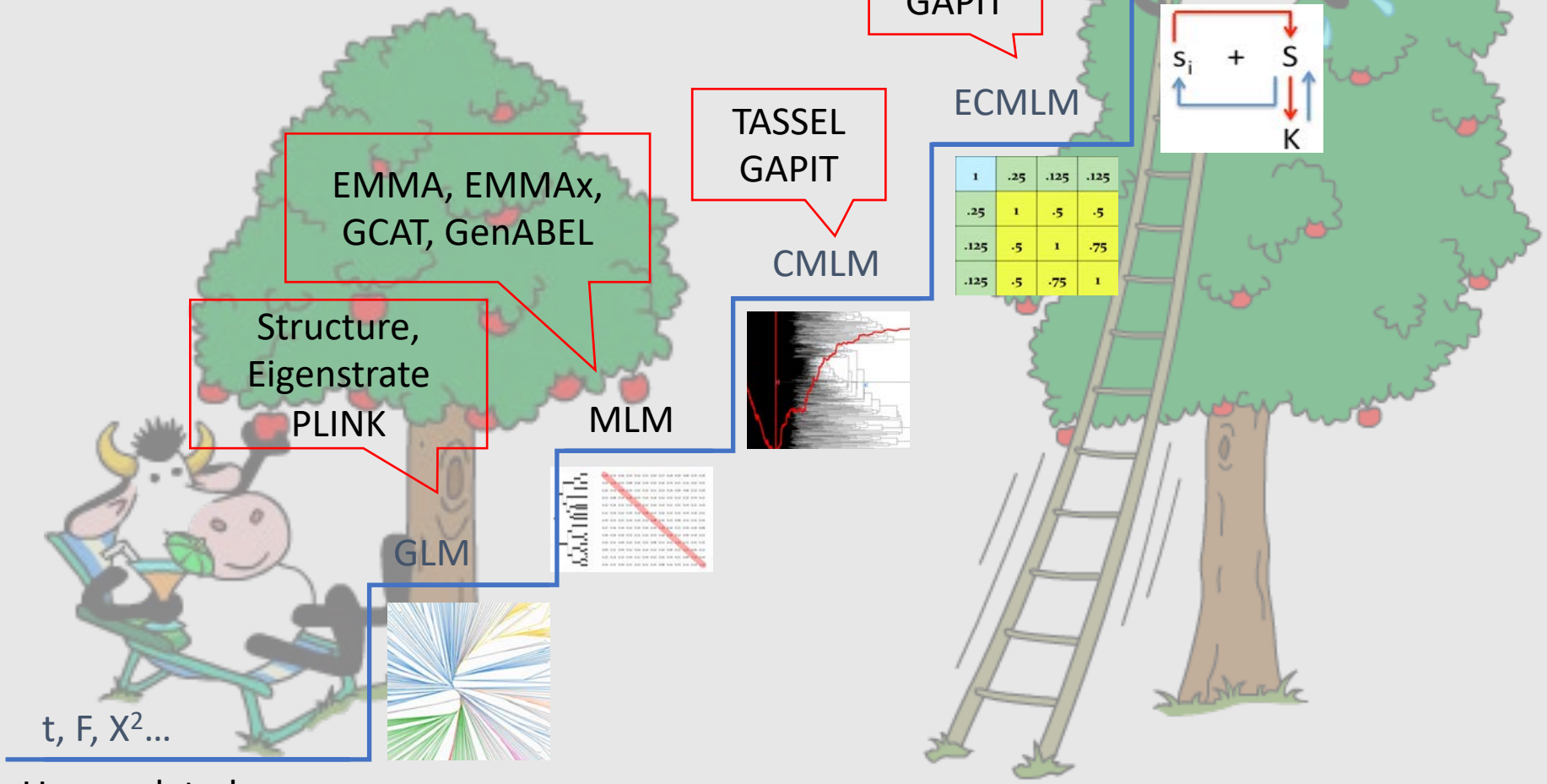
```
source("http://zzlab.net/GAPIT/gapit_functions.txt")
#Import demo data
myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#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"))
```



Incorporating GWAS in GS



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

65th ISI World Statistics Congress 2025

The Hague

05 October 2025 - 09 October 2025

Statistics Concourse Of Machine Learning And Artificial Intelligence

Organiser

Prof. Zhiwu Zhang


Thank you for your attention!
World Class Scenery, Research, Education, & Extension