UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

A scalable Bayesian mixed model approach for GWAS and genomic prediction

Jicai Jiang[1], Li Ma[2], Paul M. VanRaden[3], Jeffrey R. O'Connell[1]

[1]University of Maryland School of Medicine, Baltimore, MD

[2]University of Maryland, College Park, MD

[3]AGIL-ARS-USDA, Beltsville, MD

2019 Interbull Meeting, Cincinnati, Ohio

June 22, 2019

Introduction

Big-data challenges

Biobank-scale samples

Big *N*>0.5 million

Sequence variants

Big *M*>10 million

Recent advances

Variance component estimation	• BOLT-REML (MC-EM-REML) • RHE-reg (Randomized Haseman-Elston regression)
GWAS	• BOLT-LMM ($\chi^2_{LMM} = c \cdot \chi^2_{LM}$) • SAIGE (Like BOLT but better for binary traits) • GCTA-fastGWA (Sparsifying GRM)
Genomic prediction	• ssGBLUP • BayesRv2 (One of the fastest MCMC-based)

Issues

- $\chi^2_{LMM} = c \cdot \chi^2_{LM}$ in BOLT and SAIGE for GWAS
  - Correction factor *c* may have big variation.
  - Loss of accuracy in test statistics

- Genomic prediction
  - Scalable and accurate Bayesian approaches are lacking.

Methods

SNP-set Genomic Prediction (SSGP)

SNP grouping

$$\vec{y} = \mathbf{X}\vec{b} + \sum_{h=1}^{p} \mathbf{K}_h \vec{u}_h + \vec{e}$$

$$b_j \sim N(0, \sigma_b^2),\ j = 1, \cdots, q$$

SNP weighting

$$\vec{u}_h \sim MVN(\mathbf{0}, \mathbf{W}_h \sigma_{u_h}^2),\ h = 1, \cdots, p$$

Prior 1

$$\sigma_{u_h}^2 \sim \text{Inv-Gamma}(a_{u_h}, b_{u_h}),\ h = 1, \cdots, p$$

No grouping
Bayes A

$$\vec{e} \sim MVN(0, \mathbf{R}\sigma_e^2)$$

$$\sigma_e^2 \sim \text{Inv-Gamma}(c_0, d_0)$$

Prior 2

$$\sigma_{u_h} \sim \text{Half-Cauchy}(A_{u_h}),\ h = 1, \cdots, p$$

No grouping
Horseshoe

Mean field approximation

$$P(\boldsymbol{b}, \boldsymbol{u}_\cdot, \sigma^2_\cdot | D) \approx Q(\boldsymbol{b}, \boldsymbol{u}_\cdot, \sigma^2_\cdot) = q(\boldsymbol{b}) \prod_{h=1}^{P} q(\boldsymbol{u}_h) \prod_{h=1}^{P} q\left(\sigma^2_{u_h}\right) q(\sigma^2_e)$$

Consider proximity-based SNP grouping of equal size (*S*) …

S=1: BayesA

S=*M*: GBLUP

$D_{\mathrm{KL}}(Q||P) \downarrow$ as $S \uparrow$

Prediction accuracy of *P* may ↑ as *S* ↓

We want smaller $D_{\mathrm{KL}}(Q||P)$ and higher-accuracy *P*.

A small *S* may work well for genomic prediction.

Variational inference (VI)

When variance components are not known

Marker effect estimates are biased.

When variance components are known

VI is equivalent to block-wise Gauss-Seidel method.

Marker effect estimates are BLUP.

Association testing

$$V = \prod_{h=1}^{P} K_h W K_h' \sigma_{u_h}^2 + R\sigma_e^2$$

$$\chi^2_{LMM} = \frac{(z'V^{-1}\tilde{y})^2}{z'V^{-1}z}$$

- Note $V^{-1}\tilde{y} = R^{-1}\sigma_e^{-2}\hat{e}$.
- We compute $\hat{e}$ by block-wise Gauss-Seidel method.
- We compute $V^{-1}z$ similarly.

$$V_h = K_h W K_h' \sigma_{u_h}^2 + R\sigma_e^2$$

$\chi^2_h = \frac{(z'V^{-1}\tilde{y})^2}{z'V_h^{-1}z}$ for z in group h.

- $\chi^2_{LMM} = c \cdot \chi^2_h$
- $S \in [1000, 5000]$

Fast and accurate approximation!

Time complexity

- Scaling linearly in group size (*S*), animals (*N*), and markers (*M*)

Results

Dairy bull data for genomic prediction

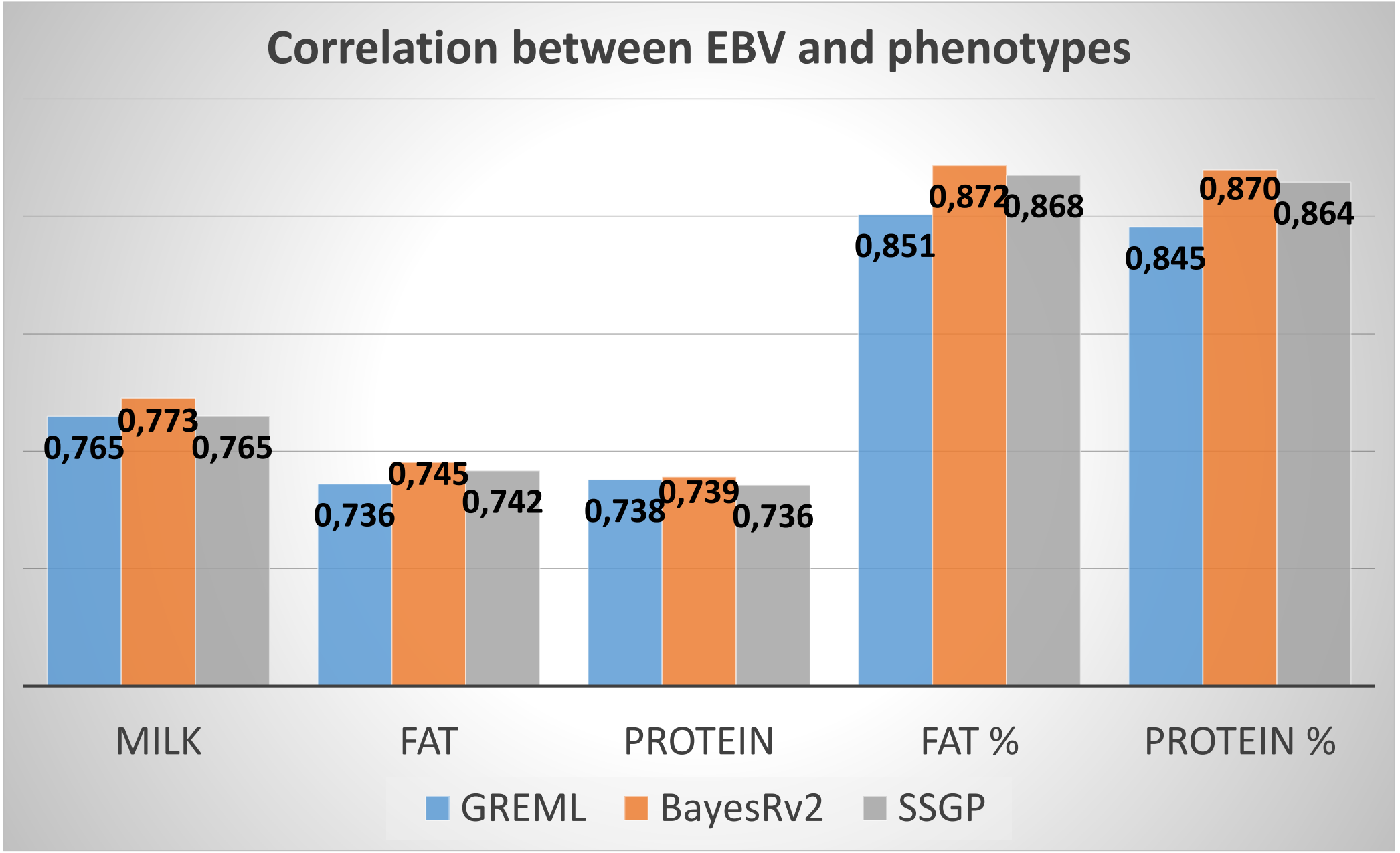
20K old bulls as training

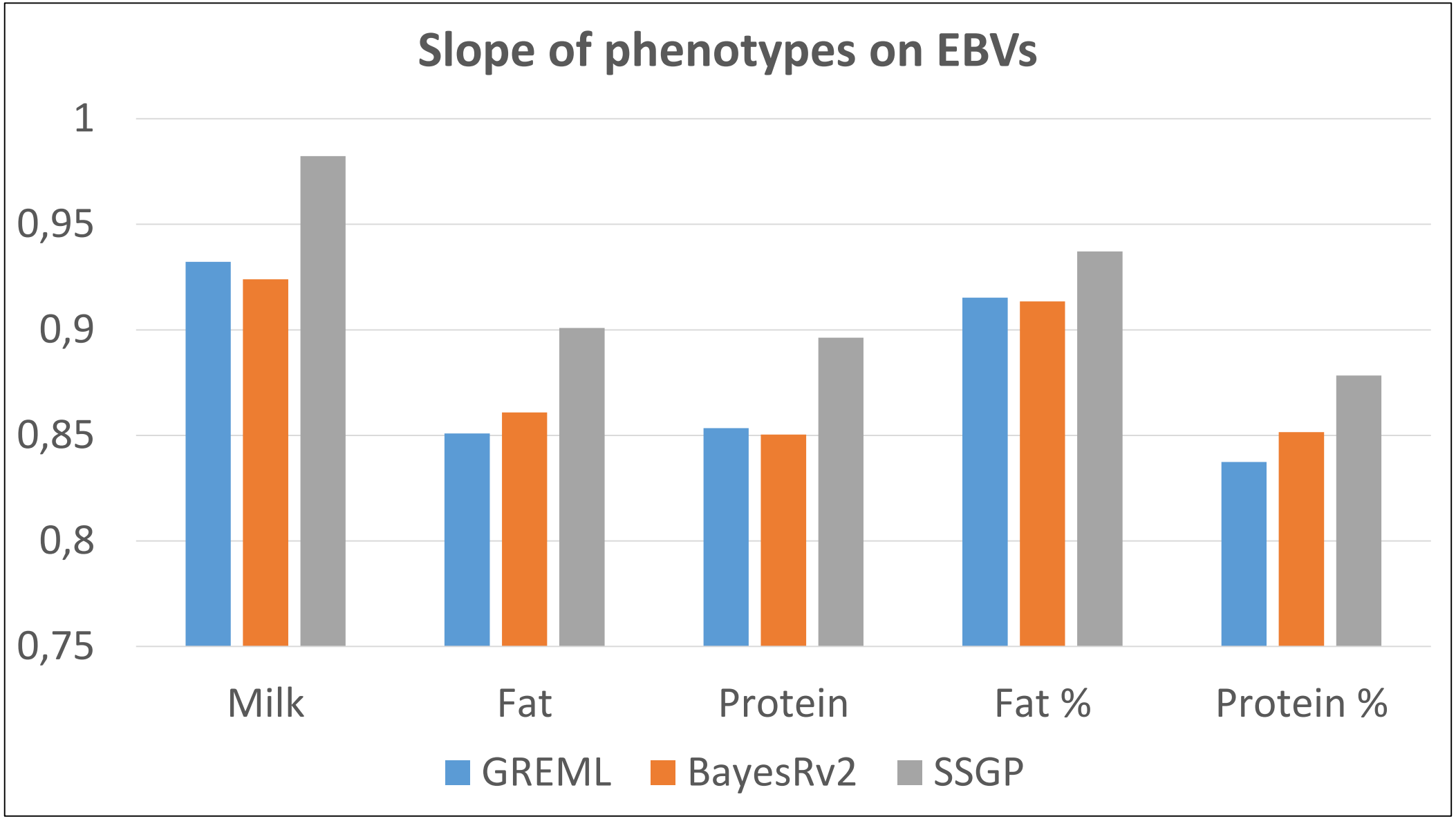
5K young bulls as validation

54K SNPs

S=10 and half-Cauchy prior used in SSGP

GCTA-GREML and BayesRv2 as benchmark


Correlation between EBV and phenotypes
0,765
0,773
0,765
0,736
0,745
0,742
0,738
0,739
0,736
0,851
0,872
0,868
0,845
0,870
0,864
MILK
FAT
PROTEIN
FAT %
PROTEIN %
GREML
BayesRv2
SSGP


Slope of phenotypes on EBVs
1
0,95
0,9
0,85
0,8
0,75
Milk
Fat
Protein
Fat %
Protein %
GREML
BayesRv2
SSGP

Time with single core on Intel Xeon E5-2680

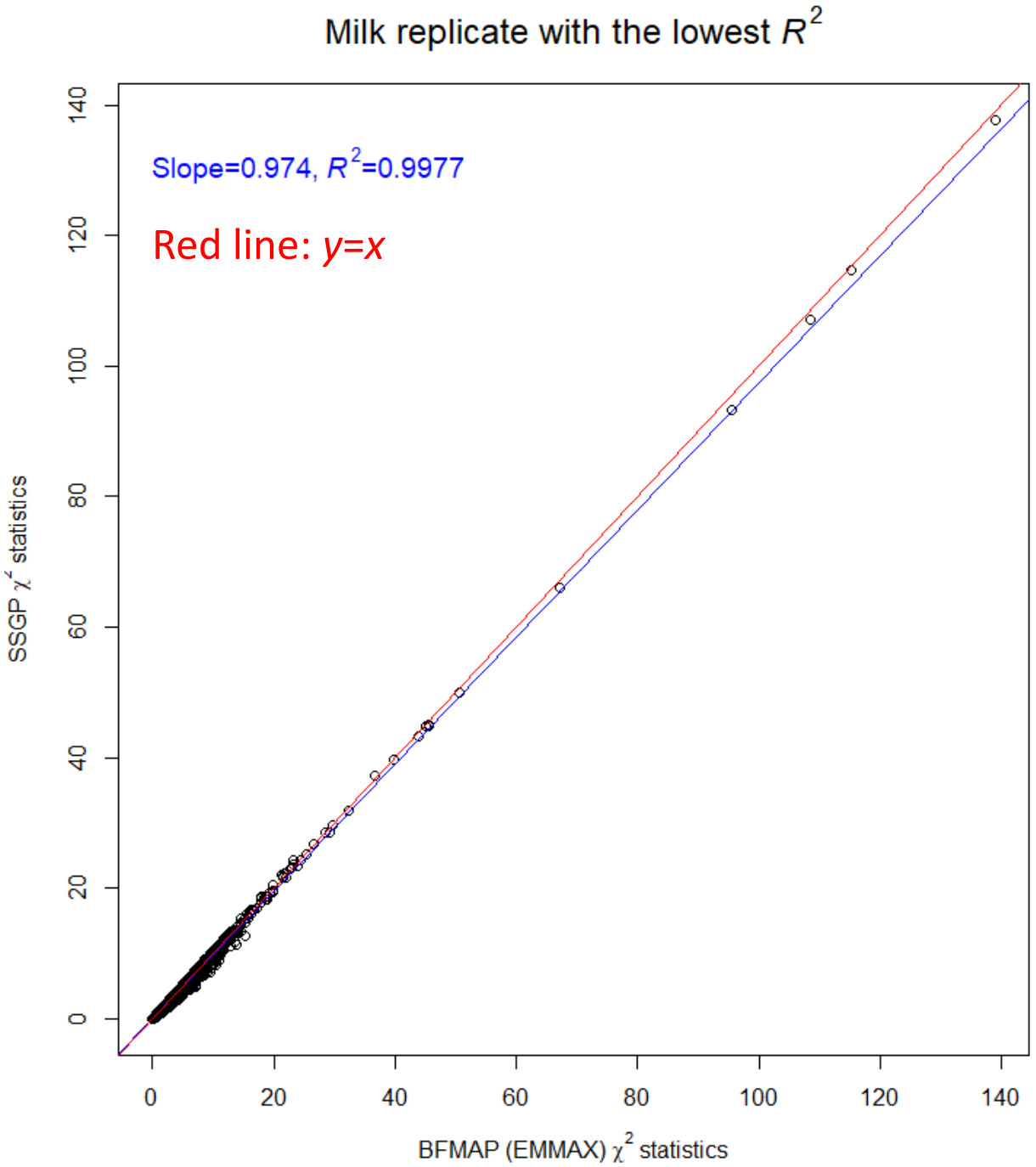
~3 min. by SSGP

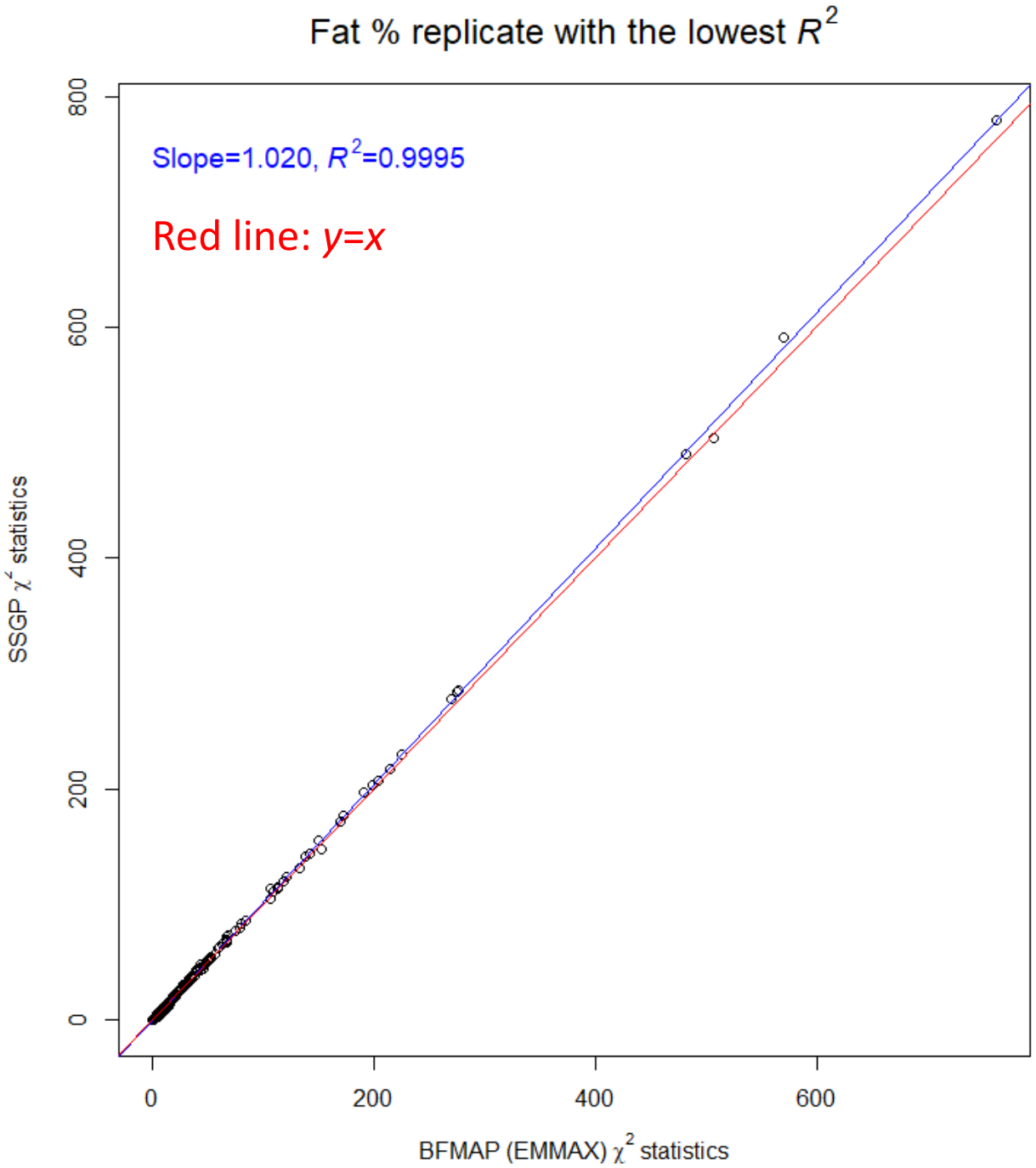
~80 min. by BayesRv2

Cow data for GWAS

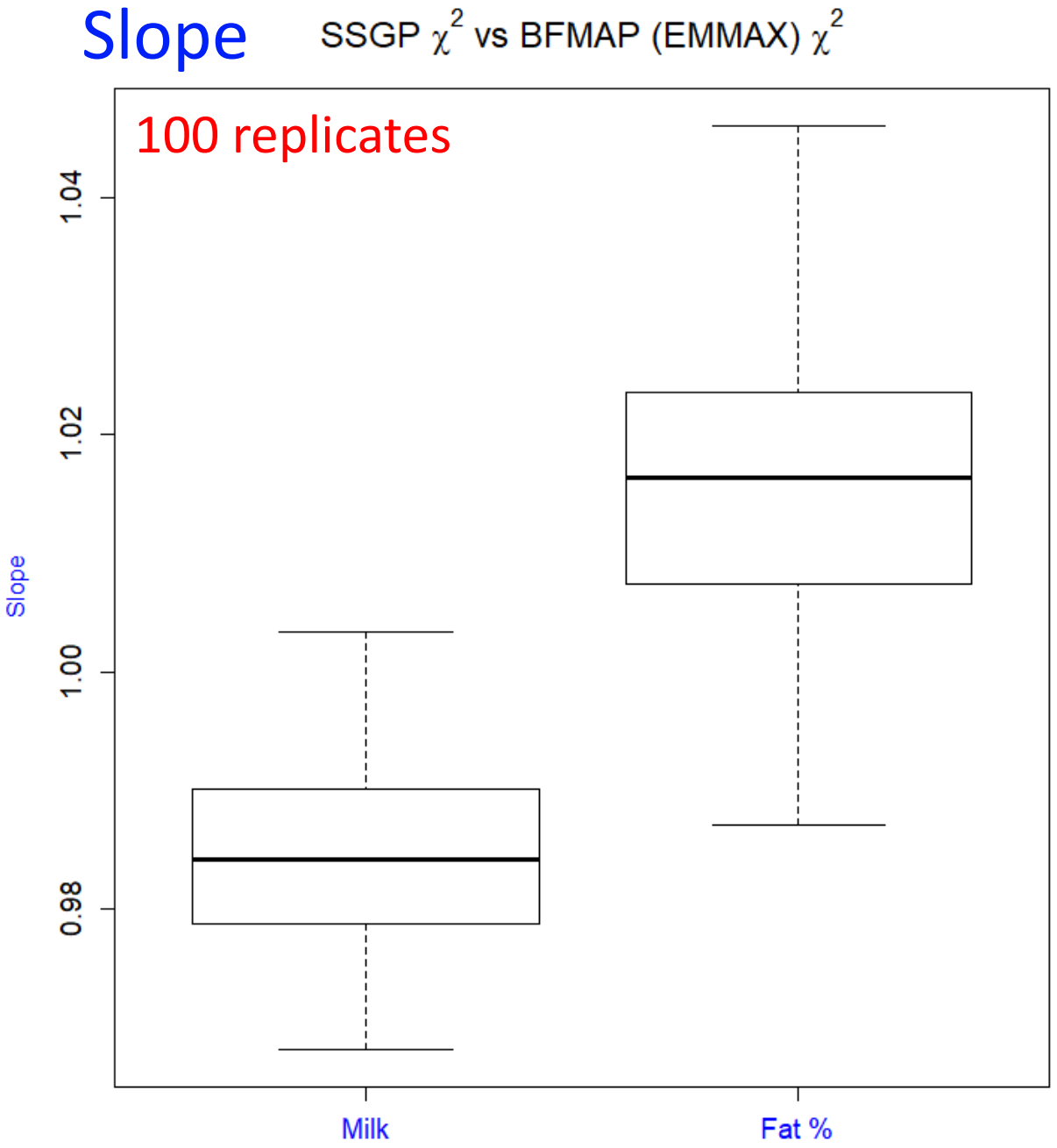
- 300K cows with yield deviations
- 60K SNPs
- SAIGE as benchmark

- 10K cows randomly sampled from 300K
  - 100 replicates
  - BFMAP (like EMMAX but 15X faster) as benchmark
  - Slope and R^2 of lm($\chi^2_{\mathrm{SSGP}} \sim \chi^2_{\mathrm{BFMAP}} - 1$) for 60K SNPs

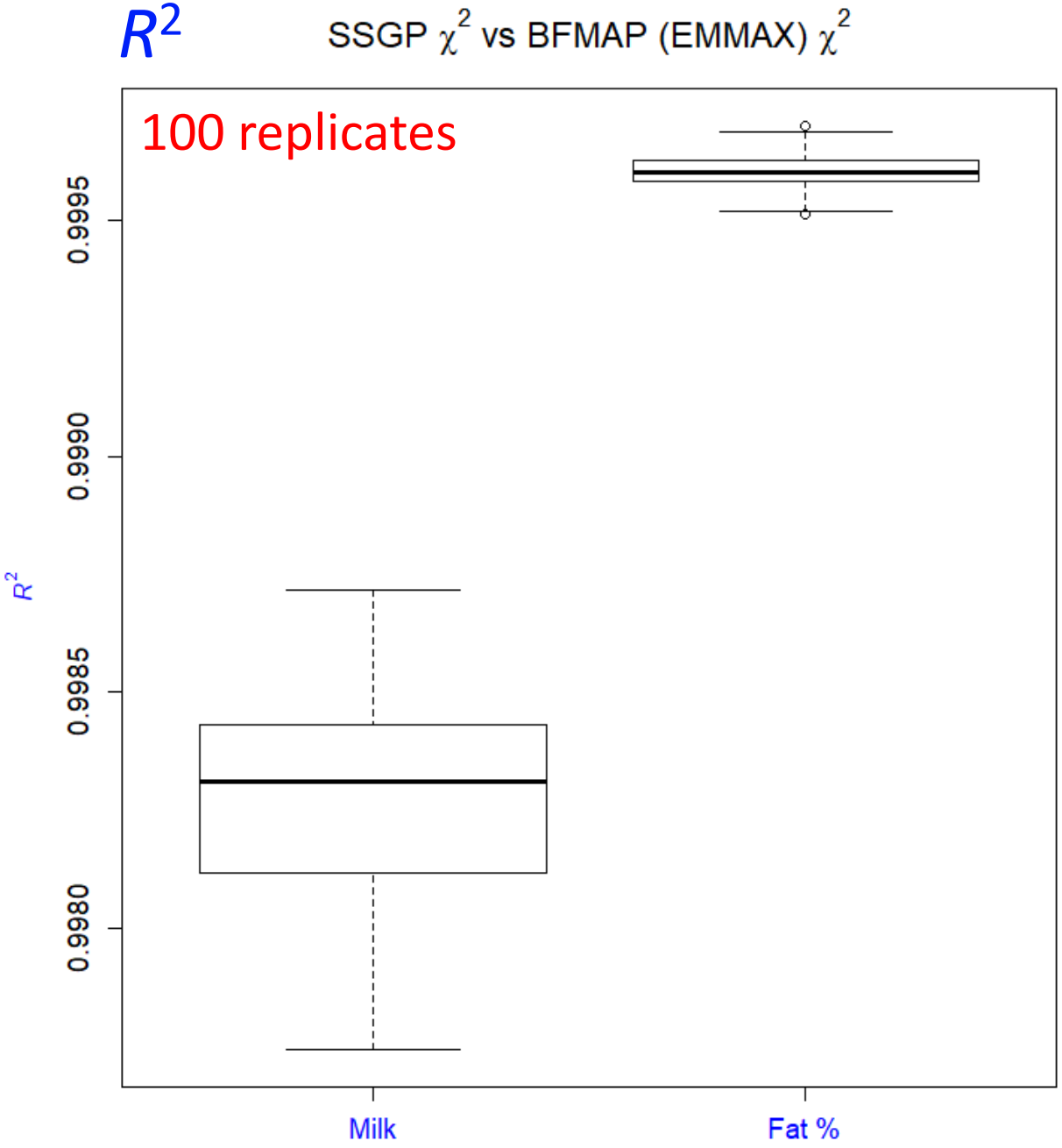

Milk replicate with the lowest R^2
Slope=0.974, R^2=0.9977
Red line: $y=x$
SSGP χ^2 statistics
BFMAP (EMMAX) χ^2 statistics
0
20
40
60
80
100
120
140

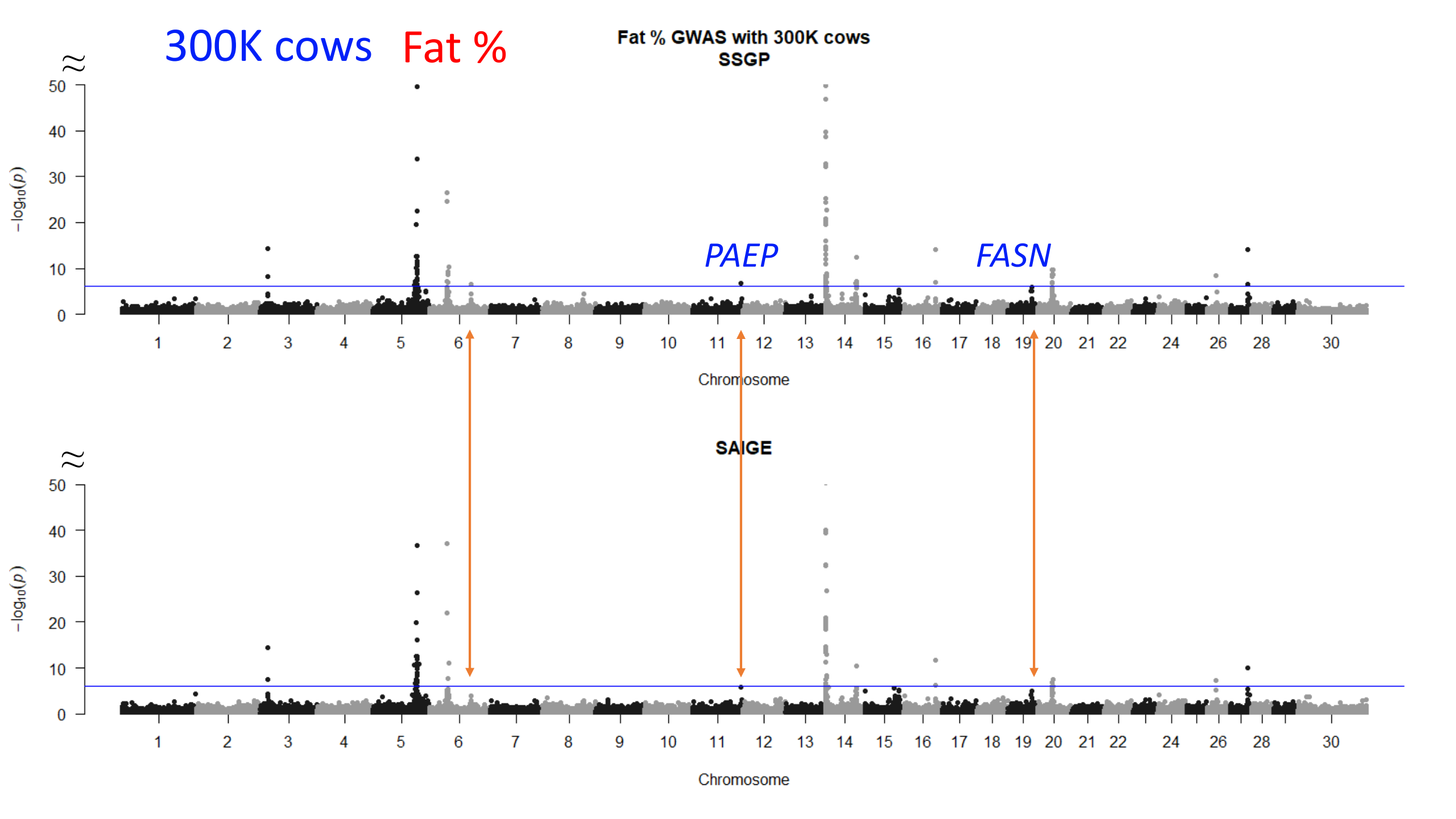

Fat % replicate with the lowest R^2
Slope=1.020, R^2=0.9995
Red line: $y=x$
SSGP χ^2 statistics
BFMAP (EMMAX) χ^2 statistics
0
200
400
600
800

Slope


SSGP χ^2 vs BFMAP (EMMAX) χ^2
100 replicates
Slope
1.04
1.02
1.00
0.98
Milk
Fat %

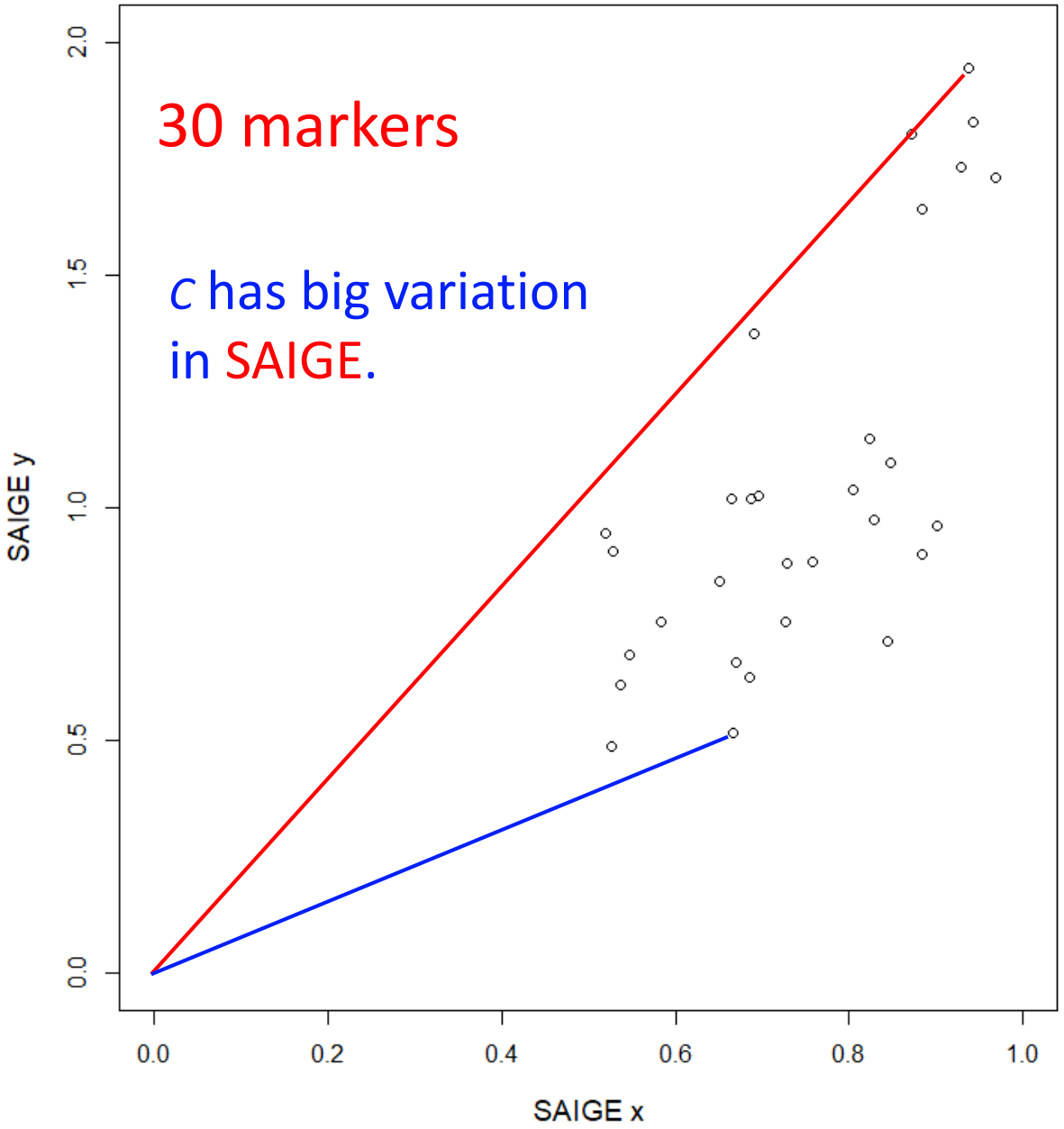
R^2


SSGP χ^2 vs BFMAP (EMMAX) χ^2
100 replicates
R^2
0.9995
0.9990
0.9985
0.9980
Milk
Fat %


300K cows Fat %
Fat % GWAS with 300K cows
SSGP
PAEP
FASN
SAIGE
−log10(p)
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 26 28 30
0 10 20 30 40 50

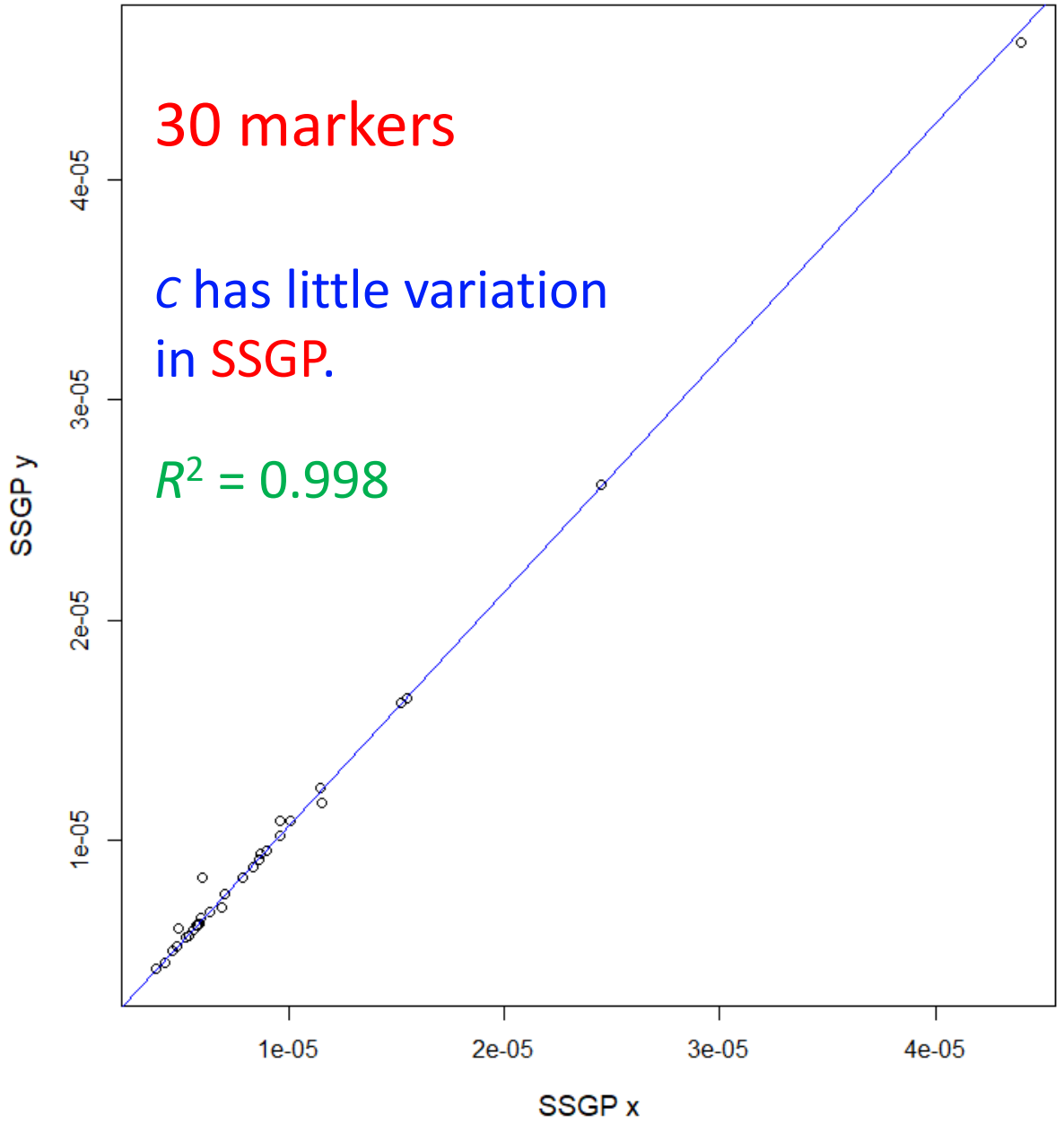
SAIGE

SAIGE $\hat{c}$ = mean(y/x)



SSGP

SSGP $\hat{c}$ = slope of lm(y~x-1)



Time on MacBook Pro (Intel i9) for 300K cows

~2.7 hours for GWAS by SSGP

>200 hours by SAIGE

Estimating the correction factor is time-consuming.

Summary

- SSGP can be applied to various types of samples.
  - Mostly unrelated, like UK Biobank
  - Highly related, like dairy cattle
  - Admixed samples
- SSGP is accurate for GWAS and for genomic prediction.
- SSGP is fast.
  - 1 million animals and 60K SNPs: <10 hours for GWAS and <5 hours for computing SNP effects on standard hardware.
- SSGP can be applied to sequence GWA.
  - Reasonable increase in computation compared to linear regression

Acknowledgements

- Funding NHLBI U01 HL137181-01

Software

- SSGP
  - https://sites.google.com/view/ssgp
- BFMAP
  - https://jiang18.github.io/bfmap/
  - GWA is currently not available in the online version.

Contact

- Jicai Jiang
  - jicai.jiang@gmail.com