



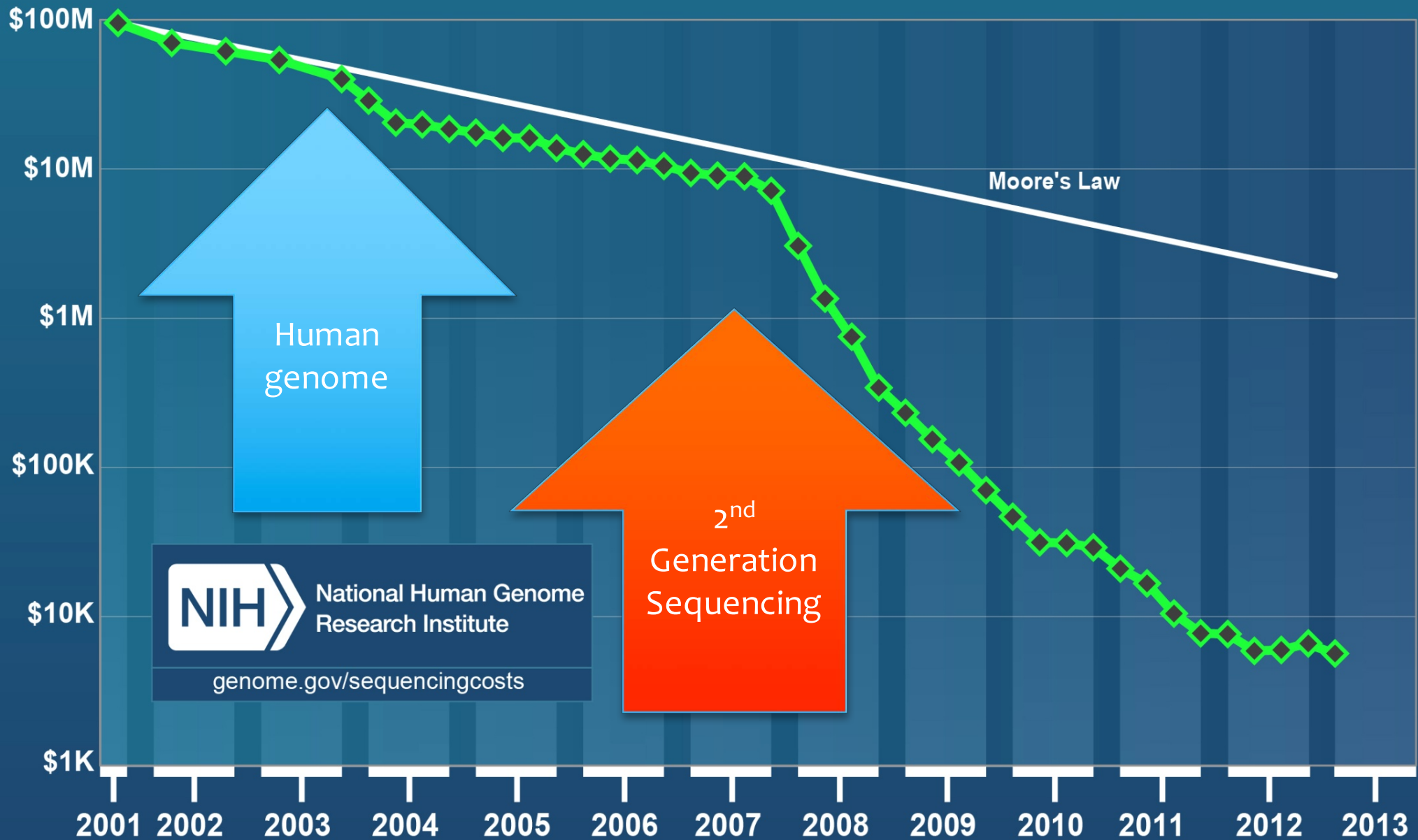
Saving the Babies from the Bathwater in Genome-Wide Association Studies

Zhiwu Zhang

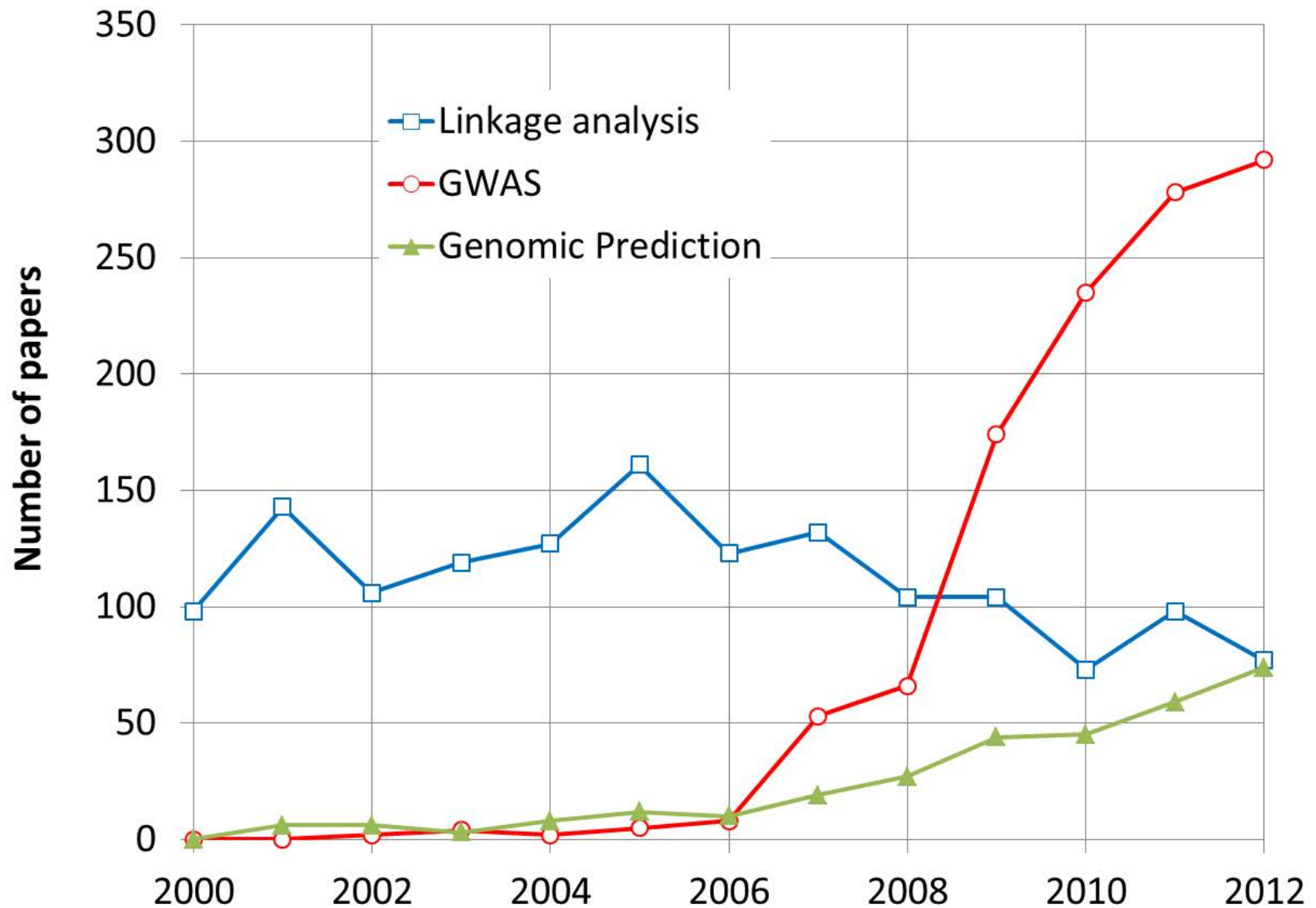
Washington State University



Biomedical Innovations out-paced Computer Innovations



More Research on GWAS and GS



By May 31, 2013

Genome-wide association studies of 14 agronomic traits in rice landraces

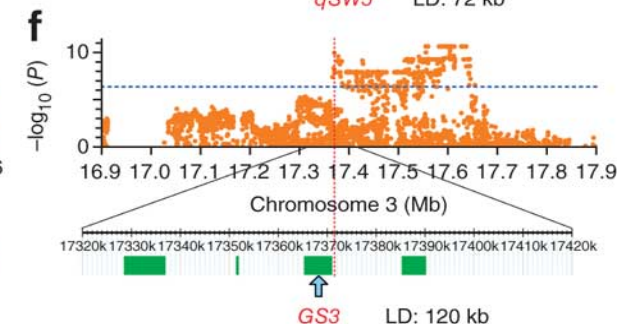
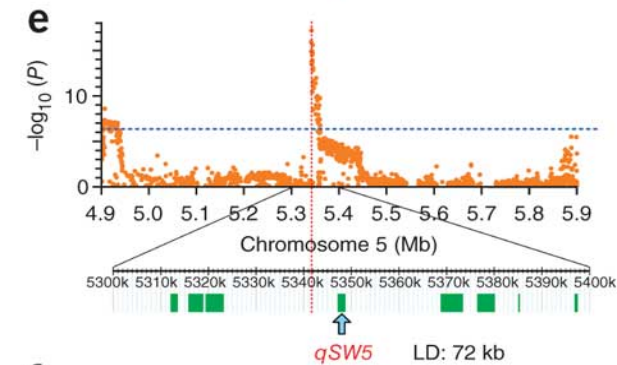
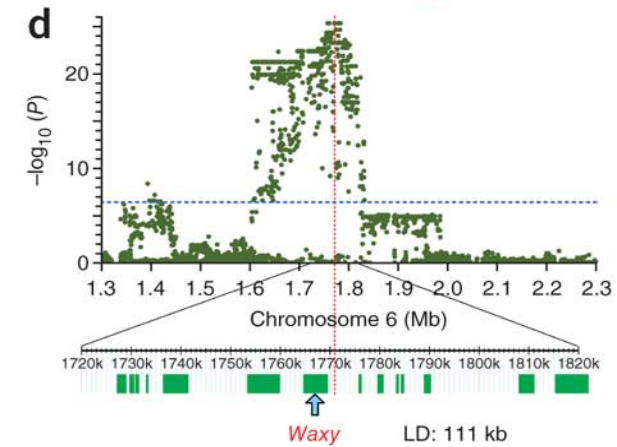
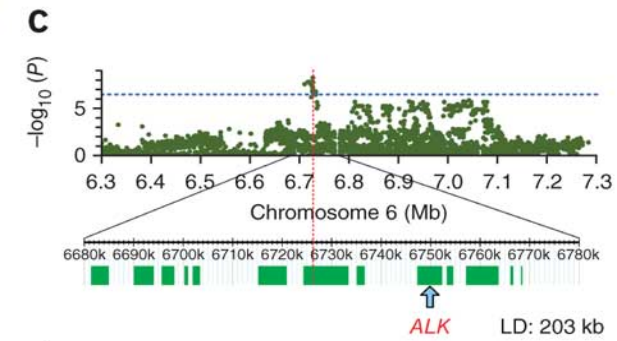
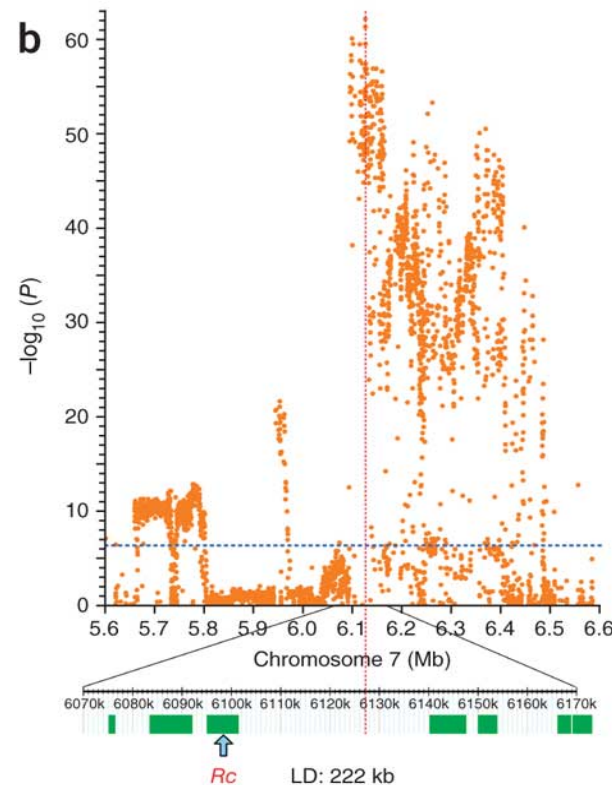
Xuehui Huang, Xinghua Wei, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, Chuanrang Zhu, Tingting Lu, Zhiwu Zhang, Meng Li, Danlin Fan, Yunli Guo, Ahong Wang, Lu Wang, Liuwei Deng, Wenjun Li, Yiqi Lu, Qijun Weng, Kunyan Liu, Tao Huang, Taoying Zhou, Yufeng Jing, Wei Li, Zhang Lin, Edward S Buckler, Qian Qian, Qi-Fa Zhang, Jiayang Li & Bin Han [Show fewer authors](#)

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

Nature Genetics **42**, 961–967 (2010) | doi:10.1038/ng.695

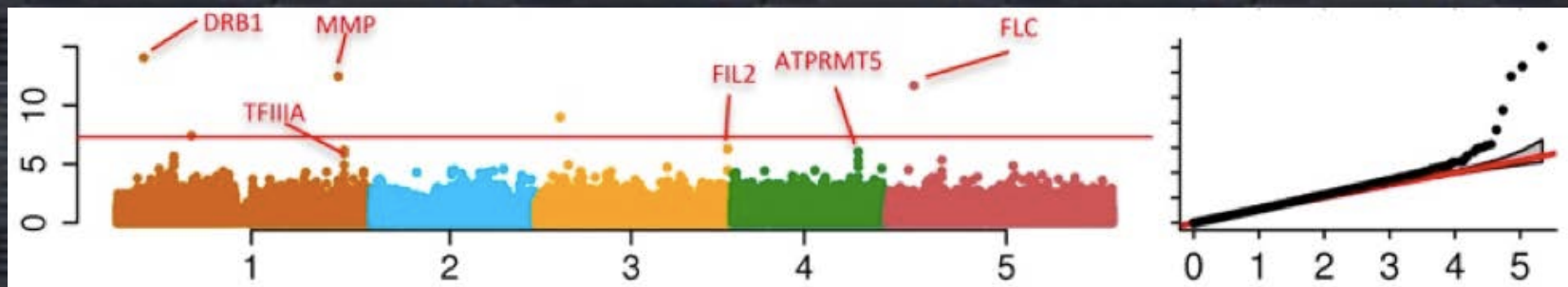
Received 10 May 2010 | Accepted 27 September 2010 | Published online 24 October 2010

- * As fast as one season
- * 50~300 kb resolution



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





EMMAX



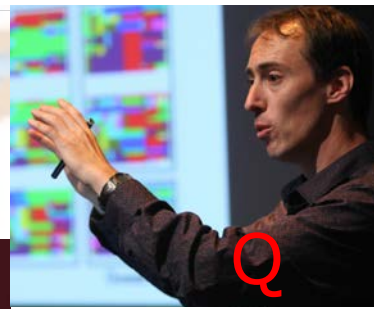
EMMA



PC+K



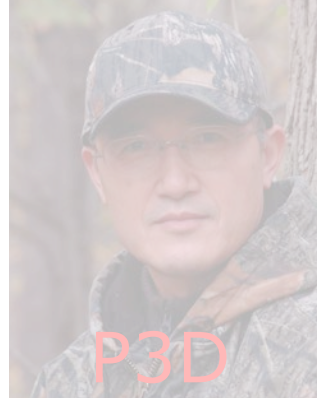
PC



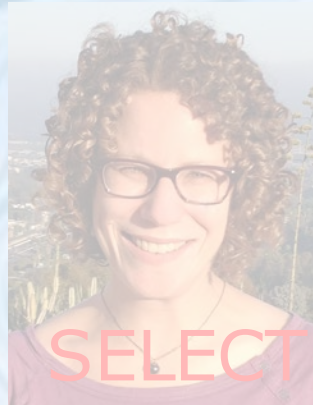
Q



Q+K



P3D



SELECT



MLMM



CMLM



GCTA



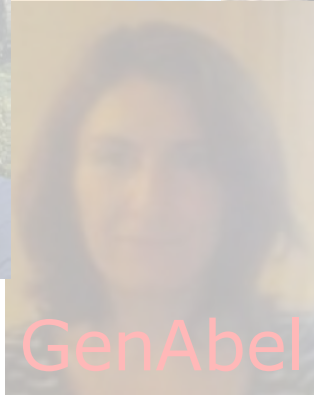
ECMLM



FST-LMM



GEMMA



GenAbel

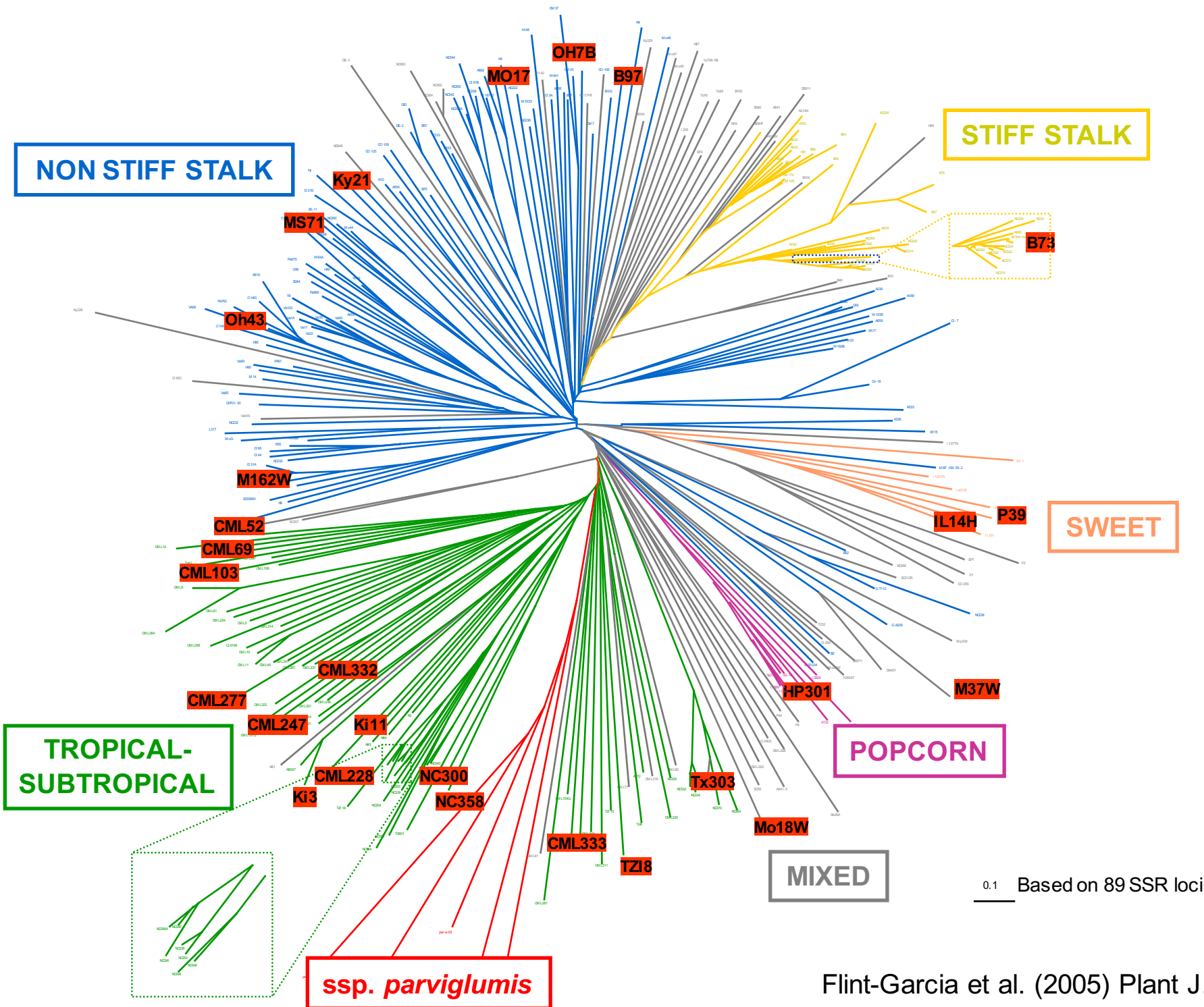


FarmCPU

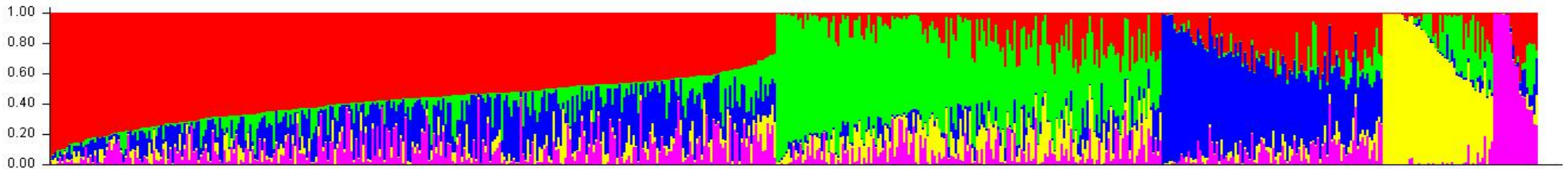


BLINK

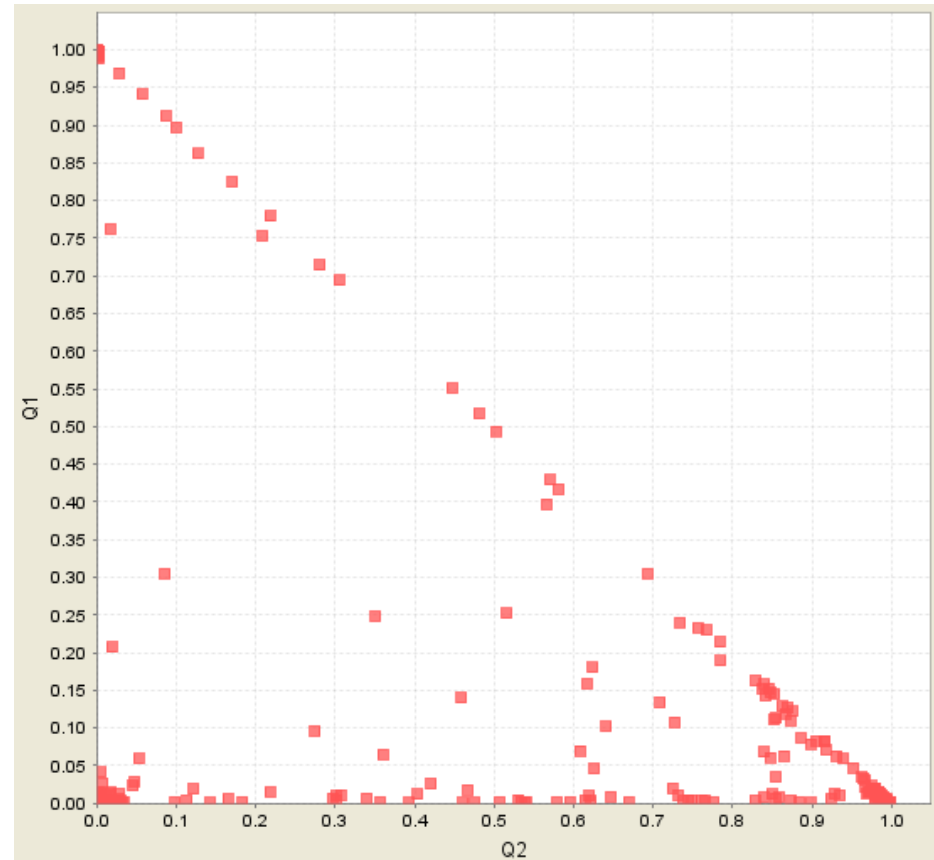
Stream



Population structure



Lines	Q1	Q2	Q3
33-16	0.014	0.972	0.014
38-11	0.003	0.993	0.004
4226	0.071	0.917	0.012
4722	0.035	0.854	0.111
A188	0.013	0.982	0.005
A214N	0.762	0.017	0.221
A239	0.035	0.963	0.002
A272	0.019	0.122	0.859
A441-5	0.005	0.531	0.464
A554	0.019	0.979	0.002
A556	0.004	0.994	0.002
A6	0.003	0.03	0.967
A619	0.009	0.99	0.001



GLM for GWAS

Phenotype on individuals

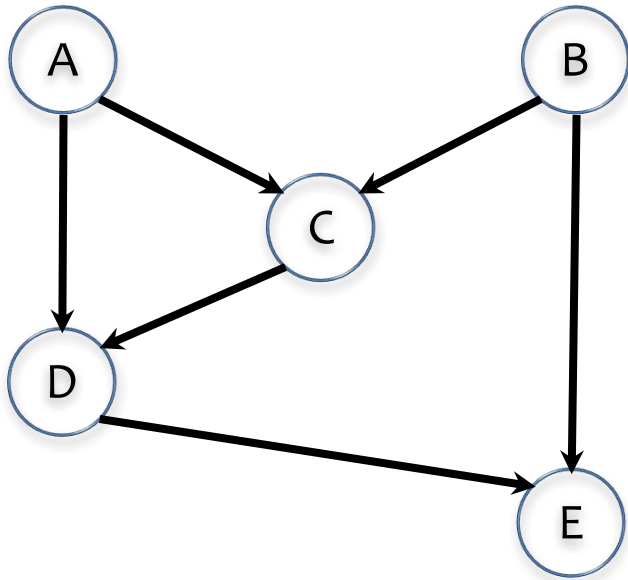
Population structure

$$Y = \text{SNP} + Q \text{ (or PCs)} + e$$

(fixed effect) (fixed effect)

General Linear Model
(GLM)

Additive numerator relationship



Individual	Father	Mother
A		
B		
C	A	B
D	A	C
E	D	B

	A	B	C	D	E
A	1	0	0.5	0.75	0.375
B	0	1	0.5	0.25	0.625
C	0.5	0.5	1	0.75	0.625
D	0.75	0.25	0.75	1.25	0.75
E	0.375	0.725	0.625	0.75	1.125

Diagonals = 1 + F

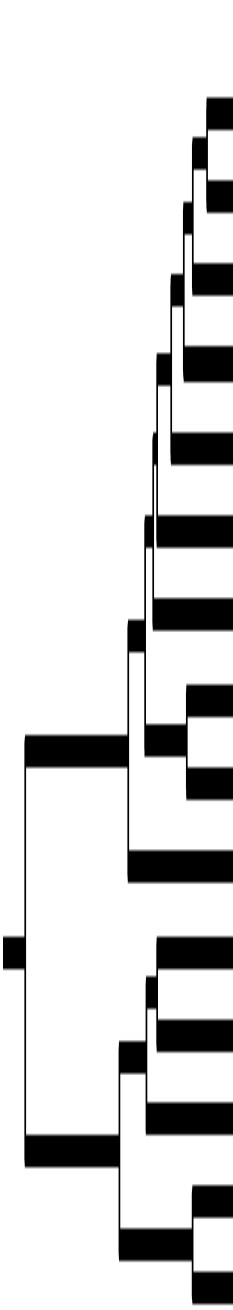
Marker based kinship

- * Proportion of shared allele
- * Average across markers

Marker	1	2	3	4	5	Average
Individual 1	AA	AA	AA	BB	AB	
Individual 2	AA	AB	BB	BB	AB	
Similarity	1	0.5	0	1	0.5	0.6

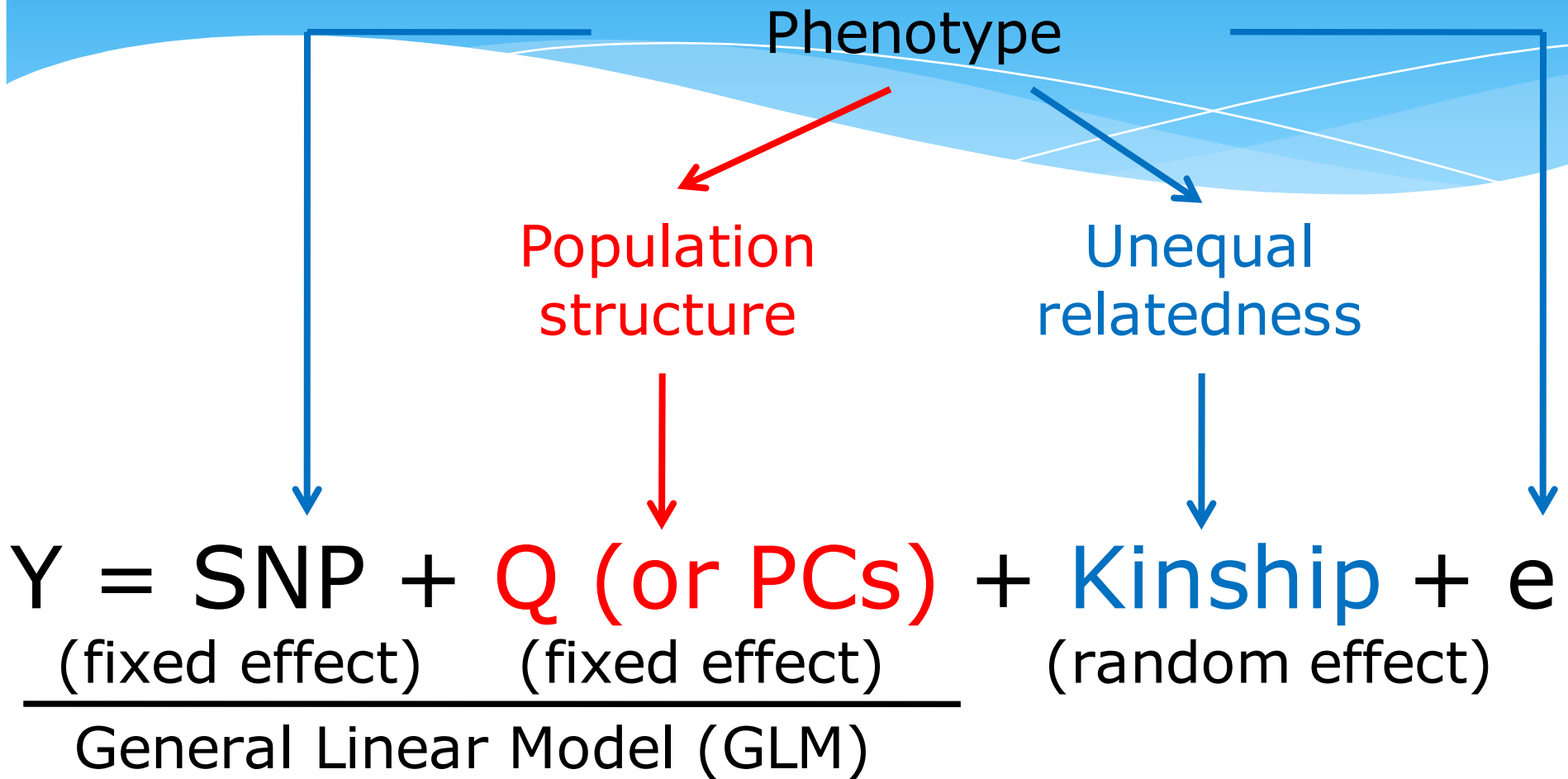
Maximum similarity: 1

Kinship matrix (re-scaled)



taxa	B73Htrhm	B73	NC326	A680	R229	B109	A679	NC328	NC310	N192	NC368	NC306	NC294	NC324	NC314
B73Htrh	1.00	1.60	1.49	1.49	1.42	1.34	1.32	1.37	1.33	1.26	1.01	1.11	1.09	0.79	0.86
B73	1.60	1.00	1.50	1.51	1.43	1.34	1.34	1.39	1.34	1.27	1.01	1.11	1.08	0.79	0.87
NC326	1.49	1.50	1.00	1.41	1.34	1.26	1.22	1.31	1.34	1.20	1.02	1.14	1.04	0.86	0.91
A680	1.49	1.51	1.41	1.00	1.34	1.26	1.36	1.32	1.27	1.23	0.97	1.07	1.02	0.76	0.83
R229	1.42	1.43	1.34	1.34	1.00	1.21	1.18	1.24	1.18	1.16	0.90	1.00	1.03	0.78	0.86
B109	1.34	1.34	1.26	1.26	1.21	1.00	1.11	1.15	1.12	1.12	0.91	0.96	1.01	0.76	0.83
A679	1.32	1.34	1.22	1.36	1.18	1.11	1.00	1.21	1.10	1.07	0.94	1.05	0.98	0.74	0.76
NC328	1.37	1.39	1.31	1.32	1.24	1.15	1.21	1.00	1.42	1.22	1.02	1.19	1.09	0.88	0.86
NC310	1.33	1.34	1.34	1.27	1.18	1.12	1.10	1.42	1.00	1.18	1.02	1.14	1.05	0.96	0.95
N192	1.26	1.27	1.20	1.23	1.16	1.12	1.07	1.22	1.18	1.00	0.85	0.91	0.88	0.68	0.75
NC368	1.01	1.01	1.02	0.97	0.90	0.91	0.94	1.02	1.02	0.85	1.00	1.20	1.08	0.97	0.91
NC306	1.11	1.11	1.14	1.07	1.00	0.96	1.05	1.19	1.14	0.91	1.20	1.00	1.07	1.09	1.04
NC294	1.09	1.08	1.04	1.02	1.03	1.01	0.98	1.09	1.05	0.88	1.08	1.07	1.00	1.02	0.99
NC324	0.79	0.79	0.86	0.76	0.78	0.76	0.74	0.88	0.96	0.68	0.97	1.09	1.02	1.00	1.48
NC314	0.86	0.87	0.91	0.83	0.86	0.83	0.76	0.86	0.95	0.75	0.91	1.04	0.99	1.48	1.07

MLM for GWAS



Mixed Linear Model (MLM)

(Yu et al. 2005, Nature Genetics)

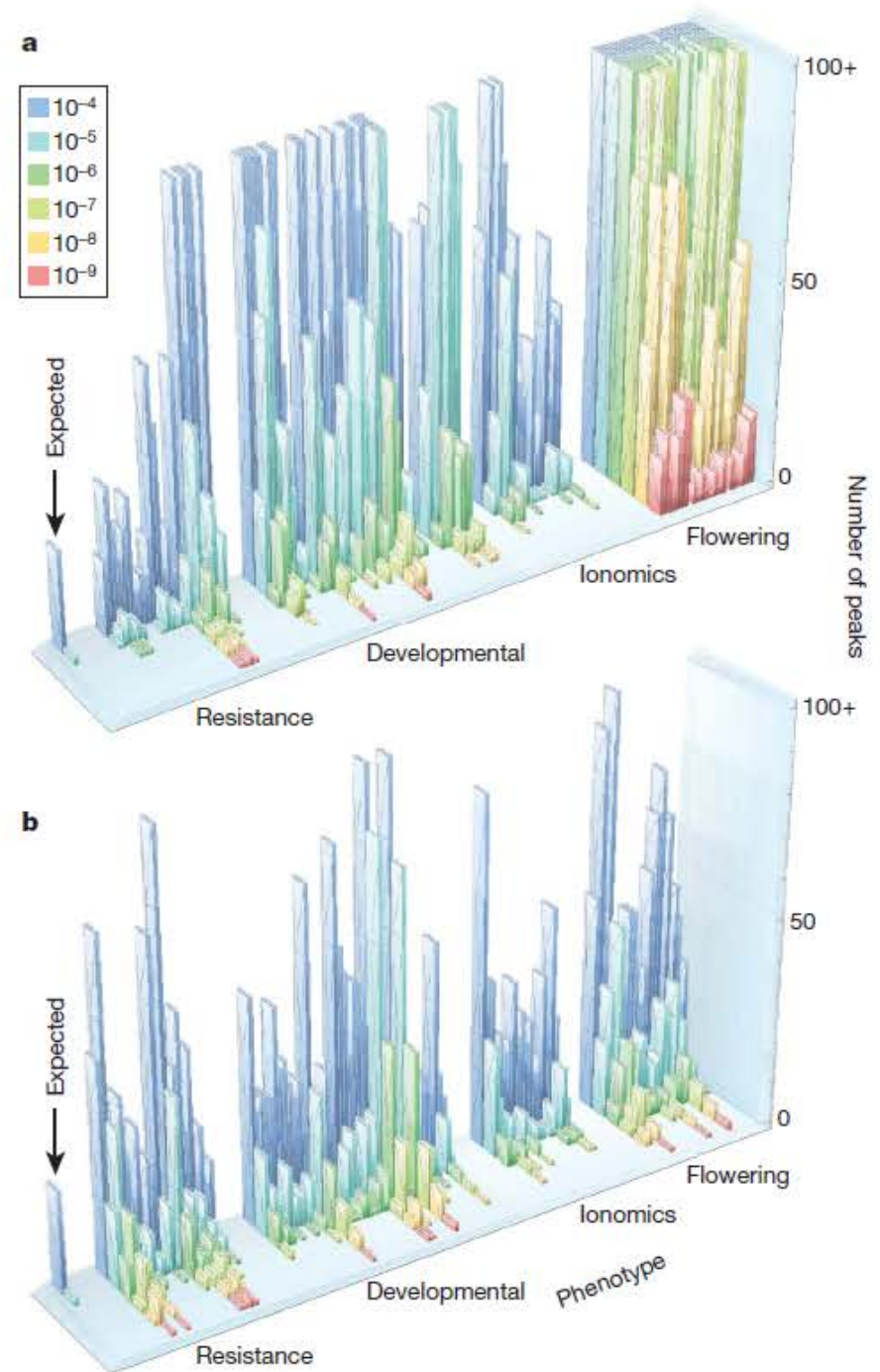
Atwell et al Nature 2010

a, No correction test

b, Correction with MLM



Magnus Nordborg

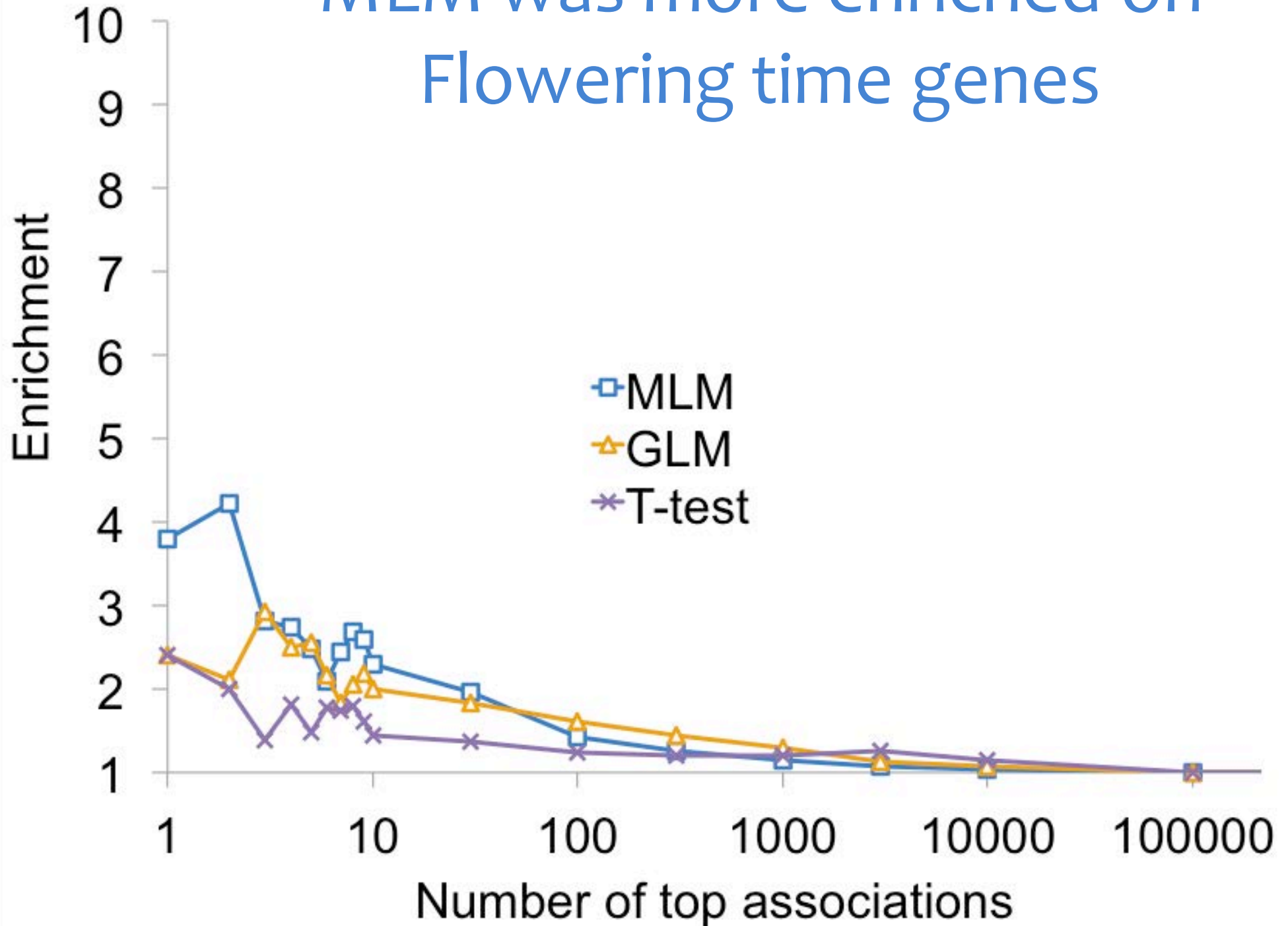


GWAS does work well for traits associated with structure

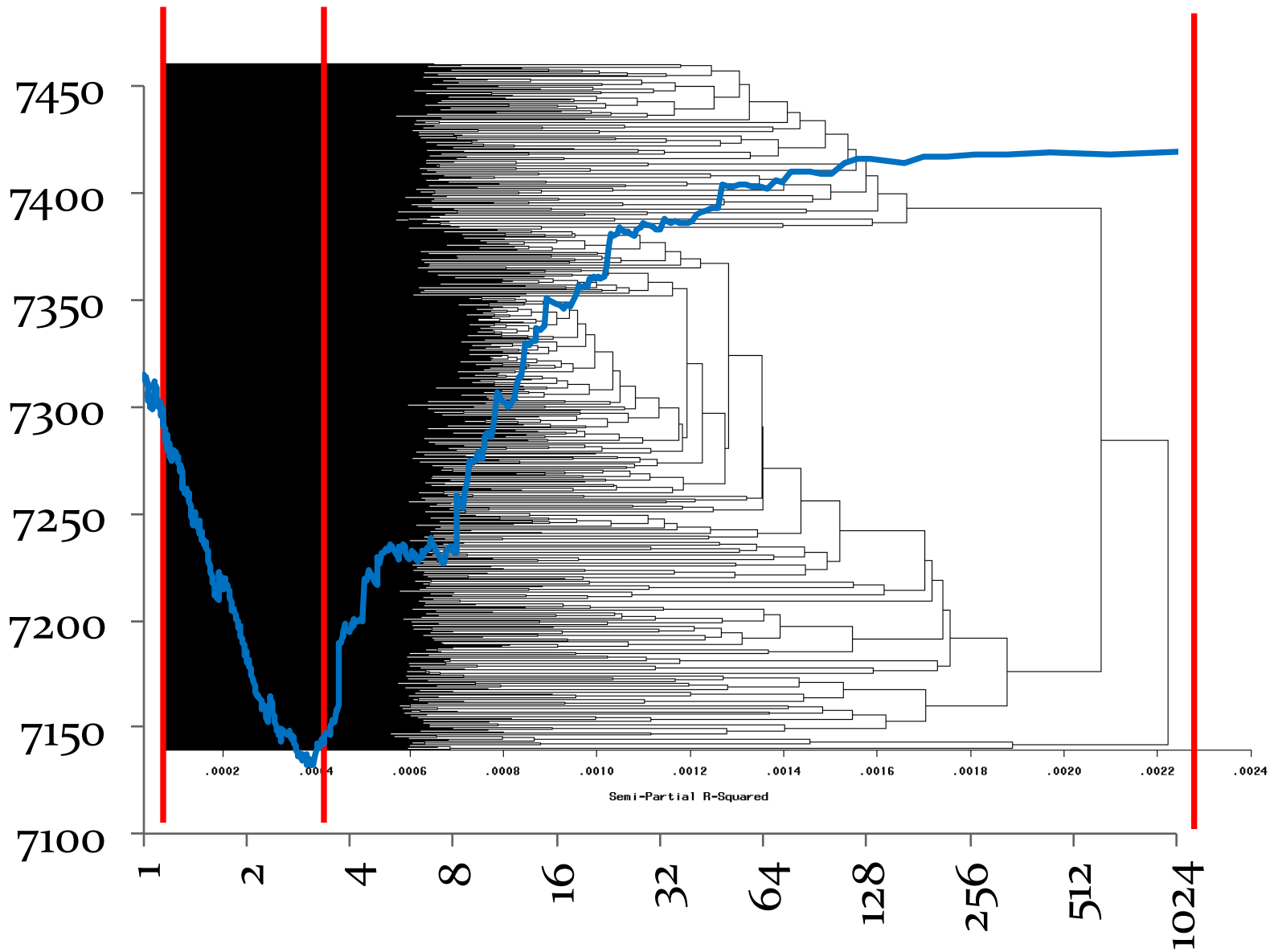
Queen + King



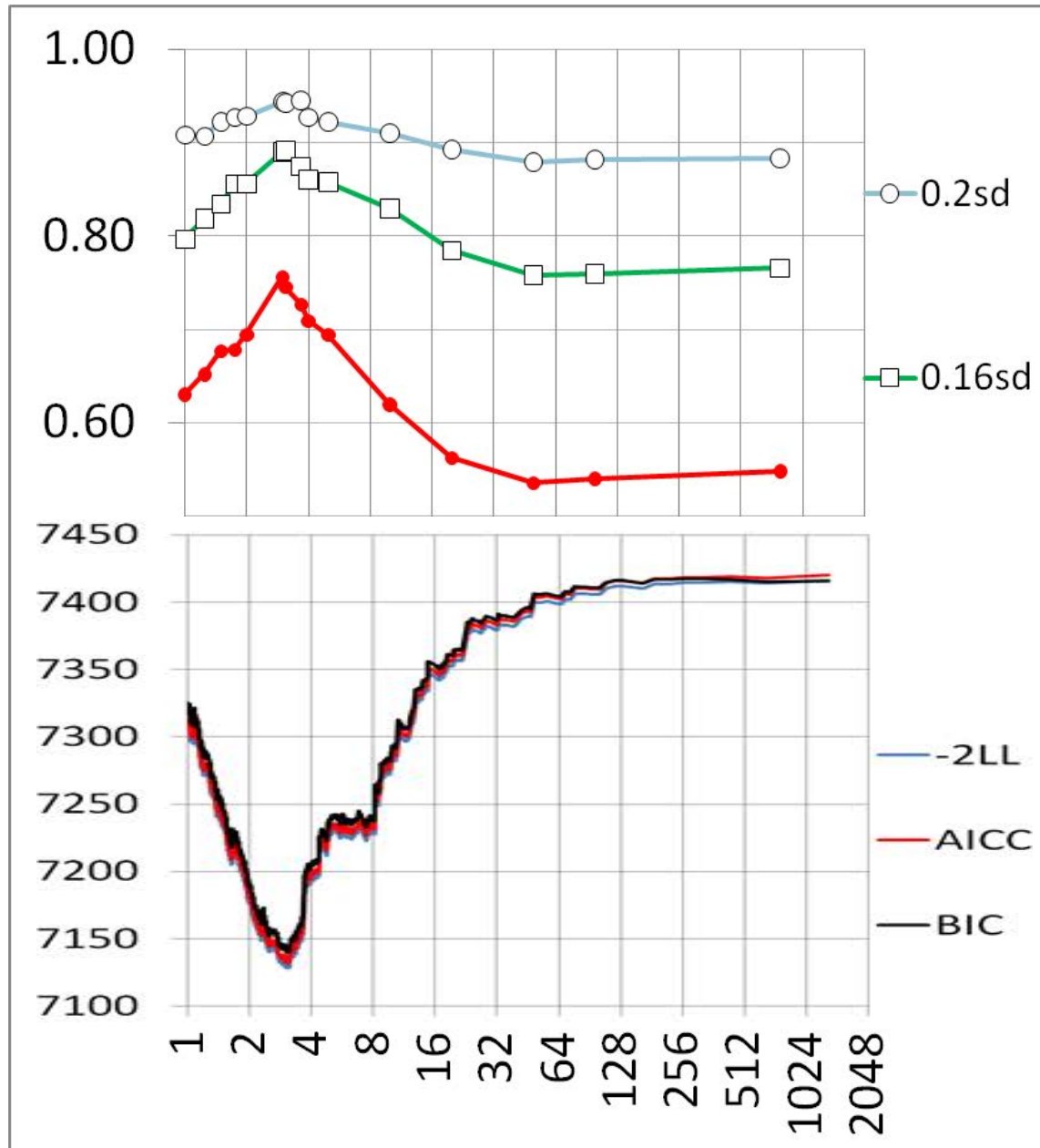
MLM was more enriched on Flowering time genes



Group by kinship



Compressed MLM improves power

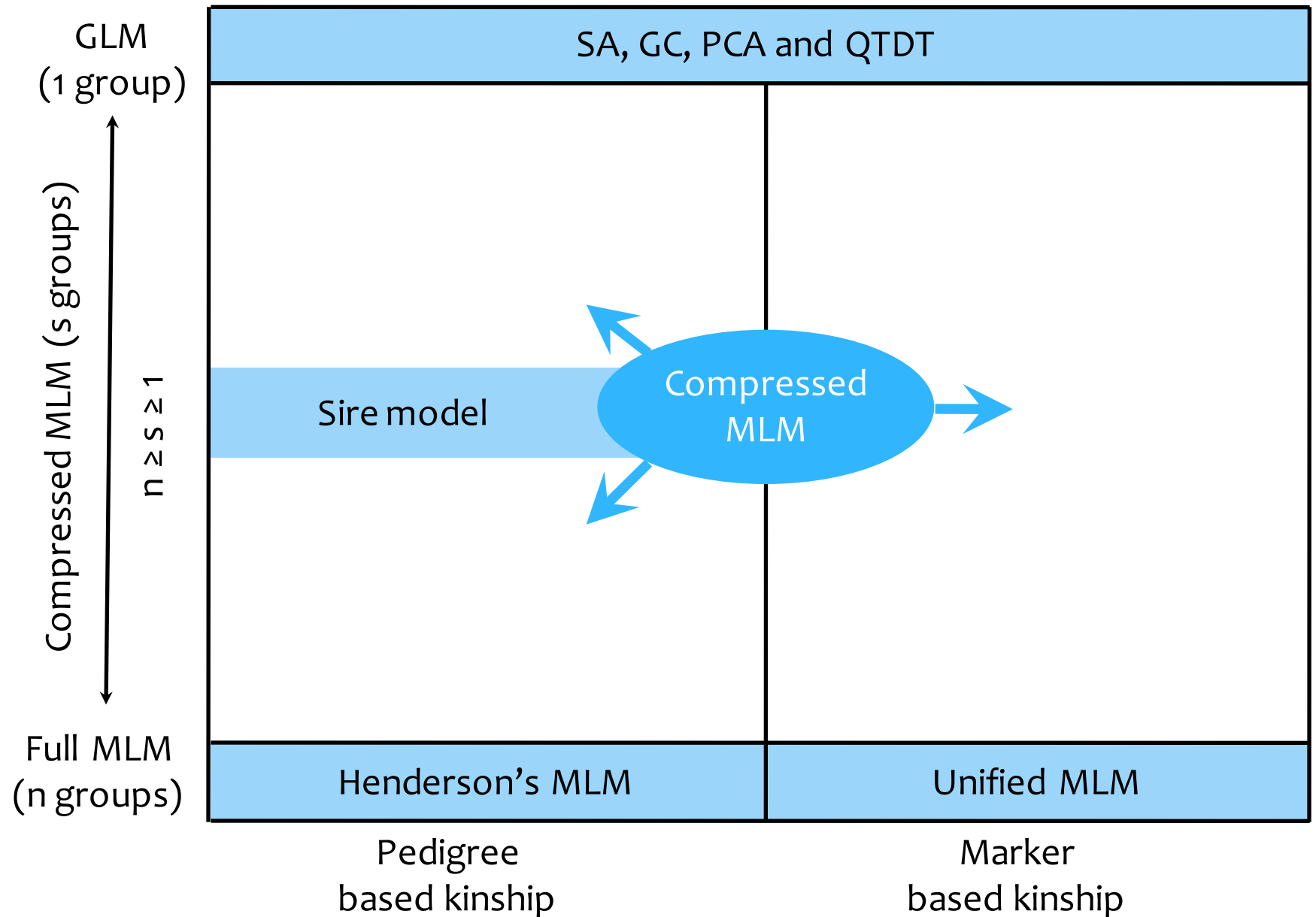


Zhang et al, Nature Genetics, 2010

Average number of individuals per group

Compressed MLM is more general

Zhang et al, Nature Genetics, 2010



Enriched Compressed MLM

Kinship: Among individuals -> among groups

1	.25	.125	.125
.25	1	.5	.5
.125	.5	1	.75
.125	.5	.75	1

Average

1	.167
.167	.72

Maximum

1	.25
.25	1

Minimum

Median

⋮

Statistical power improvement



Meng Li

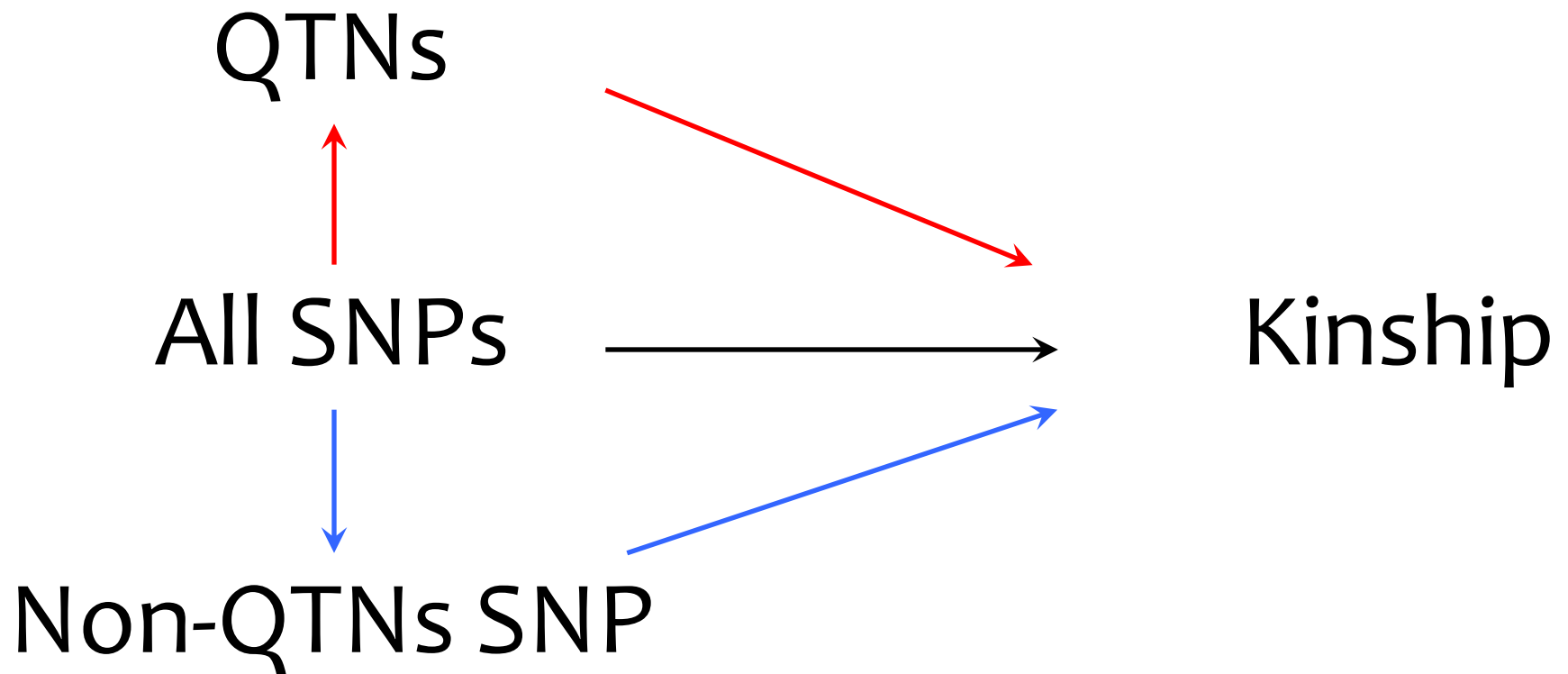
Method shift	Human	Dog	Maize	Arabidopsis
GLM to MLM	3.6%	13.8%	10.1%	29.6%
MLM to compression	4.0%	14.2%	7.6%	2.5%
Compression to group kinship	6.4%	13.3%	2.9%	2.6%

Li et al, BMC Biology, 2014

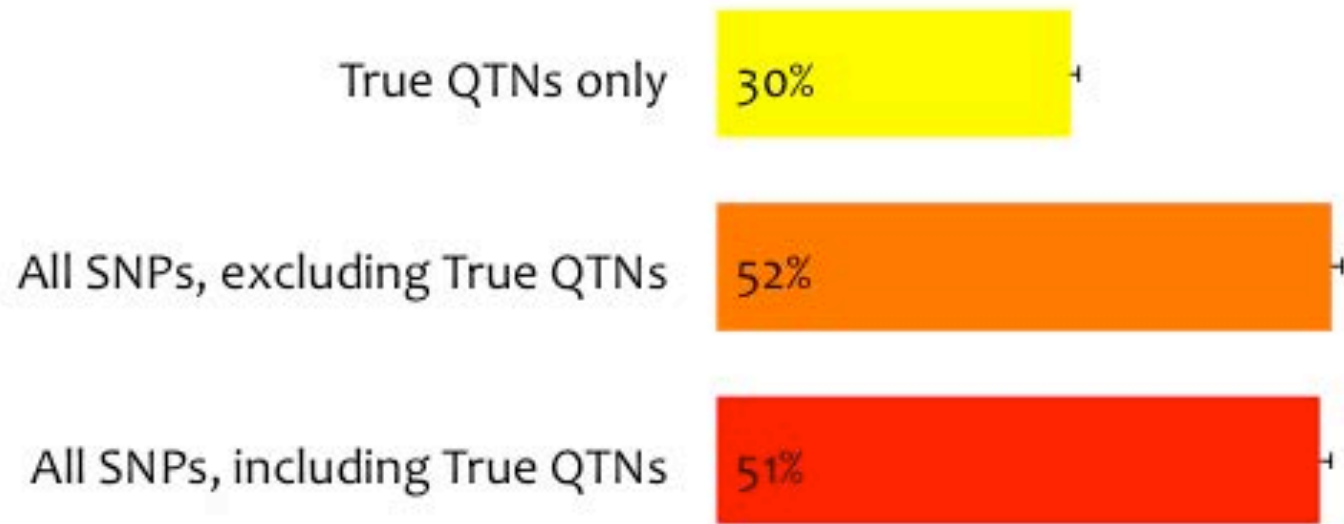
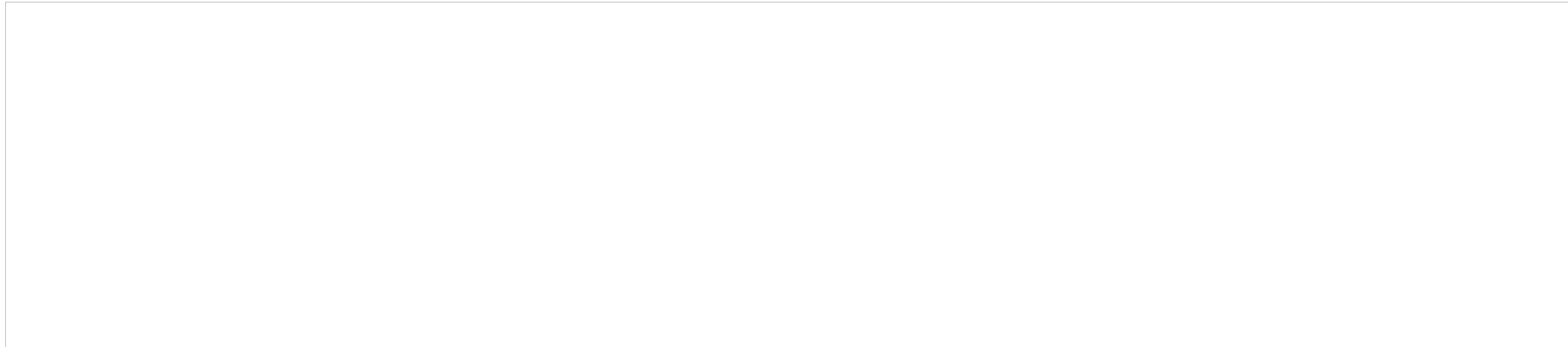
	Sensitive				Resistance			
	S1	S2	S3	S4	R1	R2	R3	R4
S1	1	1	1	1	0	0	0	0
S2	1	1	1	1	0	0	0	0
S3	1	1	1	1	0	0	0	0
S4	1	1	1	1	0	0	0	0
R1	0	0	0	0	1	1	1	1
R2	0	0	0	0	1	1	1	1
R3	0	0	0	0	1	1	1	1
R4	0	0	0	0	1	1	1	1

Adding additional markers bluer the picture

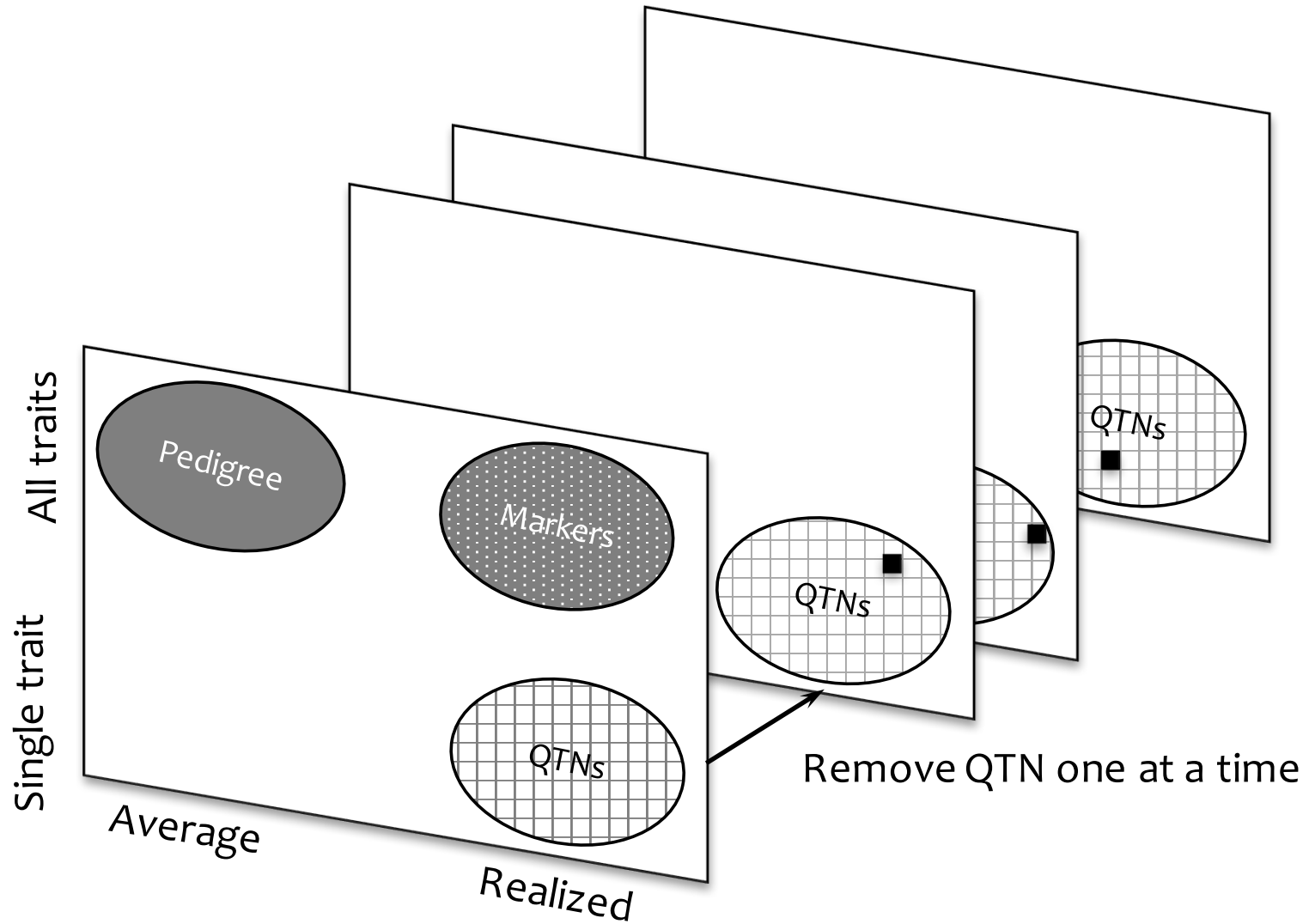
Derivation of kinship



Statistical power of kinship from



Kinship evolution



Statistical power of kinship from

True QTNs except the one tested

66%

True QTNs only

30%

All SNPs, excluding True QTNs

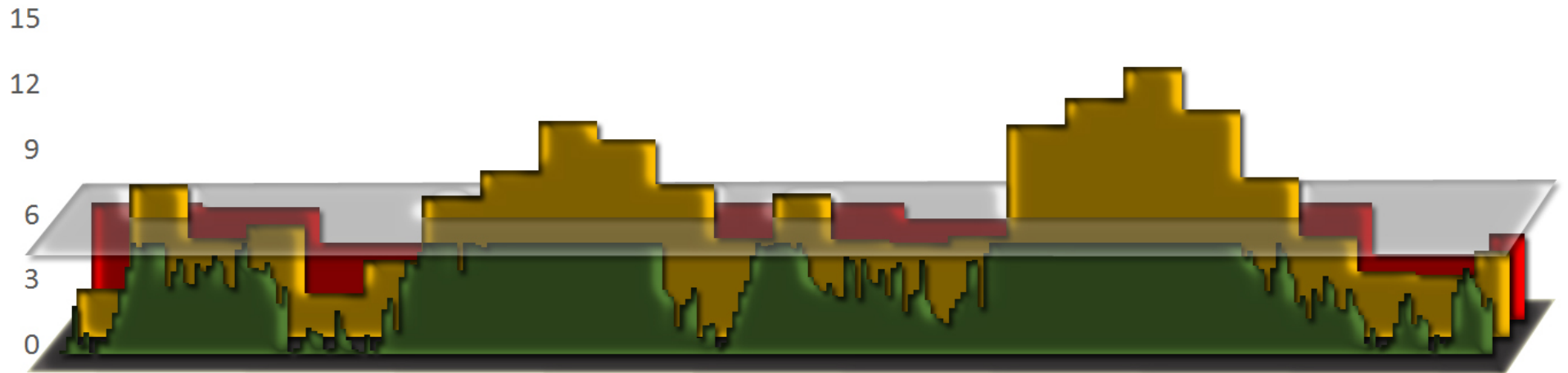
52%

All SNPs, including True QTNs

51%

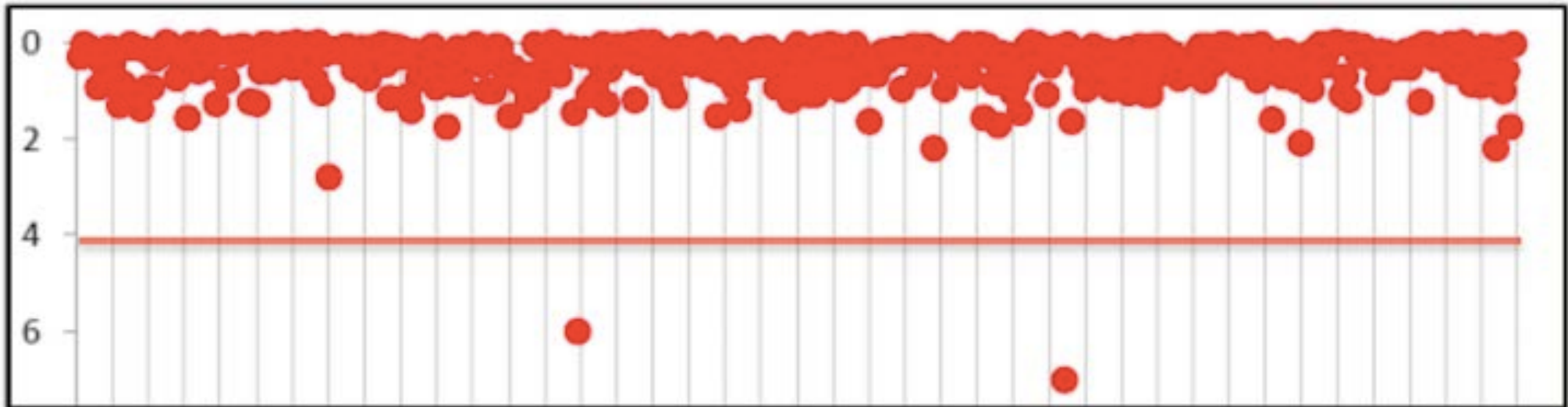


Bin approach

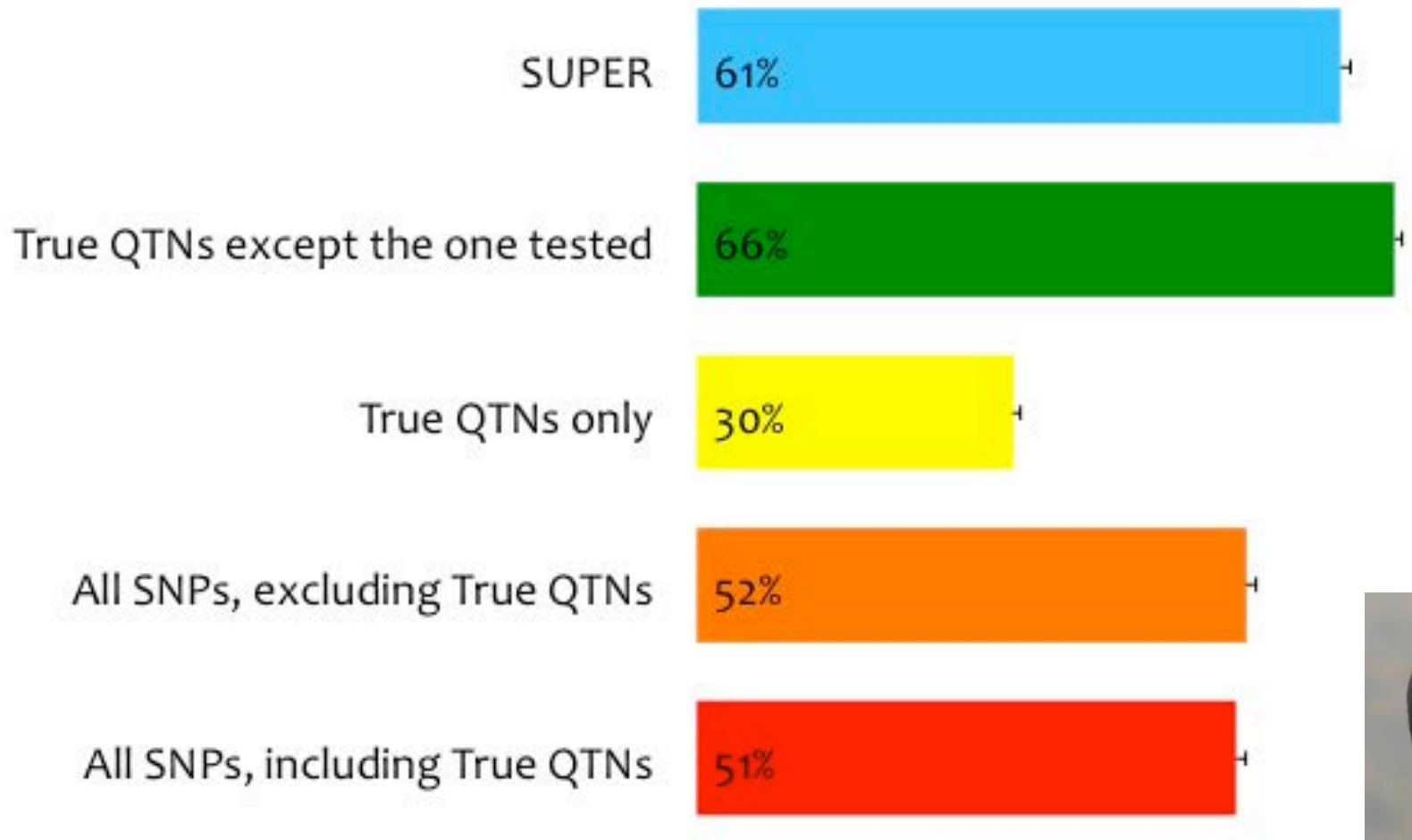


Mimic QTN-1

- * 1. Choose t associated SNPs as QTNs each represent an interval of size s .
- * 2. Build kinship from the t QTNs
- * 3. Optimization on t and s
- * 4. For a SNP, remove the QTNs in LD with the SNP, e.g. $R^2 > 1\%$
- * 5. Use the remaining QTNs to build kinship for testing the SNP

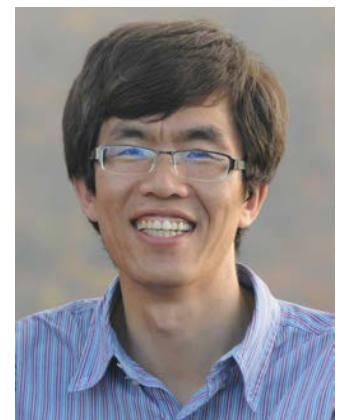


Statistical power of kinship from



SUPER

(Settlement of kinship Under Progressively Exclusive Relationship)



Qishan Wang

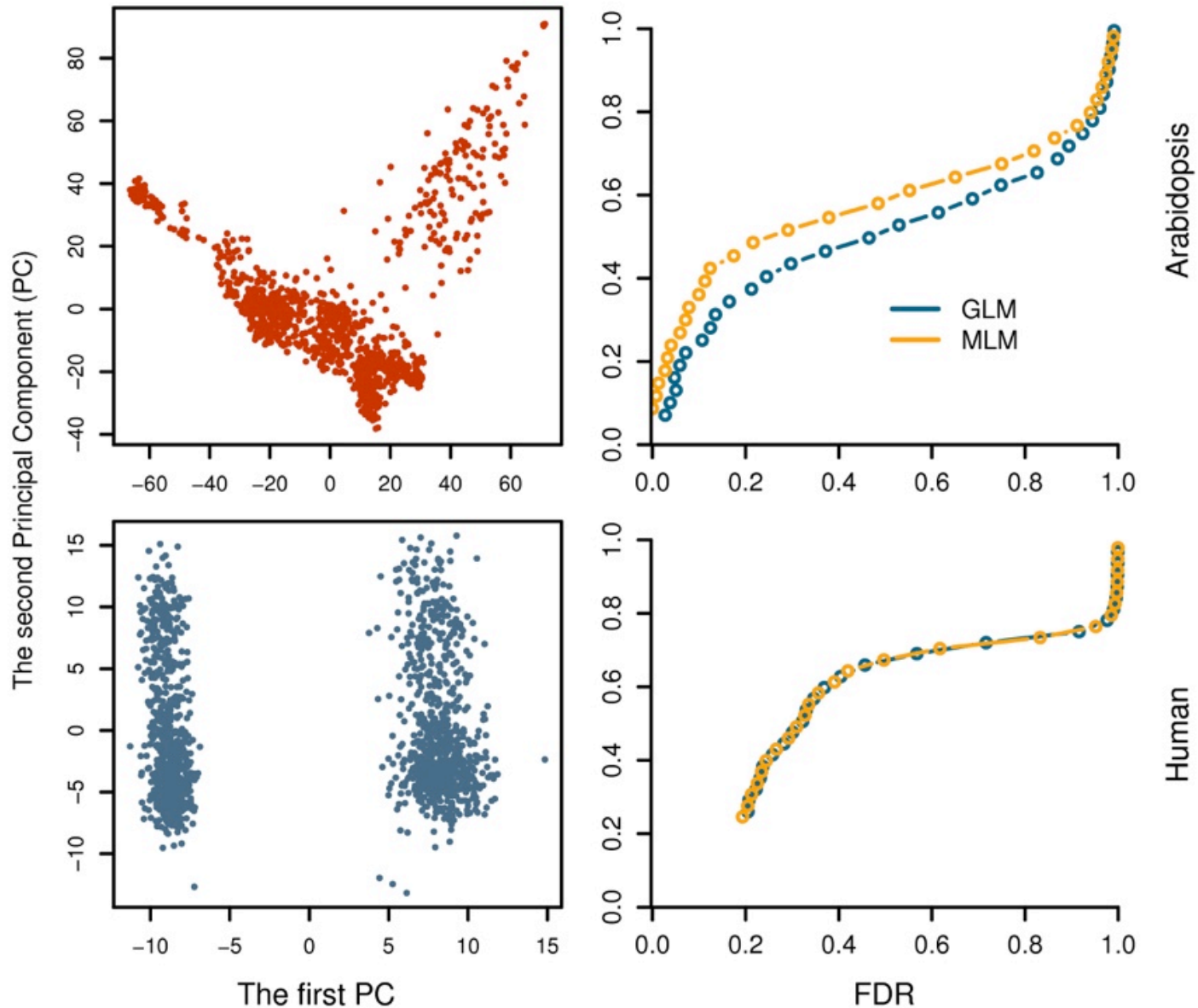
PLoS One, 2014

Usage of Software Packages

Software	Leading Authors	Corresponding authors	Language	Released	Citation
PUMA	Gabriel E. Hoffman	Jason G. Mezey	C++	2013	8
TATES	Sophie van der Sluis	Sophie van der Sluis	Fortran	2013	20
GAPIT	Lipka AE	Zhang Z	R	2012	106
MLMM	Vincent S	Nordborg M	R/python	2012	69
GEMMA	Zhou X	Stephens M	C++	2012	88
FastLMM	Christoph L, Listgarten J, Heckerman D	Christoph L, Listgarten J, Heckerman D	C++	2011	104
Qxpak	M. Pérez-Enciso	M. Pérez-Enciso	Fortran	2004	141
EMMAX	Kang HM	Sabatti C & Eskin E	C++	2010	349
GCTA	Jian Y	Jian Y	C++	2011	380
GenABEL	Aulchenko YS	Aulchenko YS	R	2007	510
TASSEL	Bradbury PJ, Zhang Z, Kroon DE	Bradbury PJ	Java	2006	660

75%

Why human geneticists not go beyond PLINK?




GWAS Model Development


s_i : Testing marker

Q: Population structure

K: Kinship

S: Pseudo QTNs


 Adjustment on marker

 Adjustment on covariates

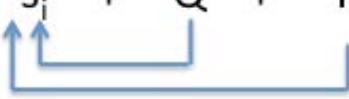
t test

$$y = s_i + e$$

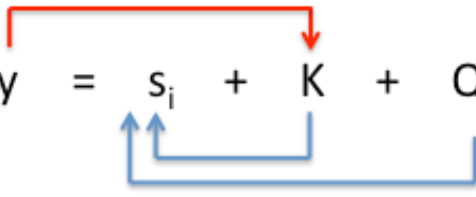
GLM

$$y = s_i + Q + e$$


MLM

$$y = s_i + Q + K + e$$


CMLM

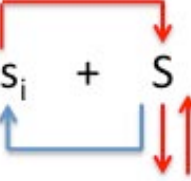
$$y = s_i + K + Q + e$$


MLMM

$$y = s_i + S + Q + K + e$$


FarmCPU

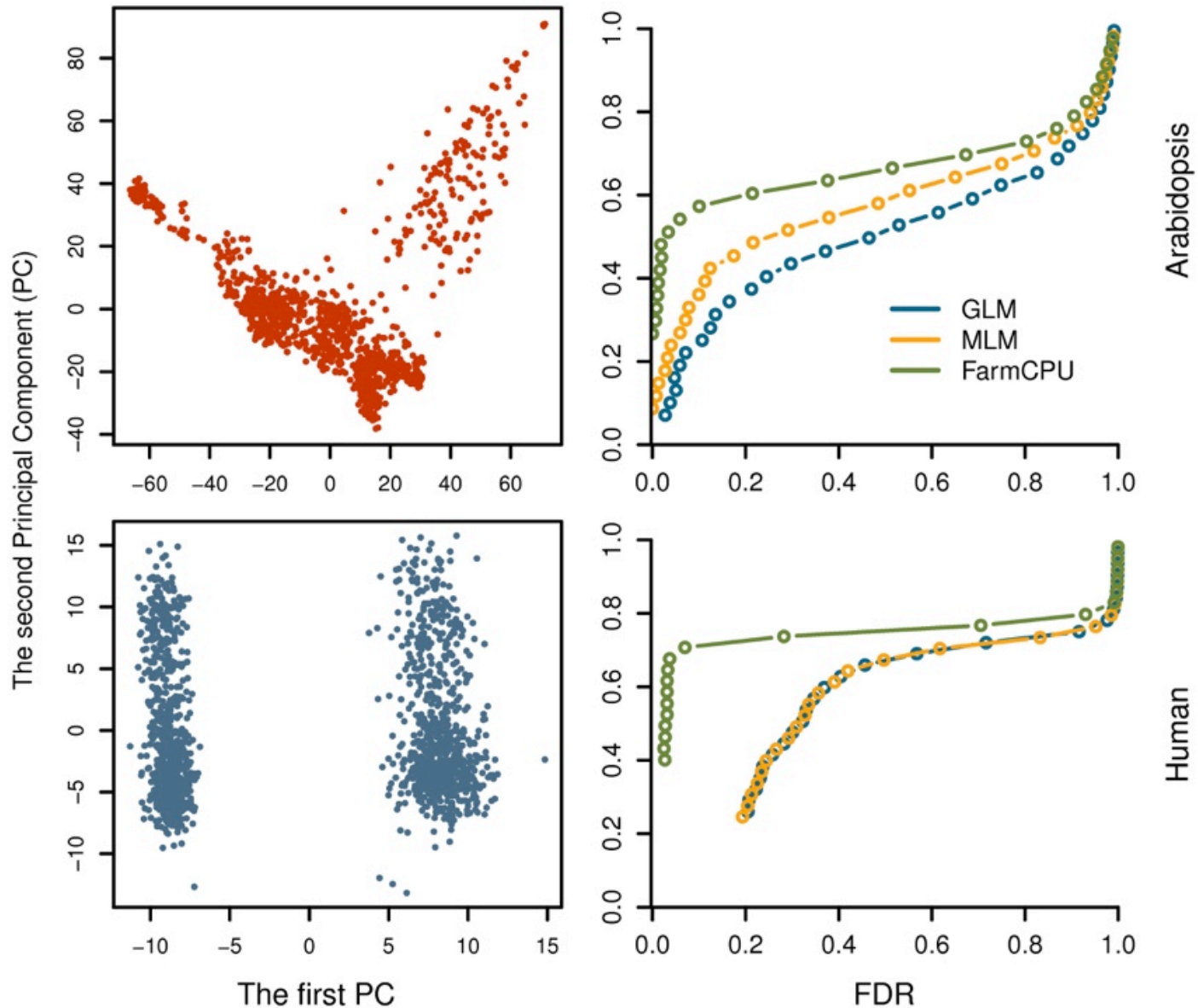
$$y = s_i + S + e$$

$$y = K + e$$


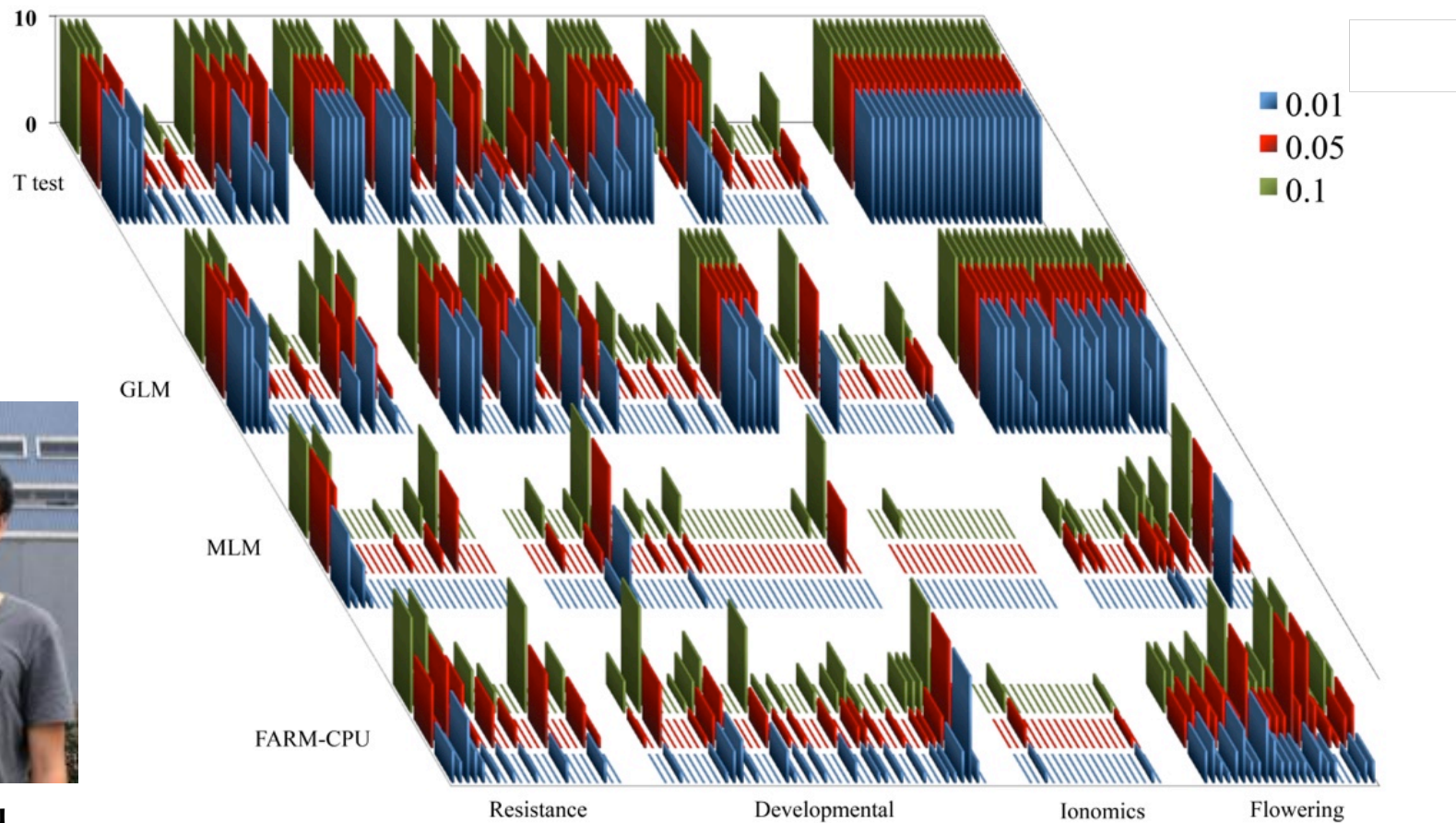
It is time for human geneticists to move forward



Xiaolei Liu

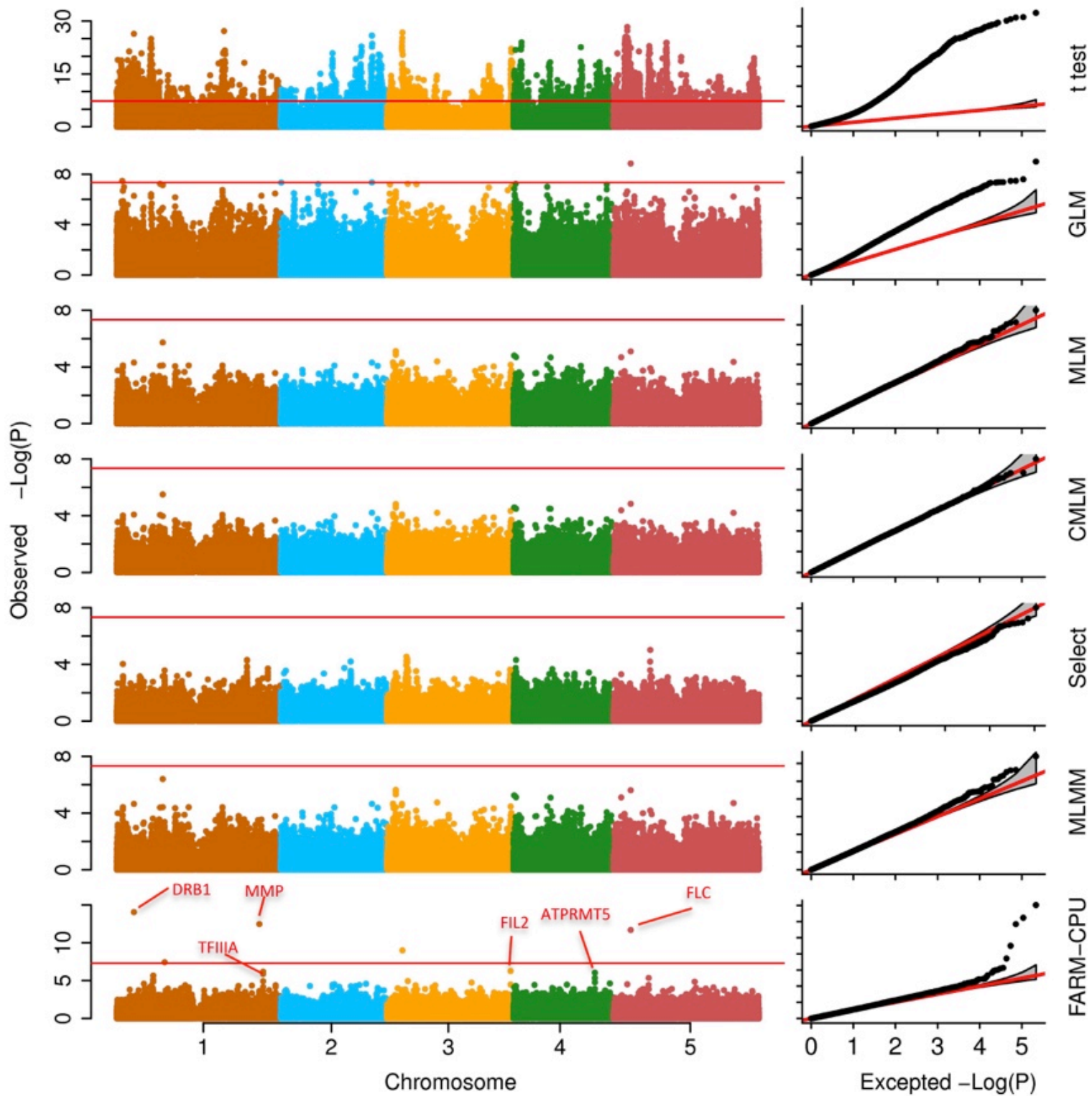


Re-analysis of *Arabidopsis* data

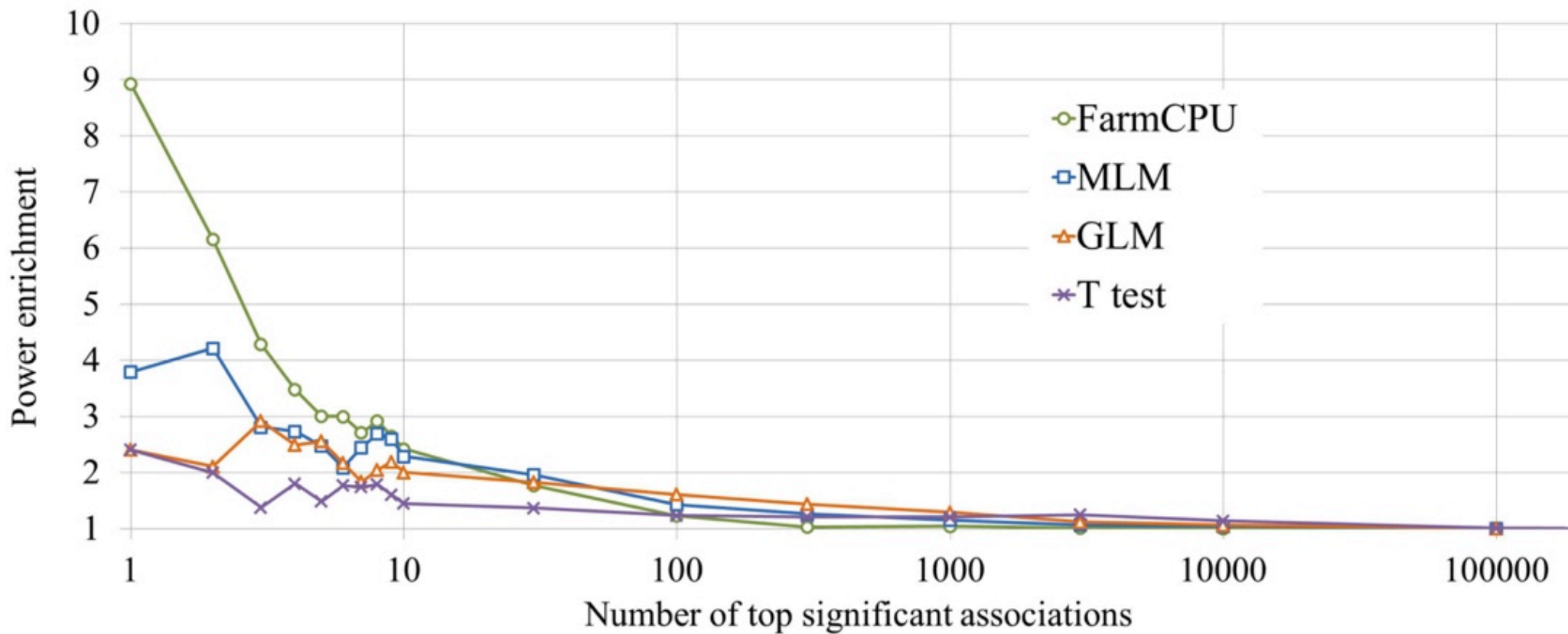


Xiaolei Liu

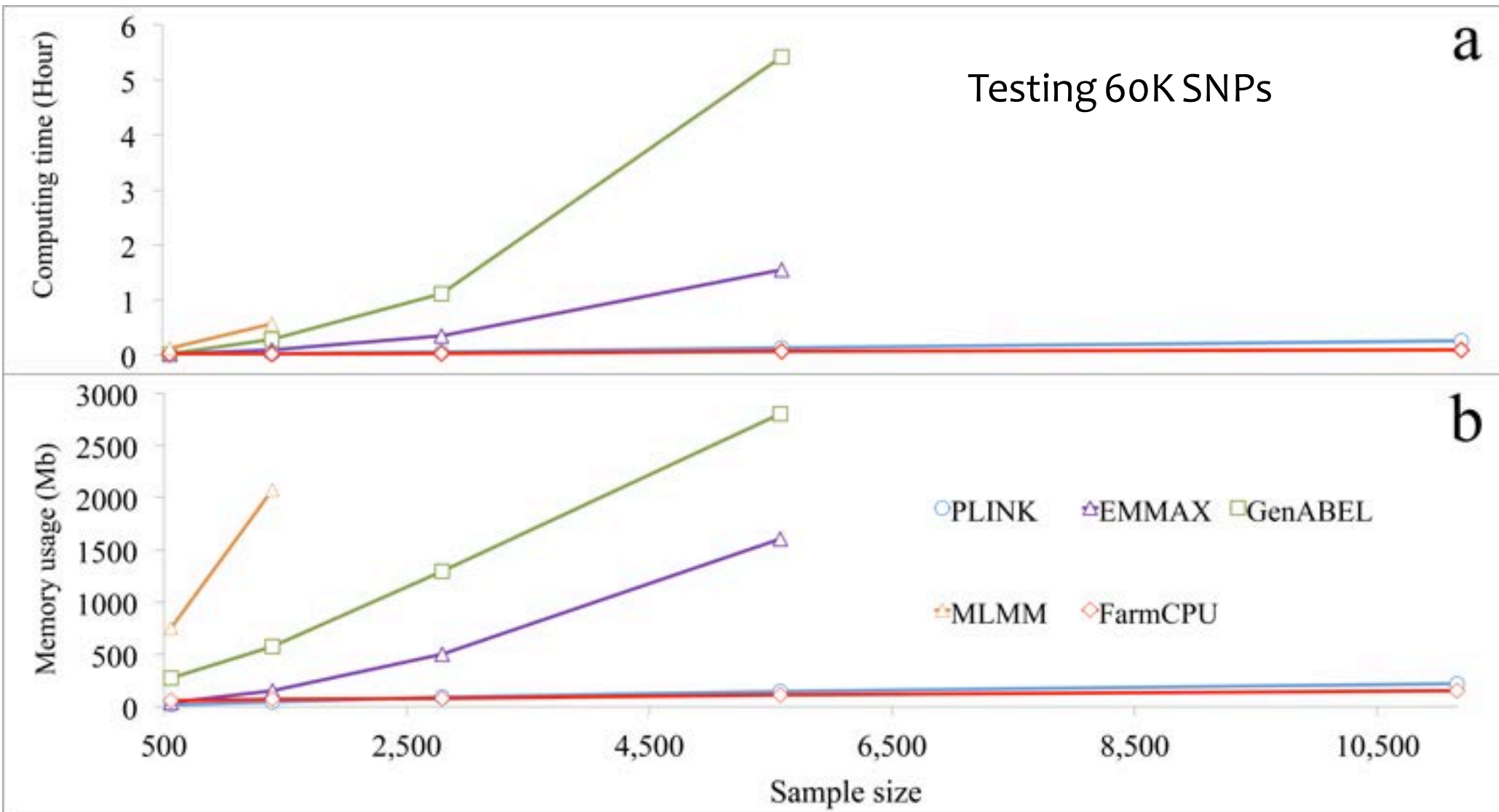
Associations on flowering time



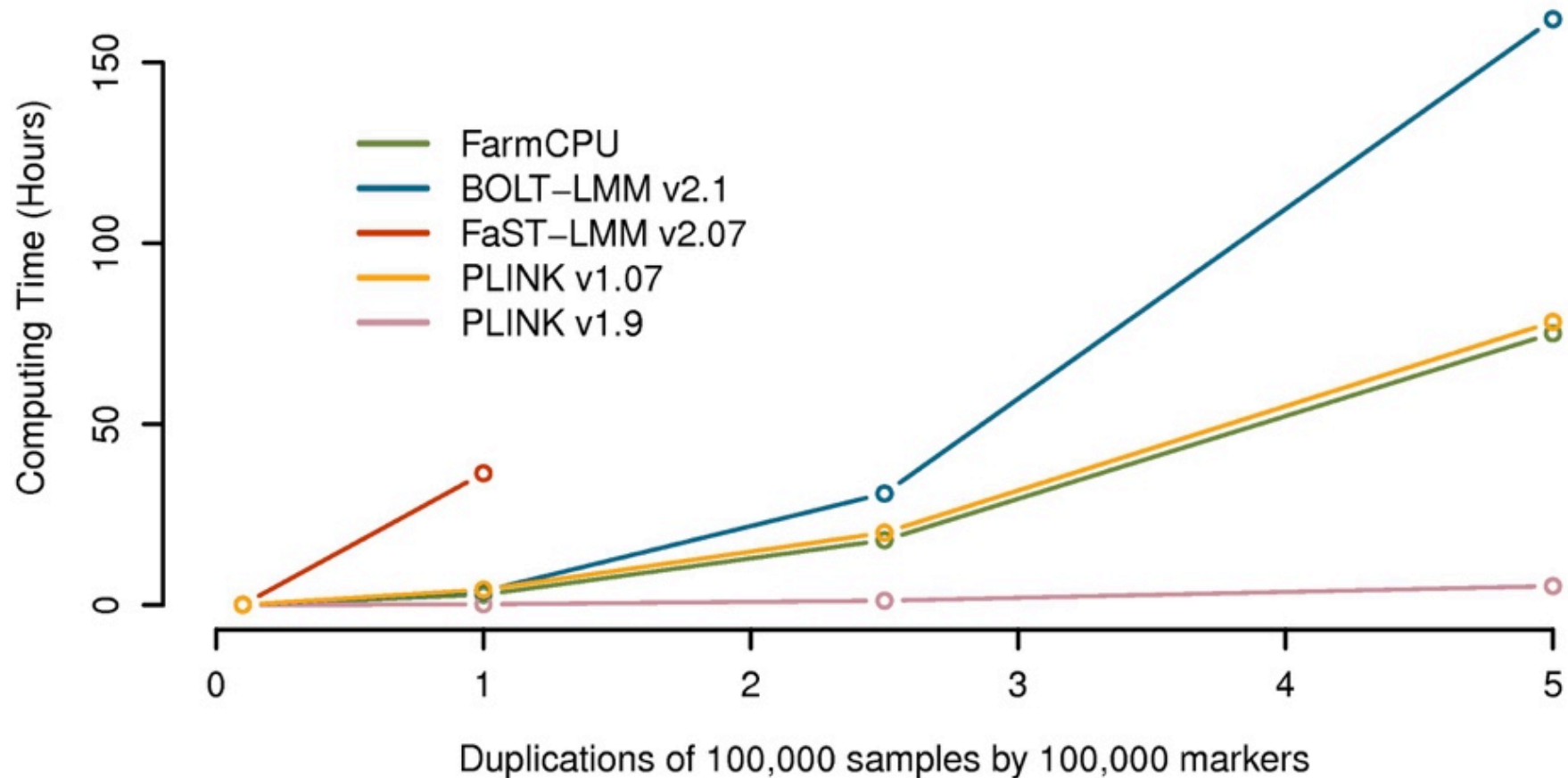
Flowering time genes enriched



FarmCPU is computing efficient



Half million individuals, half million SNPs three days



But, PLINK new version is faster

Monday, March 28, 2016

Highlight

PLoS Genetics published a new GWAS method of [FarmCPU](#) that dramatically increases the chances to map genes underlying human diseases and agriculturally important traits. The [FarmCPU R package](#) reduces computing time from weeks to days.

Capital Press reported Dr. Zhiwu Zhang as [Cyber breeder](#) along with business breeders, field breeders and molecular breeders to improve wheat varieties.

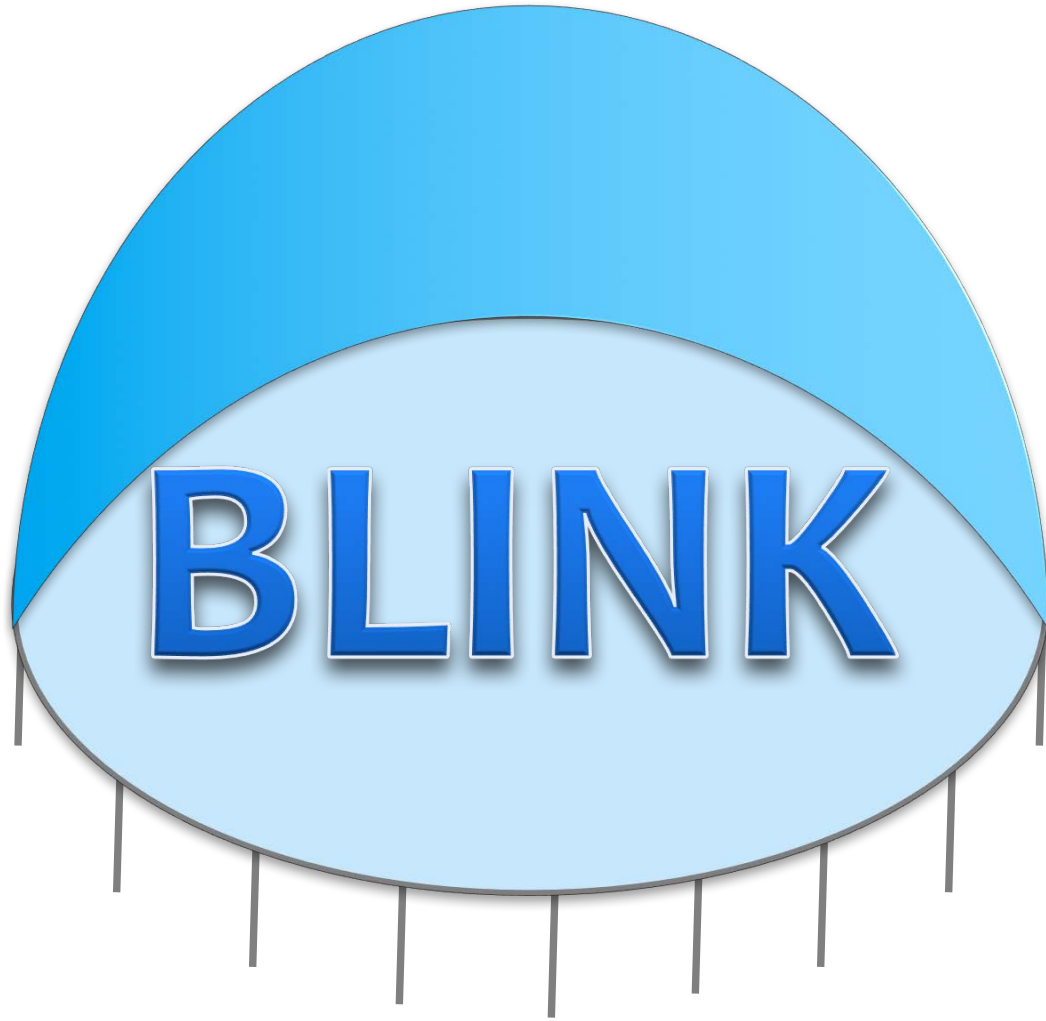
Our mission

Develop innovative, cutting-edge statistical methods and computing tools to advance genomic research toward the sustainability of food production and healthcare management.

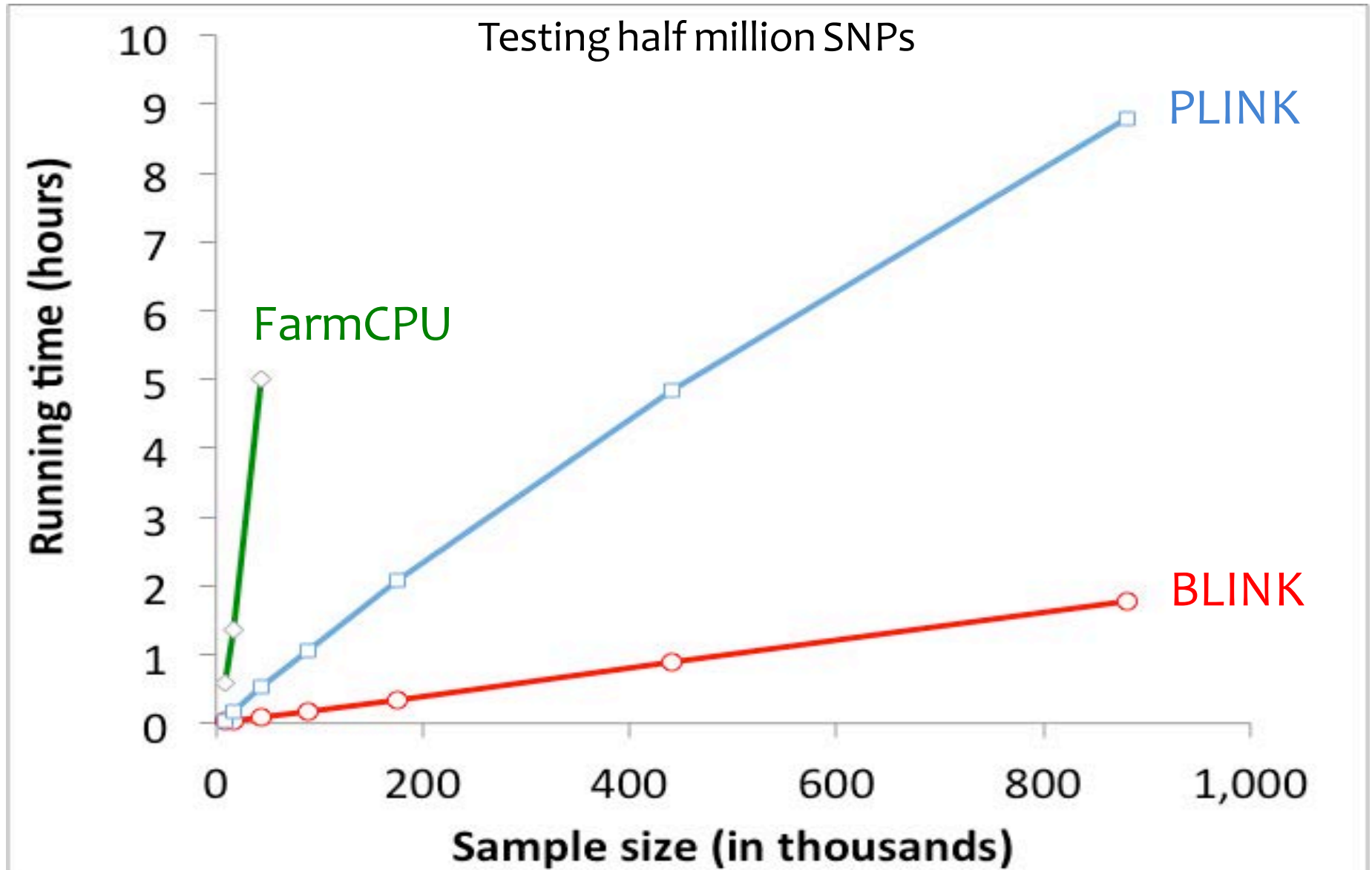
ZZLab.Net

<http://zzlab.net/GAPIT/data>



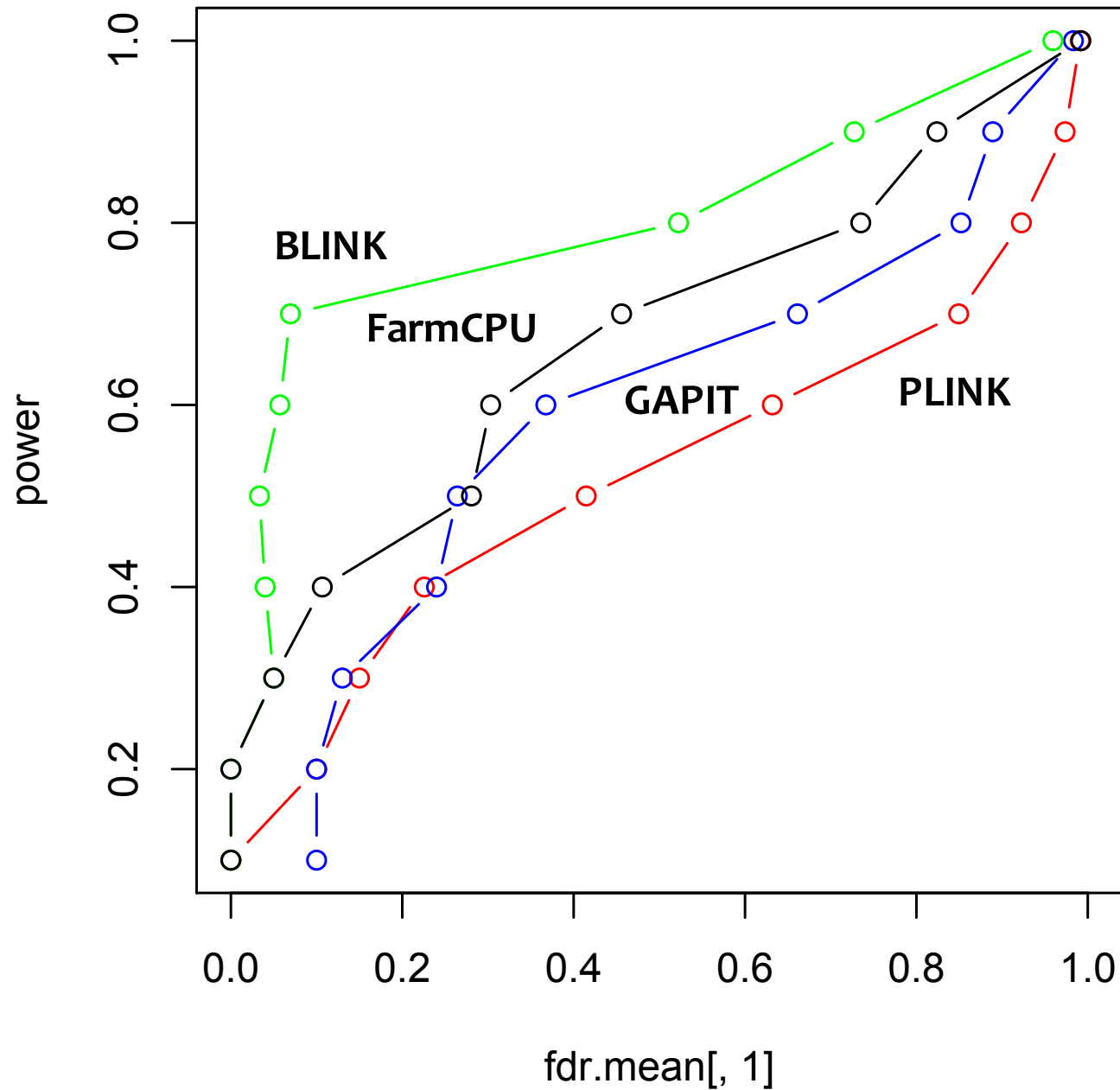


BLINK is super fast

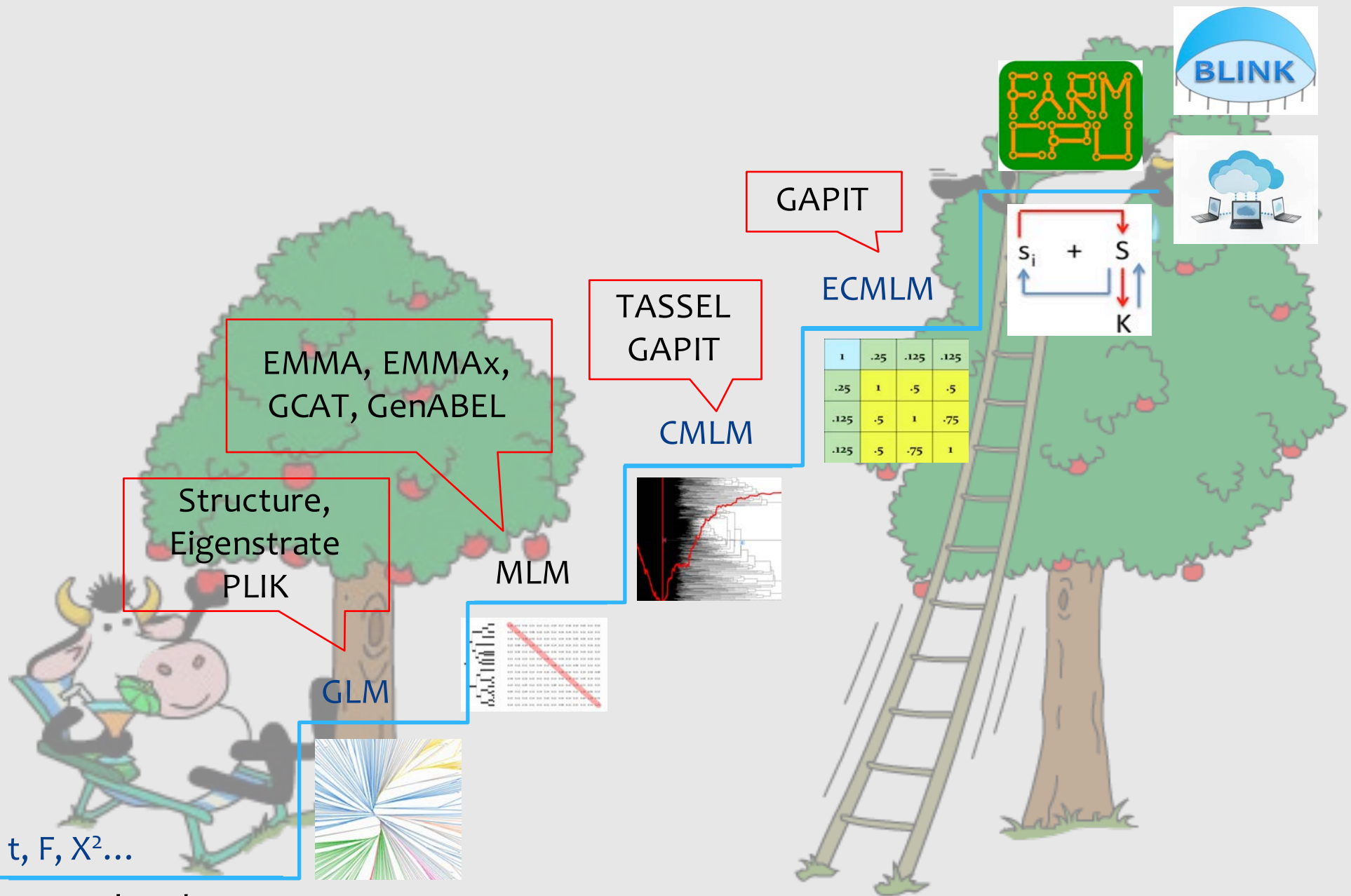


BLINK is 5X > PLINK, 200X > FarmCPU

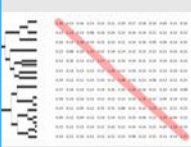
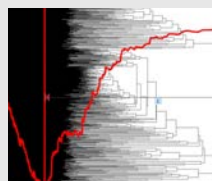
http://zzlab.net/GAPIT/data/Workshop_Iowa.R



Ladder for high hanging fruits



1	-.25	-.125	-.125
-.25	1	-.5	-.5
-.125	-.5	1	-.75
-.125	-.5	-.75	1



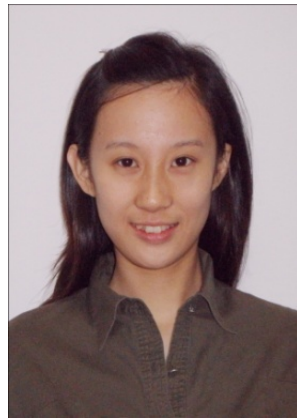
t, F, X²...

Uncorrelated or equally correlated

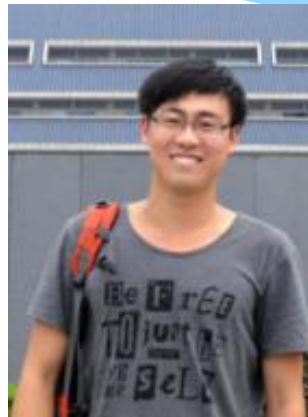
Acknowledgements



Alex Lipka



Meng Li



Xiaolei Liu



Meng Huang



Xiahui Yuan



ECMLM

