

SNP effects from ssGBLUP using BLUPF90 (postGSf90)

Daniela Lourenco
BLUPF90 TEAM – 08/2024



UNIVERSITY OF
GEORGIA

College of Agricultural &
Environmental Sciences

*Animal Breeding and
Genetics Group*



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Equivalence between GBLUP and SNP-BLUP

GBLUP

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{G}^{-1}\lambda_1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

↓
GEBV

$$\text{Var}(\mathbf{u}) = ?$$

$$\text{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$$

SNP-BLUP (Ridge Regression)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

↓
SNP effects

$$\mathbf{u} = \mathbf{Z}\mathbf{a}$$

$$\text{Var}(\mathbf{u}) = ?$$

$$\text{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$$

Are GBLUP and SNP-BLUP equivalent?

- Assumption of GBLUP: $\text{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$
- In SNP-BLUP: $\mathbf{u} = \mathbf{Z}\mathbf{a}$

$$\mathbf{u} = \mathbf{Z}\mathbf{a}$$

$$\text{Var}(\mathbf{u}) = \text{Var}(\mathbf{Z}\mathbf{a})$$

$$\text{Var}(\mathbf{u}) = \mathbf{Z} \text{Var}(\mathbf{a}) \mathbf{Z}'$$

$$\text{Var}(\mathbf{u}) = \mathbf{Z}\mathbf{Z}'\sigma_a^2$$

$$\sigma_a^2 = \frac{\sigma_u^2}{2 \sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\text{Var}(\mathbf{u}) = \mathbf{Z}\mathbf{Z}' \frac{\sigma_u^2}{2 \sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\text{Var}(\mathbf{u}) = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum_{i=1}^{SNP} p_i(1-p_i)} \sigma_u^2$$

Genomic
relationship matrix
VanRaden (2008)

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2 \sum_{i=1}^{SNP} p_i(1-p_i)}$$

$$\text{Var}(\mathbf{u}) = \mathbf{G}\sigma_u^2$$



GBLUP assumption!!!



GBLUP and SNP-BLUP are equivalent!

If we can get \mathbf{u} ($\mathbf{u} = \mathbf{Za}$) from SNP-BLUP, we can get \mathbf{a} from GBLUP!



Review

Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90

Daniela Lourenco ^{1,*}, Andres Legarra ², Shogo Tsuruta ¹ , Yutaka Masuda ¹, Ignacio Aguilar ³ 
and Ignacy Misztal ¹

<https://www.mdpi.com/2073-4425/11/7/790>

ssGBLUP and ssSNP-BLUP are also equivalent!

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1} \frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

ssGBLUP

Misztal et al. (2009)
Legarra et al. (2009)
Aguilar et al. (2010)
Christensen & Lund (2010)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{M} & \mathbf{X}'_n\mathbf{Z}_n \\ \mathbf{M}'\mathbf{Z}'\mathbf{X} & \mathbf{M}'\mathbf{Z}'\mathbf{Z}\mathbf{M} + \mathbf{I} \frac{\sigma_e^2}{\sigma_a^2} & \mathbf{M}'_n\mathbf{Z}'_n\mathbf{Z}_n \\ \mathbf{Z}'_n\mathbf{X}_n & \mathbf{Z}'_n\mathbf{Z}_n\mathbf{M}_n & \mathbf{Z}'_n\mathbf{Z}_n + \mathbf{A}^{nn} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{M}'\mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'_n\mathbf{y}_n \end{bmatrix}$$

ssSNPBLUP or ssBR

Fernando et al. (2014)
Liu et al. (2014)
Mantysaari & Strandén (2016)



J. Dairy Sci. 101:10082–10088
<https://doi.org/10.3168/jds.2018-14913>

© 2018, The Authors. Published by FASS Inc. and Elsevier Inc. on behalf of the American Dairy Science Association®.
This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Short communication: Genomic prediction using different single-step methods in the Finnish red dairy cattle population

H. Gao,[†] M. Koivula,[‡] J. Jensen,^{*} I. Strandén,[‡] P. Madsen,^{*} T. Pitkänen,[‡] G. P. Aamand,[‡] and E. A. Mantysaari[‡]

^{*}Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

[†]Nordic Cattle Genetic Evaluation, DK-8200 Aarhus, Denmark

[‡]Natural Resources Institute Finland (Luke), FIN-31600 Jokioinen, Finland

We confirmed that regular ssGBLUP and ssBR with an extra polygenic effect led to the same predictions.

SNP effects in ssGBLUP

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1}\lambda_1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

$$\hat{\mathbf{a}} = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$$

Matrix of SNP content
Genomic relationship matrix

α = blending parameter for \mathbf{G}

$$b = 1 - \frac{\lambda}{2}$$

$$\lambda = \frac{1}{n^2} \left(\sum_i \sum_j \mathbf{A}_{22ij} - \sum_i \sum_j \mathbf{G}_{ij} \right)$$

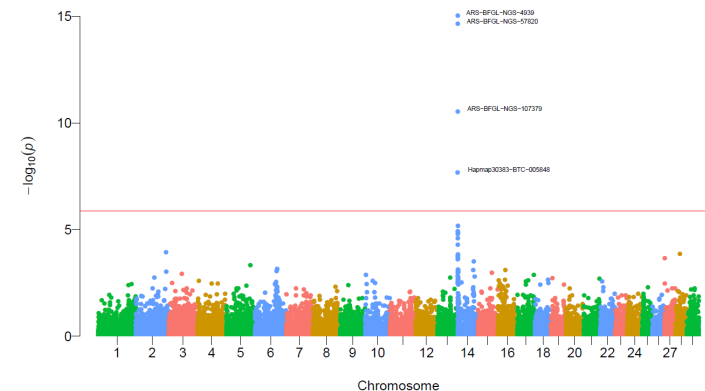
What can we do with SNP effects?

1) Predictions for animals not included in the evaluation

Indirect Predictions

Indirect Genomic Predictions

2) Genome-Wide Association Studies (GWAS)



1) Indirect Predictions

- Interim evaluations
 - Between official runs
- Not all genotyped animals are in the evaluations
 - Animals with incomplete pedigree increase bias and lower R^2
- Commercial products
 - e.g., GeneMax -> genomic testing for non-registered animals

1) Indirect Predictions

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1}\lambda_1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix} \quad \Longrightarrow \quad \hat{\mathbf{a}} = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{Z}'\mathbf{G}^{-1} \hat{\mathbf{u}}$$

$$\text{Indirect Prediction: } \mathbf{IP} = \mathbf{u}_m^* = \mathbf{Z}\hat{\mathbf{a}}$$

1) Indirect Predictions

Indirect Prediction: $\mathbf{u}_m^* = \mathbf{Z}\hat{\mathbf{a}}$

└─ Fine if comparing among animals with IP

- Not fine if comparing IP with GEBV ($\hat{\mathbf{u}}$) from the main evaluation
 - Need to put IP in the pedigree scale

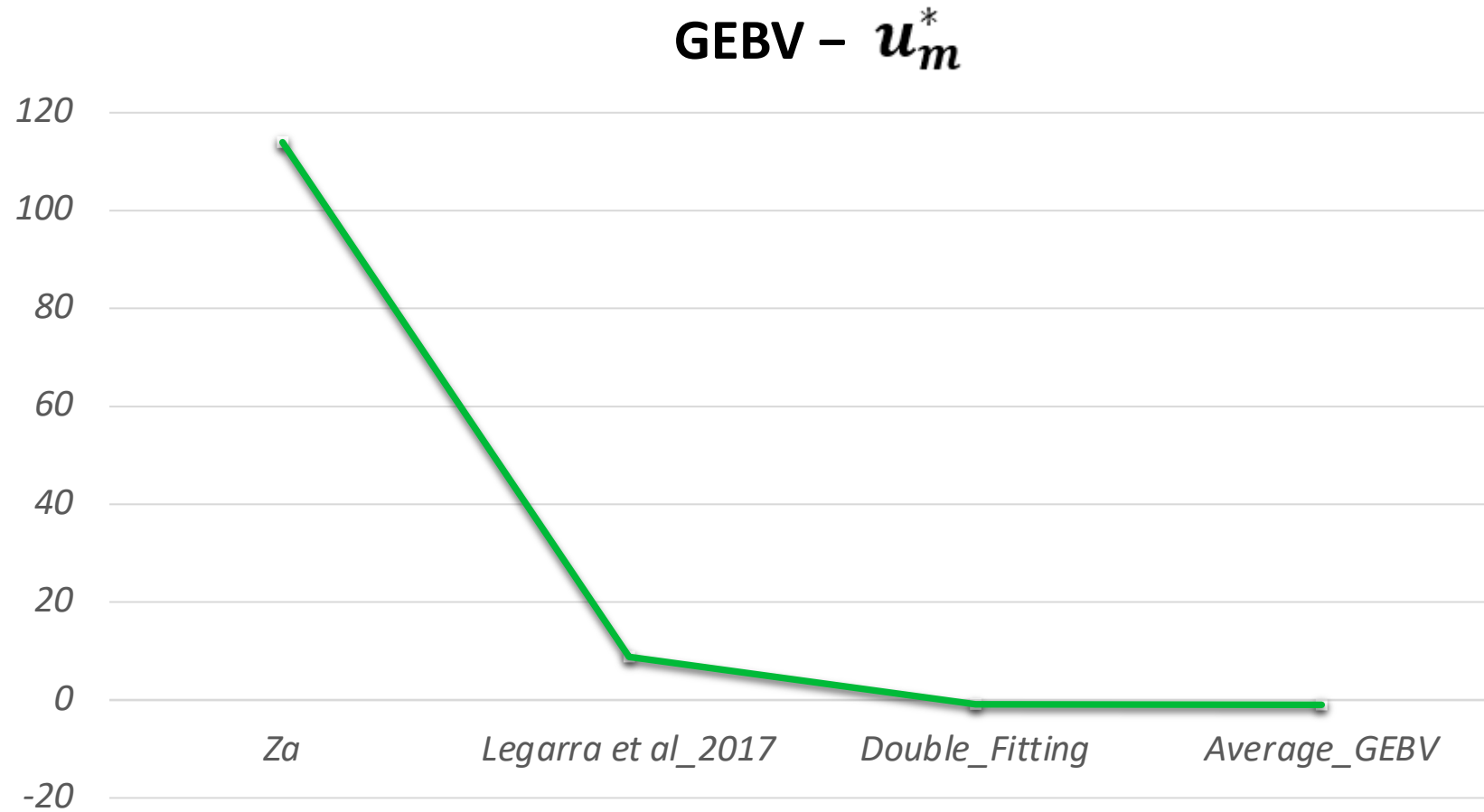
$$\mathbf{u}_m = \hat{\boldsymbol{\mu}} + \mathbf{u}_m^*$$

└─ $\hat{\boldsymbol{\mu}} = \alpha\lambda\mathbf{1}'\mathbf{G}^{-1}\hat{\mathbf{u}}$

α = blending parameter for \mathbf{G}

$$\lambda = \frac{1}{n^2} \left(\sum_i \sum_j \mathbf{A}_{22ij} - \sum_i \sum_j \mathbf{G}_{ij} \right)$$

1) Indirect Predictions



How to compute Indirect predictions

1) Pedigree + phenotypes + genotypes

2) **renumf90**

3) `preGSf90` to save clean files

4) **blupf90+** (with clean files: `OPTION no_quality_control`)

- Good practice to save time: `OPTION saveGInverse + OPTION saveA22Inverse`

5) **postGSf90** (with clean files)

- BLUPF90 family software to compute SNP effects (+more)
- Same parameter file as `blupf90+`
- Good practice to save time: `OPTION readGInverse + OPTION readA22Inverse`

Output from postGSf90

snp_sol

<http://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90>

contains solutions of SNP and weights

- 1: trait
- 2: effect
- 3: SNP
- 4: Chromosome
- 5: Position
- 6: SNP solution
- 7: weight

snp_pred

- 1st line: model, tuning, blending information
- 2nd line: Trait/effect info
- AF in 10 columns
- $\mu_{\hat{}}$, $\text{var}_{\hat{\mu}}$
- SNP effects

How to compute Indirect Predictions

6) `predf90`

- Reads `snp_pred`
- Have to provide a SNP file for the new genotyped animals to receive IP
 - same SNP as in the clean file

```
predf90 --snpfile newgen.txt --use_mu_hat
```

- The last statement adds the base, so that we have: $\mathbf{u}_m = \hat{\boldsymbol{\mu}} + \mathbf{u}_m^*$

Output from pred90

SNP_predictions

Animal ID SNP call rate Indirect Predictions

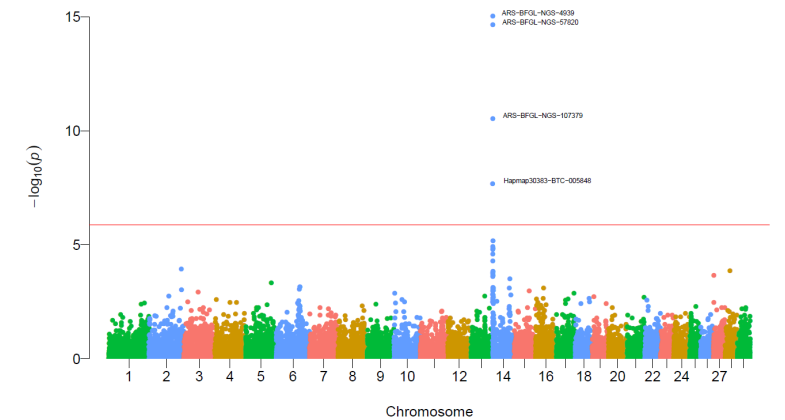
```
UGA50014    1.00    0.17414457
UGA50016    1.00    0.72332874E-01
UGA50042    1.00    1.0016705
UGA50058    1.00    0.17190497
UGA50060    1.00    0.98674759E-01
UGA50065    1.00    -0.60623702E-01
UGA50073    1.00    -0.17860851
UGA50077    1.00    -0.21597147
UGA50079    1.00    -0.69586390
UGA50084    1.00    1.0600574
UGA50085    1.00    -0.28602412
UGA50088    1.00    -0.12758011
```

pred90 can also compute accuracy of indirect predictions

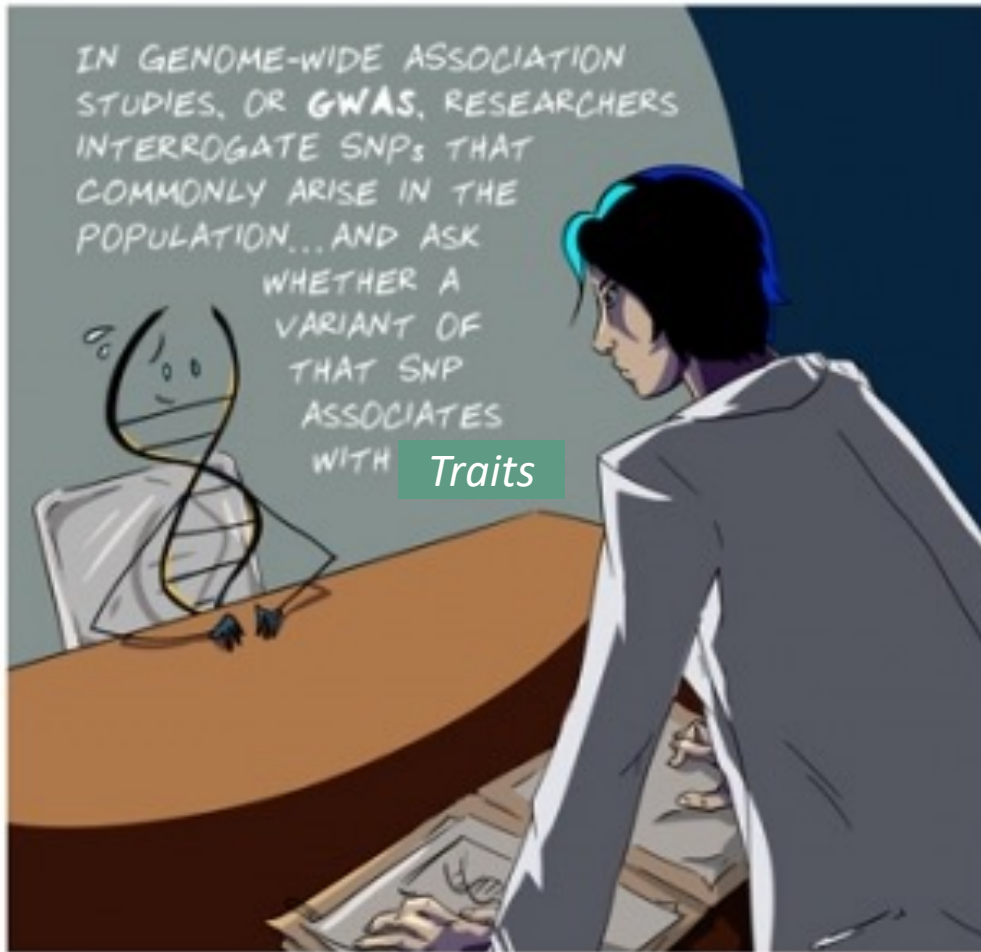
```
OPTION snp_p_value    #in blup90+
OPTION snp_var        #in postGS90
--acc                  #in pred90
```



2) Genome-wide Association Studies



Genome-wide association



YOU COULD THINK OF IT AS IF SOME SNPs ARE CARRYING A TINY CAMPAIGN SIGN SUGGESTING WHICH GENE THEY'RE ASSOCIATED WITH.

Adapted from:
<https://www.broadinstitute.org/visuals/explainer-genome-wide-association-studies>

Current standard for GWAS

- Single marker regression with **G** to compensate for relationships
 - $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{z}_i\mathbf{a}_i + \mathbf{u} + \mathbf{e}$
 - **z**: gene content {0,1,2}
 - **a**: SNP effect

What are we testing?

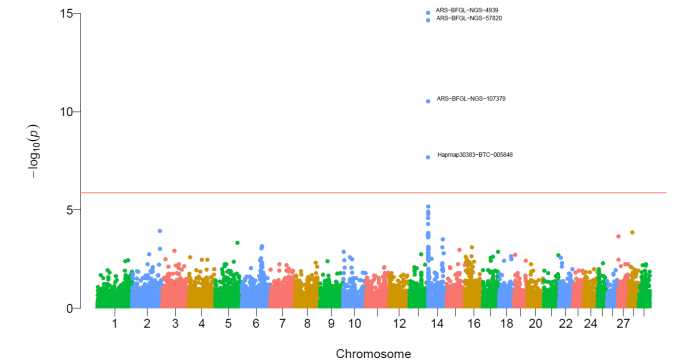
H_0 : genotypic classes do not differ in phenotype for a given SNP

H_1 : genotypic classes differ in phenotype for a given SNP

Example: do resistant and susceptible individuals have different genotypes at a given SNP?

Current standard for GWAS

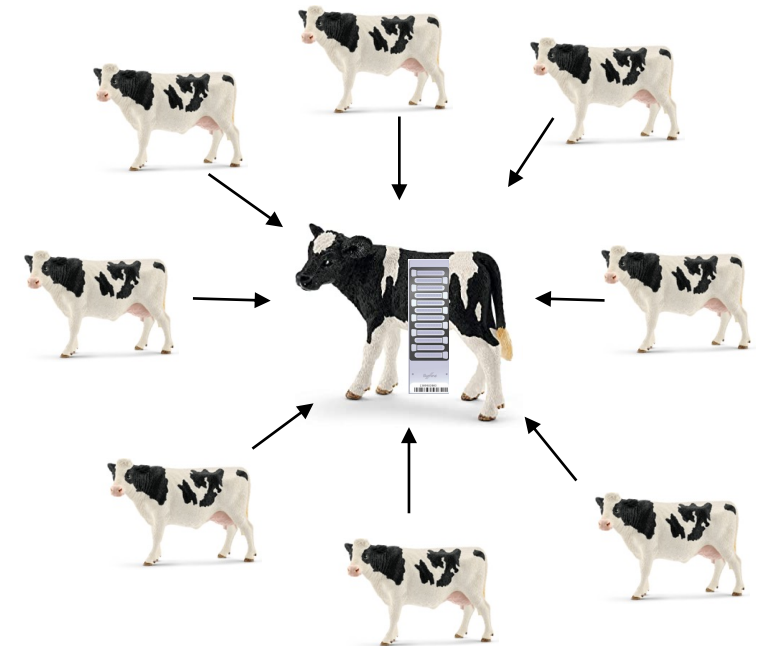
- Single marker regression with **G** to compensate for relationships
- Estimate SNP effects
- Get p-values as $pval_i = 2 \left(1 - \Phi \left(\left| \frac{\hat{a}_i}{sd(\hat{a}_i)} \right| \right) \right)$
- Apply Bonferroni to correct for multiple testing



- **Assumption: Genotyped individuals have phenotypes**

GWAS in livestock populations

- Most animals are non-genotyped
- Animals may not have phenotypes
- Some traits are sex-limited
 - milk, fat, protein
- Single marker regression
 - Only genotyped animals with phenotypes
 - Deregressed EBV
- Need a method that fits the livestock data
 - ssGWAS



Single-step GWAS (historical)

SNP
effects

GEBVs

$$\hat{\mathbf{a}} = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$$

VanRaden 2008
Stranden and Garrick 2009
Wang et al. 2012

a) Quadratic SNP variance (Falconer & Mackay, 1996)

$$d_i = \hat{a}_i^2 2p_i(1-p_i)$$

b) Nonlinear SNP variance (VanRaden, 2008)

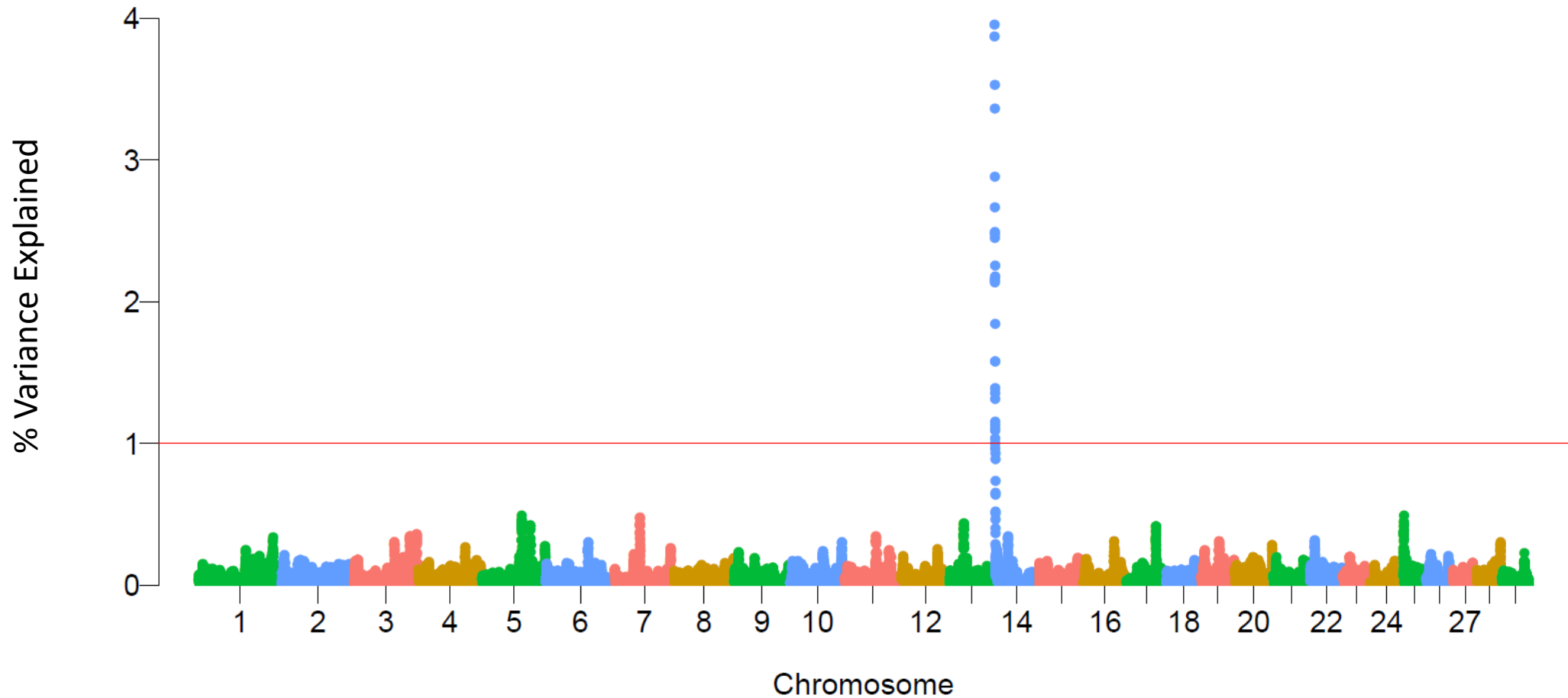
$$d_i = 1.125 \frac{|\hat{a}_i|}{sd(\hat{\mathbf{a}})} - 2$$

Single-step GWAS

Fat – US Holsteins

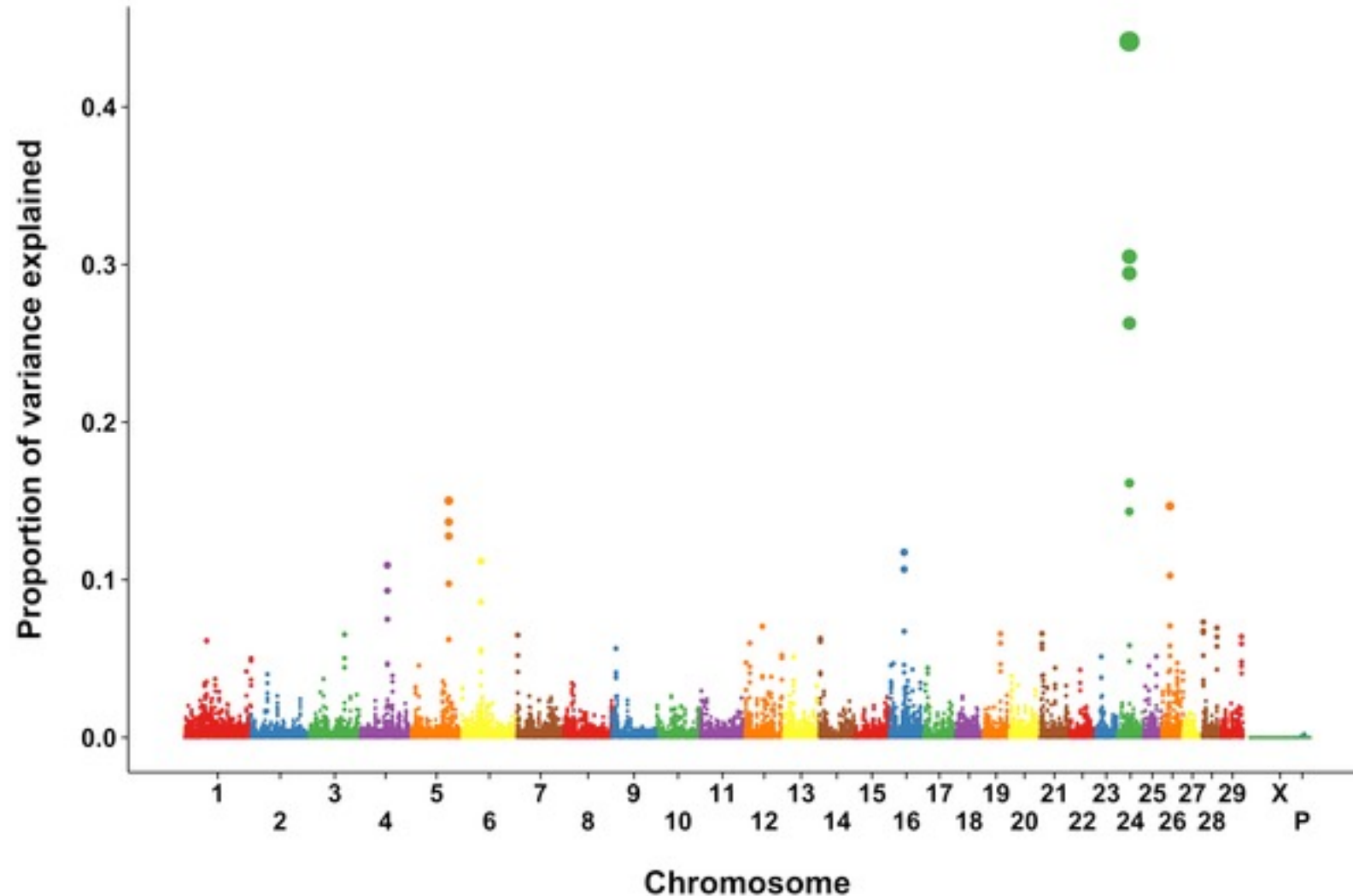
No P-value!!!

Manhattan plot of Variances



Single-step GWAS

Figure 2. Proportion of SNP variance explained by 5-SNP moving windows for rectal temperature from a **single-step GBLUP** analysis



No P-value!!!

Can we have p-values in ssGWAS?

Gualdrón Duarte et al. *BMC Bioinformatics* 2014, 15:246
<http://www.biomedcentral.com/1471-2105/15/246>



METHODOLOGY ARTICLE

Open Access

Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations

Jose L. Gualdrón Duarte¹, Rodolfo J.C. Cantet¹, Ronald O. Bates², Catherine W. Ernst², Nancy E. Raney² and Juan P. Steibel^{2,3*}

Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods

Chunyu Chen,¹ Juan P. Steibel, and Robert J. Tempelman
Department of Animal Science, Michigan State University, East Lansing, Michigan 48824
ORCID ID: 0000-0002-7833-6730 (R.J.T.)

Aguilar et al. *Genet Sel Evol* (2019) 51:28
<https://doi.org/10.1186/s12711-019-0469-3>

SHORT COMMUNICATION

Open Access

Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle

Ignacio Aguilar¹, Andres Legarra^{2*}, Fernando Cardoso^{3,4}, Yutaka Masuda⁵, Daniela Lourenco⁵ and Ignacy Misztal⁵

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics
doi: 10.1111/age.12378

Meta-analysis of genome-wide association from genomic prediction models

Y. L. Bernal Rubio^{*†}, J. L. Gualdrón Duarte^{*}, R. O. Bates^{*}, C. W. Ernst^{*}, D. Nonneman[‡], G. A. Rohrer[‡], A. King[‡], S. D. Shackelford[‡], T. L. Wheeler[‡], R. J. C. Cantet^{†§} and J. P. Steibel^{*¶}



J. Dairy Sci. 101:3140–3154
<https://doi.org/10.3168/jds.2017-13364>
© American Dairy Science Association[®], 2018.

Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency

Y. Lu,^{*} M. J. Vandehaar,^{*} D. M. Spurlock,[†] K. A. Weigel,[‡] L. E. Armentano,[‡] E. E. Connor,[§] M. Coffey,[#] R. F. Veerkamp,^{||} Y. de Haas,^{||} C. R. Staples,[¶] Z. Wang,^{**} M. D. Hanigan,^{††} and R. J. Tempelman^{*†}



P-values in ssGWAS

1) Factorize and Invert LHS of ssGBLUP with YAMS (Masuda et al., 2014)

2) Solve the MME for $\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix}$ using the sparse Cholesky factor

3) Extract coefficients for genotyped animals ($\mathbf{C}^{u_2 u_2}$) from LHS⁻¹

4) Obtain individual prediction error variance of SNP effects:

$$Var(\hat{a}_i) = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{z}'_i \mathbf{G}^{-1} (\mathbf{G}\sigma_u^2 - \mathbf{C}^{u_2 u_2}) \mathbf{G}^{-1} \mathbf{z}_i \frac{1}{2\sum p_i(1-p_i)} \alpha b$$

(Gualdron-Duarte et al., 2014)

5) Backsolve GEBV to SNP effects (\hat{a}): $\hat{a} = \alpha b \frac{1}{2\sum p_i q_i} \mathbf{Z}' \mathbf{G}^{-1} \hat{u}$

$$6) p\text{-value}_i = 2 \left(1 - \Phi \left(\left| \frac{\hat{a}_i}{sd(\hat{a}_i)} \right| \right) \right)$$

Φ is the cumulative standard normal function

blupf90+



Ignacio
Aguilar



Andres
Legarra

postGSf90



Yutaka
Masuda

How to run ssGWAS with p-values in BLUPF90

- After renumf90 and preGSf90 to save clean files:
 - blupf90+ to estimate GEBV
 - OPTION SNP_file `snp.dat_clean`
 - OPTION map_file `mrkmap.txt_clean`
 - OPTION saveGInverse
 - OPTION saveA22Inverse
 - OPTION snp_p_value
 - OPTION no_quality_control
 - postGSf90 to backsolve GEBV to SNP effect
 - OPTION SNP_file `snp.dat_clean`
 - OPTION map_file `mrkmap.txt_clean`
 - OPTION readGInverse
 - OPTION readA22Inverse
 - OPTION snp_p_value
 - OPTION windows_variance X #if need variance explained by X SNP
 - OPTION no_quality_control

Output from postGSf90

chr_{sn}_pval

contains data to create plot by GNUPLOT

- 1: trait
- 2: effect
- 3: $-\log_{10}(\text{p-value})$
- 4: SNP
- 5: Chromosome
- 6: Position in bp

Pft1e2.gnuplot

Pft1e2.R

chr_{sn}

contains data to create plot by GNUPLOT

- 1: trait
- 2: effect
- 3: values of SNP effects to use in Manhattan plots $\rightarrow [\text{abs}(\text{SNP}_i)/\text{var}(\text{SNP})]$
- 4: SNP
- 5: Chromosome
- 6: Position

Sft1e2.gnuplot

Sft1e2.R

Output from postGSf90

```
chr SNP var
```

contains data to create plot by GNUPLOT

- 1: trait
- 2: effect
- 3: variance explained by n adjacents SNP
- 4: SNP
- 5: Chromosome
- 6: Position

Vft1e2.gnuplot

Vft1e2.R

Output from postGSf90

snp_sol

contains solutions of SNP and weights

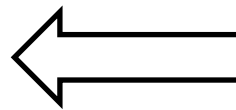
- 1: trait
- 2: effect
- 3: SNP
- 4: Chromosome
- 5: Position
- 6: SNP solution
- 7: weight

if `OPTION windows_variance` is used

- 8: variance explained by n adjacents SNP.

if `OPTION snp_p_value` is used

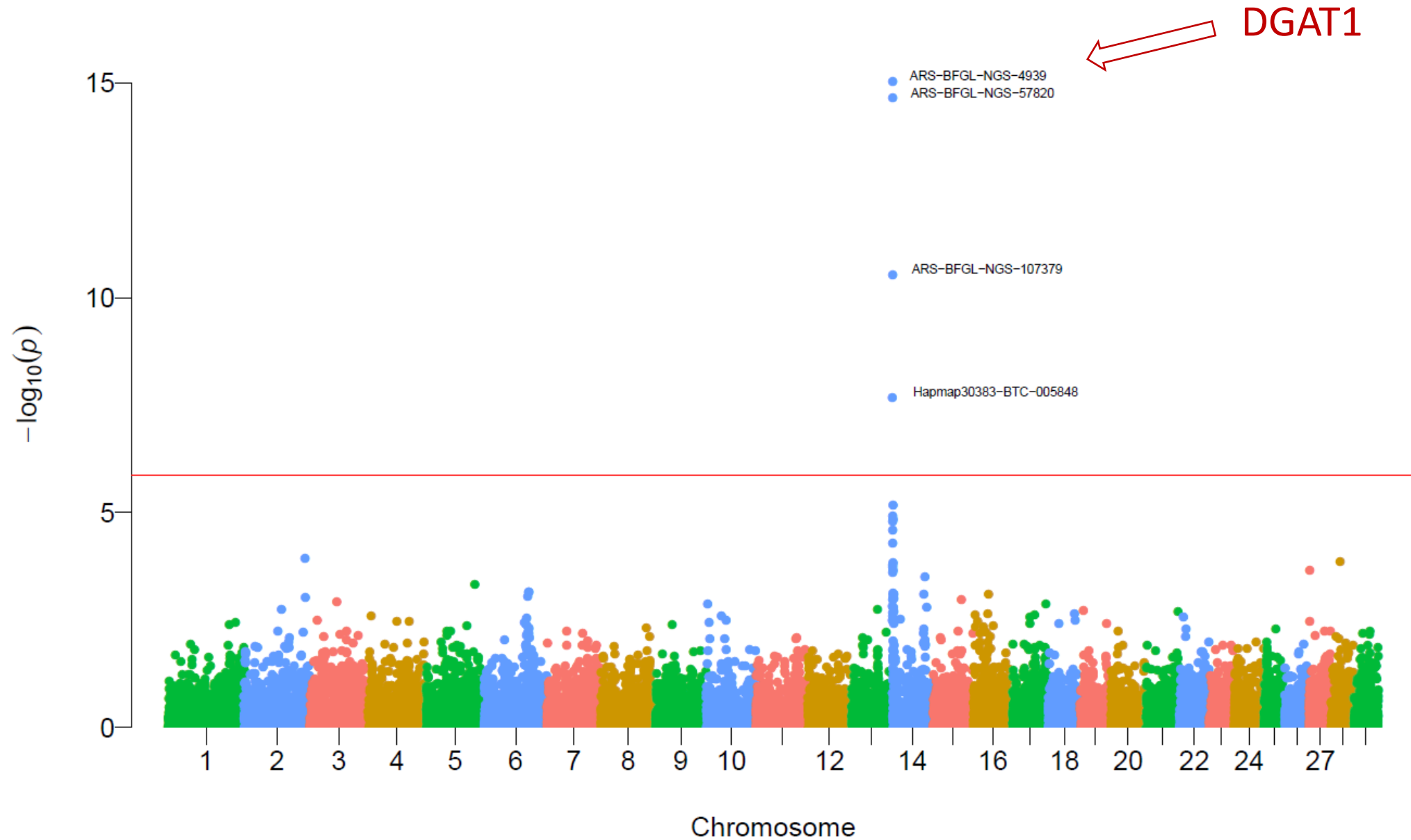
- 9: variance of the SNP solution (used to compute the p-value)



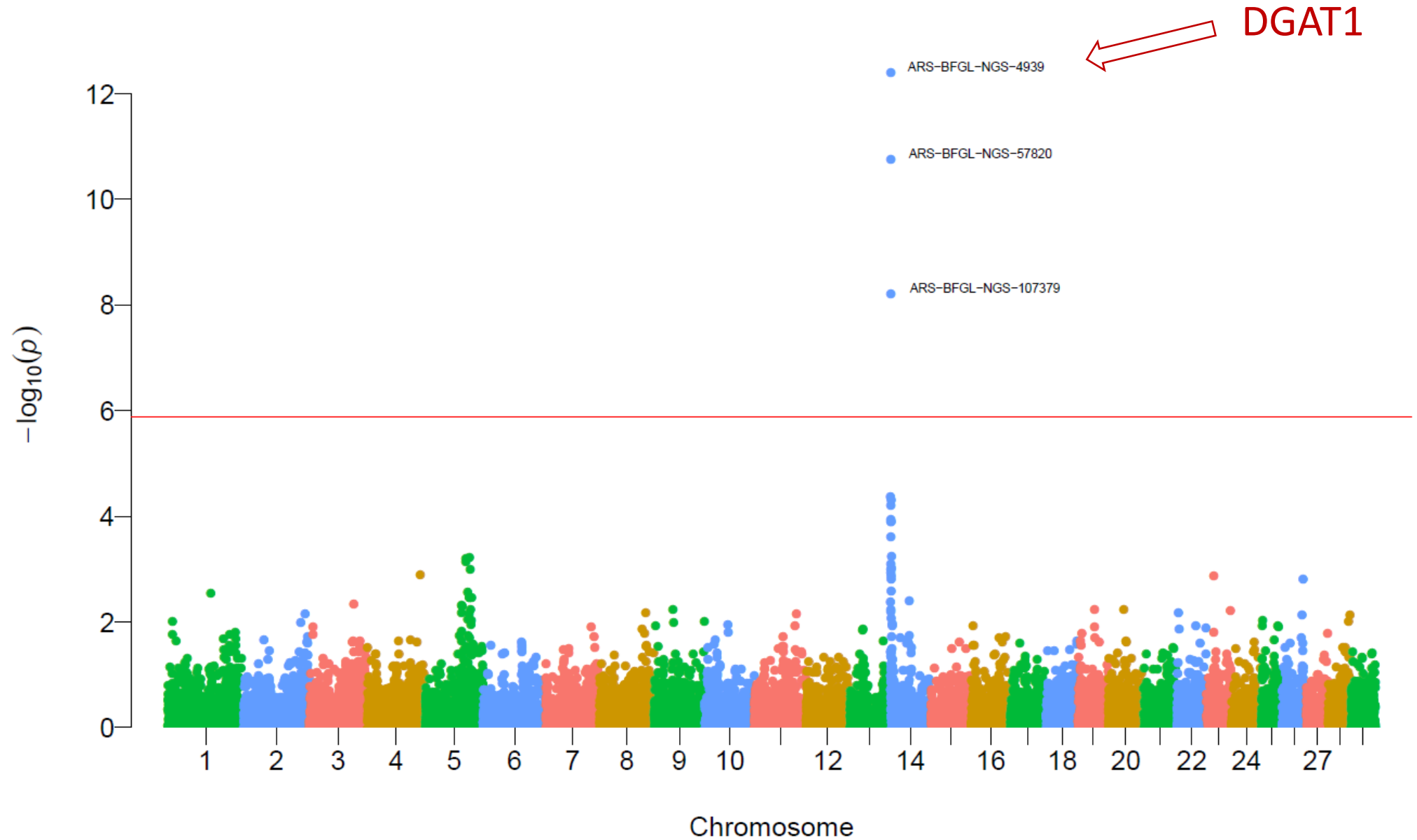
P-values in ssGWAS for US Holsteins

- US HOL 2009 data: milk, fat, protein
- Single-trait models
 - 10k genotyped bulls
 - 752k records for 100k daughters
 - 303k animals in ped

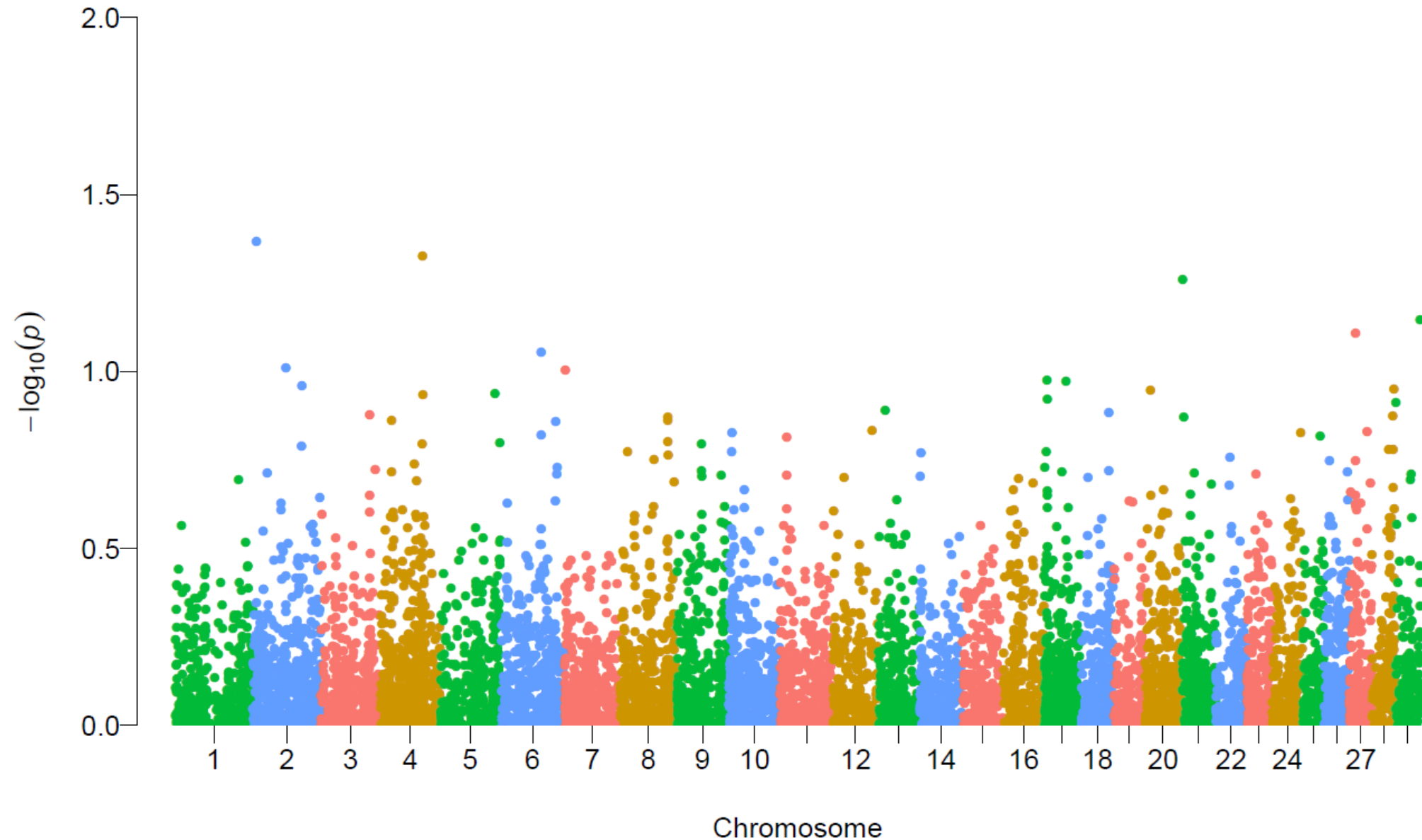
P-values in ssGWAS - Milk



P-values in ssGWAS - Fat

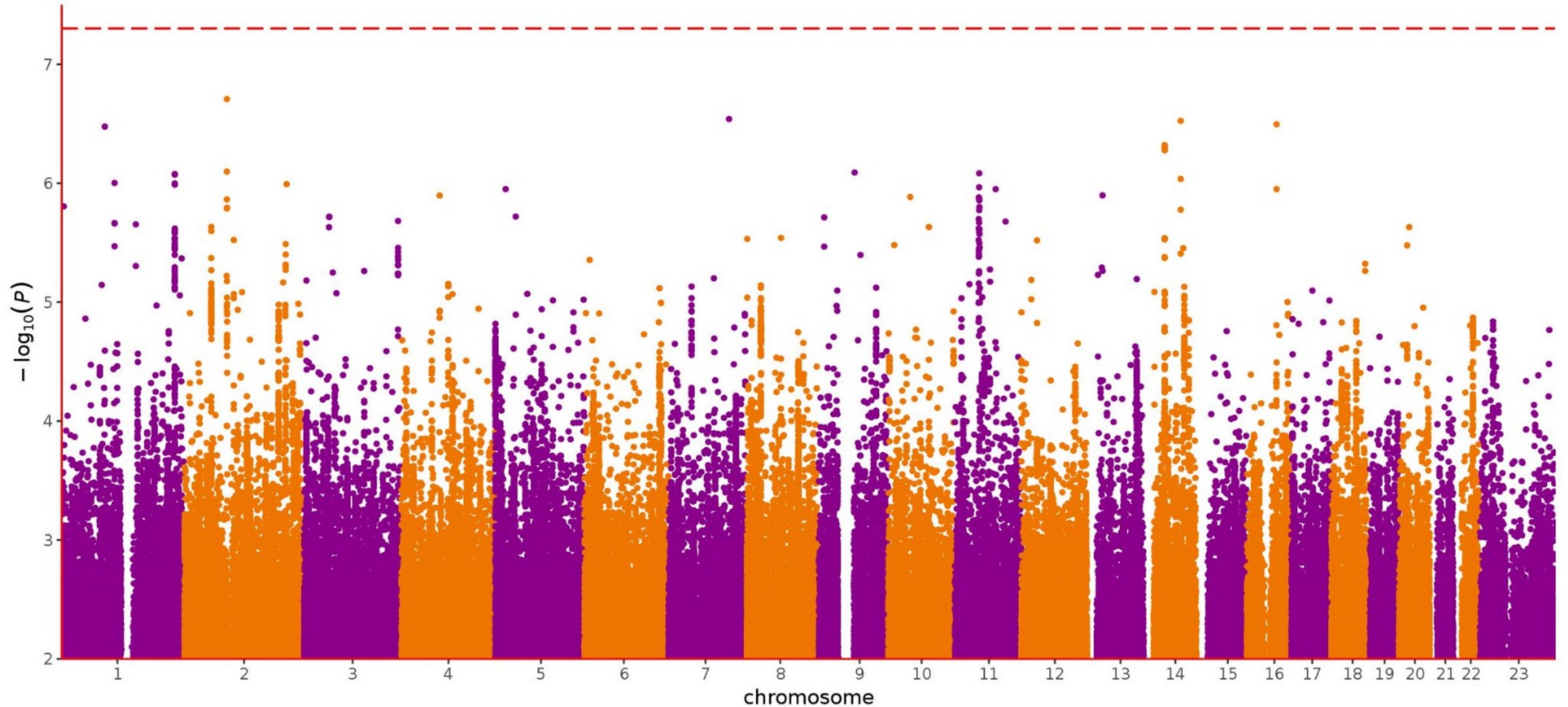


P-values in ssGWAS - Protein

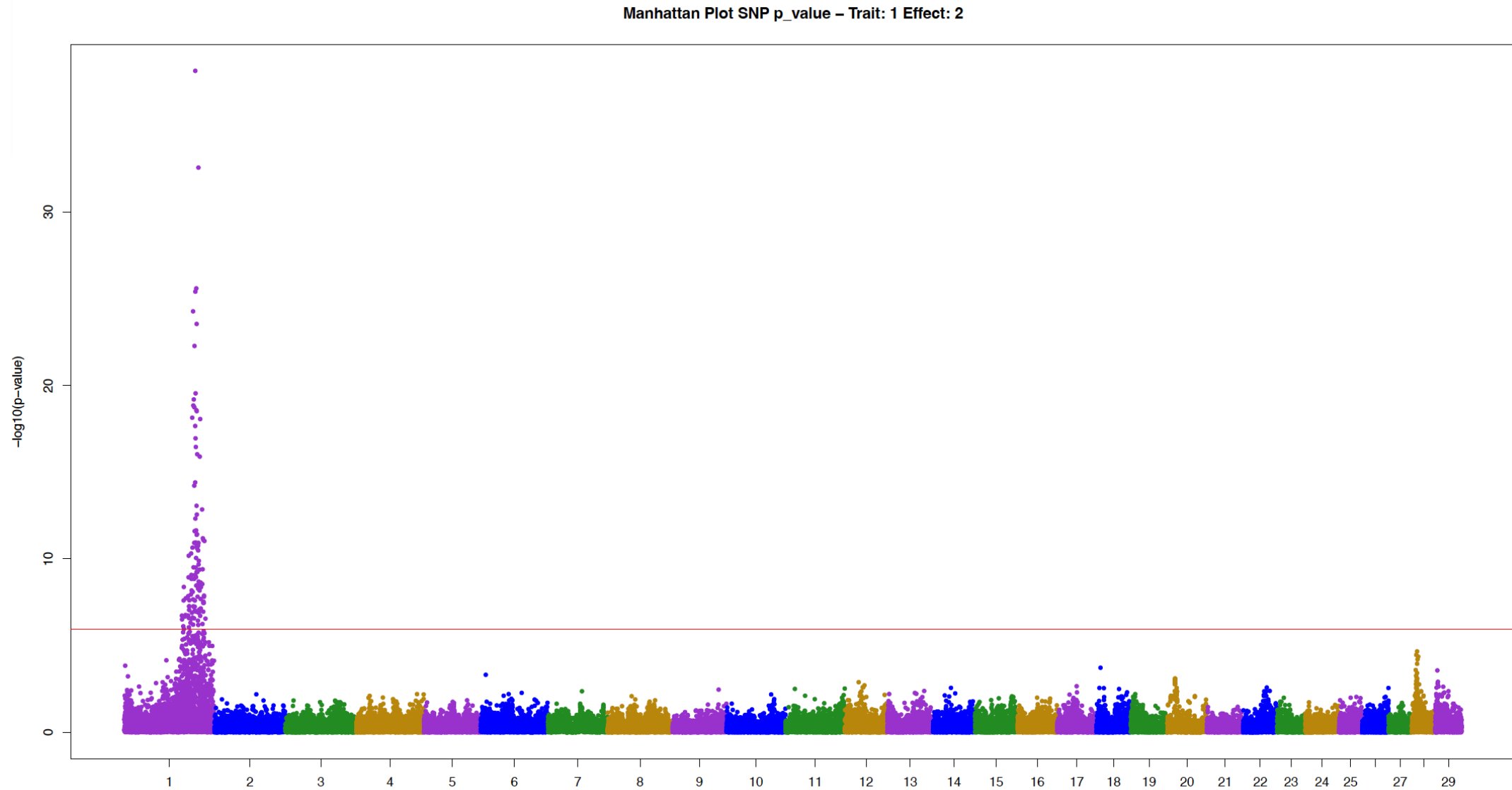


Non-significant hits

Work/job satisfaction
N=82190

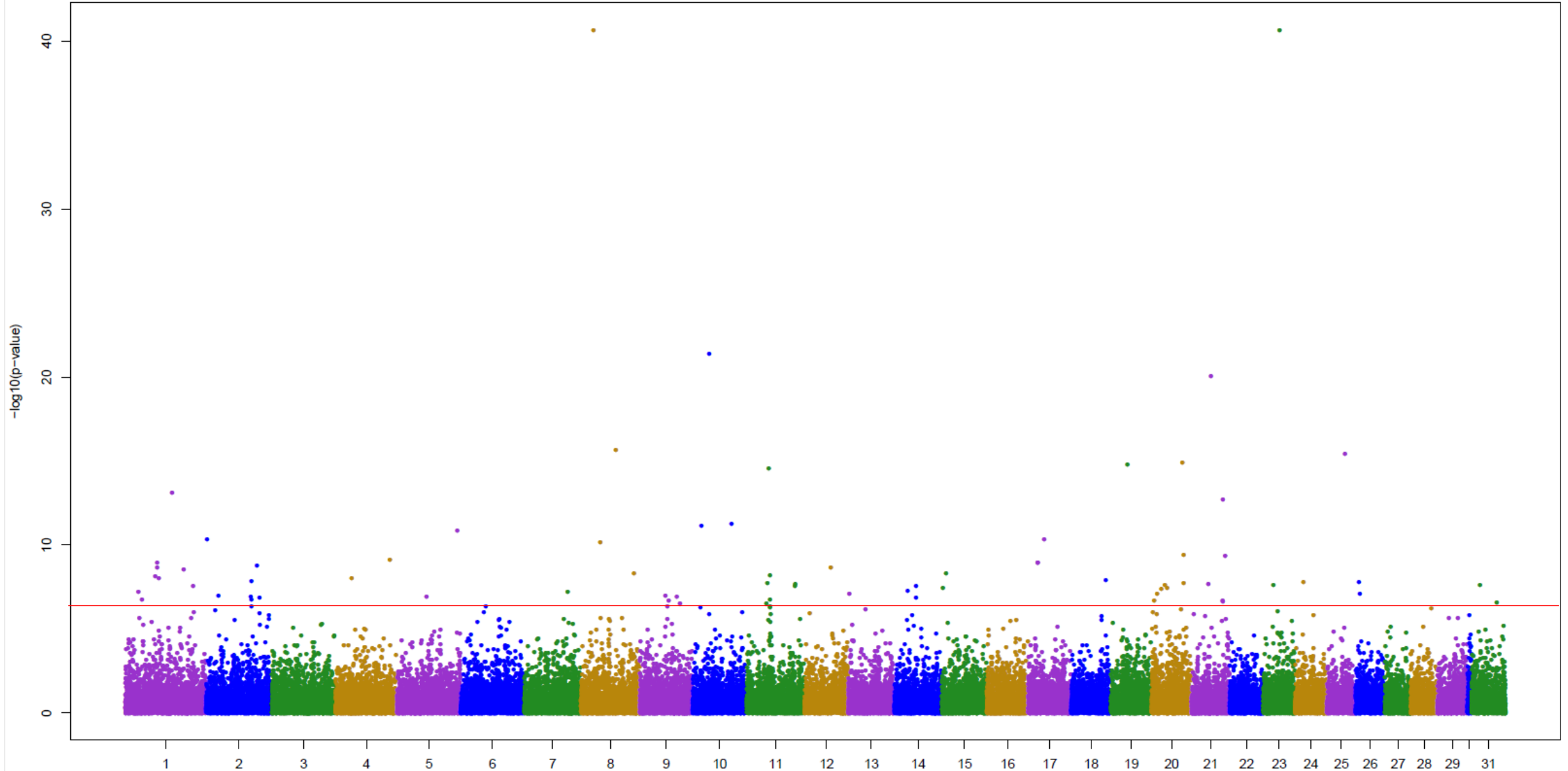


Manhattan plots we want to see



Manhattan plots we do NOT want to see

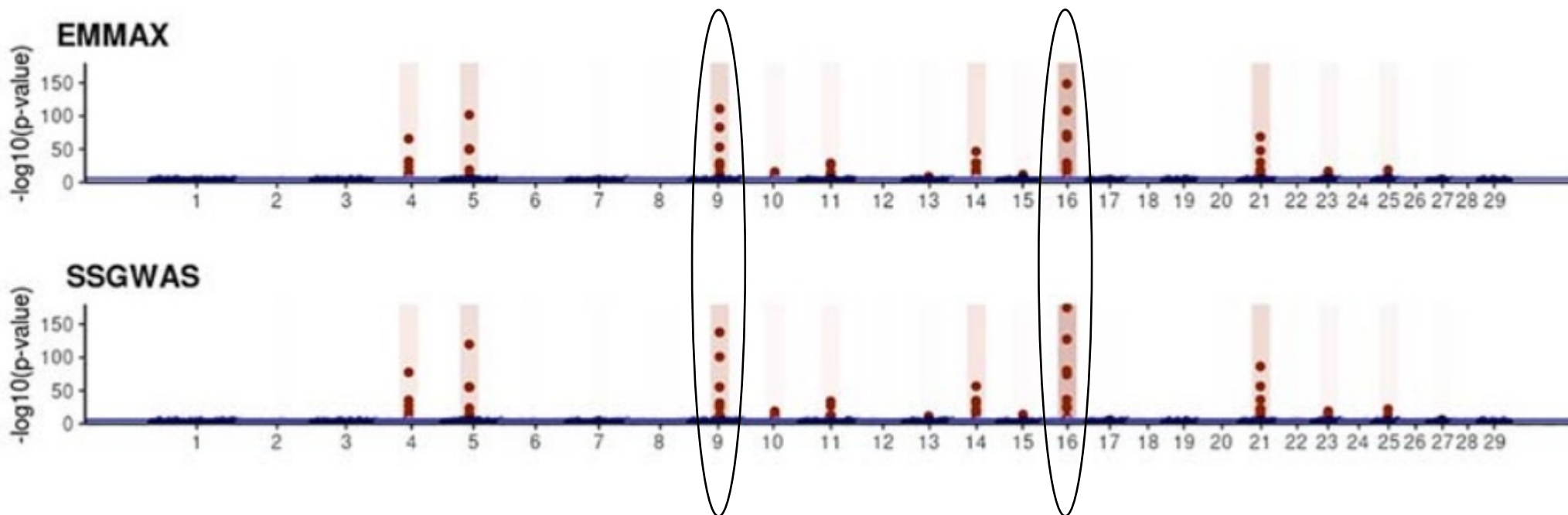
Manhattan Plot SNP p_value - Trait: 1 Effect: 5



ssGWAS vs. EMMAX

- Simulated dairy population (1 QTN per CHR)

14k genotyped sires
Deregressed EBV
(10 daughters)

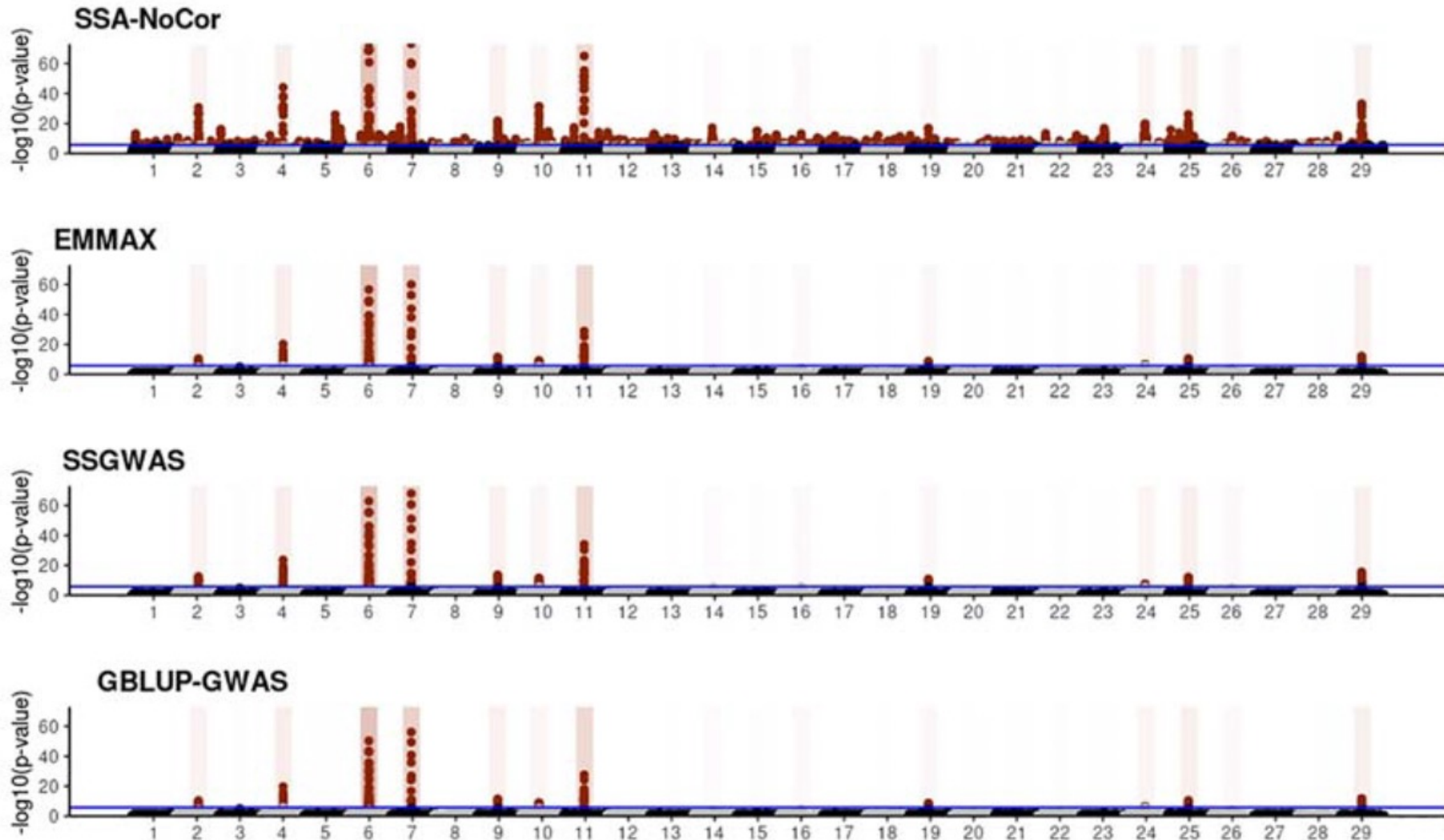


14k genotyped sires
500k Pedigree
250k phenotypes

Association	EMMAX (Khang et al., 2010)	ssGWAS (Aguilar et al., 2019)
True Positive	55.2 ^a (3.7)	61.6 ^a (8.7)
False Positive	0.0	0.0

ssGWAS vs. EMMAX

- Simulated fish population (1 QTN per CHR)



ssGWAS

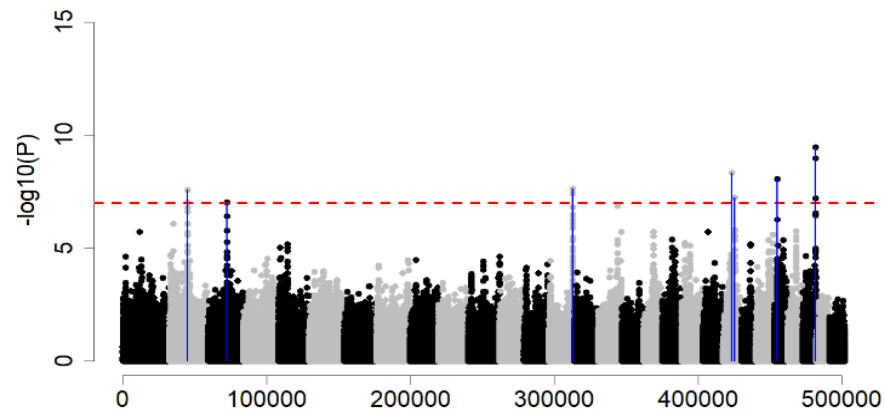
- ssGWAS works!!!
- Heavy computations
- Soft limit is the same as REML
 - 10k genotyped animals
 - 1M animals in pedigree
 - 1M phenotypes
- Limited amount of information in ssGWAS

GWAS vs. amount of information

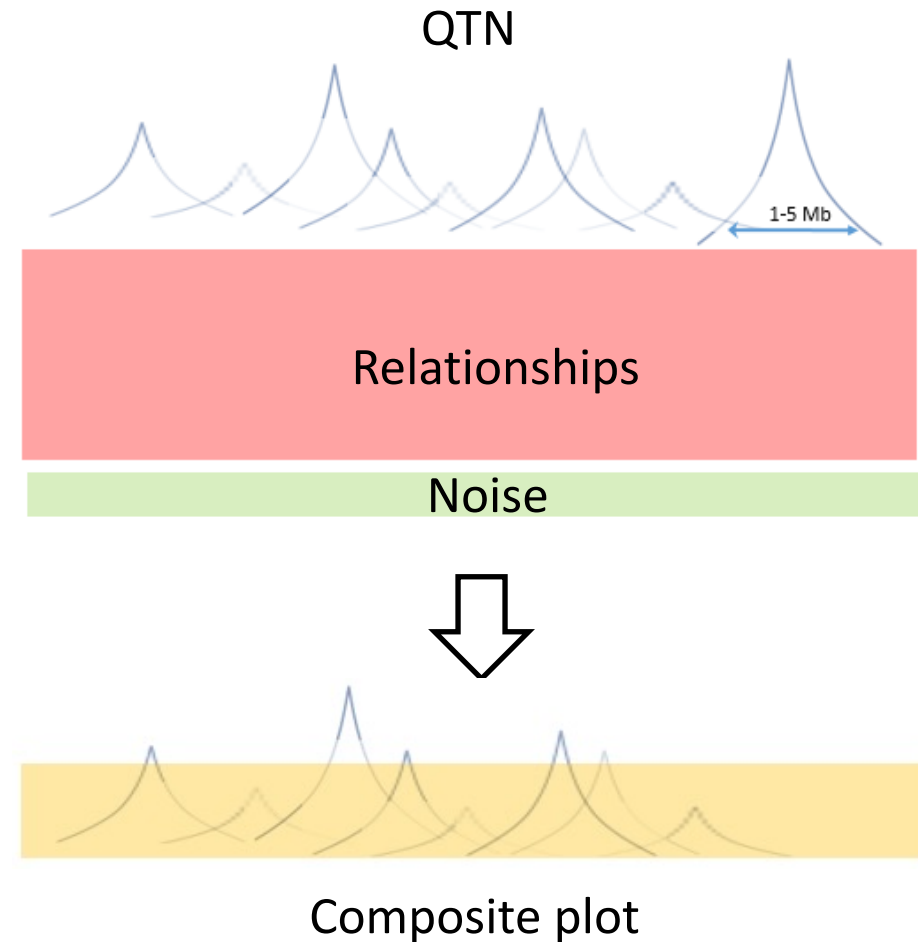
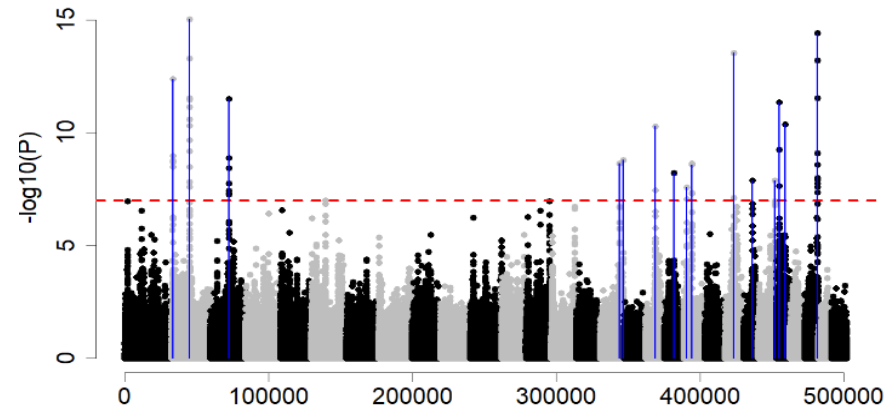
- Amount of information to identify causative variants

Ne=200 QTN=2000

Eig98 N=15,200



All N=30,000



Jang et al.
(2023)

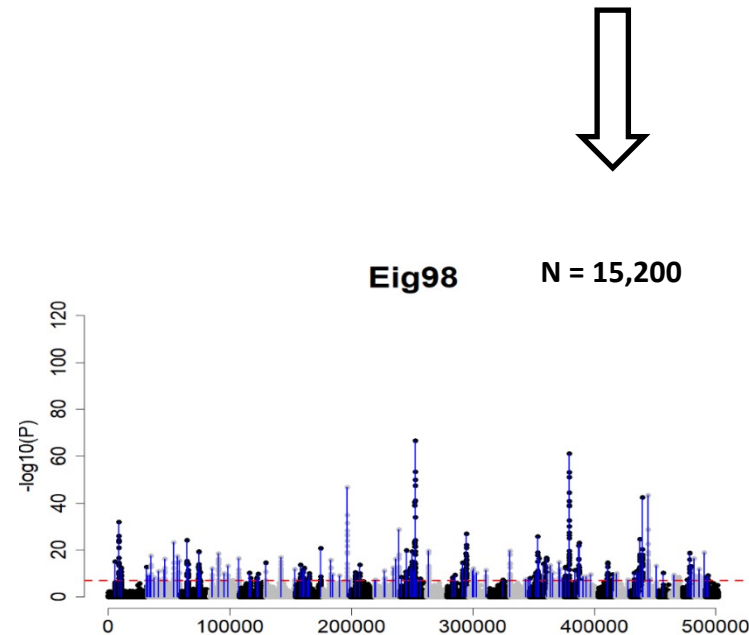
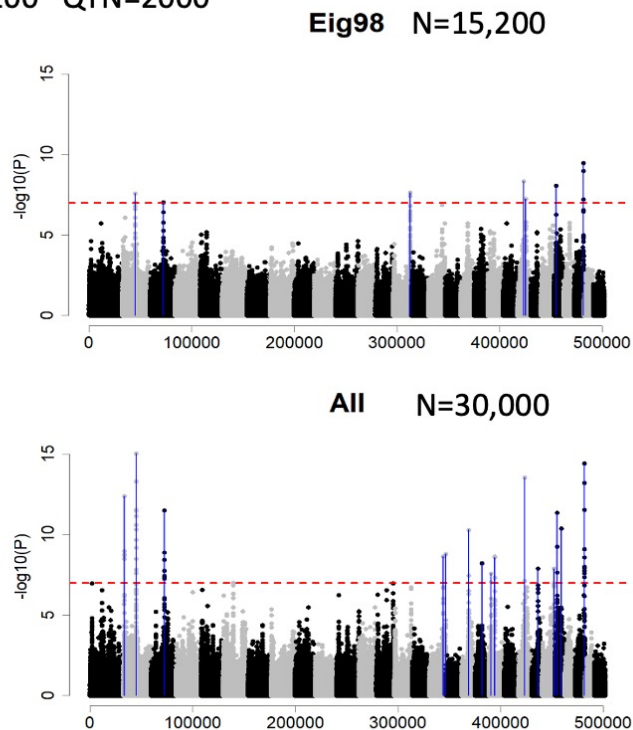
GWAS vs. amount of information

- Amount of information to identify causative variants
 - Animal with lots of information
 - GEBV accuracy ~ 0.99
 - GEBV backsolved to SNP effects
 - GWAS resolution with sample size = Me = Eig98 animals with almost perfect accuracy

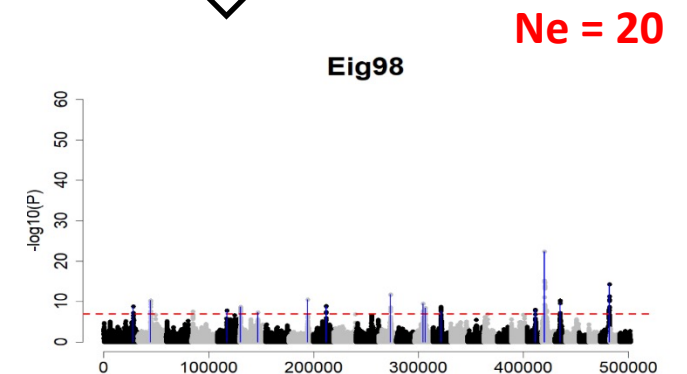


Jang et al.
(2023)

$N_e=200$ $QTN=2000$



Lots of records for each
genotyped animal



Ne vs. Segments

Theory of junctions Fisher (1949)

$$E(Me) = 4N_e L$$

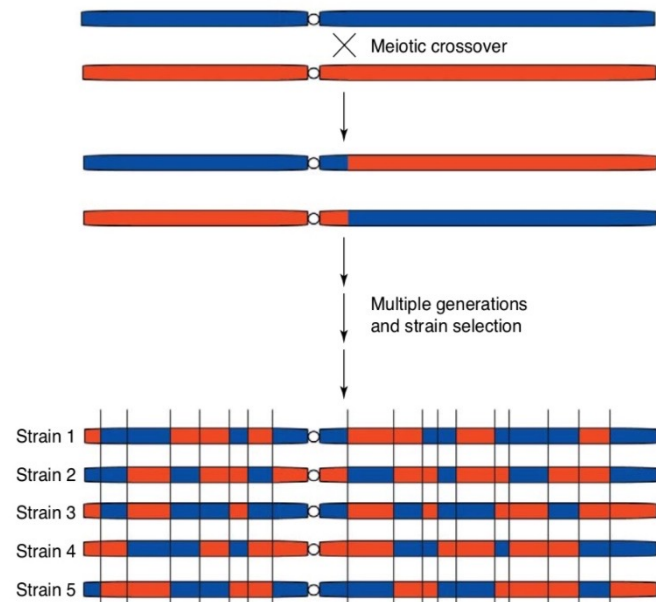
Stam (1980)

Points where the founder chromosome of origin changes

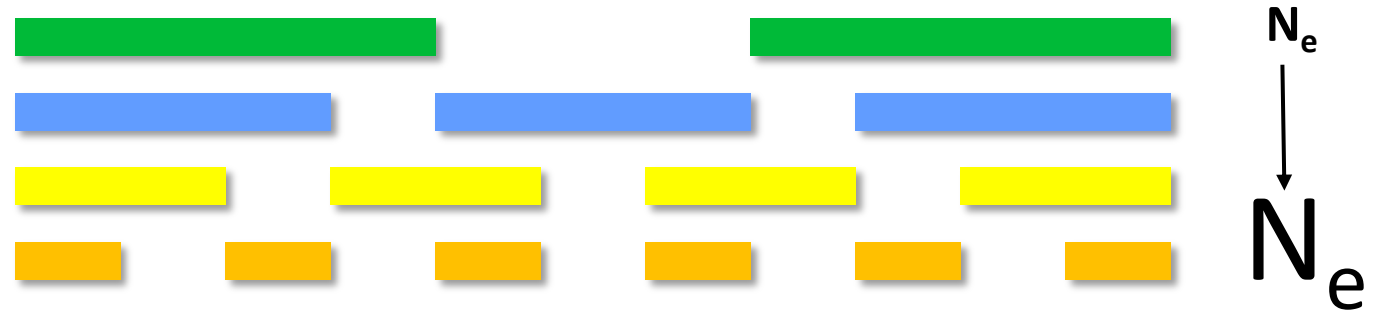
Me – Independent chromosome segments

N_e – Effective population size

L – Length of genome in Morgans



Cuppen (2005)

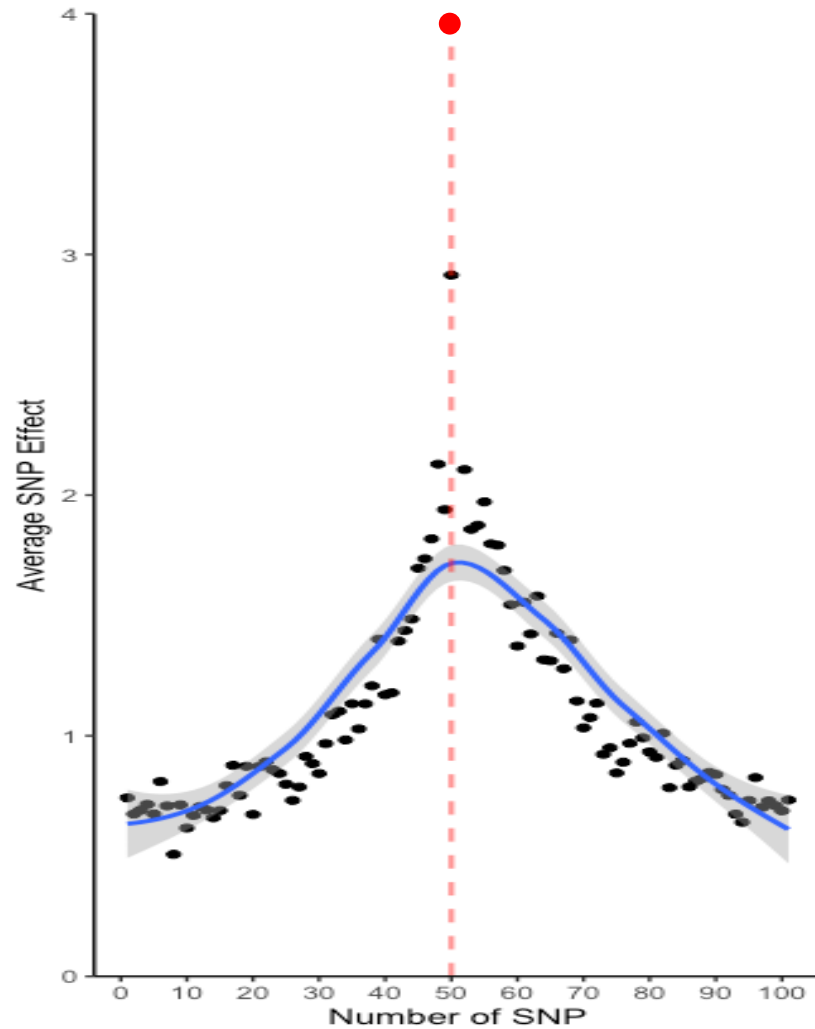


Limited # segments → limited dimensionality

Misztal (2016)

Finding causative SNP

Ne = 60
Animals = 6000



Single nucleotide polymorphism profile for quantitative trait nucleotide in populations with small effective size and its impact on mapping and genomic predictions

Ivan Pocrnic ^{1,*†} Daniela Lourenco ¹ Ignacy Misztal ^{1,*}

N_e vs. ability to find SNP

