Confusion and Prospection of Genomic Prediction to Incorporate 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

Genomic study

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

} Backward

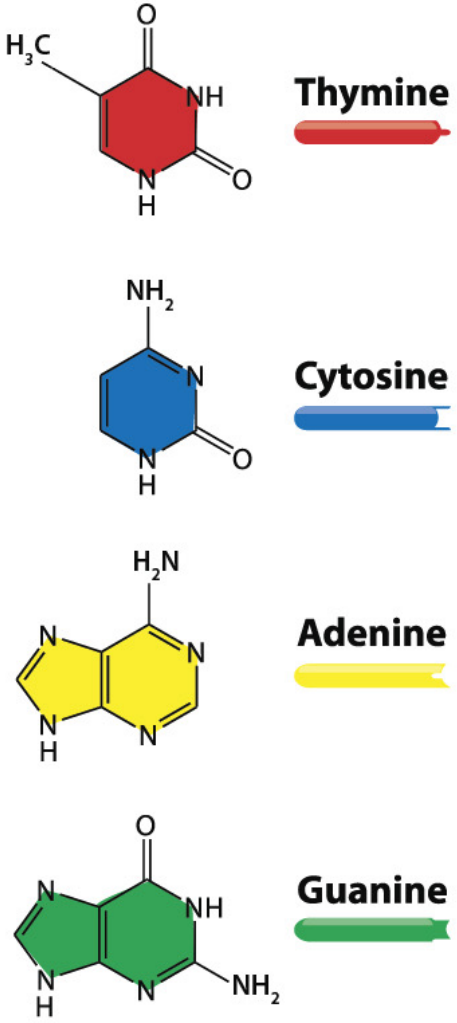
} Forward

❖Prediction
- MAS
- GS
- GWAS+GS
- AI

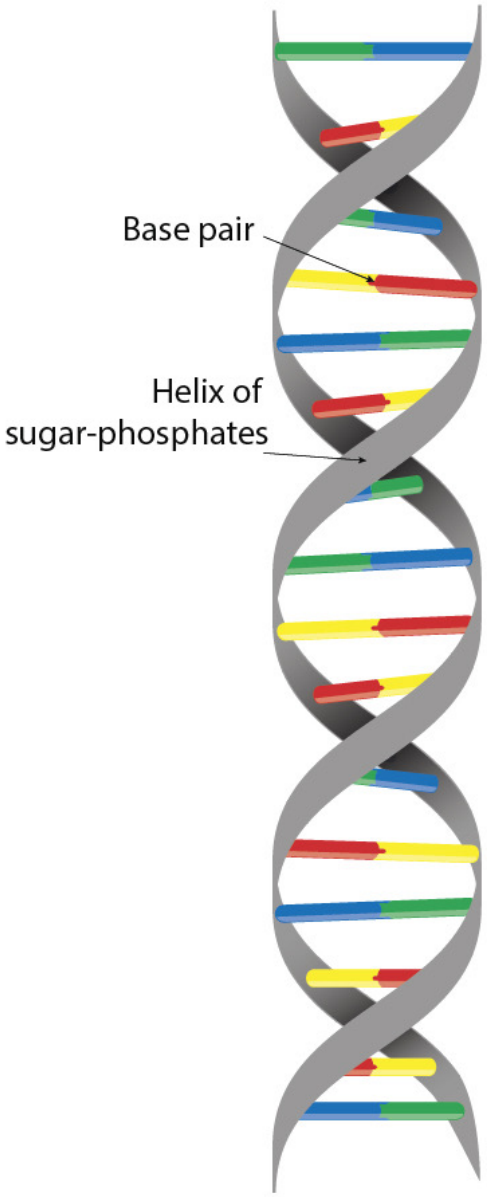
Outline

- MAS to GS
- Prediction assessment
- Muddy water
- Hidden overfitting
- GWAS+GS

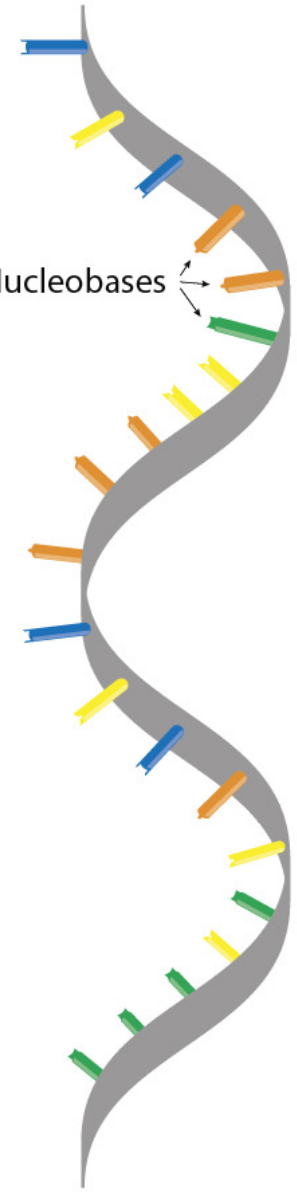
James Watson and Francis Crick DNA Model (Nobel Prize, 1962)



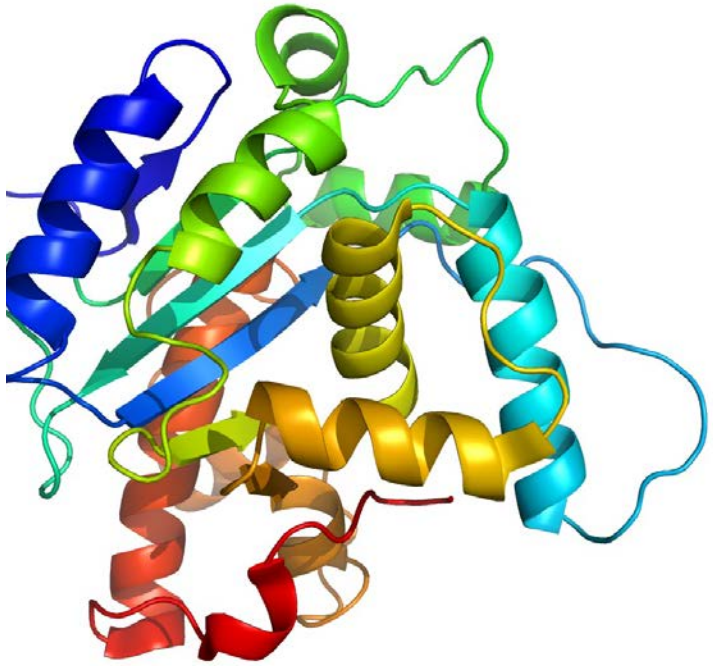
Nucleobases of DNA



DNA
Deoxyribonucleic acid



RNA
Ribonucleic Acid



Genotypes

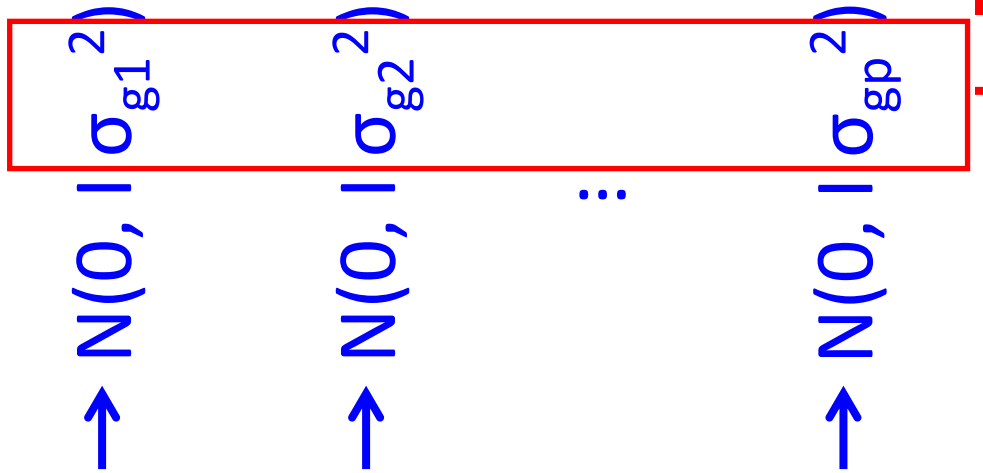
AA: 0

AT: 1 ★

TT: 2 ★★

Genomic Prediction

Prior distributions



→ RR, Bayes A, B, …

Bottom up

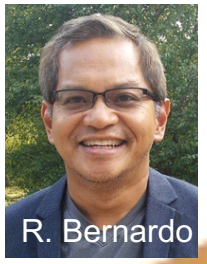
$y = x_1 g_1 + x_2 g_2 + \ldots + x_p g_p + e$

Top Down

$U \sim N(0, K \sigma_a^2)$

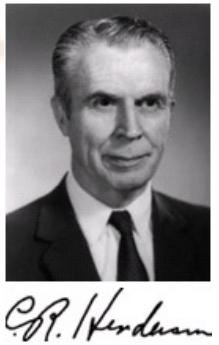
$y = Zu + e$ gBLUP, ssBLUP, cBLUP, sBLUP …

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



RFLP kinship in maize
Crop Science
1994

MAS using BLUP
GSE, 1989



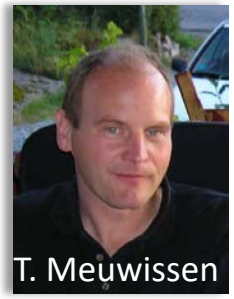
BLUP
Biometrics,1975

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

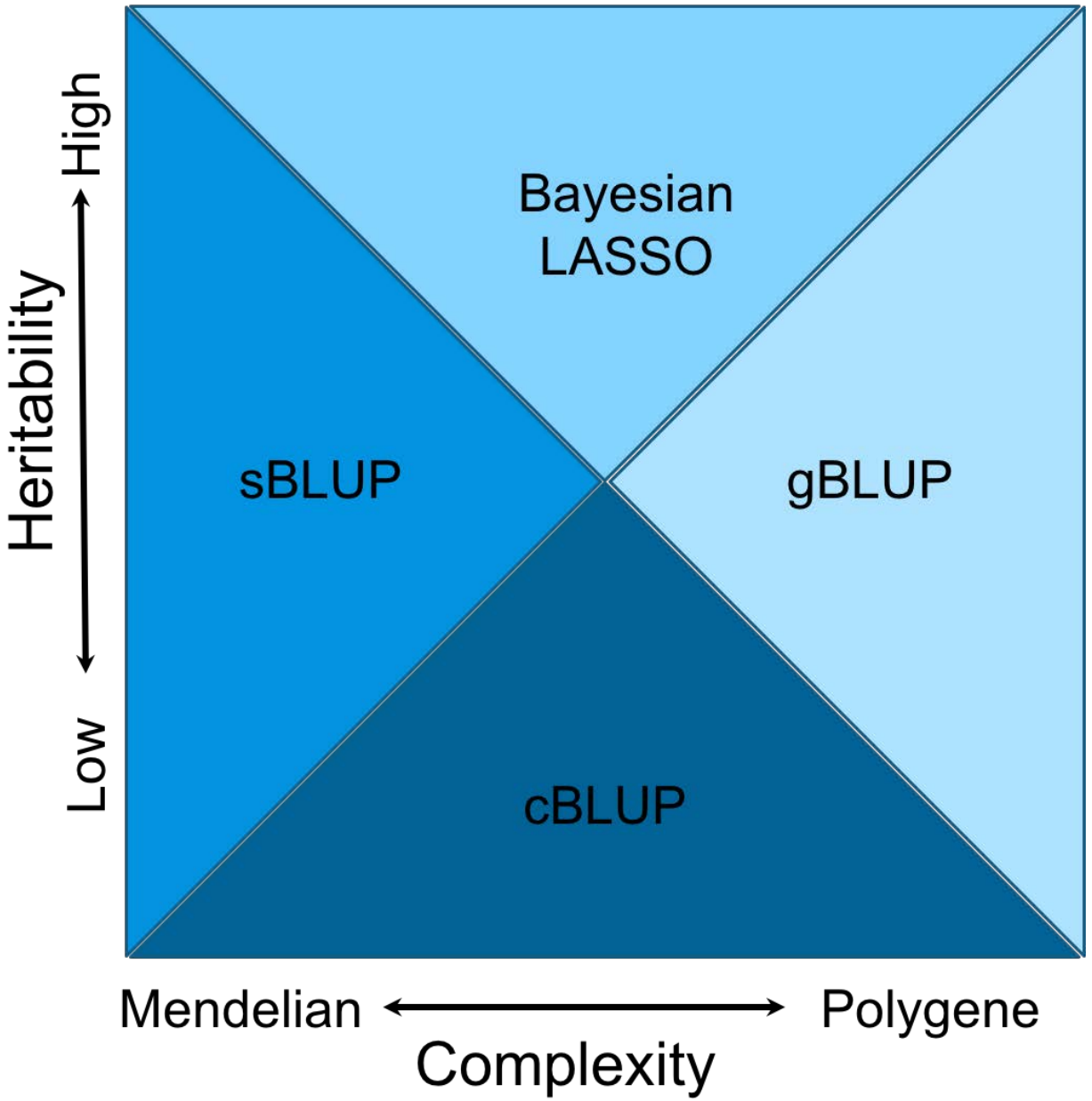
cBLUP

sBLUP

Super and compression
Heredity
2018

Genomic prediction

Domains

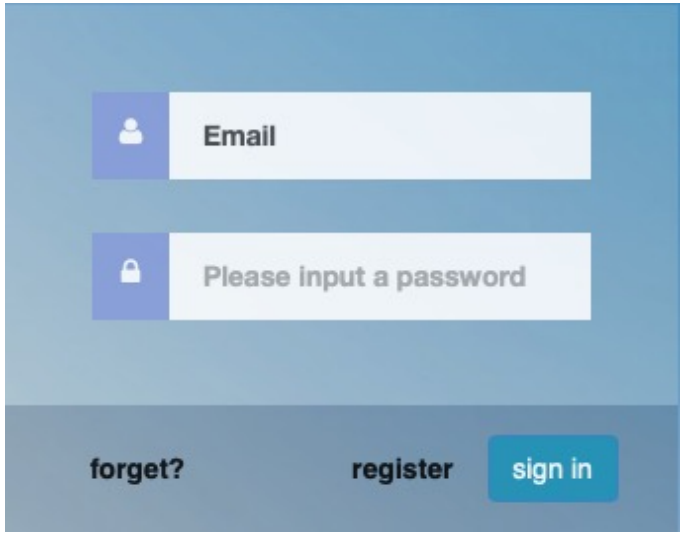


Jiabo Wang

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



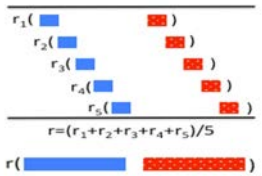
Select the best method using machine learning

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

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

Outline

- MAS to GS
- Prediction assessment
- Muddy water
- Hidden overfitting
- GWAS+GS



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 $P = 0.05$

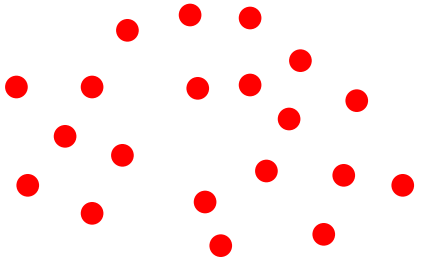
[a] Number of individuals in the biparental population

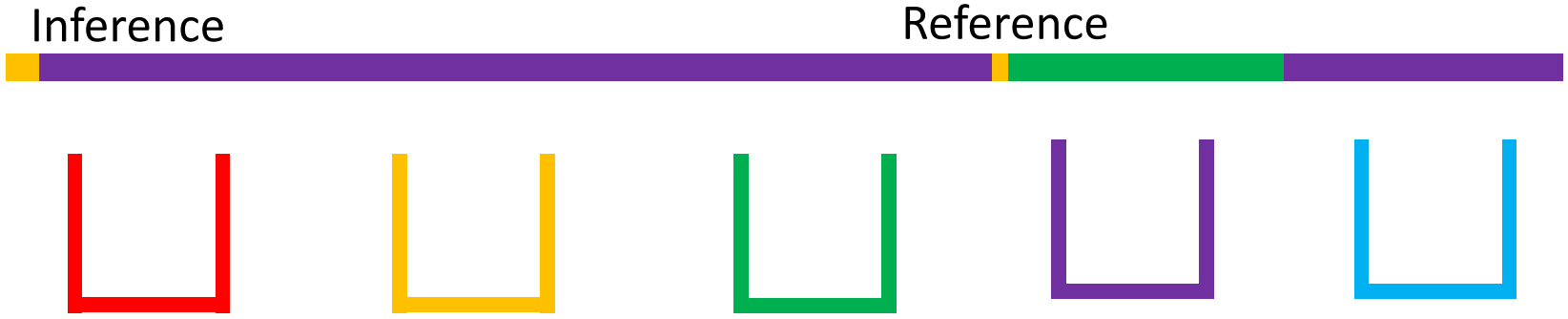
[b] Number of polymorphic SNPs in the biparental population

[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.

Five fold Cross validation



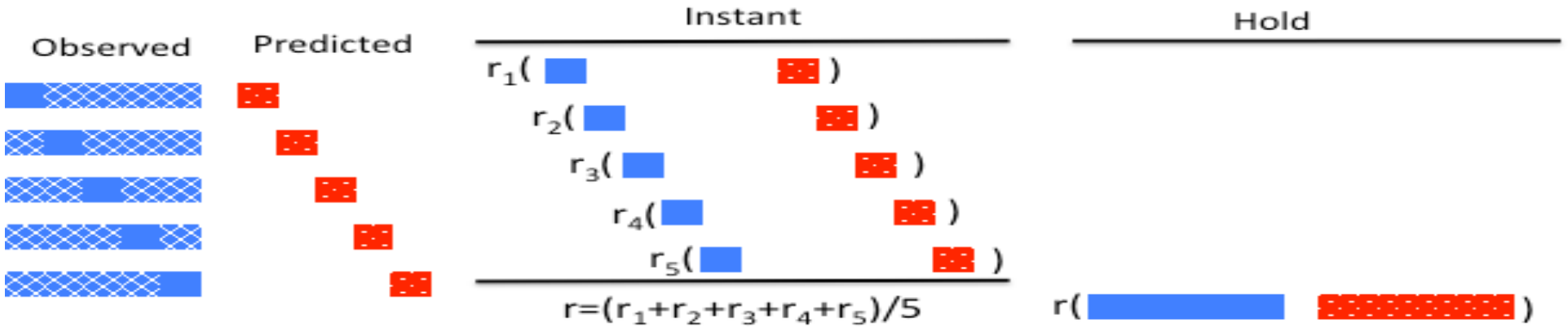


By Yao Zhou

Jack Knife: extreme case: K=N

- N: number of individuals
- K: number of folds
- Leave-one-out cross-validation
- Inference (training) contain only one individuals
- Not possible to calculate correlation between observed and predicted within inference
- Evaluation of accuracy must be hold until every individuals receive predictions.
- Resampling is not available

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

Generation of random phenotypes

taxa	PZB00859.1	PZA01271.1	PZA03613.2		PZA03402.1	PZA03429.1
4226	2	0	0		0	2
4722	2	2	0		0	2
33-16	2	0	0		0	2
38-11	2	2	0		0	2
A188	0	0	0		0	2
A214N	2	0	2		0	2
A239	0	0	2		0	2
A272	0	0	2		0	2
A441-5	2	0	0		0	2
A554	2	2	2		0	2
A556	2	0	0		0	2
A6	0	0	2		0	2
A619	2	2	0		0	2
A632	2	0	2		0	2
A634	2	0	2		0	2
A635	2	0	2		0	2
A641	0	0	2		0	2
A654	1	2	2		0	2
A659	0	2	2		0	2
A661	2	2	2		0	2
A679	2	0	2		0	2
A680	2	0	2		0	2
A682	2	0	2		0	2
AB28A	0	2	2		0	2
B10	0	0	0		0	2
B103	0	0	0		0	2
B104	2	0	2		0	2
B105	2	0	2		0	2
B109	2	0	0		0	2
B115	2	2	0		0	2
B14A	2	0	2		0	2
B164	2	2	2		0	2
B2	0	0	0		0	2
B37	0	0	0		0	2
B46	2	2	0		0	2
B52	2	0	2		0	2
B57	2	2	2		0	2
B64	2	2	0		0	2
B68	2	0	2		0	2

Genotype

Taxa	dpoll
4226	59.5
4722	71.5
33-16	64.5
38-11	68.5
A188	62
A214N	69
A239	61
A272	70
A441-5	67.5
A554	66
A556	65
A6	80.5
A619	61
A632	61
A634	59
A635	64
A641	66
A654	64
A659	58.5
A661	59
A679	66
A680	65.5
A682	57.5
AB28A	78
B10	69
B103	57.5
B104	64.5
B105	68
B109	64
B115	65.5
B14A	63.5
B164	58
B2	70
B37	65.5
B46	69
B52	70
B57	65
B64	68.5
B68	71.5

Phenotype

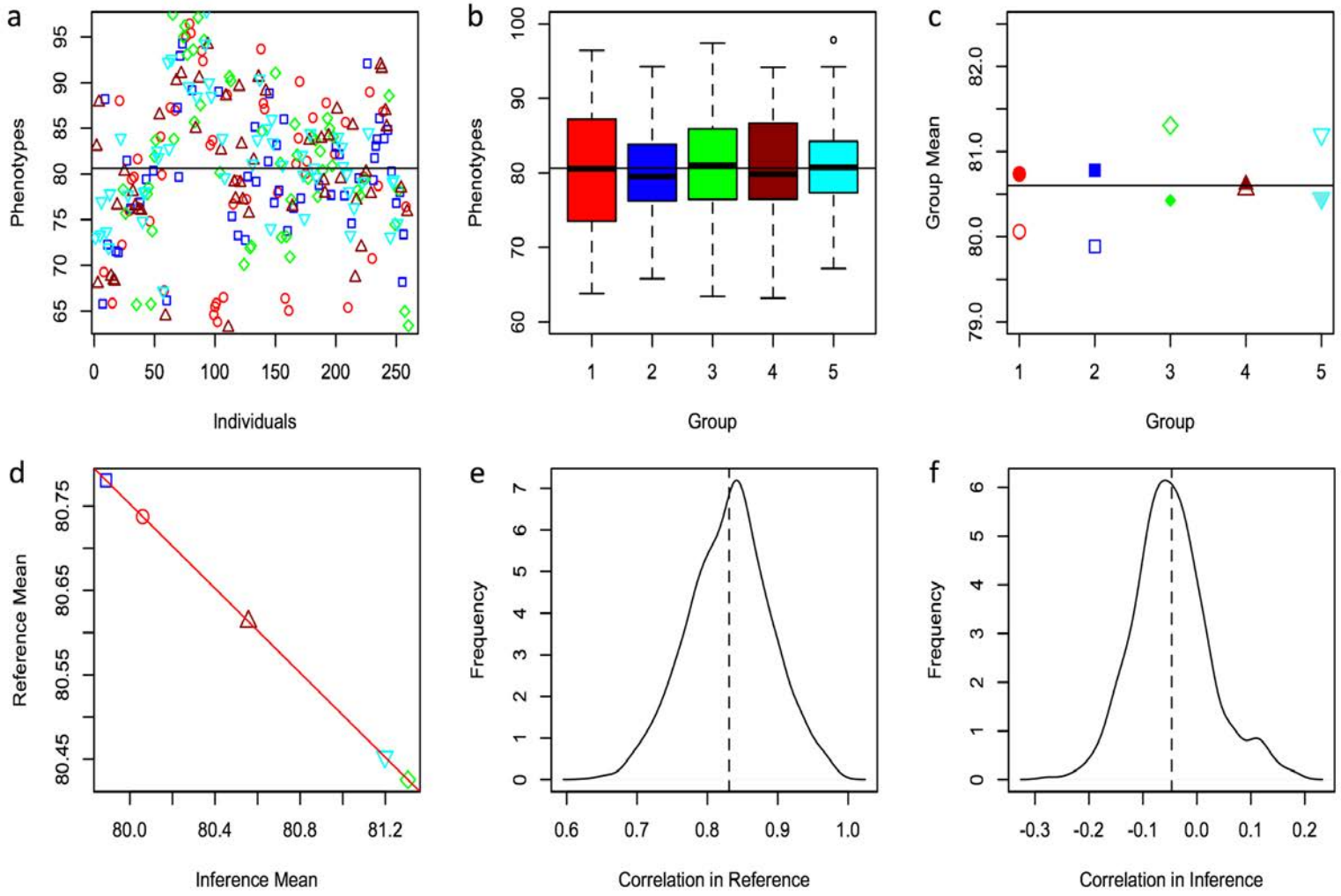
RND	dpoll
0.86856367	59.5
0.02012598	71.5
0.79231414	64.5
0.40900666	68.5
0.62047235	62
0.28593425	69
0.97570212	61
0.37095398	70
0.51322382	67.5
0.27095609	66
0.98654092	65
0.95714296	80.5
0.25612284	61
0.36130694	61
0.6319869	59
0.90208838	64
0.68898235	66
0.22075234	64
0.31518438	58.5
0.66947598	59
0.03617112	66
0.05122774	65.5
0.56681507	57.5
0.5145458	78
0.54586645	69
0.40921337	57.5
0.90235341	64.5
0.19812172	68
0.5190116	64
0.17744906	65.5
0.29742291	63.5
0.70530031	58
0.89535821	70
0.22439474	65.5
0.67132952	69
0.16697539	70
0.42612281	65
0.33780376	68.5
0.02590336	71.5

Random #

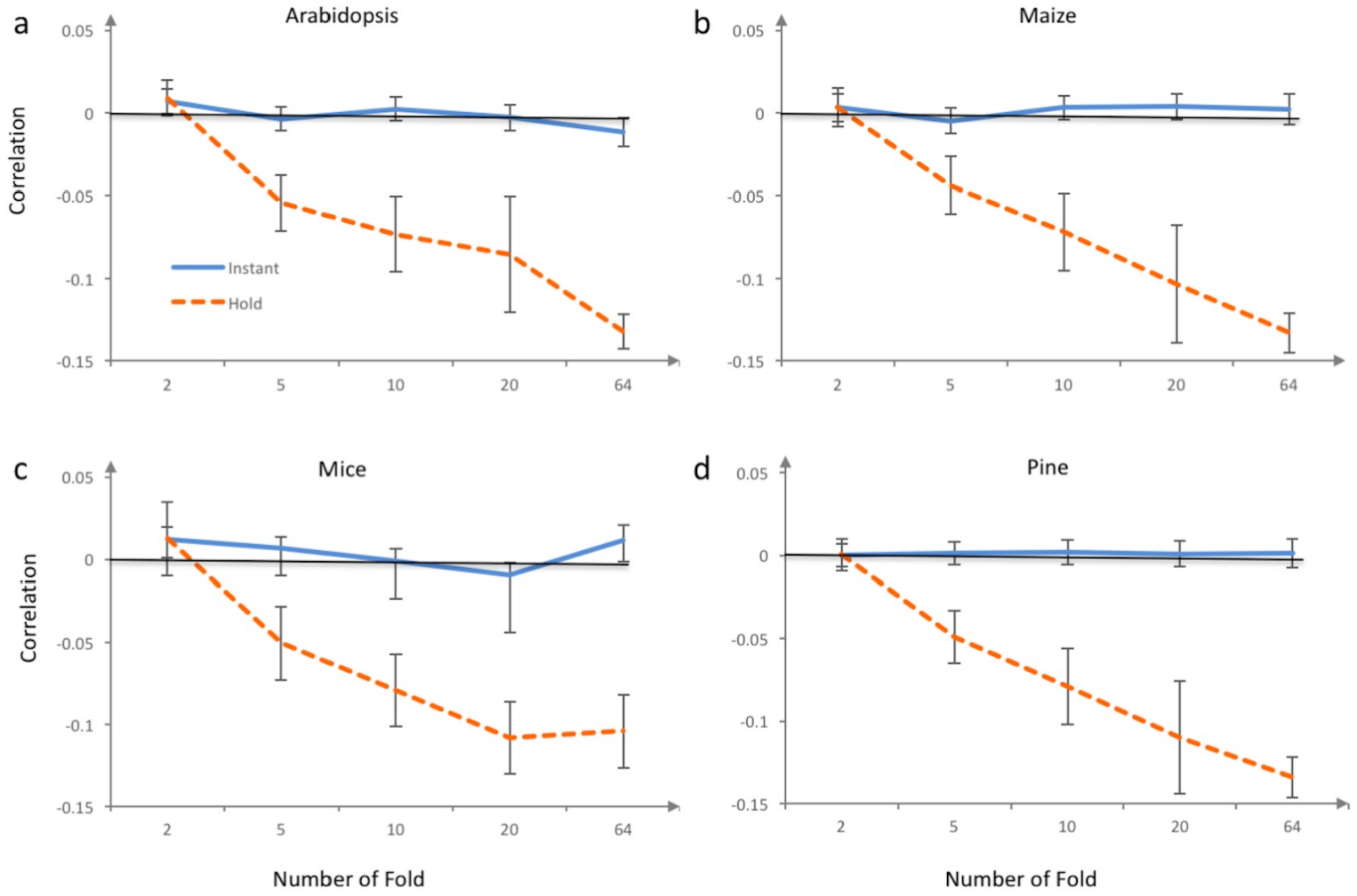
RND	dpoll
0.00092097	64.5
0.00411271	82
0.01222961	63
0.01408883	72
0.01429854	77
0.01724495	66
0.01792515	62.5
0.01813216	66
0.02012598	71.5
0.02419609	65
0.02515667	72.5
0.02590336	71.5
0.02893704	69.5
0.03223024	63
0.03380974	68.5
0.03617112	66
0.04770943	67
0.04913043	74
0.05122774	65.5
0.05374715	74.5
0.05535521	70.5
0.05623595	75
0.05878553	69.5
0.06133497	78
0.07643858	63.5
0.07654744	68
0.0775073	60
0.07755255	61.5
0.08022356	85
0.08312494	67.5
0.08381389	65.5
0.0845373	70
0.08650617	68
0.08773562	67
0.09315311	72.5
0.09870341	62.5
0.10257672	67
0.11579583	65.5
0.11955247	67

Random Phenotype

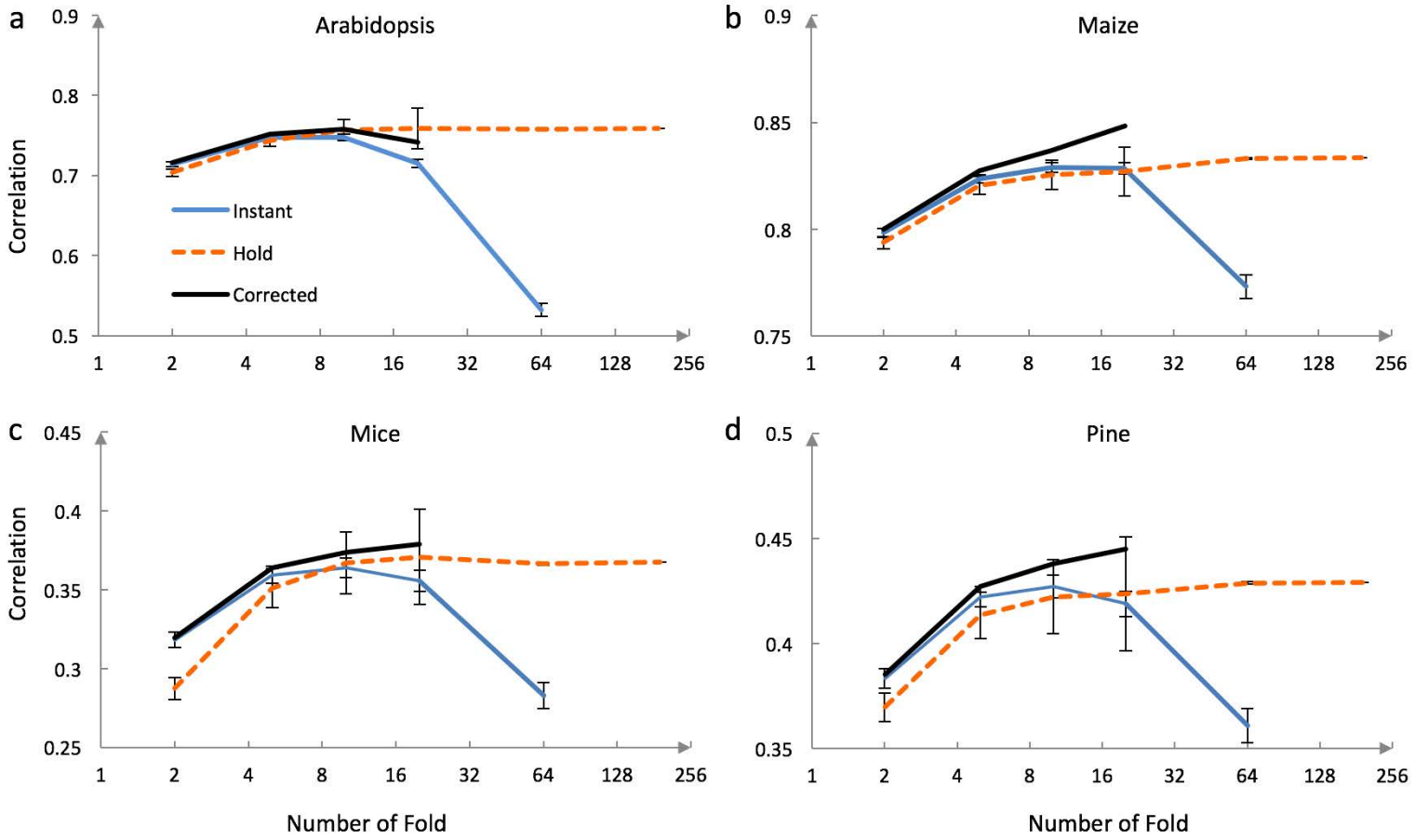
Artefactual negative hold accuracy



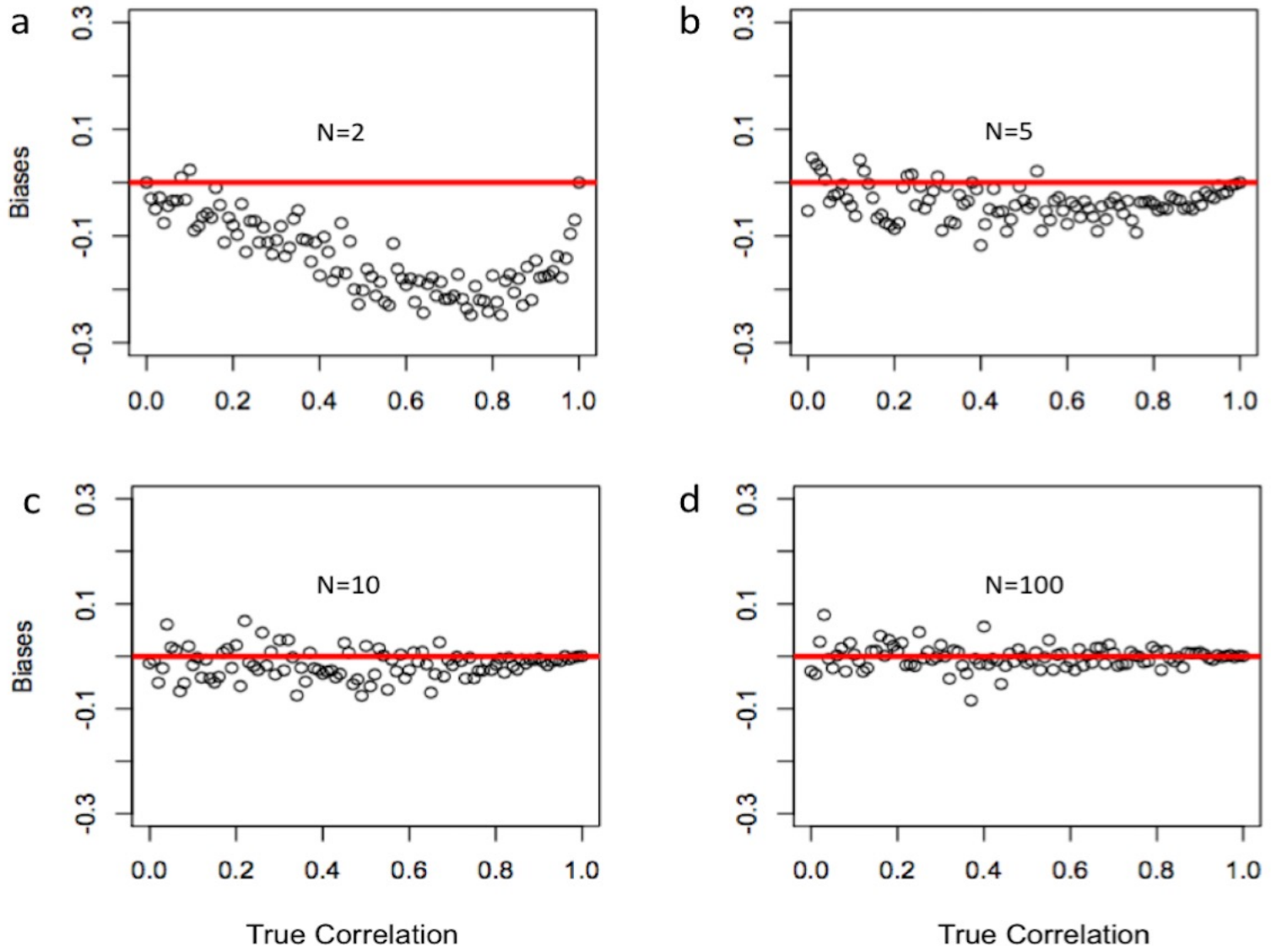
Hold bias relates to number of fold



Problem of instant accuracy

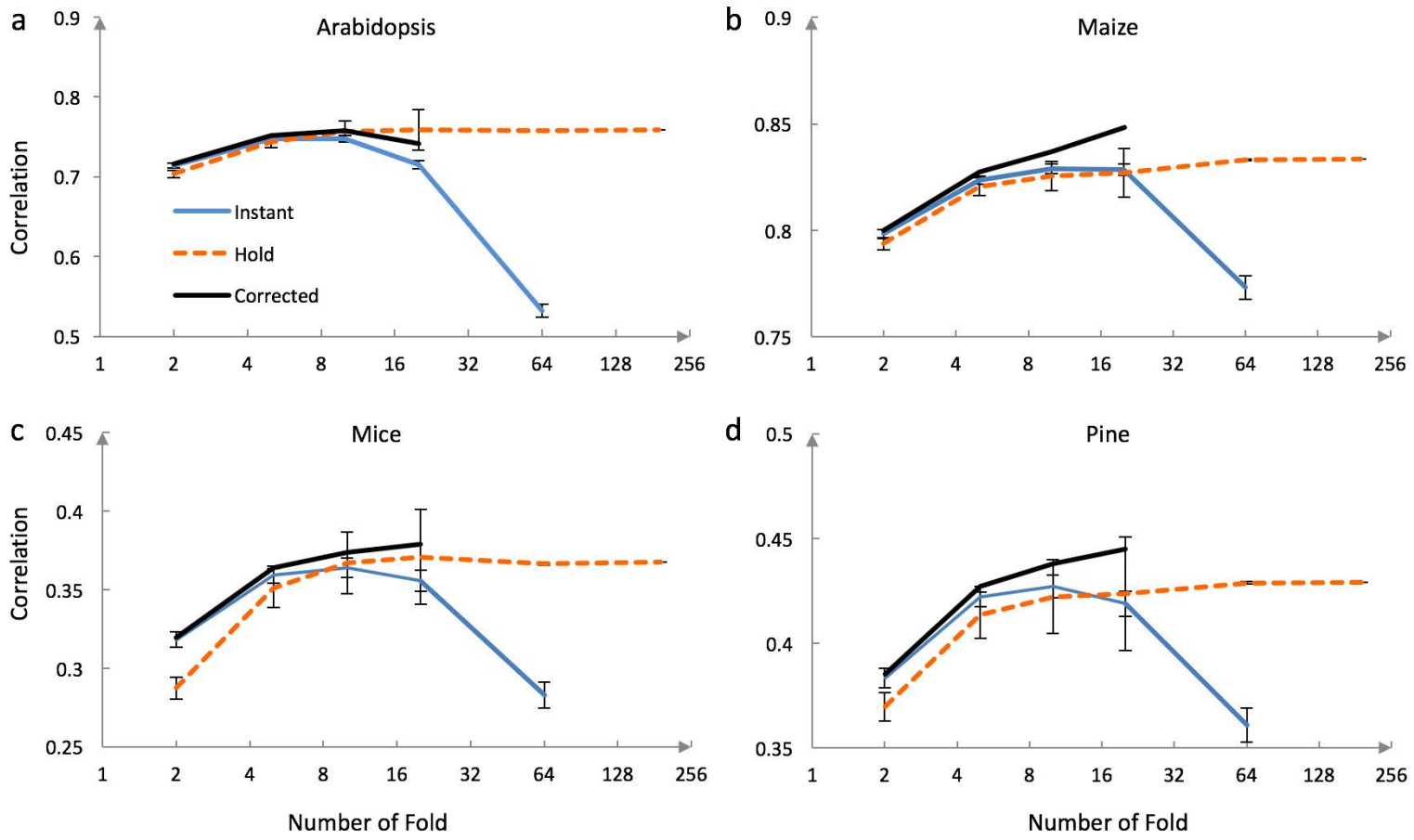


Small sample causes bias



Correction of instant accuracy

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



Summary on instant and hold accuracy

- Both are are downward biased
- Instance accuracy is correctable
- Bias correction of hold accuracy: not found yet
- True accuracy might be under estimated, especially using hold accuracy on leaving one out
- When true accuracy is low, negative hold accuracy could appear at unpresented frequency and magnitude

Outline

- MAS to GS
- Prediction assessment
- Muddy water
- Hidden overfitting
- GWAS+GS

Crop Science

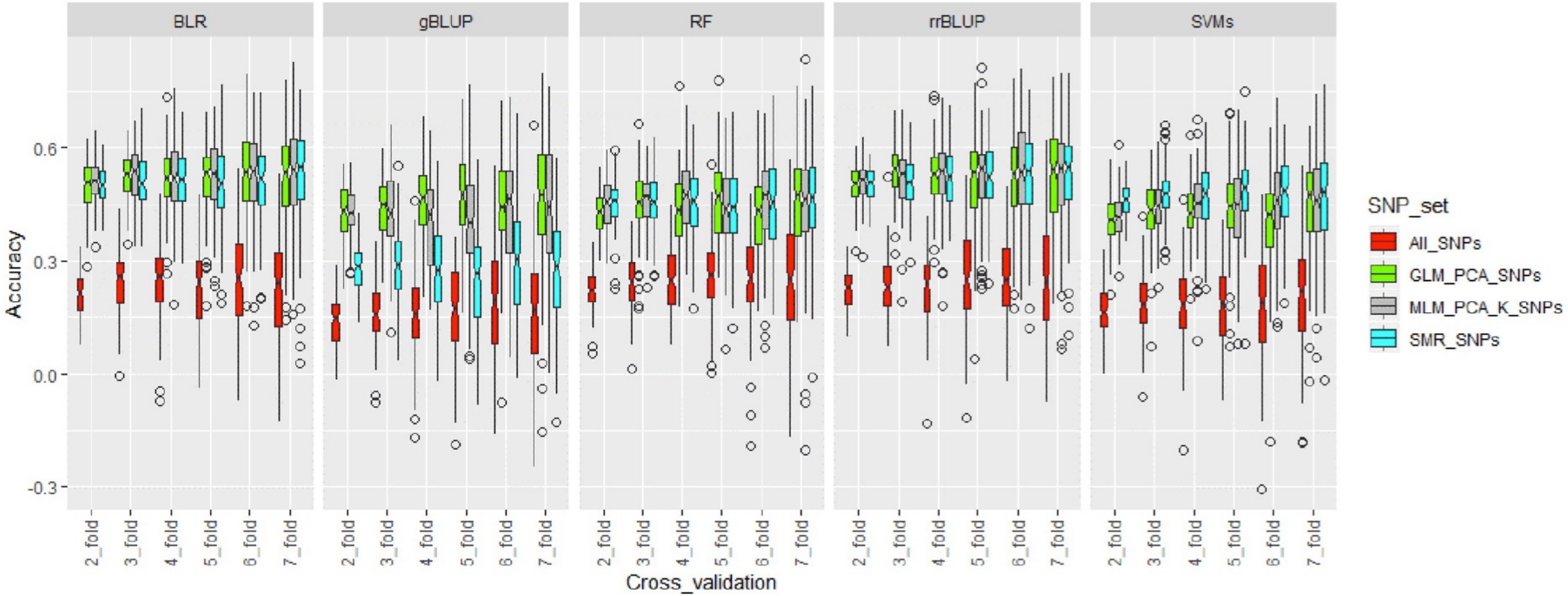
Crop Breeding & Genetics

Genomewide Selection when Major Genes Are Known

Rex Bernardo

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

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)

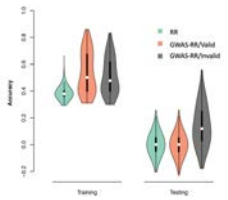
RRBLUP, best linear unbiased prediction using ridge-regression; GBLUP, genomic best linear unbiased prediction using VanRaden G matrix; WGBLUP, weighted GBLUP; 1-d-a and 1-d-r, simplex dominant models; 2-d-a and 2-d-r, duplex dominant models; d-gen, diploidized general; d-add, diploidized additive.

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).

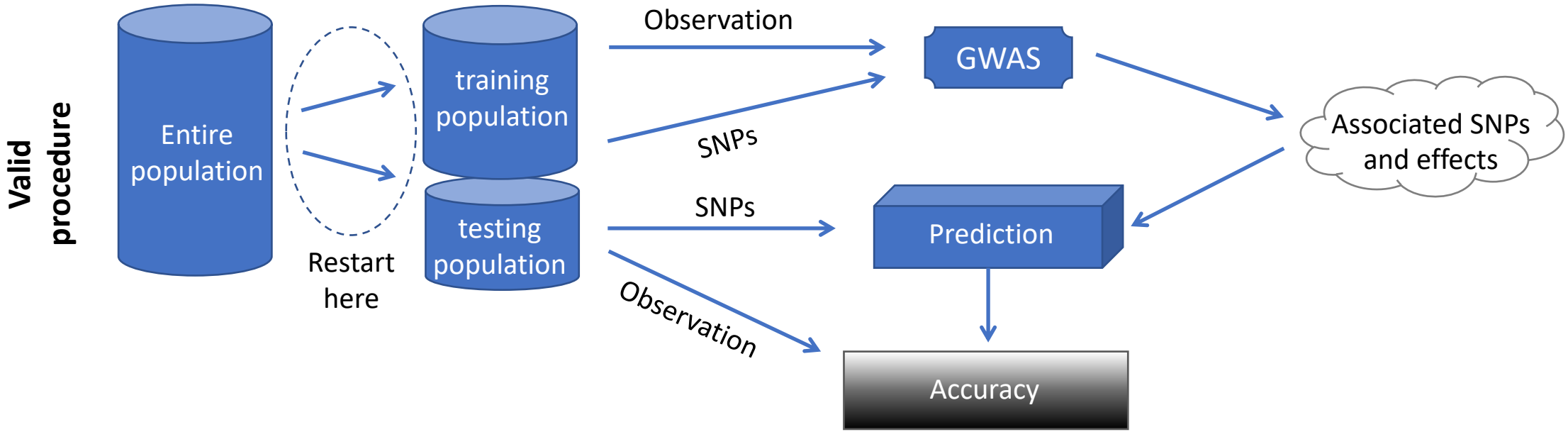
Higher prediction accuracies in all 13 agronomic traits of potato were obtained using the WGBLUP model with SNP $-\log_{10}$ *p*-values derived from the additive GWASpoly model. Traits of glucose, tuber length, or tuber shape showed accuracies higher than 0.9. It is important to point out that traits 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).

Outline

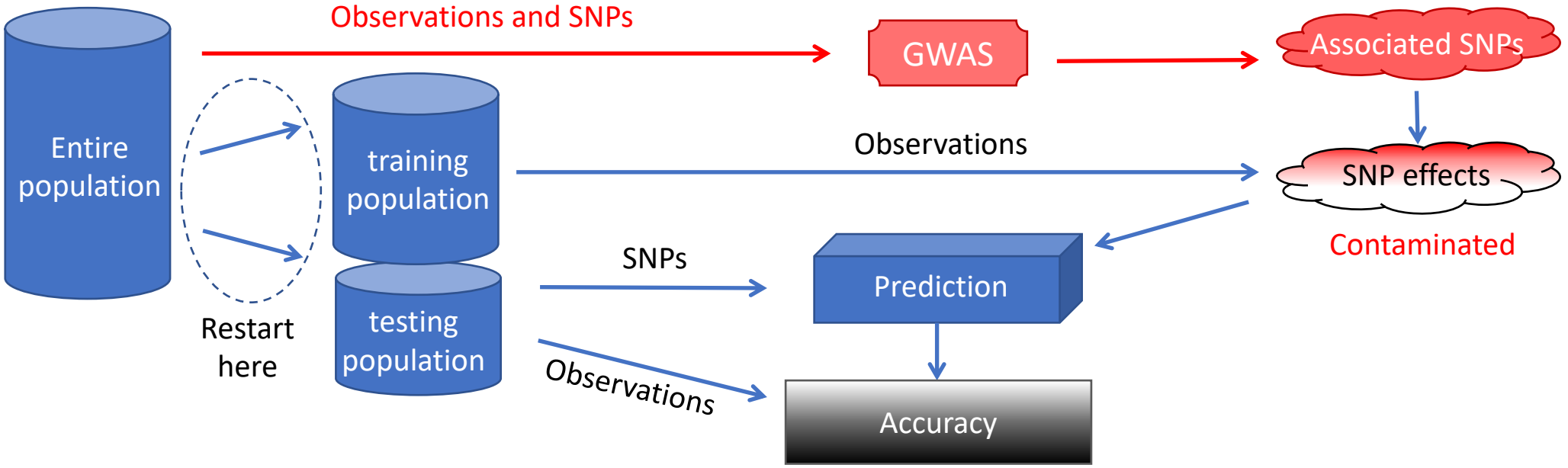
- MAS to GS
- Prediction assessment
- Muddy water
- Hidden overfitting
- GWAS+GS

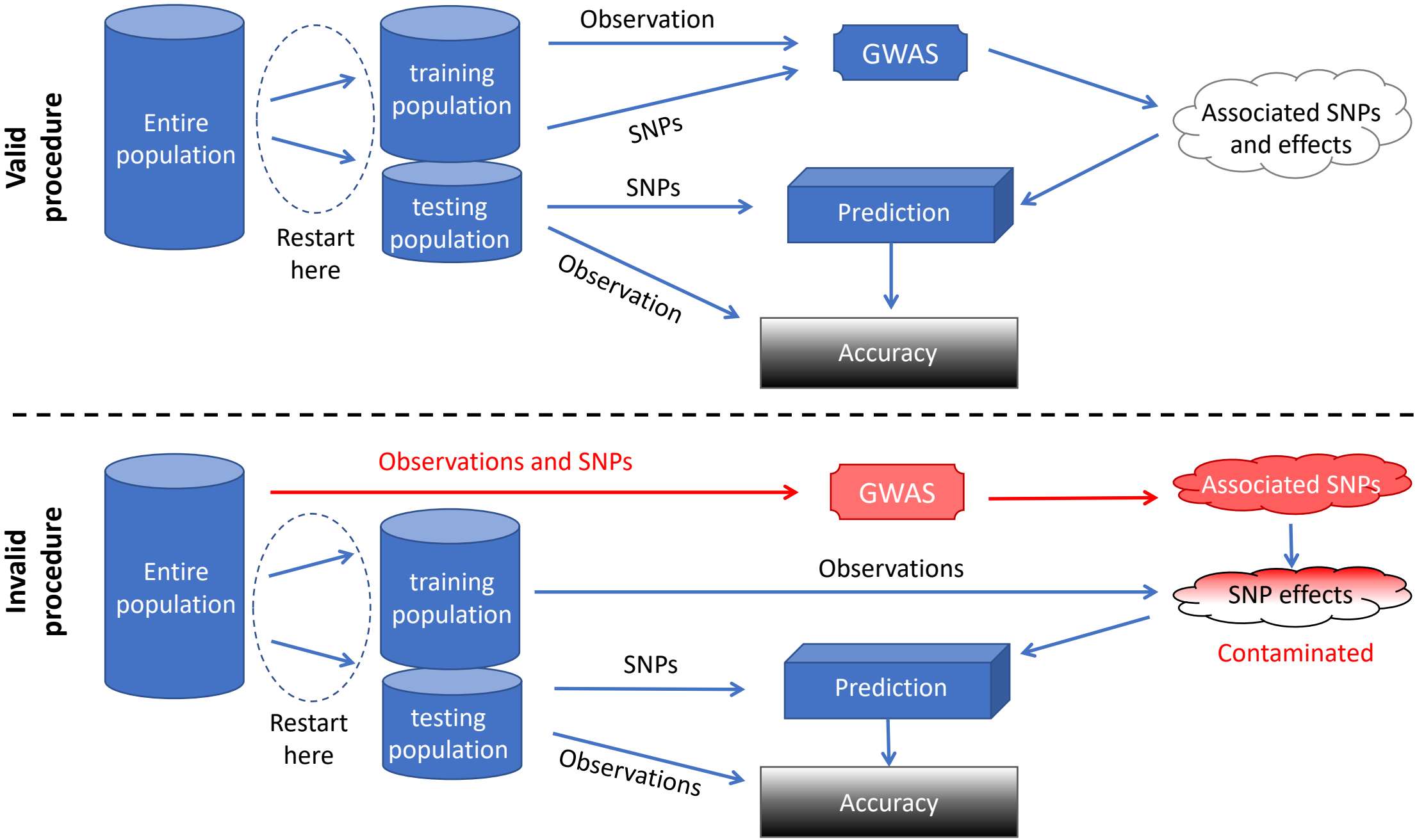


Valid procedure

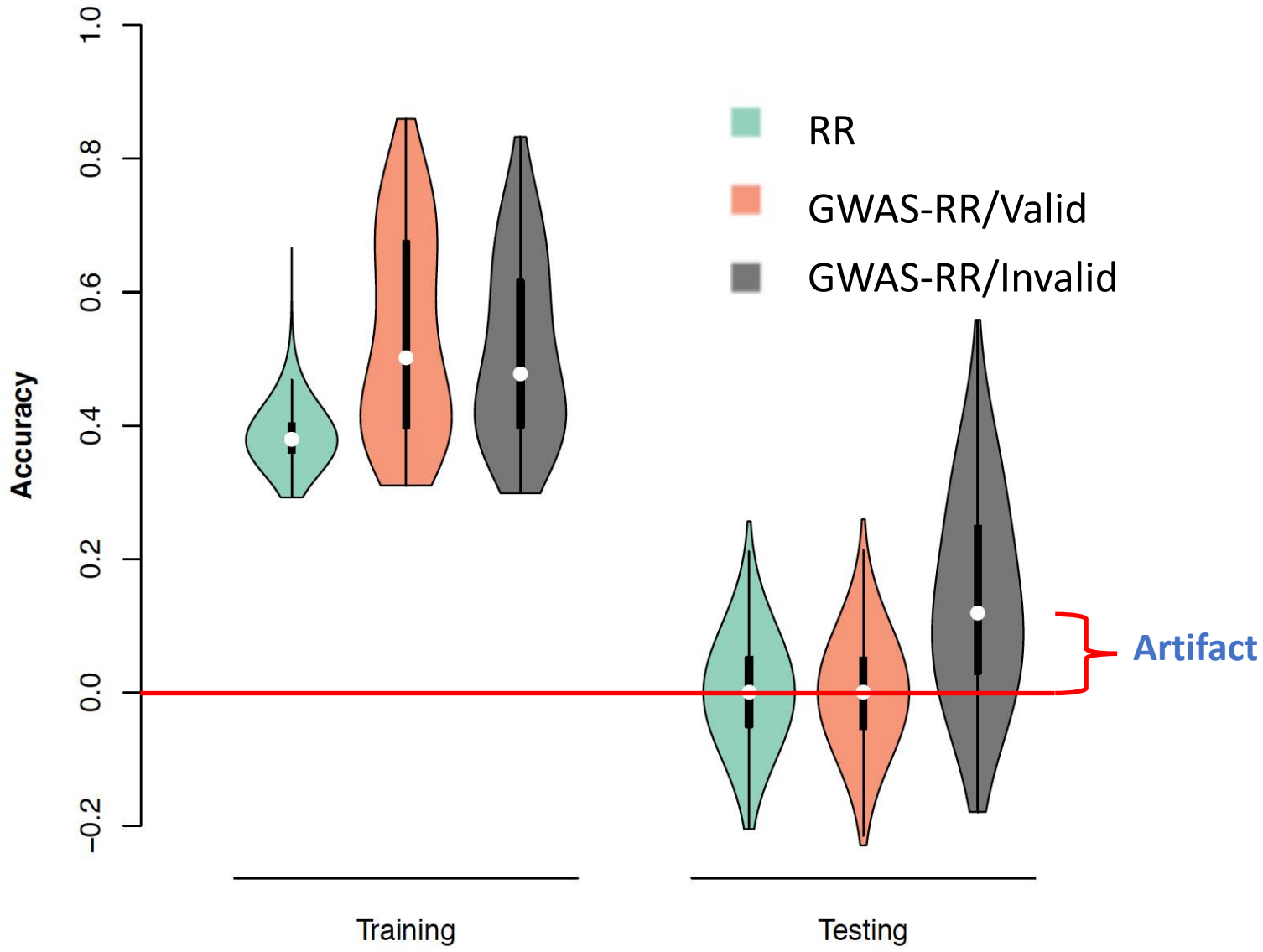


Invalid procedure




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.

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

Outline

- MAS to GS
- Prediction assessment
- Muddy water
- Hidden overfitting
- GWAS+GS


EMMAx

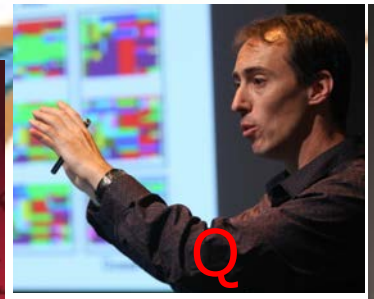

EMMA


PC+K


PC


CMLM


Q+K


Q


GWAS


P3D


GCTA


SELECT


MLMM

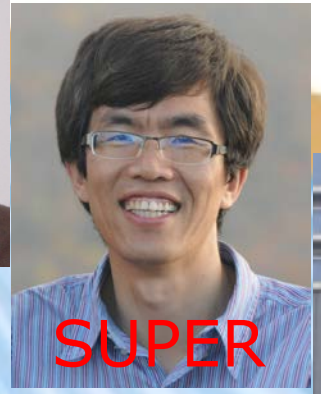
WAS Stream


FST-LMM


GEMMA


GenAbel


ECMLM


SUPER


FarmCPU


BLINK

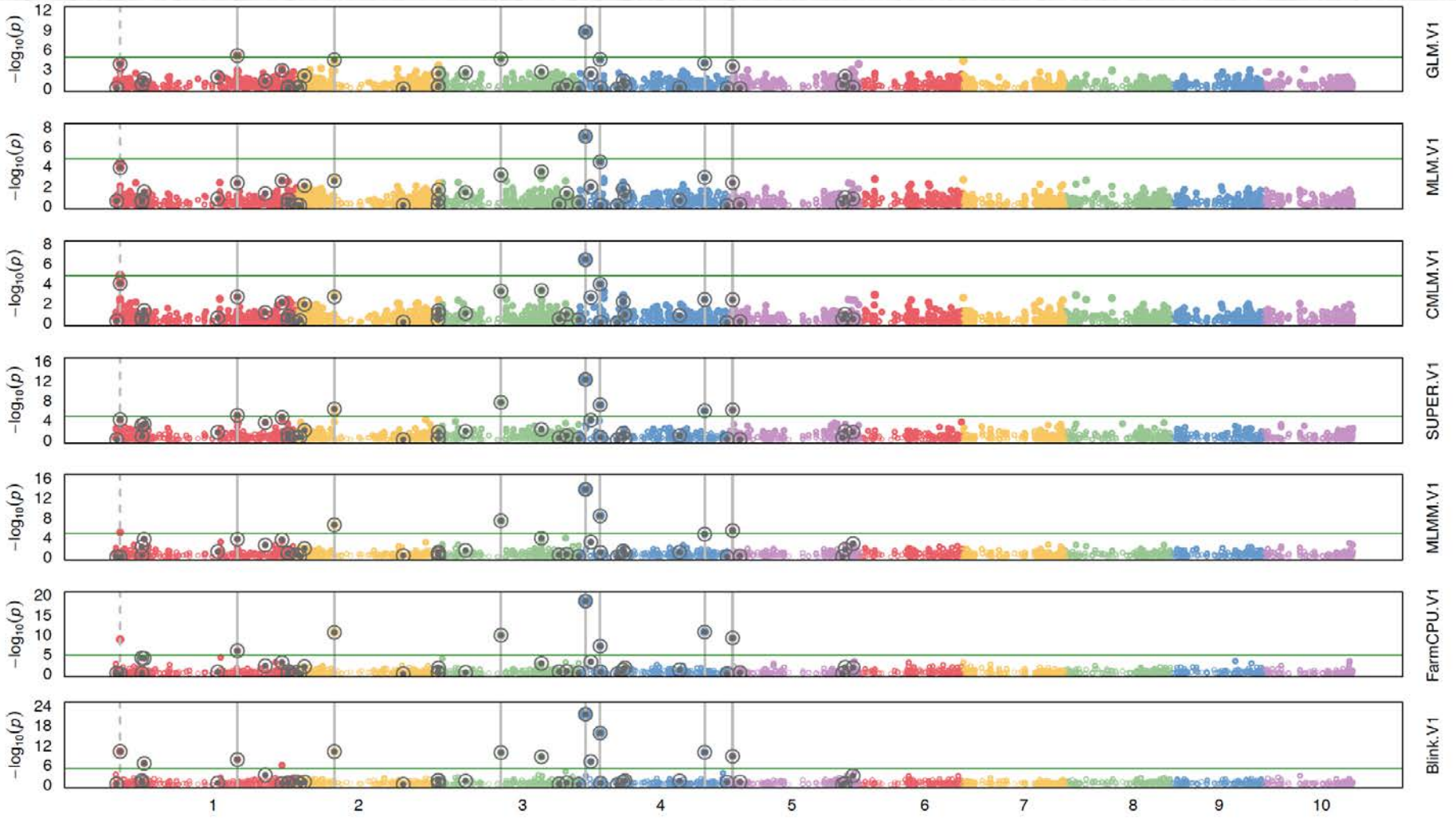
Zhiwu Zhang Laboratory
for Statistical Genomics

Home People Publication Research Teaching Software Outreach Jobs

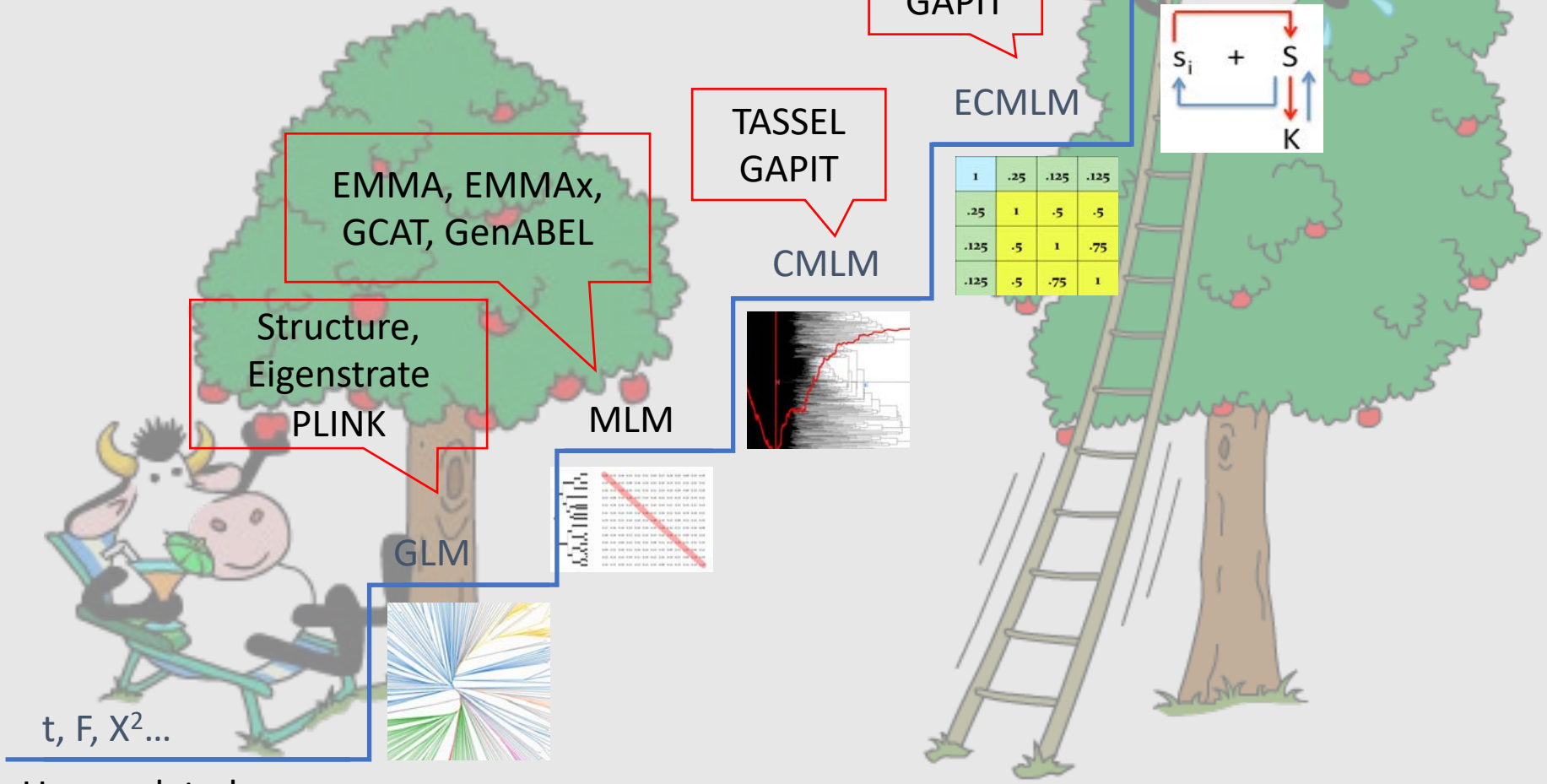
```
source("http://zzlab.net/GAPIT/gapit_functions.txt") #Import demo data
myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#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"))
```

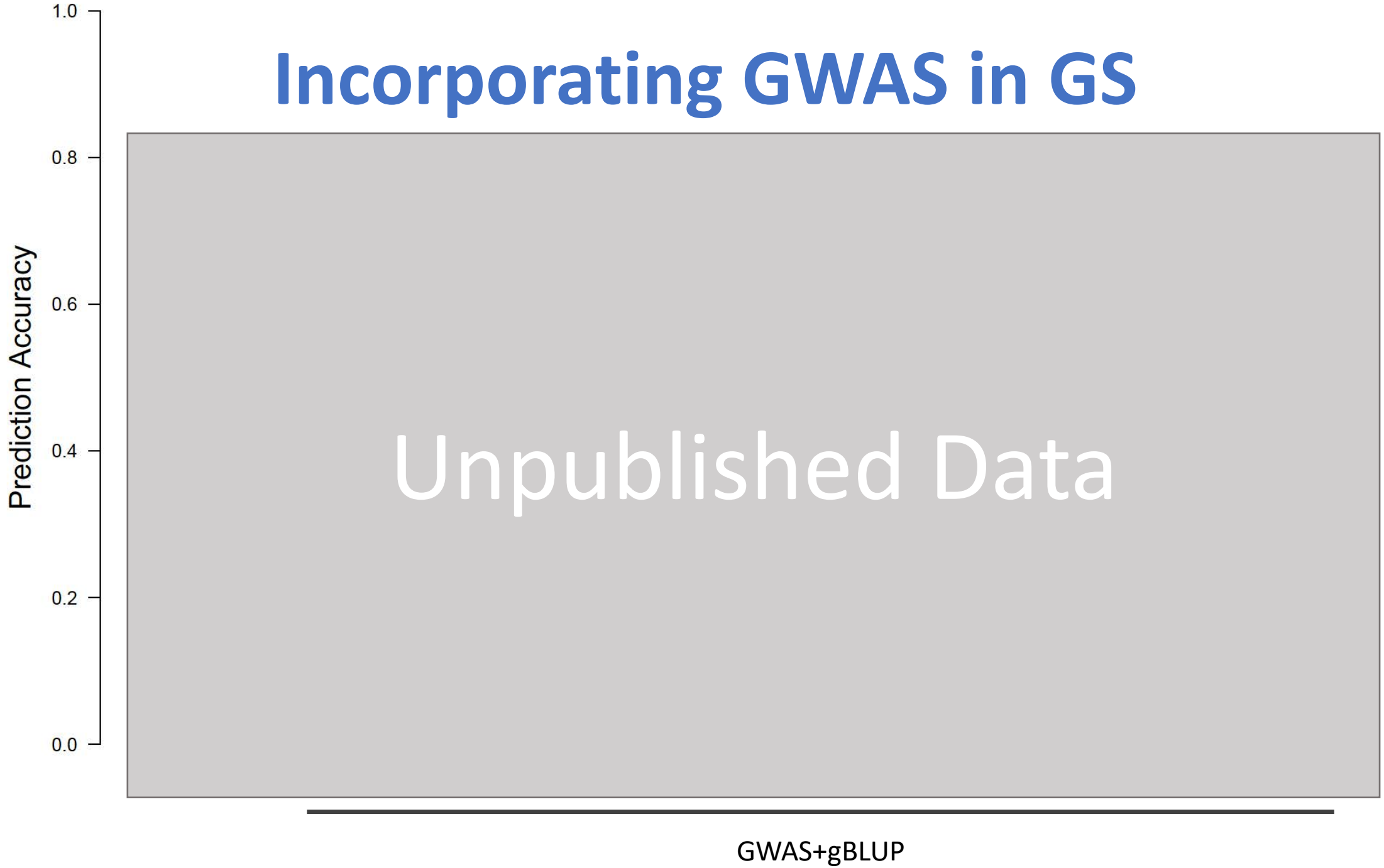


GAPIT


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

Incorporating GWAS in GS



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

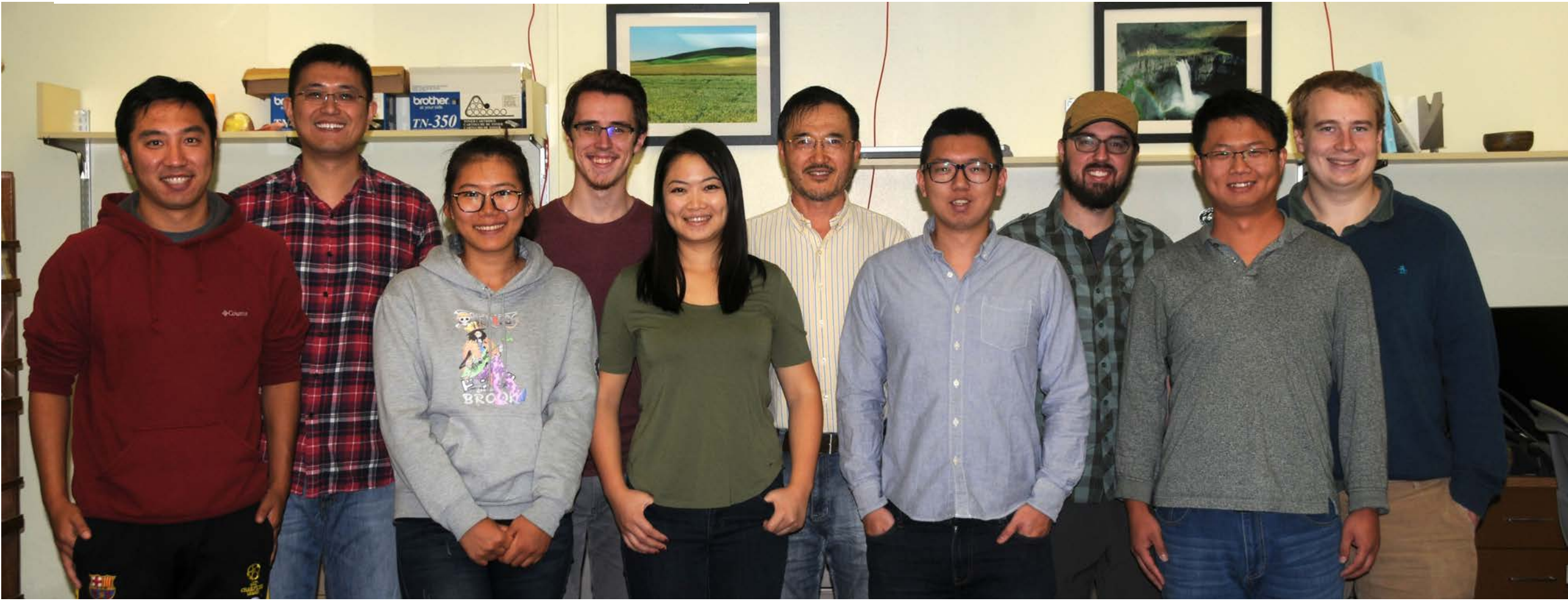
NSF

M.J. MURDOCK CHARITABLE TRUST

WSU

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!