SNP effects and weights from ssGBLUP using BLUPF90

Daniela Lourenco

UGA Team – 08/2019

SNP effect and weights in ssGBLUP

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W+H^{-1}\lambda} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

Matrix of SNP weights

Matrix of SNP content

$$\hat{\boldsymbol{a}} = \sigma_u^{-2}\, \sigma_a^2\, \mathbf{D}\, \mathbf{Z'G}^{-1}\hat{\mathbf{u}}$$

Genomic relationship matrix

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

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

b) Nonlinear A SNP weights (VanRaden, 2008)

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

SNP effect in ssGBLUP

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W+H^{-1}\lambda \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Matrix of SNP weights

Matrix of SNP content

$$\hat{a} = \sigma_u^{-2}\,\sigma_a^2\,\mathbf{D}\,\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$$

Genomic relationship matrix

- What else to do with weights SNP effects?

  1) Indirect predictions for animals not included in the evaluation

$$\mathbf{DGV} = \mathbf{Z}\hat{a}$$

Weights or variances for SNP in ssGBLUP

- What to do with weights or variance for SNP?

  1) Single-step Genome-Wide Association Studies - ssGWAS

     proportion of σ_u^2 explained by SNP

  2) Weighted single-step GBLUP - WssGBLUP

     different σ_a^2 for each SNP when constructing **G**

Weighted single-step GBLUP - WssGBLUP

Weights for SNP in ssGBLUP

- ssGBLUP
  - Same weights for SNP

$$G = \frac{ZZ'}{2 \sum p_i(1-p_i)}$$

- WssGBLUP
  - Different weights for SNP

$$G = \frac{ZDZ'}{2 \sum p_i(1-p_i)}$$

- Weights may increase accuracy of GEBV
- If SNPs explain high %variance

SNP weighting in ssGBLUP: WssGBLUP

- Wang et al. (2012):

"Iterative method needs convergence"

1) Set $\mathbf{D}_t = \mathbf{I}$ and $\mathbf{G}_t = \dfrac{\mathbf{ZDZ'}}{2\sum p_i(1-p_i)}$

2) Compute GEBV using ssGBLUP approach

Diagonal matrix of weights

3) Compute SNP effects as $\hat{a} = \lambda \mathbf{D} \, \mathbf{Z'G}^{-1}\widehat{\mathbf{GEBV}}$

4) Calculate SNP weight

5) Normalize $\mathbf{D}_{(t+1)}$

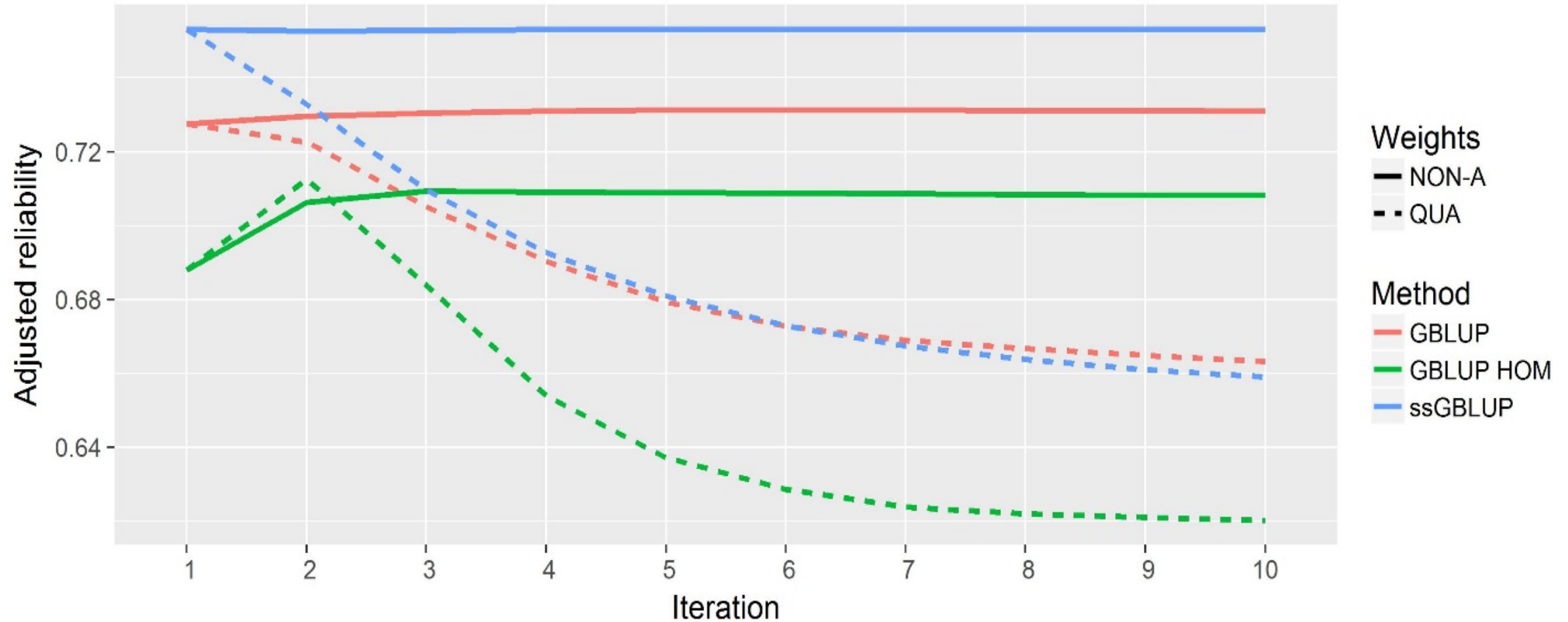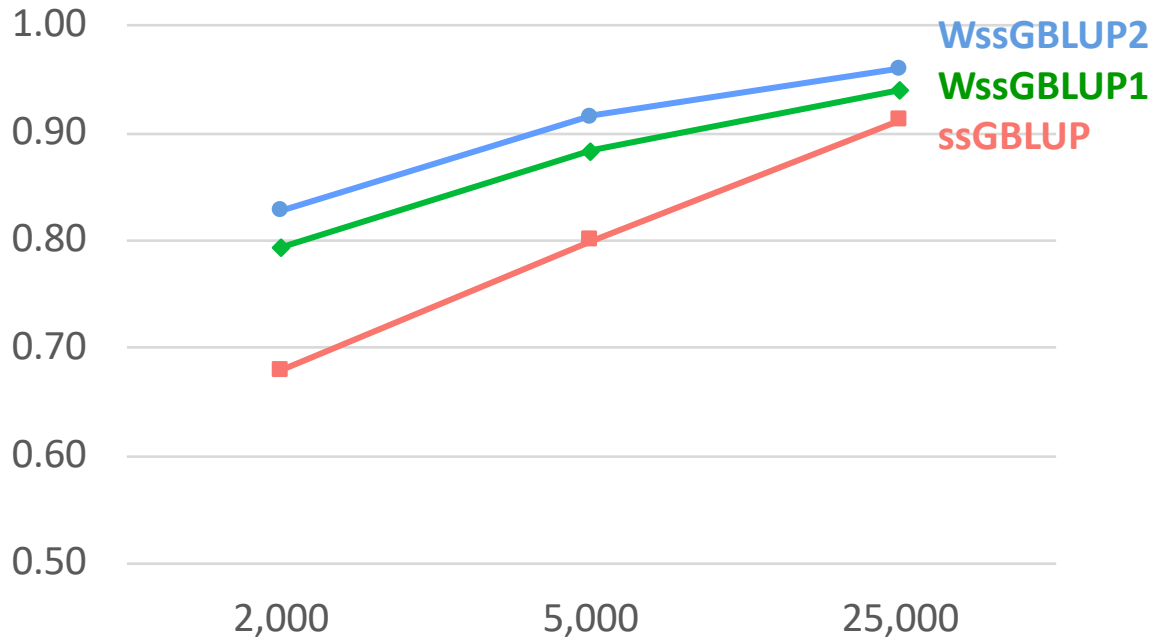6) $\mathbf{G}_{(t+1)} = \dfrac{\mathbf{ZD}_{(t+1)}\mathbf{Z'}}{2\sum p_i(1-p_i)}$

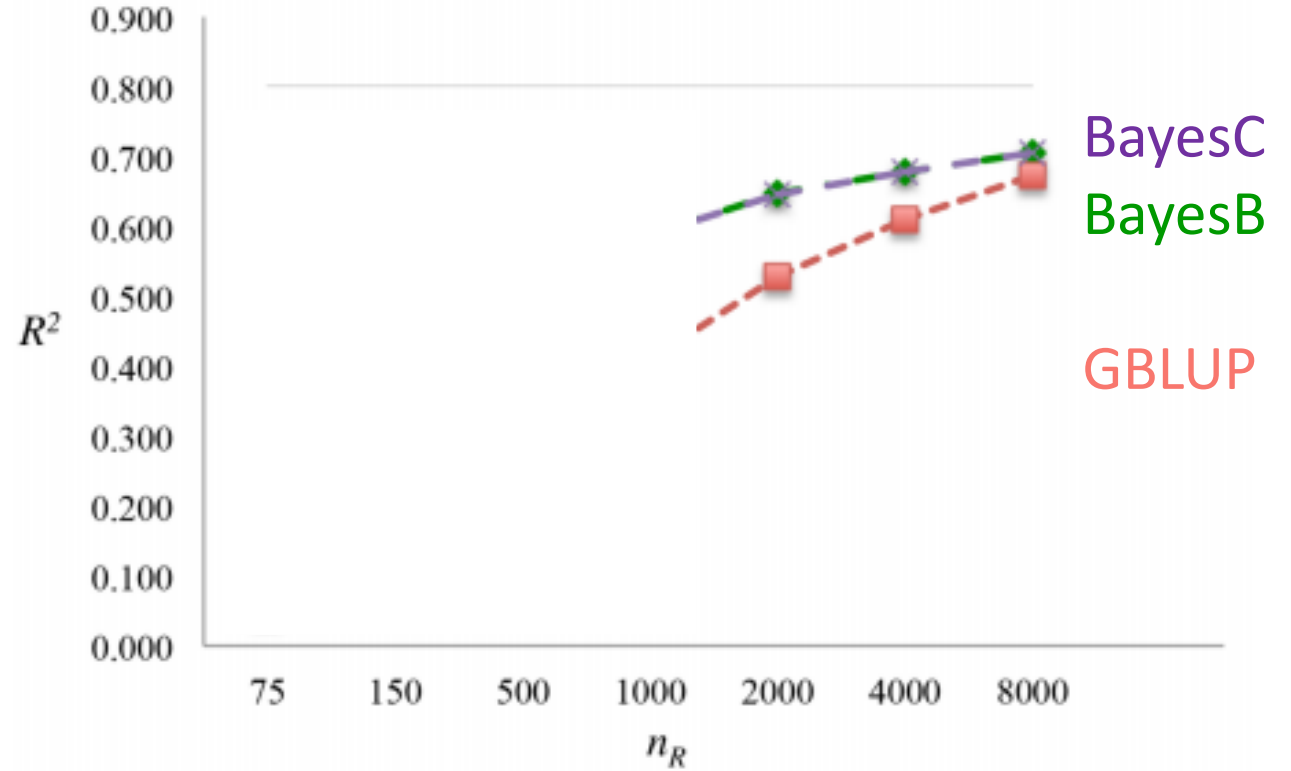# Convergence for nonlinear A and quadratic weight

WssGBLUP for large populations

10 QTN

Lourenco et al. (2017)

Karaman et al. (2016)

How to compute SNP effect and weight in BLUP90?

- After renumf90 and preGSf90 to save clean files:

    - blupf90 to estimate GEBV
        - `OPTION SNP_file snp.dat`
        - `OPTION chrinfo mrkmap.txt`
        - `OPTION saveGInverse`
        - `OPTION saveA22Inverse`

    - postGSf90 to backsolve GEBV to SNP effect
        - `OPTION SNP_file snp.dat`
        - `OPTION chrinfo mrkmap.txt`
        - `OPTION readGInverse`
        - `OPTION readA22Inverse`
        - `OPTION which_weight nonlinearA`

OPTION which_weight nonlinearA

`OPTION which_weight` nonlinearA

This option assumes the default constant (CT) is 1.125. To change the constant value to reflect a distribution closer to

normal, use a CT value closer to 1:

`OPTION which_weight` nonlinearA 1.05

By default, the maximum change in SNP variance is limited to 5, which is calculated as $CT^{(5-2)}$ and returns a value of

1.4238 with CT=1.125. If this limit is to be changed to 10, the following option can be used, where the value provided (x)

is the result of the expression $CT^{(x-2)}$. As an example, if CT is 1.05 and x is 10, the value provided to the option should be

1.4775:

`OPTION SNP_variance_limit` 1.4775

How to run WssGBLUP in BLUPF90

- After renumf90 and preGSf90 to save clean files:

  - Blupf90 to estimate GEBV
    - `OPTION SNP_file snp.dat`
    - `OPTION chrinfo mrkmap.txt`
    - `OPTION saveGInverse`
    - `OPTION saveA22Inverse`
    - `OPTION weightedG w.txt     #vector of weights`

  - postGSf90 to backsolve GEBV to SNP effect
    - `OPTION SNP_file snp.dat`
    - `OPTION chrinfo mrkmap.txt`
    - `OPTION readGInverse`
    - `OPTION readA22Inverse`
    - `OPTION which_weight nonlinearA`
    - `OPTION weightedG w.txt     #vector of weights`
    - `OPTION windows_variance 1`

How to run WssGBLUP for 3 iterations in BLUPF90

```
awk 'BEGIN { for (i==1;i<45000;i++) print 1}' > w.txt    # number of lines = number of SNP

for j in {1..3}
    do
      echo blup.par | blupf90 | tee blup.log1_$j
      cp solutions solutions1_$j
      echo post.par | postGSf90 | tee post.log1_$j
      cp snp_sol snp_sol1_$j
      cp w.txt w.txt_$j
      awk '{ if ($1==1) print $7}' snp_sol > w.txt
      mkdir plot1_$j
      cp chrsnp plot1_$j/chrsnp
      cp chrsnpvar plot1_$j/chrsnpvar
      rm chrsnp chrsnpvar snp_sol solutions
    done

rm Gi A22i
```

How to run WssGBLUP for 3 iterations and multi-trait models in BLUPF90

- Although the model can be multi-trait, there is only one **G**

  - Only one set of weights can be used

- To estimate correct weights for each trait in a multi-trait model:

  - Add an option in postGSf90

    ```
    OPTION postgs_trt_eff x1 x2
    ```

  - x1 is the trait you are interested (number of the trait)

  - x2 is the effect (number of effect in this case)

- Run once for each trait or effect of interest

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.

Single-step GWAS

Genome-Wide Association Studies

Current standard for GWAS

- Run single marker regression with **G** to compensate for relationships
  - $y = Xb + za + u + e$
  - **z:** gene content {0,1,2}
  - a: marker effect

- Get p-values from $pval_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$

- Apply Bonferroni to correct for multiple testing

$$\text{Bonferroni correction} = \frac{0.05}{Number\ of\ SNP}$$

- 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

Single-step GWAS

SNP effects

GEBVs

VanRaden 2008
Stranden and Garrick 2009
Wang et al. 2012

$$\widehat{\boldsymbol{a}} = \lambda \mathbf{Z}'\mathbf{G}^{-1}\widehat{\boldsymbol{u}}$$

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

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

b) Nonlinear A SNP weights (VanRaden, 2008)

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

How to run ssGWAS in BLUPF90

- After renumf90 and preGSf90 to save clean files:

  - Blupf90 to estimate GEBV   Do not run iterations for ssGWAS!
    - `OPTION SNP_file snp.dat`
    - `OPTION chrinfo mrkmap.txt`
    - `OPTION saveGInverse`
    - `OPTION saveA22Inverse`
    - `OPTION weightedG w.txt      #vector of weights`

  - postGSf90 to backsolve GEBV to SNP effect
    - `OPTION SNP_file snp.dat`
    - `OPTION chrinfo mrkmap.txt`
    - `OPTION readGInverse`
    - `OPTION readA22Inverse`
    - `##OPTION which_weight nonlinearA`
    - `OPTION weightedG w.txt      #vector of weights`
    - `OPTION windows_variance 1`

postGSf90 options

```
OPTION SNP_moving_average n
```

Solutions for SNP effects will be by moving average of n adjacents SNPs.

```
OPTION windows_variance n
```

Calculates the variance explained by n adjacents SNPs.

When this option is used, the sum of variance explained by n adjacent SNPs (column 8 of snp_sol or column 3 of chrsnpvar) is not 100%.
This is because moving variance is used. If windows size is 20, the proportion of variance assigned to SNP 1 is calculated from SNP 1 to 20, for SNP 2 it goes from 2 to 21, for SNP 3 it goes from 3 to 22, and so forth.
A file called windows_variance has variance that sums to 100% in column 9.

```
OPTION windows_variance_mbp n
```

Calculates the variance explained by n Mb window of adjacents SNPs.

```
OPTION which_weight x
```

Generates a weight variable w to be used in the creation of a weighted genomic relationship matrix G=ZDZ'

- 1: $w = y^2 * (2(p(1-p)))$
- 2: $w = y^2$
- 3: experimental with the degree of brief
- 4: $w = C^{**}(abs(y)/sqrt(var(y's))-2)$ from VanRaden et al. (2009)
- nonlinearA: same as 4

postGSf90 options

```
OPTION Manhattan_plot
```

Plot using GNUPLOT the Manhattan plot (SNP effects) for each trait and correlated effect.

```
OPTION Manhattan_plot_R
```

Plot using R the Manhattan plot (SNP effects) for each trait and correlated effect.
`pdf` images are created: *manplot_St1e2.pdf*, but other formats can be specified.
Note: *t1e2* corresponds to trait 1, effect 2.

```
OPTION Manhattan_plot_R_format <format>
```

Control the format type to create images in R
`format` values accepted:

- pdf (default)
- png
- tif

```
OPTION plotsnp <n>
```

Control the values of SNP effects to use in Manhattan plots

- 1: plot regular SNP effects: abs(val)
- 2: plot standardized SNP effects: abs(val/sd) (default)

Output from postGSf90

`chrsnp`

contains data to create plot by GNUPLOT

- 1: trait
- 2: effect
- 3: values of SNP effects to use in Manhattan plots
- 4: SNP
- 5: Chromosome
- 6: Position
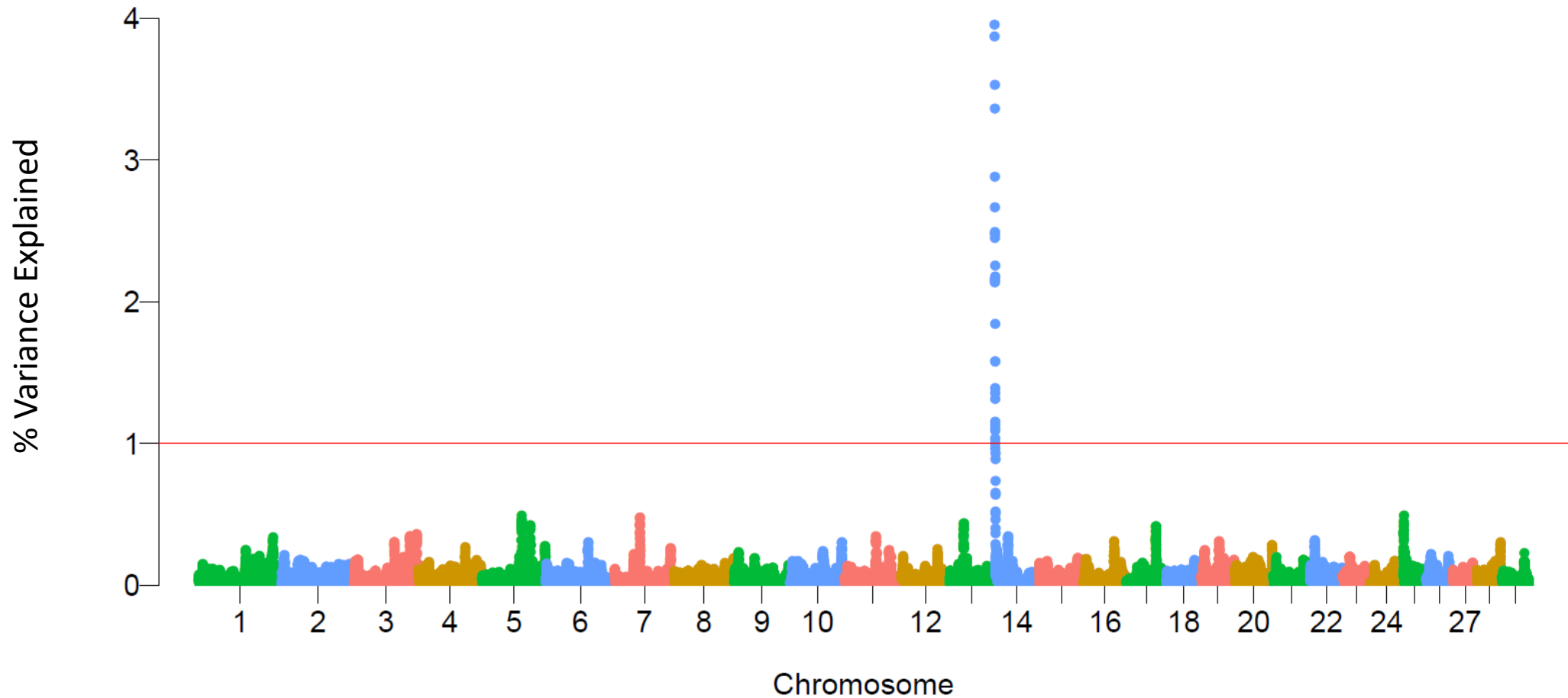
`chrsnpvar`

contains data to create plot by GNUPLOT

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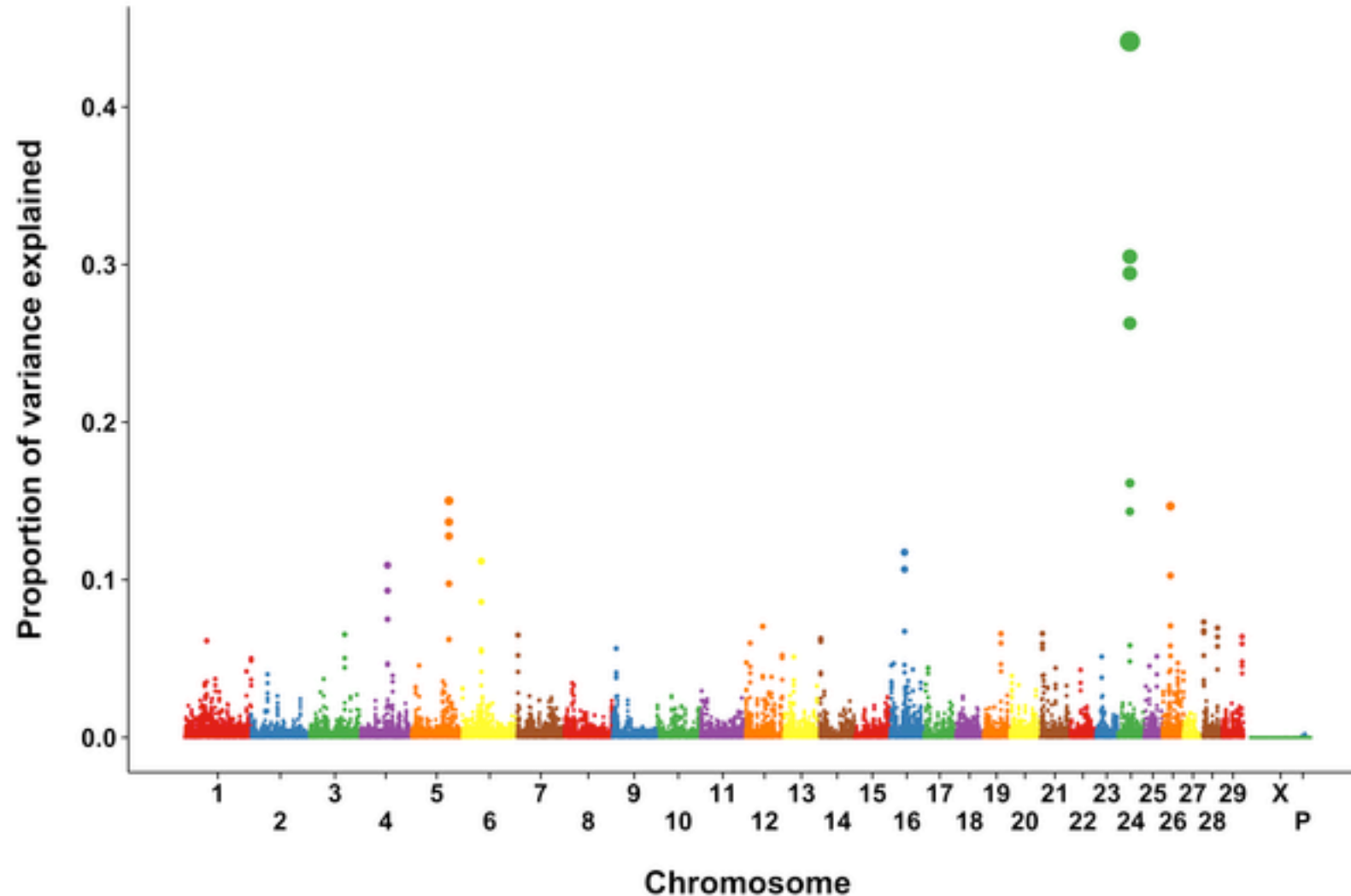- 6: Position

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!!!

Dikmen S, Cole JB, Null DJ, Hansen PJ (2013) Genome-Wide Association Mapping for Identification of Quantitative Trait Loci for Rectal Temperature during Heat Stress in Holstein Cattle. PLOS ONE 8(7): e69202. https://doi.org/10.1371/journal.pone.0069202
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0069202

Can we have p-values in ssGWAS?

BMC Bioinformatics

METHODOLOGY ARTICLE   **Open Access**

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

Jose L Gualdrón Duarte[1], Rodolfo JC Cantet[1], Ronald O Bates[2], Catherine W Ernst[2], Nancy E Raney[2] and Juan P Steibel[2,3*]

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

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*[¶]

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.)

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*[1]

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

blupf90

Ignacio
Aguilar

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

4) Obtain individual prediction error variance of SNP effects:

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

(Gualdron-Duarte et al., 2014)

postGSf90

Yutaka
Masuda

5) Backsolve GEBV to SNP effects ($\hat{a}$): $\hat{a} = \frac{1}{2\sum p_i q_i} \mathbf{Z}' \mathbf{G}^{-1} \hat{u}$
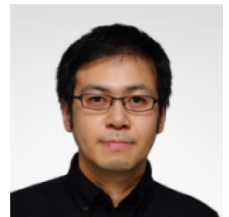
6) p-value$_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$

Φ is the cumulative standard normal function

Andres
Legarra

OPTION in blupf90 and postGSf90

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 Aguilar[1], Andres Legarra[2]*, Fernando Cardoso[3,4], Yutaka Masuda[5], Daniela Lourenco[5] and Ignacy Misztal[5]

- Single option for both programs

    OPTION snp_p_value

- Output

| trait	effect	-log10(p-value)	SNP	CHR	POS
1	5	0.6467097526	1	1	120183
1	5	0.3510786763	2	1	135098
1	5	0.3606678137	3	1	158820
1	5	0.2585950992	4	1	183040
1	5	0.6969161959	5	1	208728
1	5	1.7646253513	6	1	267940
1	5	1.0802326921	7	1	278952
1	5	0.6819748588	8	1	290690
1	5	1.0131137254	9	1	309487
1	5	0.0038533074	10	1	393248

P-values in ssGWAS for US Holsteins

- US HOL 2009 data: milk, fat, protein

- Single-trait models

  - 10k genotyped bulls

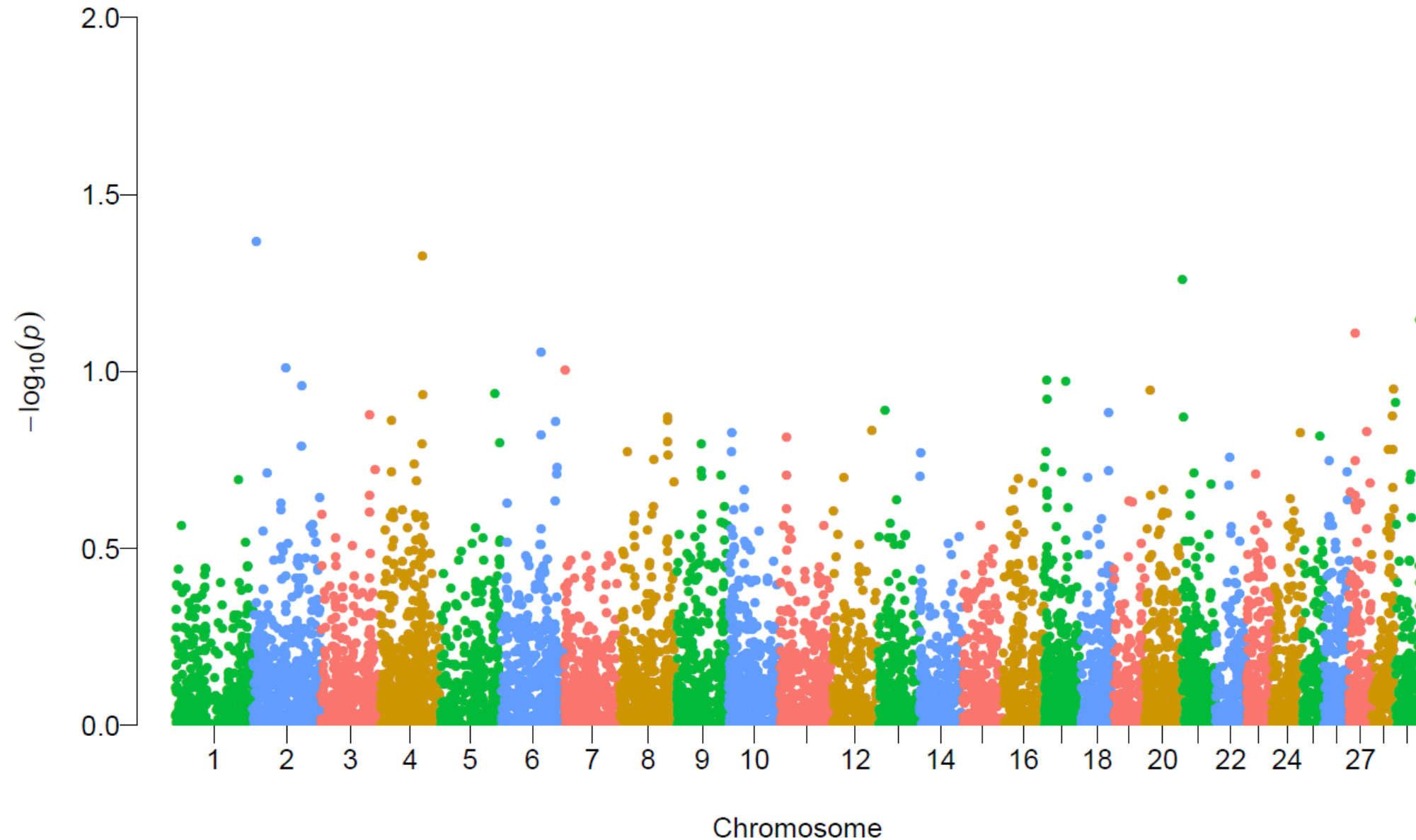 - 752k records for 100k daughters

  - 303k animals in ped

P-values in ssGWAS - Milk
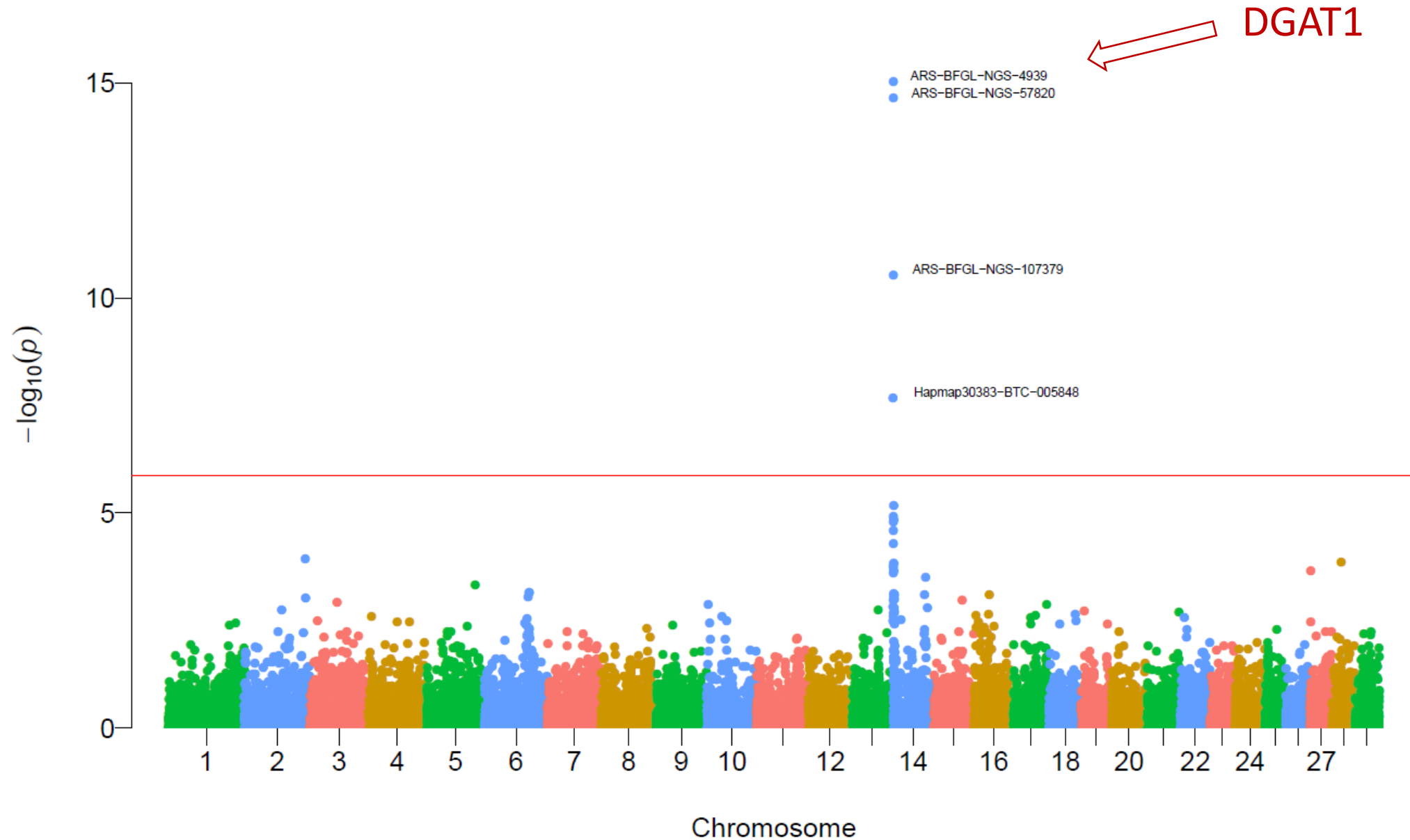
P-values in ssGWAS - Fat
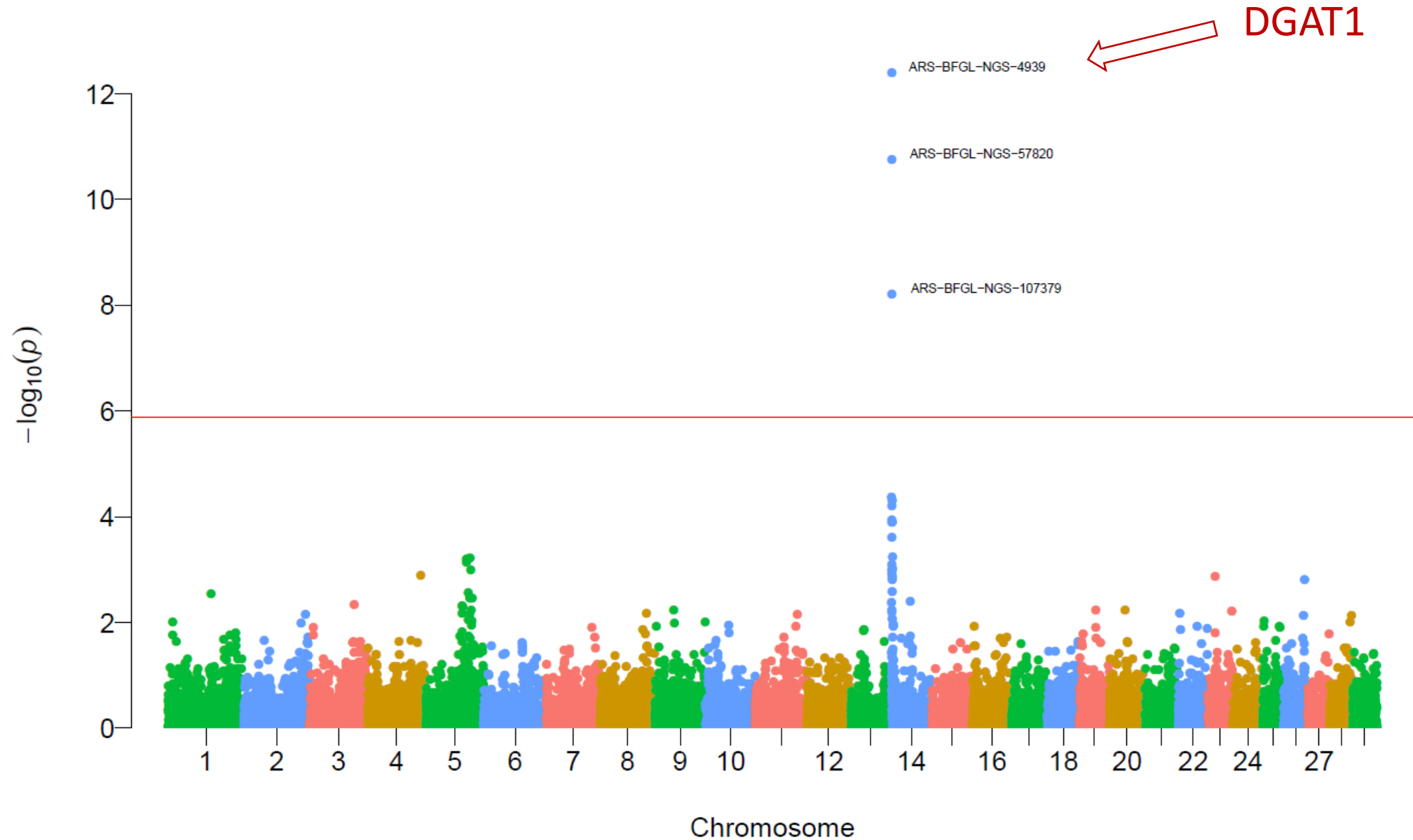
P-values in ssGWAS - Protein

How to run ssGWAS with p-values in BLUPF90

- Should not use iterations!
- After renumf90 and preGSf90 to save clean files:

    - Blupf90 to estimate GEBV
        - `OPTION SNP_file snp.dat`
        - `OPTION chrinfo mrkmap.txt`
        - `OPTION saveGInverse`
        - `OPTION saveA22Inverse`
        - `OPTION snp_p_value`

    - postGSf90 to backsolve GEBV to SNP effect
        - `OPTION SNP_file snp.dat`
        - `OPTION chrinfo mrkmap.txt`
        - `OPTION readGInverse`
        - `OPTION readA22Inverse`
        - `##OPTION which_weight nonlinearA`
        - `OPTION windows_variance 1`
        - `OPTION snp_p_value`

Output from postGSf90

```
chrsnp_pval
```

contains solutions of SNP and weights

- 1: trait
- 2: effect
- 3: -log10(p-value)
- 4: SNP
- 5: Chromosome
- 6: Position in bp

nce.ads.uga.edu/wiki

BLUPF90 Family of Programs

Now with support for genomic selection

Ignacy Misztal and collaborators, University of Georgia

BLUPF90 family of programs is a collection of software in Fortran 90/95 for mixed model computations in animal breeding. The goal of the software is to be as simple as with a matrix package and as efficient as in a programming language. For general description, see a 📄 paper from the CCB'99 workshop or see a 📄 paper on BGF90 at 7th WCGALP.

For variance component estimation, the family offers choices for simple and complicated models; see paper 📄 "Reliable computing in estimation of variance components" . From 2009 the programs are successively modified for genomic selection using a 📄 single-step approach (or ssGBLUP) by Ignacio Aguilar and Shogo Tsuruta.

For support, join 🌐 blupf90 group at yahoo.com.

Troubleshooting

⚠️ If the software crashes with segmentation fault, please change settings in your operating system. See FAQ:Segmentation fault for details. Also, The FAQ pages provide useful suggestions and solutions.

Headline

- History
- Modules
- Condition of use
- Distribution / Download
- Documentation / Manual / Tutorial
- Application program details
- Support
- FAQ
- Tricks / Tips
- To Do
- Courses
- Sample data
- Undocumented options

35