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

Jose L. Gualdron Duarte, Rodolfo, J.C. Cantet, Ronald O. Bates, Catherine W. Ernst, Nancy E. Raney and Juan P. Steibel.

Yeni Liliana Bernal Rubio. PhD.
QuantGen Group
Department of Epidemiology and Biostatistics
Michigan State University

Outline

Background

Datasets and Methods

- Real dataset – Simulation
- Proposal: Test of association using variance of each SNP effect.

Results and Discussion

- Solution of genomic predictions efficiently.
- Identification of genome segments associated with target traits.

gwaR: GWA from GBLUP model

Conclusions

- Useful method for identification of significant genomic regions.
- Reduces false positives.
- Useful for Meta-analysis of GWA.

Background

Availability of HD genotypes of SNP markers

&

Phenotypic data for complex traits

1. GEBVs for genomic evaluation

2. Estimates of effects of genomic regions in GWAS.

Genomic selection (Meuwissen et al., 2001)

SNP effects $\quad y_i = \mu + X_{ij} g_j + e_i$ ➡ GEBV based on $\boxed{\sum X_{ij} \hat{g}_j}$

An equivalent model (Stranden & Garrick, 2009)

$$y = X\beta + Zg + e$$

With $Zg = a$

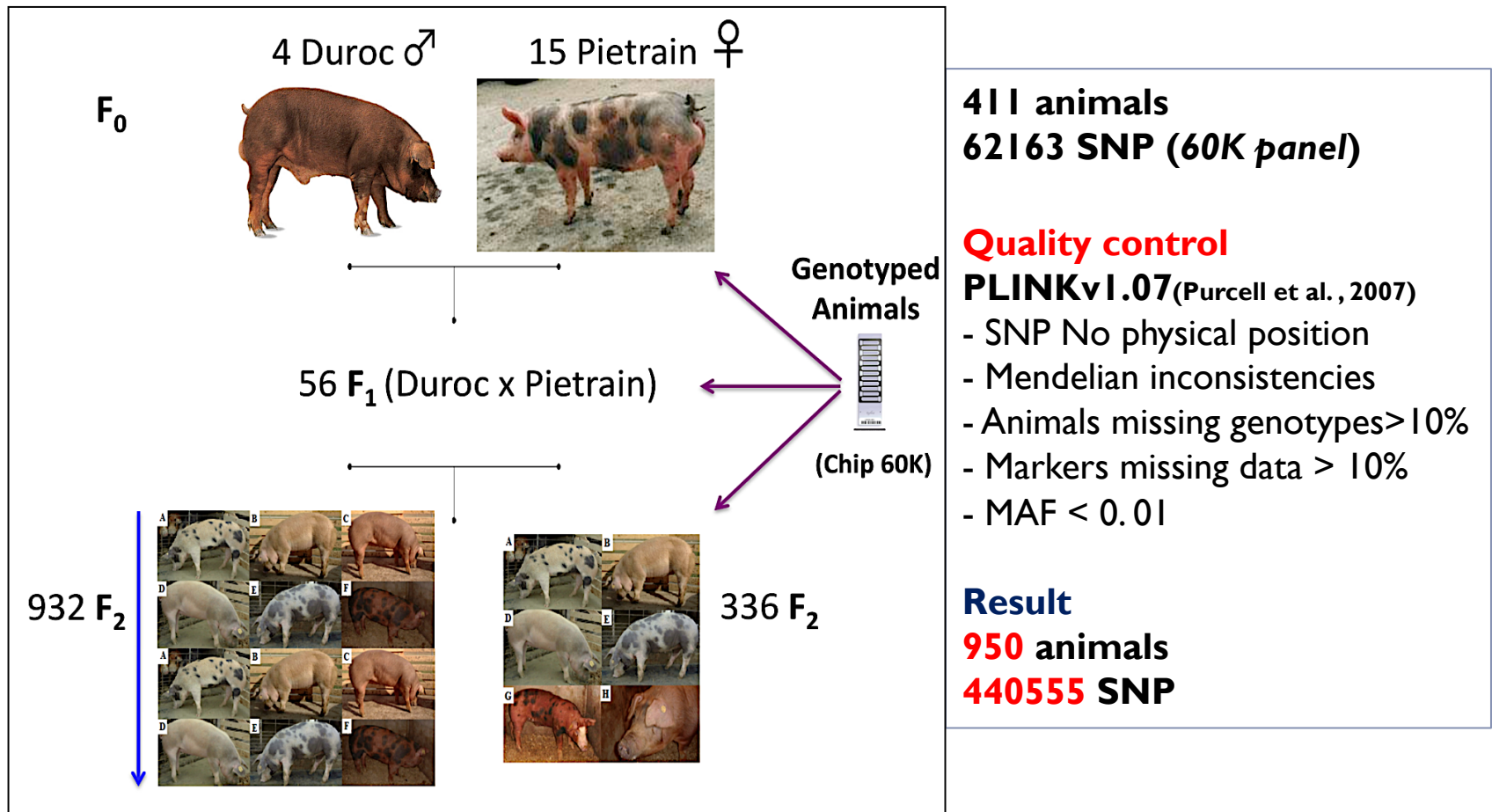Increasing number of GWAS based on mixed models and multiple testing

Models are difficult when number of individuals and SNP effects are large.

Objectives

Use a linear transformation of GEBV to estimate the SNP effects, and test those effects considering the variance of the same effects.

Locate genome segments displaying strong association for economically relevant traits.

Dataset

4 Duroc ♂ 15 Pietrain ♀

F_0

56 F_1 (Duroc x Pietrain)

932 F_2 336 F_2

Genotyped Animals

(Chip 60K)

411 animals
62163 SNP (60K panel)

Quality control
PLINKv1.07(Purcell et al., 2007)
- SNP No physical position
- Mendelian inconsistencies
- Animals missing genotypes>10%
- Markers missing data > 10%
- MAF < 0.01

Result
950 animals
440555 SNP

TRAIT: Measures of 13 week tenth rib back fat (bf10_13wk, mm).

Methods

Estimation of genomic relationship matrix $\quad G = ZZ'$

Prediction model

Centered animal model $\qquad y = X\,\beta + \,a + e$
$$a \sim N(0, G\sigma_a^2)$$
$$e \sim N(0, I\sigma_e^2)$$

Equivalent model
(Stranden & Garrick, 2009):
$$y = X\,\beta + \,Zg + e$$

SNP effects $\longleftarrow$

Given that $a = Zg$:

$$G\,\sigma_A^2 = \mathrm{Var}(a) = \mathrm{Var}\,(Z\,g) = Z\,\mathrm{Var}\,(g)\,Z'$$
$$= ZZ'\,\sigma_g^2$$

Necessary conditions for models to be equivalent
$$\begin{cases} G = ZZ' \\ \sigma_A^2 = \sigma_g^2. \end{cases}$$

Methods

Estimates and variance of SNP effects

$$\text{BLUP}(\hat{g}) = Z' \, G^{-1} \, \hat{a}$$

And:

$$\text{Var}(\hat{g}) = \text{Var}(Z' G^{-1} \hat{a})$$

$$\text{Var}(\hat{g}) = Z' \, G^{-1} Z \, \sigma_A^2 - Z' \, G^{-1} \, C^{aa} \, G^{-1} \, Z$$

Where:

$$C^{aa} = \sigma_e^2 \left(I - X(X'X)^{-1}X' + G^{-1}\lambda \right)^{-1}, \quad \lambda = \frac{\sigma_e^2}{\sigma_A^2}$$

Methods

Standardization of SNP effects

$$Var\left(\widehat{g}_j\right)$$

$$SNP_{ej} = \frac{\hat{g}_j}{\sqrt{\mathrm{Var}\left(\hat{g}_j\right)}}$$

Standardization of SNP

$$PEV\left(\widehat{g}_j\right)$$

$$SNP_{epj} = \frac{\hat{g}_j}{\sqrt{\mathrm{Var}\left(g_j\right) - \mathrm{Var}\left(\hat{g}_j\right)}}$$

P-values and genome screening

$$p\text{-value}_j = 2\left(1 - \Phi\left(\left|SNP_{ej}\right|\right)\right)$$
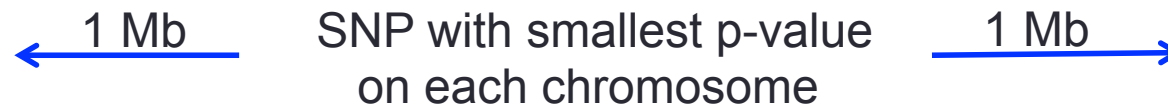
Methods

SNP effects and tests obtained by a single marker model

EMMA
(Kang et al., 2008)

$$y = X\beta + z_{ij}b_{ij} + a + e$$

One marker at a time

$$a \sim N(0, G\sigma_a^2)$$

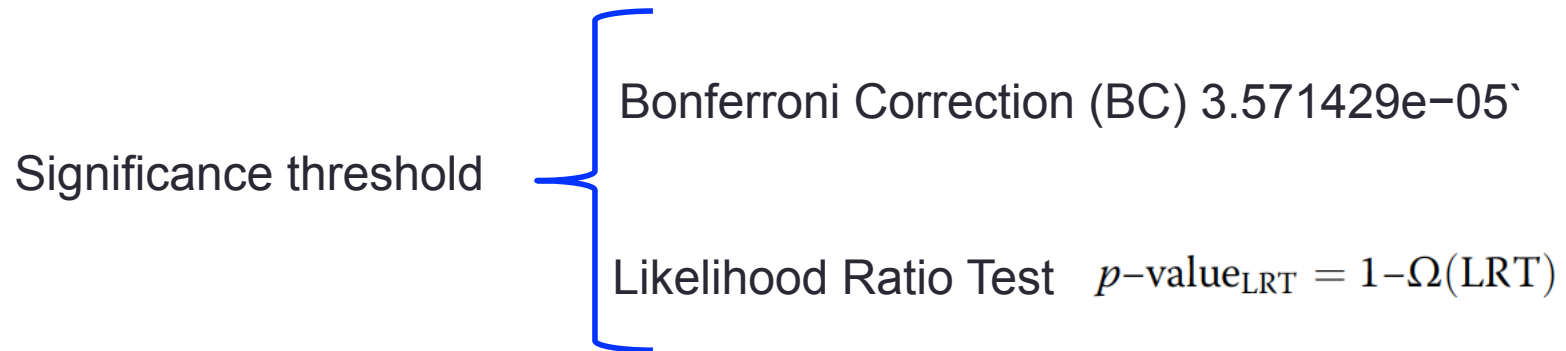**Proportion of variance explained by segments with large effect**

$\xleftarrow{\quad 1\text{ Mb} \quad}$ SNP with smallest p-value $\xrightarrow{\quad 1\text{ Mb} \quad}$
on each chromosome

The model fitted

$$y = X\beta + a_1 + a_2 + e$$

SNP in segment $\longleftarrow$ $\longrightarrow$ All SNP except
those in a_1

$$a_1 \sim N\left(0, \ G_1\sigma_{A_1}^2\right)$$

$$a_2 \sim N\left(0, \ G_2\sigma_{A_2}^2\right)$$

Methods

Significance threshold

Bonferroni Correction (BC) 3.571429e−05`

Likelihood Ratio Test $p\text{-value}_{\mathrm{LRT}} = 1-\Omega(\mathrm{LRT})$

Simulation

Plasmode simulation: compare SNP_{ej} and SNP_{epj}; test for the effect to be equal to zero.

Two scenarios: 1) Dependency within chromosomes, keeping LD

2) Independency, LE between markers

Results

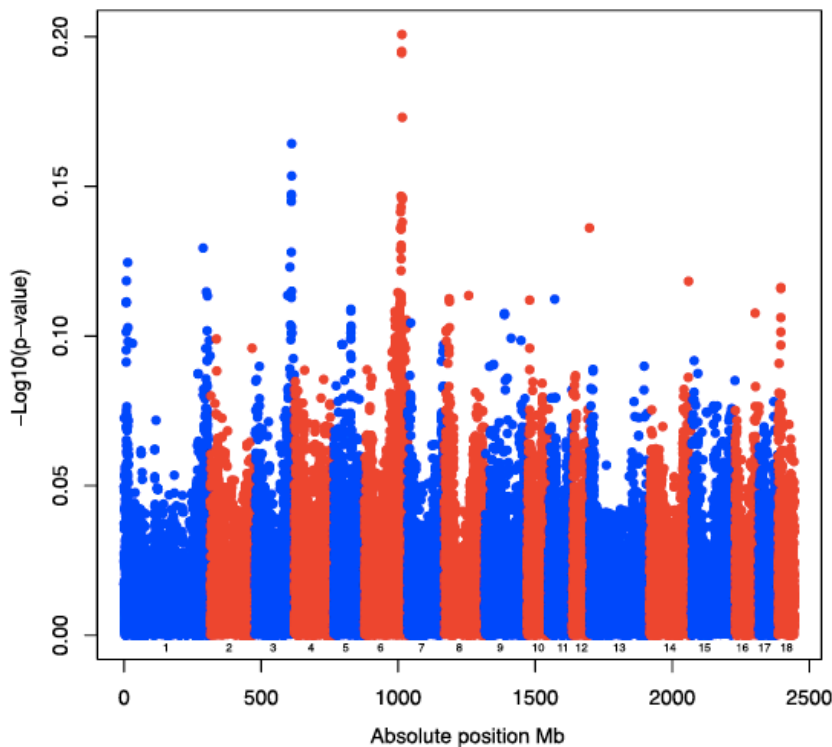**Genome screening:** Manhattan Plot for trait 13-week tenth rib backfat (mm)
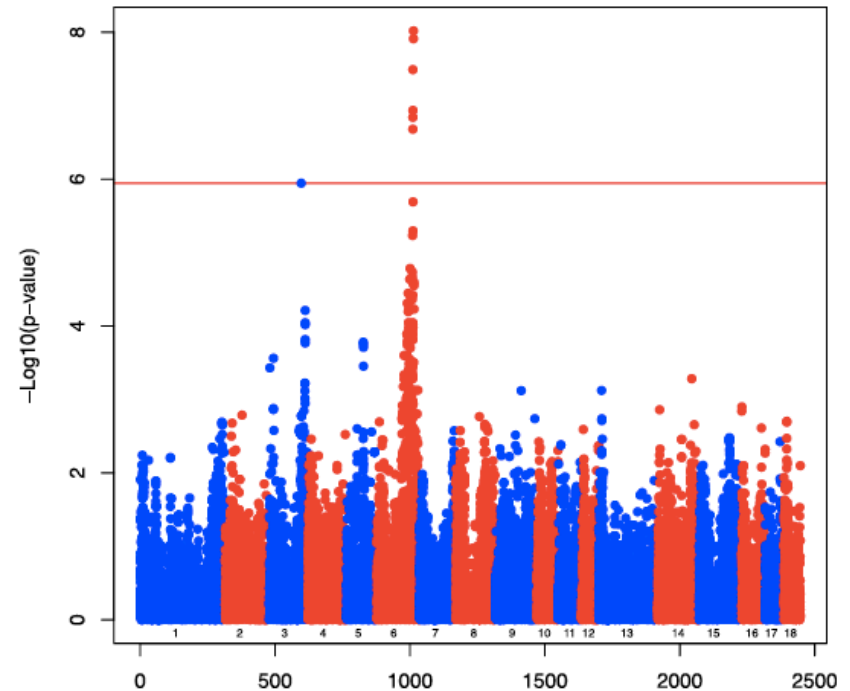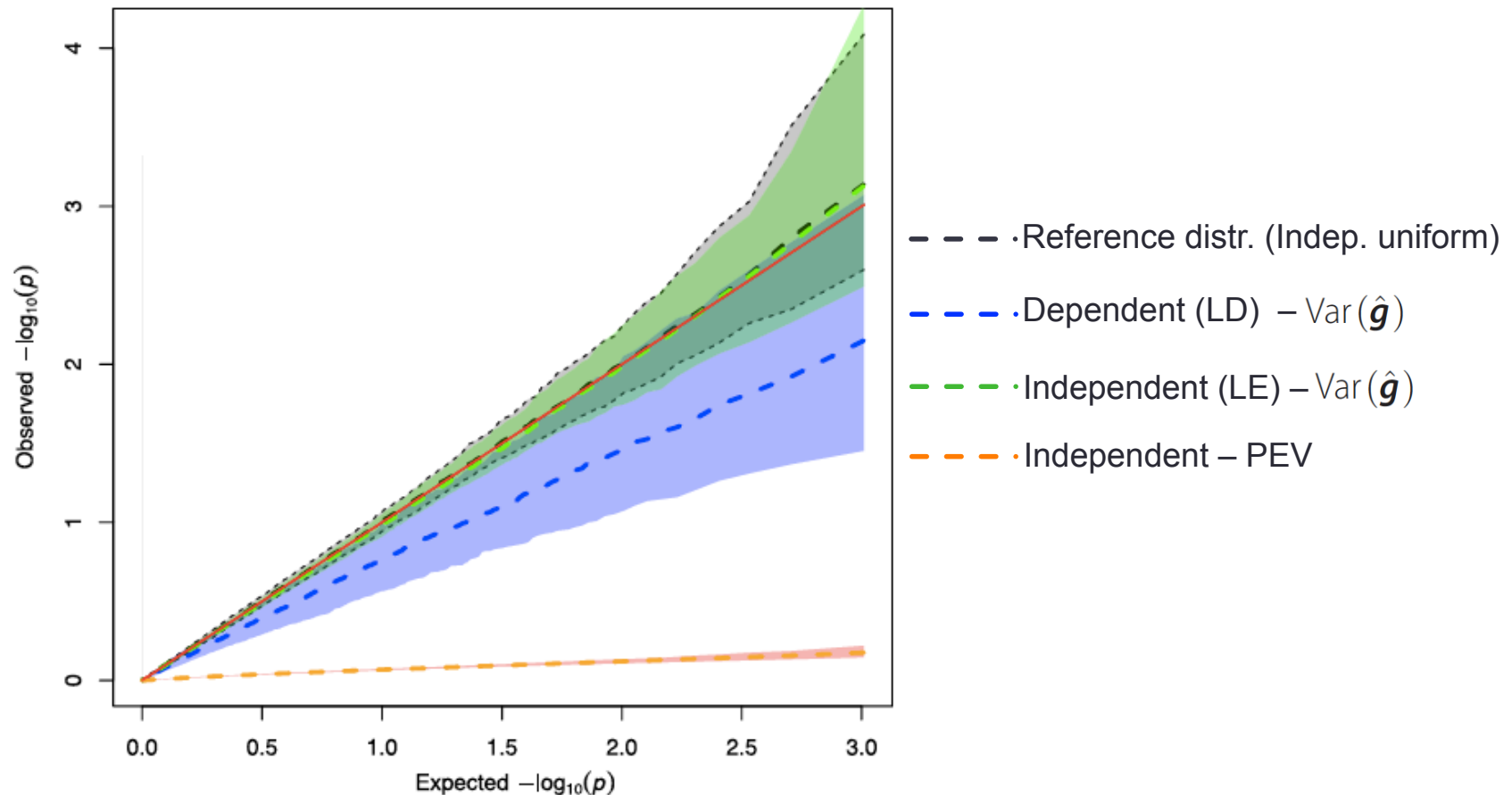

Figure 1. Standardization SNP_{epj}
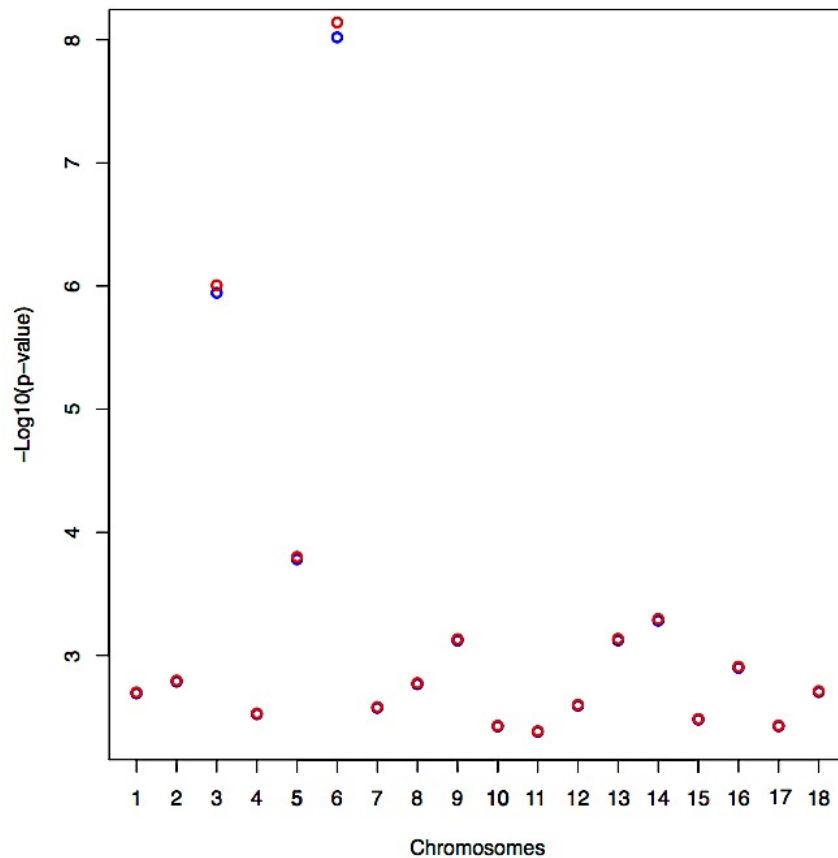
Figure 2. Standardization SNP_{ej}

Results

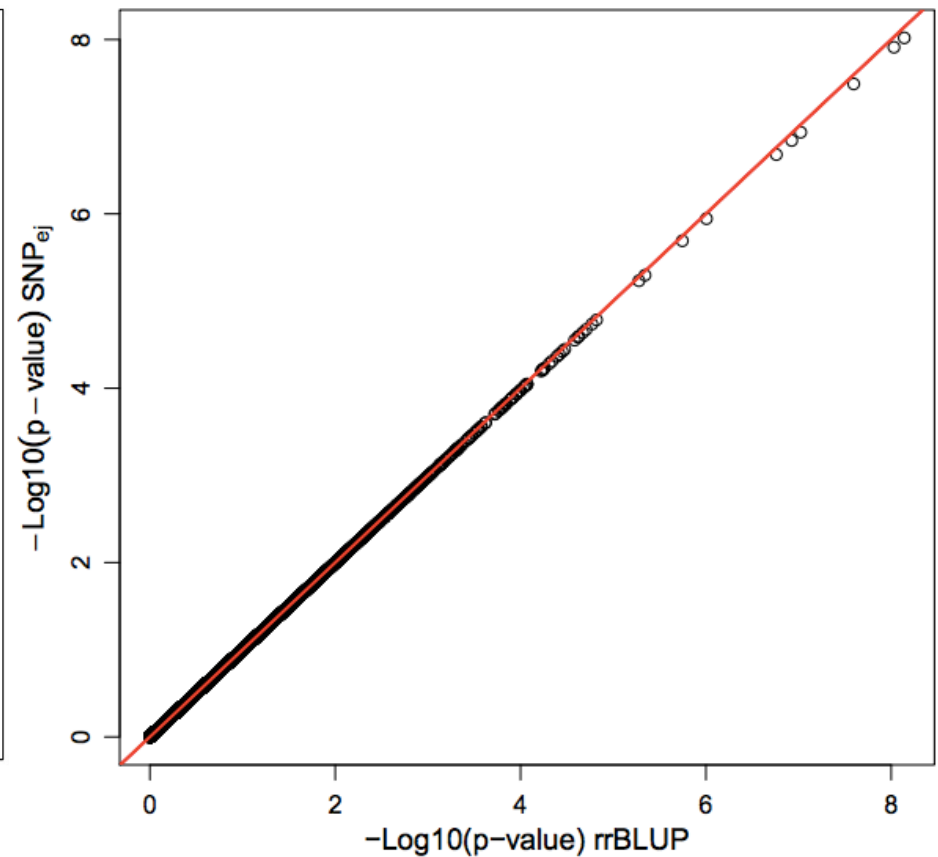**Quantil - quantil plot of observed and expected –log(p-values) obtained by simulation.**

Results

Highest -log10(*p*-values) by standardization SNPej (blue circles) and EMMA (red circles)

Dispersion plot of -Log10(*p*-values) by standardization SNPej and EMMA

Results

Test of segment effect

Table 1 SNP selected by smallest p-value per chromosome

SNP-name	Chromosome	Position Mb	$-\log_{10}$(p-value)	$|\hat{g}|$
ALGA0104402	6	136.08	8.02	0.77
H3GA0010564	3	119.34	5.95	0.48
ALGA0032063	5	61.37	3.78	0.42

m1: $y = X\beta + a + e$
m2: $y = X\beta + a1 + a2 + e$

Table 2 Variance components and LogLikelihood for models with or without the segment

Seg-chromosome	6	3	5
SNP – $\log_{10}$(p-value)	8.02	5.94	3.78
Lk_m1	−1227.938	−1227.938	−1227.938
Lk_m2	−1210.800	−1223.178	−1224.540
LRT	34.28	9.52	6.80
p-value$_{LRT}$	1.1×10^{-9}	6.5×10^{-4}	3.1×10^{-3}
VarE_m1	3.70	3.70	3.70
VarA_m1	2.68	2.68	2.68
VarE_m2	3.73	3.67	3.69
VarA_m2	1.95	2.42	2.55
segmVA	0.70	0.63	0.15
%segmVA	0.11	0.09	0.02

Discussion

Variance of the SNP effect

Comparison of p-values			
	$Var\left(\widehat{g}_j\right)$	$PEV\left(\widehat{g}_j\right)$	EMMA
LE	~ Uniform dist.	Very conservative	Similar to $Var\left(\hat{g}\right)$
LD	Conservative		

Discussion

Variance of SNP effects estimated under different approaches:

Wang et al. (2012)　　　　　　McClure et al. (2012)

$$\sigma^2_{A,j} = \hat{g}^2_j \, 2p_j \left(1 - p_j\right)$$

$$\left(2 \sum p_j q_j\right)^{-1} \sigma^2_A$$

<u>Proposed test:</u> Consider SNP specific standard deviation.

◆ Genome positions with notable effects are highlighted.

◆ Resulting *p*-value: appealing tool for researchers familiar with methods testing one SNP at the time.

gwaR
GWA from GBLUP models
(https://github.com/steibelj/gwaR)

GBLUP model

$$y = X\mathbf{b} + U_g + \sum_i Z_i U_i + e$$

GWA algorithm

$$\widehat{g} = Z'_g A^{-1} \widehat{u}_g$$

$$Var(\widehat{g_J}) = Z'_{jg} A^{-1} Var(\widehat{u}_g) A^{-1} Z_{jg}$$

Generalized Least Squares (GLS) model

$$y = X\beta + e$$

$$Var(e) = V = \sigma_g^2 A + \sum_i \sigma_i^2 Z_i G_i Z_i' + \sigma_e^2 R$$

The estimated SNP effects are:

$$\hat{g} = \sigma_g^2 Z_g' V^{-1} \hat{e}$$

$$\hat{e} = y - X\hat{b} = Q\,y$$

$$= (I - X(X'V^{-1}X)^{-1}X'V^{-1})y$$

$$Var(\hat{g}) = \left(\sigma_g^2\right)^2 Z_{jg}' V^{-1} Q\,V\,Q\,V^{-1} Z_{jg}$$

Also, the test statistic:

$$tg_i = \frac{\hat{g}_i}{\sqrt{\mathrm{Var}\left(\hat{g}_i\right)}}$$

Is equivalent to a test obtained from:

EMMAX

$$y = X\beta + z_{ij}b_{ij} + a + e \qquad e \sim N(0, I\sigma_e^2)$$

$$tb_i = \frac{\hat{b}_i}{\sqrt{\mathrm{Var}\left(\hat{b}_i\right)}}$$

Conclusions

- Proposed algorithm allows to perform GWA based on SNP effects linearly transformed from GEBV.

- Advantageous standardized test of SNP effects using their own variance (specific variation, computing time).

- Genome segments approach allowed for detection of genome regions responsible for sizeable fractions of the trait genetic variance.

Useful methodology for detection of specific genome regions that affect an economically relevant trait when using single or multiple populations.

References

Gualdrón Duarte et al. (2014). Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics 15:246.

Kang et al. (2008). Efficient control of population structure in model organism association mapping. Genetics. 178: 1709-1723.

McClure et al. (2012). Genome-wide association analysis for quantitative trait loci influencing Warner-Bratzler shear force in five taurine cattle breeds. Anim. Genet. 43: 662-673.

Meuwissen et al. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829.

Stranden and Garrick (2009). Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J.Dairy Sci. 92: 2971-2975.

Wang et al. (2012) Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res. 94: 73-83.

https://github.com/steibelj/gwaR