



**WORKSHOP
DE BIOINFORMÁTICA
APLICADA À GENÔMICA E
MELHORAMENTO ANIMAL**



Métodos estatísticos em seleção genômica



Fabyano Fonseca e Silva

**Prof. Adjunto IV - Dep Zootecnia – UFV
Estatística Genômica e Bioinformática**

Campo Grande, 14/07 a 15/07 de 2014

- desde o primeiro estudo sobre GWS, a comparação entre as metodologias estatísticas vem sendo amplamente explorada nas principais publicações

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes[†] and M. E. Goddard^{†,‡}

Genetics 157: 1819–1829 (April 2001)

TABLE 2

Comparing estimated *vs.* true breeding values
in generation 1003

	$r_{\text{TBV;EBV}} + \text{SE}$	$b_{\text{TBV;EBV}} + \text{SE}$
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827
BayesB	$0.848 + 0.012$	$0.946 + 0.018$

➤ **Modelo geral para GWS: Regressão múltipla**

$$y = \mathbf{1}\mu + \sum_{i=1}^m \mathbf{M}_i a_i + \mathbf{e}, \quad i=1,2,\dots,m \text{ (num. de SNPs)}$$

$$y = \mathbf{1}\mu + \mathbf{M}_1 a_1 + \mathbf{M}_2 a_2 + \dots + \mathbf{M}_m a_m + \mathbf{e}$$

$$\mathbf{e} \mid \sigma_e^2 \sim N(0, \mathbf{I}\sigma_e^2) \Rightarrow \mathbf{y} \mid \mu, a_1, \dots, a_m, \sigma_e^2 \sim N(\mathbf{1}\mu + \sum_{i=1}^m \mathbf{M}_i a_i, \mathbf{I}\sigma_e^2)$$

A diferença entre os métodos estatísticos está nas pressuposições a respeito dos efeitos de SNPs (a_i)

1) RR-BLUP (Random regression - BLUP): Meuwissen et al. (2001)

Pressupõe-se: $a_i | \sigma^2 \sim N(0, \sigma^2)$ (mesma variância para todos SNPs)

MME
$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{M} \\ \mathbf{M}'\mathbf{1} & \mathbf{M}'\mathbf{M} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{M}'\mathbf{y} \end{bmatrix}$$

OBS.: $\lambda = \sigma_e^2 / \sigma_a^2$

$$\hat{\mathbf{a}} = (\mathbf{M}'\mathbf{M} + \mathbf{I}\lambda)^{-1} \mathbf{M}'(\mathbf{y} - \mathbf{1}\mu)$$
 Output: Solução para o vetor de efeitos SNPs

$$\hat{\mathbf{u}} = \mathbf{M}\hat{\mathbf{a}}$$

Indiretamente: Solução para o vetor de EGBV

$$\sigma_u^2 = 2 \sum_{i=1}^m p_i (1 - p_i) \sigma_a^2$$
$$h^2 = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$$

OBS.: p_i é a MAF

2) GBLUP: Van Raden (2008)

Pressupõe-se: $\mathbf{u} | \sigma_u^2 \sim N(0, \sigma_u^2 \mathbf{G})$ (\mathbf{G} é a matriz de parentesco genômica)

$$y = \mathbf{1}\mu + \sum_{i=1}^m \mathbf{M}_i a_i + \mathbf{e} \Rightarrow y = \mathbf{1}\mu + \overset{\text{“Z”}}{\mathbf{I}} \left\{ \overset{\text{“u”}}{\sum_{i=1}^m \mathbf{M}_i a_i} \right\} + \mathbf{e} \Rightarrow \boxed{y = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e}}$$

Modelo equivalente

$$\text{MME} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \sigma_e^2 \text{Var}(\mathbf{u})^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

$$\text{OBS.: } \text{Var}(\mathbf{u})^{-1} = \sigma_u^{-2} \mathbf{G}^{-1} \\ \lambda = \sigma_e^2 / \sigma_u^2$$

$$\hat{\mathbf{u}} = \left(\mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\lambda \right)^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{1}\mu)$$

Output: Solução para o vetor GEBV

$$\hat{\mathbf{a}} = (\mathbf{M}'\mathbf{M})^{-1} (\mathbf{M}'\hat{\mathbf{u}})$$

Indiretamente: Solução para o vetor de efeitos SNPs

$$h^2 = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$$

OBS. GBLUP E RR-BLUP SÃO MODELOS EQUIVALENTES !

$$y = \mathbf{1}'\mu + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

GBLUP

$$\begin{bmatrix} \mathbf{N} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \sigma_e^2 \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

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

$$y = \mathbf{1}'\mu + \mathbf{I} \sum \mathbf{M}_i \mathbf{a}_i + \mathbf{e}$$

RR-BLUP

$$\begin{bmatrix} \mathbf{N} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} + \sigma_e^2 \left[\text{var} \left(\sum \mathbf{M}_i \mathbf{a}_i \right) \right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum \mathbf{M}_i \mathbf{a}_i} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{y} \end{bmatrix}$$

$$\text{var} \left(\sum \mathbf{M}_i \mathbf{a}_i \right) = \sum \text{var} \{ \mathbf{M}_i \mathbf{a}_i \} = \sum \mathbf{M}_i \mathbf{A}_i \mathbf{M}_i' = \sum \mathbf{M}_i \mathbf{M}_i' \sigma_{ai}^2 = \text{like } \mathbf{A} \sigma_g^2$$

numerator relationship matrix = \mathbf{A}

$$\begin{bmatrix} \mathbf{N} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} + \sigma_e^2 \left[\sum \mathbf{M}_i \mathbf{M}_i' \sigma_{ai}^2 \right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\sum \mathbf{M}_i \mathbf{a}_i} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{y} \end{bmatrix}$$

$$y = \mathbf{1}'\mu + \mathbf{Z}u + e \quad \text{GBLUP}$$

$$\begin{bmatrix} \mathbf{N} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \sigma_e^2 \mathbf{G}^{-1} \end{bmatrix}$$

$$y = \mathbf{1}'\mu + \mathbf{I} \sum \mathbf{M}_i \mathbf{a}_i + e \quad \text{RR-BLUP}$$

$$\begin{bmatrix} \mathbf{N} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} + \sigma_e^2 \left[\text{var} \left(\sum \mathbf{M}_i \mathbf{a}_i \right) \right]^{-1} \end{bmatrix}$$

$$\frac{1}{\sigma_u^2} A^- = V \left(\sum_{i=1}^m \mathbf{M}_i \mathbf{a}_i \right)^-$$

$$V \left(\sum_{i=1}^2 \mathbf{M}_i \mathbf{a}_i \right) = \underbrace{V(\mathbf{M}_1 \mathbf{a}_1 + \mathbf{M}_2 \mathbf{a}_2)}_{\text{OBS.: } a_i \sim N(0, \sigma_a^2)} = V(\mathbf{M}_1 \mathbf{a}_1) + V(\mathbf{M}_2 \mathbf{a}_2) = \underbrace{\mathbf{M}_1 V(\mathbf{a}_1) \mathbf{M}_1' + \mathbf{M}_2 V(\mathbf{a}_2) \mathbf{M}_2'}_{\text{OBS.: RRblup assume } V(\mathbf{a}_1) = V(\mathbf{a}_2) = \sigma_a^2}$$

$$V \left(\sum_{i=1}^2 \mathbf{M}_i \mathbf{a}_i \right) = \sigma_a^2 \underbrace{(\mathbf{M}_1 \mathbf{M}_1' + \mathbf{M}_2 \mathbf{M}_2')}_{\substack{\text{OBS.: } M = M_1 \parallel M_2 \\ MM' = M_1 M_1' + M_2 M_2'}} = \sigma_a^2 MM'$$

$$\frac{1}{\sigma_u^2} A^- = (\sigma_a^2 MM')^- \Rightarrow \frac{1}{\sigma_u^2} A^- = \frac{1}{\sigma_a^2} (MM')^- \Rightarrow A^- = \frac{\sigma_u^2}{\sigma_a^2} (MM')^-$$

$$A = \frac{\sigma_a^2}{\sigma_u^2} (MM') = \frac{\sigma_a^2}{\sum_{i=1}^m 2p_i(1-p_i)\sigma_a^2} (MM') = \frac{(MM')}{\sum_{i=1}^m 2p_i(1-p_i)}$$

➤ Softwares para RR-BLUP

pacote rrBLUP do R
GS3 (Legarra et al., 2011)
GWP (Mewissen, 2009)

➤ Softwares para GBLUP

ASREML (Gilmour et al., 2066): opção .giv
WOMBAT (Meyer, 2009): opção GINV
pacote rrBLUP do R
GVCBLUP
BLUPF90 (Miształ)

Aula prática 2

RR-BLUP e GBLUP

Nivelamento em Inferência Bayesiana

Additive Genetic Variability and the Bayesian Alphabet

Daniel Gianola,^{*,†,‡,1} Gustavo de los Campos,^{*} William G. Hill,[§] Eduardo Manfredi[‡]
and Rohan Fernando^{**}

Habier et al. *BMC Bioinformatics* 2011, **12**:186
<http://www.biomedcentral.com/1471-2105/12/186>



RESEARCH ARTICLE

Open Access

Extension of the bayesian alphabet for genomic selection

David Habier^{1*}, Rohan L Fernando¹, Kadir Kizilkaya^{1,2} and Dorian J Garrick^{2,3}

Genetics: Early Online, published on May 1, 2013 as 10.1534/genetics.113.151753

Priors in whole-genome regression: the Bayesian alphabet returns

Daniel Gianola^{*}

Motivação: Gustavo de los Campos et al. (2013) 10.1534/genetics.112.143313

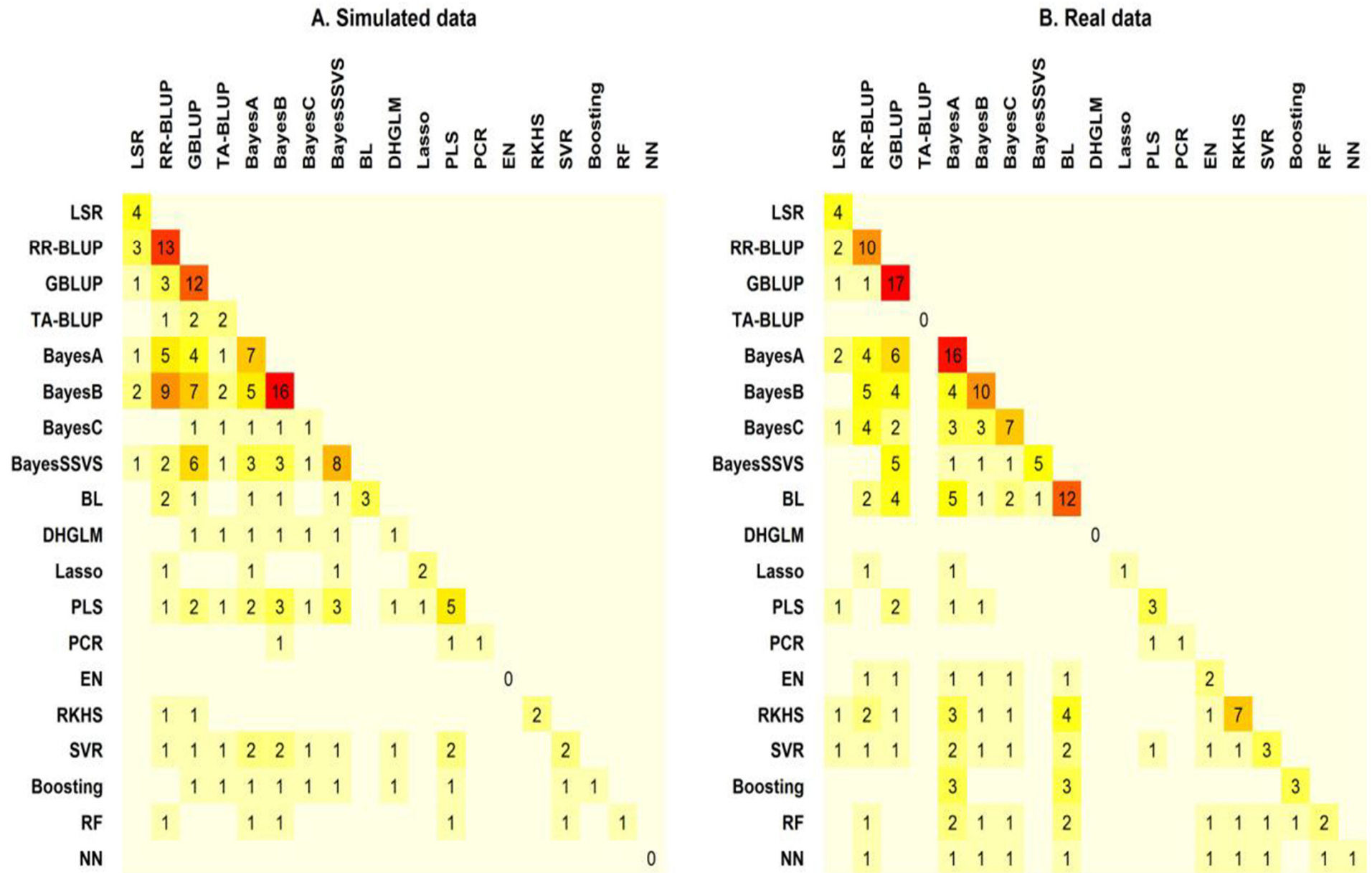


Figure 3. Number of articles reviewed comparing one or more methods using simulated (A) or real data (B). The abbreviations used for the methods are given in Table 2. The following

Motivação: Daetwyler et al. (2012) **10.1534/genetics.112.147983**

Table 5. Accuracy of prediction¹ for the wheat data, using 10-fold random cross-validation.

	Trait 1 Acc(SD)	Trait 2 Acc(SD)	Trait 3 Acc(SD)	Trait 4 Acc(SD)
BayesA1	0.524(0.098)	0.503(0.130)	0.392(0.136)	0.468(0.149)
BayesB1	0.520(0.098)	0.502(0.130)	0.391(0.136)	0.465(0.149)
BayesC	0.525(0.104)	0.503(0.130)	0.390(0.140)	0.468(0.145)
BayesA2	0.527(0.101)	0.504(0.130)	0.392(0.136)	0.469(0.150)
BayesB2	0.523(0.101)	0.502(0.130)	0.392(0.136)	0.465(0.150)
Bayesian Lasso1	0.530(0.101)	0.504(0.130)	0.393(0.136)	0.471(0.150)
GBLUP	0.518(0.149)	0.493(0.139)	0.397(0.130)	0.437(0.187)
Bayesian Lasso2	0.548(0.098)	0.502(0.139)	0.412(0.130)	0.470(0.139)

Motivação: Meuwissen et al. (2009) Genetics Selection Evolution 2009, 41:2

Table 1: The accuracy of MBayesB, fBayesB and BLUP, defined by the correlation between true and estimated breeding values in generation 1002.

<u>Method</u>	Accuracy + se	Regression + se
fBayesB	0.849 ± 0.011	1.145 ± 0.025
MBayesB	0.860 ± 0.010	0.923 ± 0.011
BLUP	0.694 ± 0.006	0.990 ± 0.009

Motivação: Meuwissen e Goddard (2010) Genetics 185:623–631 (June 2010)

TABLE 3

The accuracy of the predictions of total genetic value (\pm SE) in the TEST1 data set when the training data contained $T = 200$ individuals and GWBLUP or BayesB is used to estimate the marker effects

Data	Causative SNPs			
	GWBLUP		BayesB	
	Excluded	Included	Excluded	Included
3 QTL	0.503 ± 0.011	0.508 ± 0.011	0.938 ± 0.013	0.973 ± 0.004
30 QTL	0.491 ± 0.016	0.493 ± 0.010	0.806 ± 0.023	0.826 ± 0.019

3) Métodos Bayesianos

* RR - BLUP Bayes

$$a_i | \sigma^2 \sim N(0, \sigma^2)$$

Mesma variância para todos os SNPs

$$\sigma^2 \sim \chi^{-2}(S, \nu)$$

Software:

GWP (Meuwissen, 2009)

BAYZ (Janss, 2011)

BLR (de los Campos, 2010)

GS3 (Legarra, 2010)

$$\sigma_u^2 = 2\sigma^2 \sum_{i=1}^m p_i(1-p_i)$$

* Bayes A

$$\alpha_i | \sigma_i^2 \sim N(0, \sigma_i^2)$$

Variância específica para cada SNP

$$\sigma_i^2 \sim \chi^{-2}(S, \nu)$$

Software:

GWP (Meuwissen, 2009)

BAYZ (Janss, 2011)

$$\sigma_u^2 = 2 \sum_{i=1}^m p_i(1-p_i) \sigma_i^2$$

Distribuições condicionais completas a posteriori: Bayes A

$$\mu|ELSE \sim N \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} b_j \right), \frac{\sigma_e^2}{n} \right]$$

Média geral

$$\sigma_e^2|ELSE \sim n \left(1 + \frac{v_e}{n} \right) \left(\frac{\sum_{i=1}^n \left(y_i - \mu - \sum_{j=1}^p x_{ij} b_j \right)^2 + v_e S_e^2}{n + v_e} \right) \chi_{v_e+n}^{-2}$$

Variância residual

$$b_j|ELSE \sim N \left[\frac{\sum_{i=1}^n x_{ij} \left(y_i - \mu - \sum_{j'=1}^p x_{ij'} b_{j'} \right)}{\sum_{i=1}^n x_{ij}^2 + \frac{\sigma_e^2}{\sigma_{b_j}^2}}, \frac{\sigma_e^2}{\sum_{i=1}^n x_{ij}^2 + \frac{\sigma_e^2}{\sigma_{b_j}^2}} \right]$$

Efeito do SNP_j

$$\sigma_{b_j}^2|ELSE \sim v \left(1 + \frac{1}{v} \right) S^2 \left(\left[\frac{\left(\frac{b_j}{S} \right)^2 + v}{1 + v} \right] \right) \chi_{v+1}^{-2}$$

variância do SNP_j

* Bayes B

$$a_i | p, \sigma_{i1}^2 \sim (1 - \gamma_i) N(0, \sigma_{i0}^2 = 0) + \gamma_i N(0, \sigma_{i1}^2)$$

Variância específica para cada SNP + seleção de SNPs via mistura de dist com prop fixa

$$\left. \begin{array}{l} P(\gamma_i = 0) = p \\ P(\gamma_i = 1) = 1-p \end{array} \right\} p \text{ fixo}$$

OBS.: se $P(\gamma_i = 1) = 1 \Rightarrow$ Bayes A

$$\sigma_{i1}^2 \sim \chi^{-2}(S, \nu)$$

Software:

GWP (Meuwissen, 2009)

TABLE 2

Comparing estimated *vs.* true breeding values in generation 1003

	$r_{TBV;EBV} + SE$	$b_{TBV;EBV} + SE$
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827
BayesB	0.848 ± 0.012	0.946 ± 0.018

OBS. críticas

1) Meuwissen et al. (2001) Simularam dados sob o modelo Bayes B

2) A condição $\sigma_{i0}^2 = 0$

Não é real em estatística

$$\sigma_u^2 = 2 \sum_{i=1}^m p_i (1-p_i) \sigma_i^2$$

Distribuições condicionais completas a posteriori: Bayes B

Efeito do SNP_j

$$(\beta_j | \mu, \dots) \sim \begin{cases} N(\hat{\beta}_j, \frac{\sigma_e^2}{c_j}) & \delta_j = 1 \\ N(0, \sigma_j^2) & \delta_j = 0 \end{cases}$$

where $\hat{\beta}_j = \frac{\mathbf{X}_j' \mathbf{w}}{c_j}$ and $c_j = (\mathbf{x}_j' \mathbf{x}_j + \frac{\sigma_e^2}{\sigma_j^2})$

Variável indicadora delta

$$\Pr(\delta_j = 1 | \delta_{j-}, \beta, \pi) = \frac{h(\delta_j = 1)}{h(\delta_j = 1) + h(\delta_j = 0)}$$

where

$$h(\delta_j) = \pi^{(1-\delta_j)} (1 - \pi)^{\delta_j} \exp\left\{-\frac{(\mathbf{w} - \mathbf{x}_j \beta_j \delta_j)' (\mathbf{w} - \mathbf{x}_j \beta_j \delta_j)}{2\sigma_e^2}\right\}$$

Não se caracteriza
como dist de prob
conhecida:

**Metropolis-
Hastings**

Média geral

Variância residual

Variância do SNP_j

Idem Bayes A

Bayes A e Bayes B: alta demanda computacional pela necessidade de gerar cadeias MCMC para milhares de parâmetros

Alternativa: evitar MCMC por meio de aproximações das integrais requeridas – método acelerado

Genetics Selection Evolution



Research

Open Access

A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value

Theo HE Meuwissen^{*1}, Trygve R Solberg¹, Ross Shepherd³ and John A Woolliams^{1,2}

Software GWP – genome wide predictor

* Improved BayesB

$$a_i | p, \sigma_{i0}^2, \sigma_{i1}^2 \sim (1 - \gamma_i) N(0, \sigma_{i0}^2) + \gamma_i N(0, \sigma_{i1}^2)$$

Uma var para cada SNP + seleção de SNPs via mistura de distribuições com prop aleatória

$$\sigma_{i1}^2 \sim \chi^{-2}(S, \nu)$$

Software:
BAYZ (Janss, 2011)

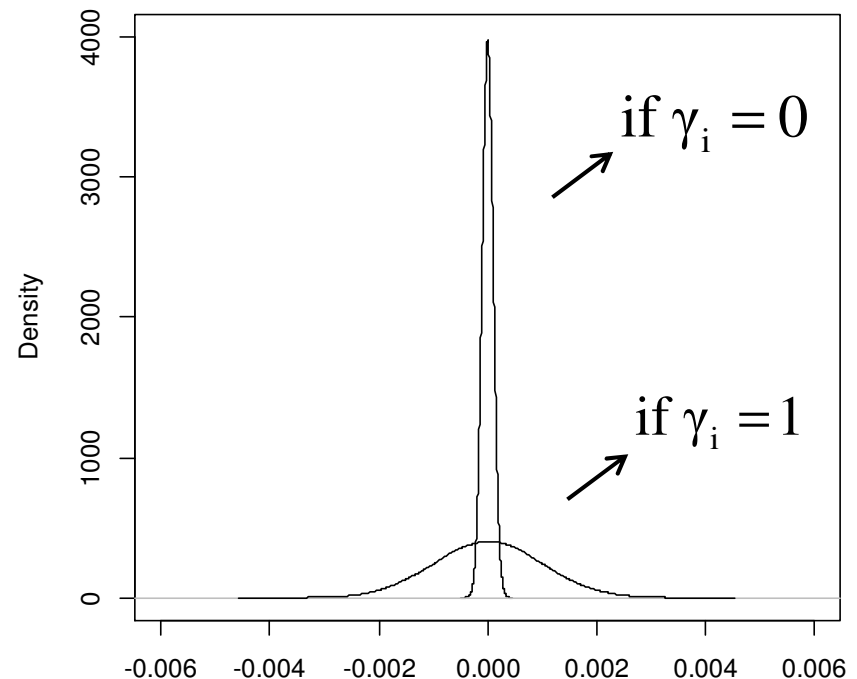
$$\gamma_i \sim \text{Bernoulli}(p) \begin{cases} \text{if } \gamma_i = 0, a_i \sim N(0, \underbrace{\sigma_{i0}^2}_{\text{small}}) \\ \text{if } \gamma_i = 1, a_i \sim N(0, \underbrace{\sigma_{i1}^2}_{\text{big}}) \end{cases}$$

$$P(\gamma_i = 0) = p, \quad p \sim \text{Beta}(\alpha, \beta)$$

$$\text{fixed variance ratio: } (\sigma_{i1}^2 / \sigma_{i0}^2) = 1,000$$

OBS.: if $P(\gamma_i = 1) = p = 1 \rightarrow \text{Bayes A}$

$$\sigma_u^2 = 2 \sum_{i=1}^m p_i (1 - p_i) \sigma_i^2$$



* BayesC π

$$a_i | \pi, \sigma^2 \sim \pi N(0, \sigma^2 = 0) + (1 - \pi) N(0, \sigma^2)$$

Mesma var todos SNPs + seleção de SNPs via mistura de distribuições com prop aleatória

$\pi \sim U(0,1)$ **Dist. Uniforme para π**

$$\sigma_u^2 = 2(1 - \pi) \sigma^2 \sum_{i=1}^m p_i (1 - p_i)$$

$$\sigma^2 \sim \chi^{-2}(S^*, v)$$

OBS. S^* é fixo e dado por:

$$S^* = \tilde{\sigma}^2 (v - 2) / v$$

$\tilde{\sigma}^2$ valor esperado a priori para var SNP

* BayesD π

$$a_i | \pi, \sigma_i^2 \sim \pi N(0, \sigma^2 = 0) + (1 - \pi) N(0, \sigma_i^2)$$

Uma var para cada SNP + seleção de SNPs via mistura de distribuições com prop aleatória

$\pi \sim U(0,1)$ **Dist. Uniforme para π**

$$\sigma_i^2 \sim \chi^{-2}(S, v)$$

$$S \sim \text{Gama}(\alpha, \beta)$$

$$\sigma_u^2 = 2(1 - \pi) \sum_{i=1}^m p_i (1 - p_i) \sigma_i^2$$

Software:

GENSEL (Iowa State Univ.)

GS3 (apenas BayesC π)

* **Bayesian LASSO** (de los Campos, 2009)

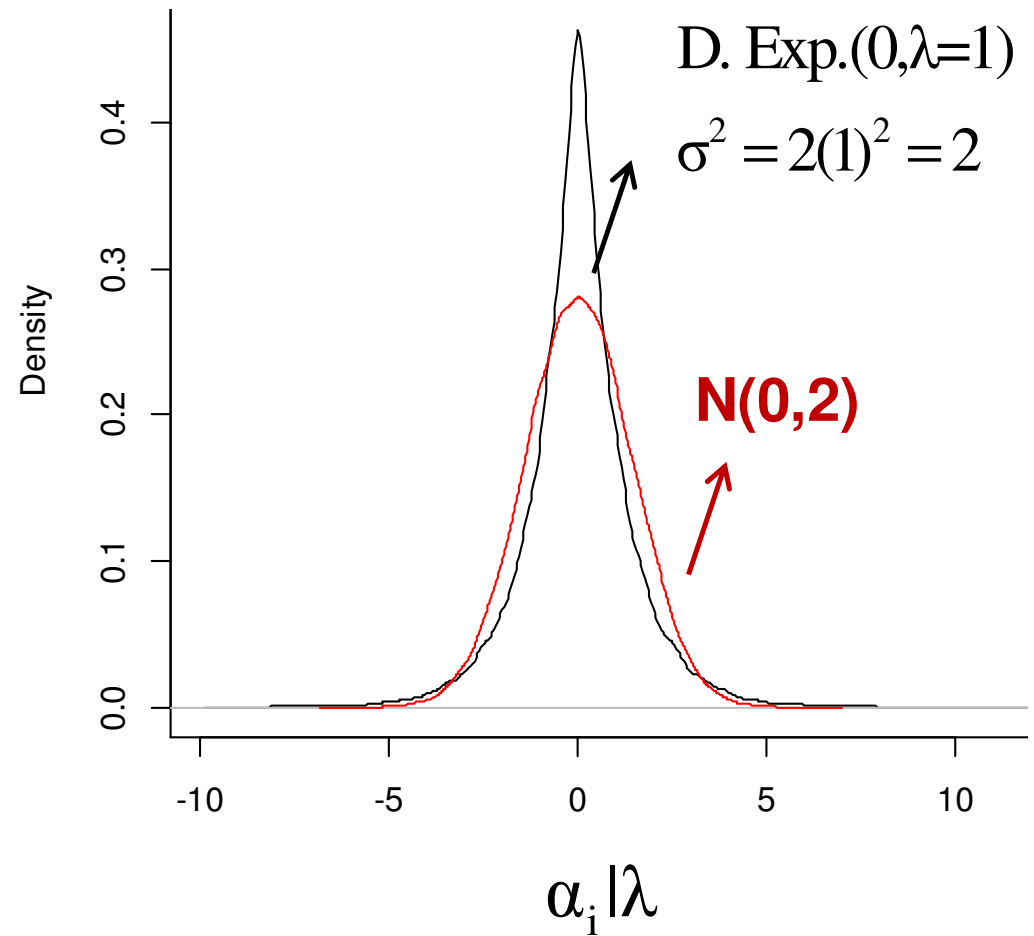
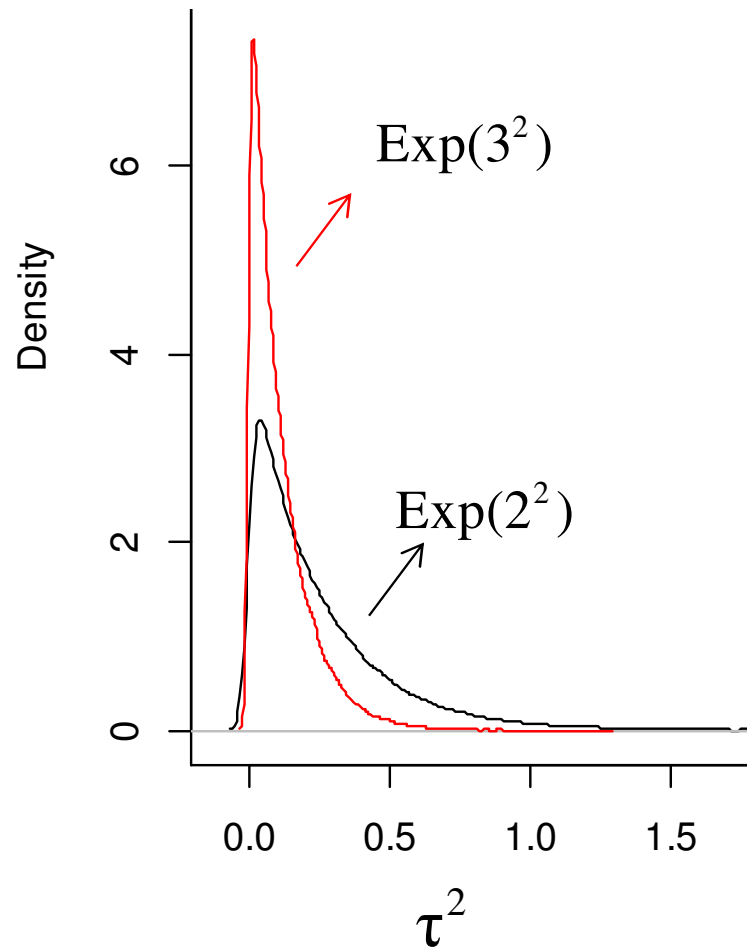
$$\left. \begin{aligned} a_i | \tau_i, \sigma_e^2 &\sim N(0, \underbrace{\tau_i^2 \sigma_e^2}_{\sigma_i^2}) \\ \tau_i^2 | \lambda^2 &\sim \text{Exp}(\lambda^2) \\ \lambda^2 &\sim \text{Gamma}(\varphi_1, \varphi_2) \end{aligned} \right\} \begin{aligned} &\int N(0, \tau_i^2 \sigma_e^2) \text{Exp}(\lambda^2) d\tau_i^2 \\ a_i | \lambda, \sigma_e^2 &\sim \text{Double-Exp.}(0, \lambda) \end{aligned}$$

**Uma var para cada SNP (baseada na var resid)
+ seleção de SNPs via regressão penalizada
(não exige fixação de π)**

Software:
BAYZ (Janss, 2011)
BGLR
(de los Campos, 2013)

$$\sigma_u^2 = \sum_{i=1}^m \underbrace{\tau_i^2 \sigma_e^2}_{\sigma_i^2} 2p_i(1-p_i)$$

$$a | \lambda, \sigma_e^2 \sim \prod_i \frac{\lambda}{2\sigma_e} \exp\left(\frac{-\lambda |a_i|}{\sigma_e}\right)$$



Silva et al. (2011) A note on accuracy of Bayesian LASSO regression in GWS
Livestock Science 142 (2011) 310–314.

Distribuições condicionais completas a posteriori: **LASSO** Bayesiano

$$p(\mu|else) \propto N\left([n\sigma_\epsilon^{-2} + \sigma_\mu^{-2}]^{-1} \sum_i y_i^* \sigma_\epsilon^{-2}, [n\sigma_\epsilon^{-2} + \sigma_\mu^{-2}]^{-1} \right)$$

$$p(\beta_{jl}|else) \propto N\left(\sigma_\epsilon^{-2} \sum_{i=1}^n x_{jli} y_i^{ooo} / \left[\sigma_\epsilon^{-2} \sum_{i=1}^n x_{jli}^2 + \sigma_\epsilon^{-2} \tau_j^{-2} \right], \left[\sigma_\epsilon^{-2} \sum_{i=1}^n x_{jli}^2 + \sigma_\epsilon^{-2} \tau_j^{-2} \right]^{-1} \right)$$

$$p(\tau_j^{-2}|else) \propto IG\left(\tau_j^{-2} \middle| \mu_j = \frac{\sigma_\epsilon \lambda}{|\beta_{jl}|}, S = \lambda^2 \right)$$

$$p(\lambda^2|else) \propto G\left(\lambda^2 \middle| p, \frac{1}{2} \sum_{j=1}^p \tau_j^2 \right)$$

$$p(\sigma_\epsilon^2|else) \propto \chi^{-2}\left(\sigma_\epsilon^2 \middle| S = S_\epsilon + \mathbf{\epsilon}'\mathbf{\epsilon}, df = df_\epsilon + n \right)$$

*** Improved Bayesian LASSO
(Legarra, 2011)**

$$a_i | \tau_i, a_i, \sigma_u^2 \sim N(0, \underbrace{\tau_i^2 \sigma_u^2}_{\sigma_i^2})$$

$$\tau_i^2 | \lambda^2 \sim \text{Exp}(\lambda^2)$$

$$\lambda^2 \sim \text{Gamma}(\varphi_1, \varphi_2)$$

Uma var para cada SNP (baseada na var genética) + seleção de SNPs via regressão penalizada

**Software:
GS3 (Legarra, 2011)**

$$\sigma_a^2 = \frac{2}{\lambda^2} = \frac{\sigma_u^2}{2 \sum_i p_i (1 - p_i)}$$

$$\sigma_u^2 = \frac{4 \sum_i p_i (1 - p_i)}{\lambda^2}$$

$$a | \lambda, \sigma_a^2 \sim \prod_i \frac{\lambda}{2\sigma_a} \exp\left(\frac{-\lambda |a_i|}{\sigma_a}\right)$$

$$\sigma_a^2 = \frac{\sigma_u^2}{2 \sum_i p_i (1 - p_i)}$$

Aula prática 3

Métodos Bayesianos em GWS

4) Regressão múltipla com redução de dimensionalidade PLS – *Partial Least Square* - Quadrados Mínimos Parciais

Quadrados Mínimos Parciais (*PartialLeastSquares* – PLS) foi introduzido por Wold em 1975, sendo considerado útil a construção de equações de predições em situações nas quais se tem um grande número de variáveis explicativas e um número relativamente pequeno de dados amostrais

OBS.: O fato de ser usado para predição de y quando $p > n$ chamou a atenção para a utilização deste método em GWS

Segundo Garthwaite(1994), o método PLS apresenta similaridades com o método de Regressão via Componentes Principais (PCR), sendo a maior diferença entre eles dada pelo fato do PCR levar em consideração apenas as variáveis explicativas na construção dos componentes, enquanto que o PLS também leva em consideração a variável dependente.



ORIGINAL ARTICLE

Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins

N. Long¹, D. Gianola^{1,2,3}, G.J.M. Rosa^{1,3} & K.A. Weigel²

mance. Therefore, the computational burden can be reduced with PCR and PLS, compared to methods that estimate all marker effects by treating them as random variables, such as Bayes A (Meuwissen *et al.* 2001) or the Bayesian least absolute shrinkage and selection operator (Lasso) (Park & Casella 2008; de los Campos *et al.* 2009).

Azevedo et al. (2014): proposta de melhora via Comp. Indep.



ORIGINAL ARTICLE

Supervised independent component analysis as an alternative method for genomic selection in pigs

C.F. Azevedo¹, F.F. Silva^{1,2}, M.D.V. de Resende^{1,3}, M.S. Lopes⁴, N. Duijvesteijn⁴, S.E.F. Guimarães², P.S. Lopes², M.J. Kelly⁵, J.M.S. Viana⁶ & E.F. Knol⁴

1 Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Brazil

2 Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Brazil

3 Embrapa Florestas, Departamento de Engenharia Florestal, Universidade Federal de Viçosa, Viçosa, Brazil

4 TOPIGS Research Center IPG, Beuningen, the Netherlands

5 Queensland Alliance for Agriculture & Food Innovation, The University of Queensland, St Lucia, QLD, Australia

6 Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, Brazil

RESEARCH

Open Access

Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance

Albart Coster^{1*}, John WM Bastiaansen¹, Mario PL Calus², Johan AM van Arendonk¹, Henk Bovenhuis¹

Table 3 Average (standard error) accuracy of MEBV for individuals in the evaluation population.

Method	unequal QTL variance			equal QTL variance		
	low nQTL	int. nQTL	high nQTL	low nQTL	int. nQTL	high nQTL
	sc. 1	sc. 2	sc. 3	sc. 4	sc. 5	sc. 6
BM	0.77 (0.009)	0.67 (0.010)	0.60 (0.012)	0.71 (0.004)	0.67 (0.005)	0.67 (0.006)
LARS	0.75 (0.009)	0.67 (0.005)	0.65 (0.004)	0.65 (0.005)	0.63 (0.006)	0.63 (0.006)
PLSR	0.66 (0.009)	0.66 (0.007)	0.67 (0.007)	0.68 (0.006)	0.67 (0.006)	0.66 (0.007)

The MEBV were calculated with methods BM, LARS and PLSR. Simulated number of QTL was low (low nQTL), intermediate (int. nQTL) or high (high nQTL). The simulated variance of every tenth QTL was 81 times larger than variance of the remaining QTL (unequal QTL variance) or equal for all QTL (equal QTL variance). The averages and standard deviations were calculated using 60 replicated simulations.

Exemplo com 2 covariáveis (2 SNPs no contexto de GWS)

Variáveis originais			Variáveis centradas na média		
Y	X_1	X_2	U_1	V_{11}	V_{12}
y_1	X_{11}	X_{21}	$y_1 - \bar{Y}$	$X_{11} - \bar{X}_1$	$X_{21} - \bar{X}_2$
y_2	X_{12}	X_{22}	$y_2 - \bar{Y}$	$X_{12} - \bar{X}_1$	$X_{22} - \bar{X}_2$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
y_n	X_{1n}	X_{2n}	$y_n - \bar{Y}$	$X_{1n} - \bar{X}_1$	$X_{2n} - \bar{X}_2$

1º Ajuste: regressões de U_1 em função de V_{11} e V_{12} (separadamente)

$$\hat{U}_{11} = \hat{b}_{11}V_{11} \quad \text{e} \quad \hat{U}_{12} = \hat{b}_{12}V_{12}$$

Construção do 1º Componente (considerando pesos iguais para cada covariável)

$$T_1 = \hat{U}_{11} + \hat{U}_{12} = \hat{b}_{11}V_{11} + \hat{b}_{12}V_{12}$$

2º Ajuste: regressões de U_1 , V_{11} e V_{12} em função de T_1 (separadamente)

$$U_1 = c_1 T_1 + e_1 \Rightarrow \hat{e}_1 = U_2 = U_1 - \hat{c}_1 \hat{T}_1$$

$$V_{11} = d_{11} T_1 + e_{11}^* \Rightarrow \hat{e}_{11}^* = V_{21} = V_{11} - \hat{d}_{11} T_1$$

$$V_{12} = d_{12} T_1 + e_{12}^* \Rightarrow \hat{e}_{12}^* = V_{22} = V_{12} - \hat{d}_{12} T_1$$

3º Ajuste: regressões de U_2 em função de V_{21} e V_{22} (separadamente)

$$\hat{U}_{21} = \hat{b}_{21} V_{21} \quad \text{e} \quad \hat{U}_{22} = \hat{b}_{22} V_{22}$$

Construção do 2º Componente (considerando pesos iguais para cada covariável)

$$T_2 = \hat{U}_{21} + \hat{U}_{22} = \hat{b}_{21} V_{21} + \hat{b}_{22} V_{22}$$

OBS.: Já foram obtidos dois componentes (T_1 e T_2)

4º Ajuste: regressões de y (variável dependente original) em função de T_1 e T_2 (simultaneamente – enfoque de regressão múltipla)

- p/ apenas um componente (T_1): $y = \alpha_0 + \alpha_1 T_1 + \varepsilon$
- p/ os dois componentes (T_2): $y = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \varepsilon$

Predição de y em função dos componentes (GEBV no contexto de GWS)

- p/ apenas um componente (T_1): $\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 T_1 = GEBV_{1comp}$
- p/ os dois componentes (T_2): $\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 T_1 + \hat{\alpha}_2 T_2 = GEBV_{2comp}$

OBS. É necessário determinar o número ótimo de componentes (Ex. EQM)

Havendo interesse em obter os coeficientes da regressão de y em função das variáveis originais (X_1 e X_2) é necessário “voltar” com os termos que originaram cada componente (T_1 e T_2), i.e., deixar a equação em função de X_1 e X_2 .

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1[\hat{b}_{11}V_{11} + \hat{b}_{12}V_{12}] + \hat{\alpha}_2[\hat{b}_{21}V_{21} + \hat{b}_{22}V_{22}]$$

$$\hat{y} = \hat{a}_0 + \hat{a}_1[\hat{b}_{11}(X_{11} - \overline{X}) + \hat{b}_{12}(X_{12} - \overline{X})] + \hat{a}_2[\hat{b}_{21}(V_{11} - \hat{d}_{11}T_1) + \hat{b}_{22}(V_{12} - \hat{d}_{12}T_1)]$$

Após algumas operações, é possível isolar os termos X_1 e X_2 :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2.$$

OBS. Sob o enfoque de GWS, β_1 e β_2 representam os efeitos dos SNPs

O algoritmo apresentado algebricamente, passo a passo, pode ser rescrito em forma matricial

Partial Least Squares (Garthwaite, 1994, JASA 89:122-127)

Step 1: Center both \mathbf{y} and \mathbf{X}

$$\mathbf{u}_1 = \mathbf{y} - \mathbf{1}\bar{y}$$

and

$$\mathbf{V}_1 = [\mathbf{x}_1 - \mathbf{1}\bar{x}_1, \mathbf{x}_2 - \mathbf{1}\bar{x}_2, \dots, \mathbf{x}_k - \mathbf{1}\bar{x}_k]$$

Step 2: In iteration i , get regression estimate of \mathbf{u}_i using \mathbf{v}_{ij} for $j = 1, 2, \dots, k$ as

$$\hat{u}_{i(j)} = \mathbf{v}_{ij} \left(\frac{\mathbf{v}_{ij}' \mathbf{u}_i}{\mathbf{v}_{ij}' \mathbf{v}_{ij}} \right)$$

Define \mathbf{t}_i as weighted mean of these regression estimates:

$$\mathbf{t}_i = \sum_{j=1}^k w_{ij} \left(\frac{\mathbf{v}_{ij}' \mathbf{u}_i}{\mathbf{v}_{ij}' \mathbf{v}_{ij}} \right) \mathbf{v}_{ij}$$

Partial Least Squares

Using

$$w_{ij} = \mathbf{v}_{ij}' \mathbf{v}_{ij}$$

\mathbf{t}_i becomes

$$\mathbf{t}_i = \mathbf{V}_i \mathbf{V}_i' \mathbf{u}_i$$

Step 3: Define \mathbf{u}_{i+1} as

$$\mathbf{u}_{i+1} = \mathbf{u}_i - \mathbf{t}_i \left(\frac{\mathbf{t}_i' \mathbf{u}_i}{\mathbf{t}_i' \mathbf{t}_i} \right),$$

which is the residual from the prediction of \mathbf{u}_i using \mathbf{t}_i , and

$$\mathbf{V}_{i+1} = \mathbf{V}_i - \mathbf{t}_i \left(\frac{\mathbf{t}_i' \mathbf{V}_i}{\mathbf{t}_i' \mathbf{t}_i} \right),$$

which are the residuals from the prediction of \mathbf{v}_{ij} using \mathbf{t}_i .
Repeat steps 2 and 3 for r iterations.

Aula prática 4

PLS em GWS

Como incorporar os resultados dos métodos apresentados em programas de melhoramento??

1) Índice de seleção (VanRaden et al., 2009)

VanRaden et al. (2009) propuseram o cálculo do GEBV como sendo um índice de seleção combinando resultados das análises tradicionais e genômicas, o qual é dado por:

$$\text{GEBV} = b_1 * \text{DGV} + b_2 * \text{PI}_s + b_3 * \text{PI}_t, \quad (1)$$

em que:

GEBV é o valor genético genômico;

DGV é o valor direto genômico obtido da análise GWS considerando apenas os indivíduos genotipados

PI_t é o valor genético predito pela análise tradicional ao se considerar todos os indivíduos da população

PI_s é o valor genético predito pela análise tradicional considerando apenas informação de parentesco entre os animais genotipados,

b_1 , b_2 e b_3 são os pesos atribuídos a cada fator que serão estimados via sistema de equação apresentado a seguir.

O seguinte sistema linear será utilizado para a obtenção dos pesos:

$$\begin{aligned} b_1 * V_{11} + b_2 * V_{12} + b_3 * V_{13} &= V_{11}, \\ b_1 * V_{12} + b_2 * V_{22} + b_3 * V_{23} &= V_{22}, \\ b_1 * V_{13} + b_2 * V_{23} + b_3 * V_{33} &= V_{33}, \end{aligned} \quad (2)$$

em que:

V_{11} , V_{22} , e V_{33} são, respectivamente, a confiabilidade (acurácia ao quadrado, r^2) de DGV, PI_t e PI_s . De acordo com VanRaden et al. (2009), define-se $V_{12} = V_{22}$, $V_{23} = V_{22}$ e $V_{13} = V_{22} + (V_{11} - V_{22})(V_{33} - V_{22})/(1 - V_{22})$.

Uma vez resolvido o sistema linear em [2], os valores obtidos para b_1 , b_2 e b_3 serão utilizados em [1] para se obter o GEBV e também na equação [3] para se obter a confiabilidade esperada para o GEBV (r^2_{GEBV}), a qual é dada por:

$$r^2_{GEBV} = b_1 * V_{11} + b_2 * V_{22} + b_3 * V_{33}. \quad (3)$$

2) Análise multicaracterística (Kachman, 2008)

Table 1. Traits categorized according to information available.

DNA Tests	Industry-collected Phenotypes	
	No	Yes
No	---	EPD _(y)
Yes	EPD _(m1)	EPD _(m2)

$$\begin{pmatrix} y \\ m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} X_y & 0 & 0 \\ 0 & X_1 & 0 \\ 0 & 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_y \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} Z_y & 0 & 0 \\ 0 & Z_1 & 0 \\ 0 & 0 & Z_2 \end{pmatrix} \begin{pmatrix} u_y \\ u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} e \\ \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

Pode ponderar pela acurácia (matriz W)

3) Single-step (Legarra et al., 2009): restrito ao G-BLUP

Legarra et al. (2009) propuseram uma abordagem baseada neste modelo tradicional, na qual a matriz de covariância dos efeitos aleatórios (também denominada de matriz de parentesco) contempla simultaneamente informações de parentesco baseadas em pedigree (matriz **A** tradicional) e baseadas em marcadores SNPs, a denominada matriz de parentesco genômico (**G**), que é dada por: $\mathbf{G} = \mathbf{M}\mathbf{M}'/2 \sum_i q_i(1-q_i)$ sendo **M** a matriz de genótipos (N linhas e p colunas, em que N é o número de animais genotipados e p é o número de marcadores) e q_i a menor frequência alélica de cada marcador i.

Dessa forma, considerando o modelo tradicional:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

em que:

y é o vetor de fenótipos (de todos os indivíduos sob análise) corrigidos para efeitos fixos e desregressados para efeitos genéticos de genitores;

μ representa a média geral;

u representa diretamente o vetor EGBV, assumindo que $\mathbf{u} \sim N(0, \mathbf{H}\sigma_u^2)$;

e representa o vetor de resíduos, assumindo que $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$;

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12} & \mathbf{G} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix},$$

\mathbf{A}_{11} é a matriz de parentesco tradicional entre os indivíduos não genotipados,

\mathbf{A}_{12} é a matriz de parentesco tradicional entre os indivíduos genotipados e não genotipados,

\mathbf{A}_{22} é a matriz de parentesco tradicional entre os indivíduos genotipados e \mathbf{G} é a matriz de parentesco genômico entre indivíduos genotipados.

Qual a melhor opção para incorporar informações genômicas?

- 1) Índice de seleção**
- 2) Multicaracterístico**
- 3) Single-Step**

Como comparar?

4 Seleção Genômica para Características Longitudinais

Embora recentemente vários métodos estatísticos tenham sido propostos para GWS, este vêm sendo aplicados apenas para características (fenótipos) pontuais, como pesos em idades específicas e/ou taxas de crescimento em certos períodos.

Assim, surge o interesse em aplicar estes métodos a características longitudinais, tais como: curvas de crescimento, de progresso de doença, de lactação, de produção de ovos, de digestibilidade, dentre outras. De forma geral, análises de GWS para estas características permitem estimar o valor genômico dos animais para qualquer tempo de interesse, assim como os efeitos de marcadores.

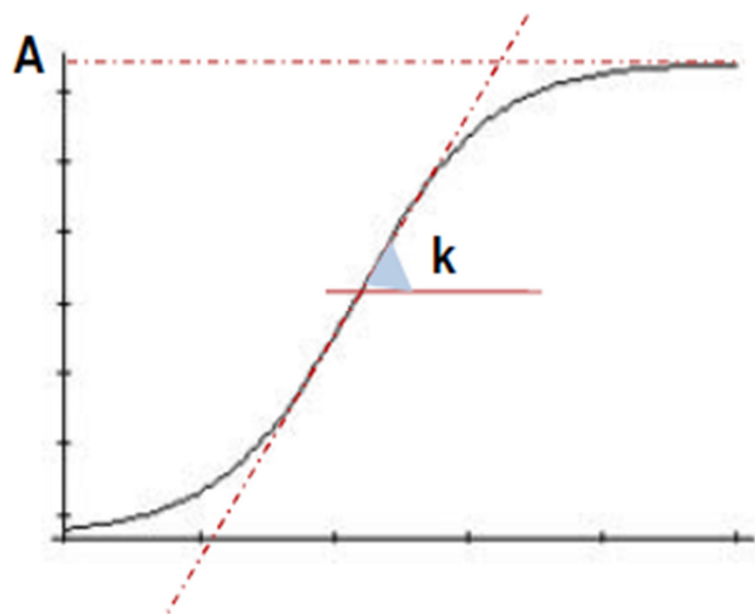
Para ilustrar uma análise GWS para características longitudinais, serão utilizados dados simulados utilizados provenientes do QTLMAS2009 (Workshop of Quantitative Trait Loci Mapping and Marker Assisted Selection - <http://www.qtlmas2009.wur.nl/UK/>) realizado na Holanda em abril de 2009.

O conjunto de dados consiste de 1000 indivíduos com informações completas de 453 marcadores SNPs, aleatoriamente distribuídos sobre 5 cromossomos. Tais dados também foram analisados por Rocha (2011) e Pong Wong (2010), e constam de 5 informações longitudinais de crescimento para cada indivíduo. As trajetórias longitudinais serão descritas pelo modelo Logístico já apresentado no módulo I.

modelo de regressão não-linear

- Utilizados quando os modelos de regressão polinomiais não se ajustam bem a trajetória observada
- Apresentam parâmetros com interpretação biológica (assíntota e taxa de crescimento)
- Muito utilizado em análises de curvas de crescimento, de lactação, de produção de ovos, de absorção de nutrientes, de produção em geral, etc...

•Ex. Modelos de curvas de crescimento



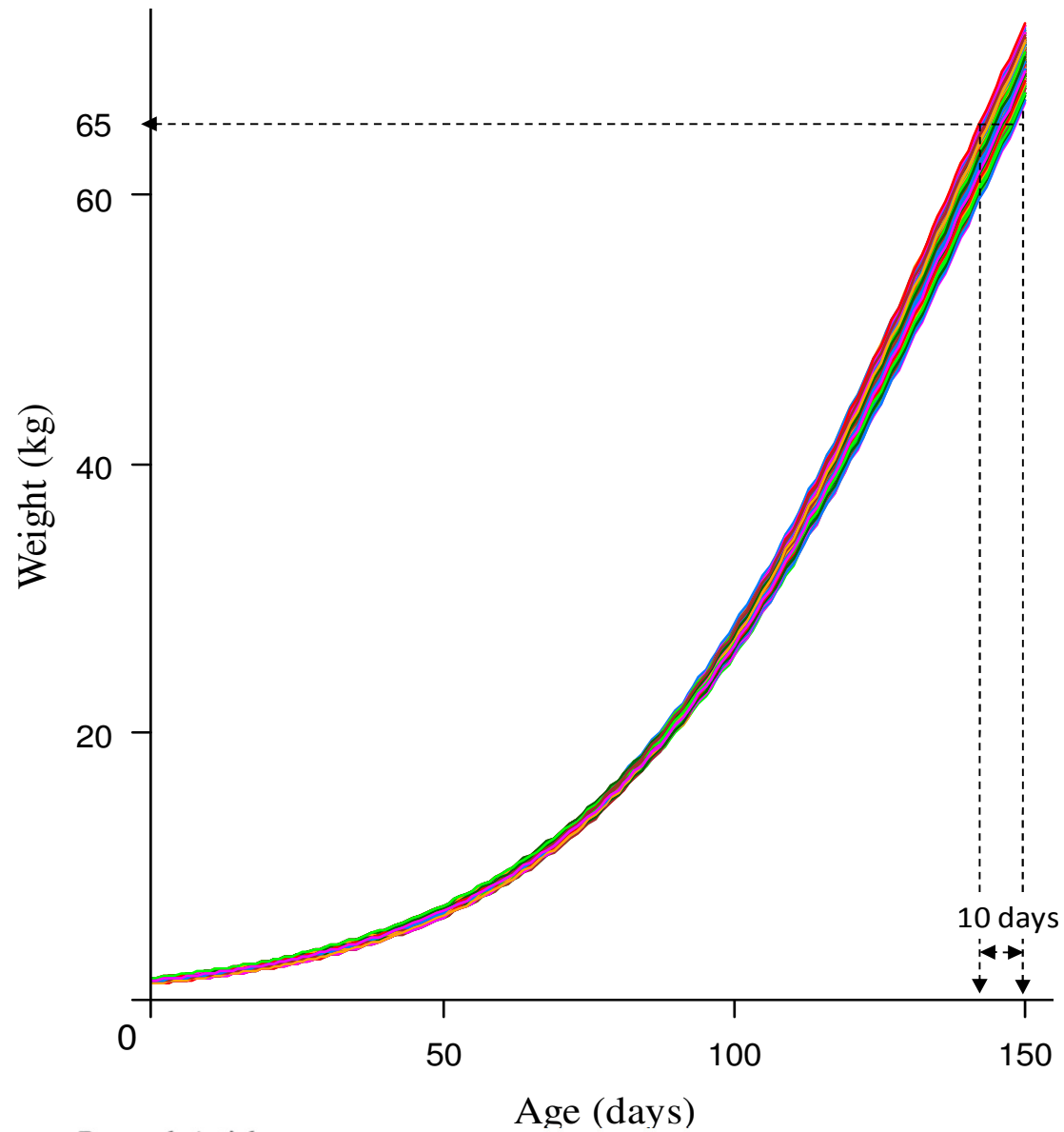
A: valor máximo atingido (peso assintótico)

K: velocidade com que A é atingido
(medida de precocidade)

b: não tem interpretação biológica
(algumas vezes pode ser fixado)

Quadro 1. Modelos não lineares utilizados para o ajuste das curvas de crescimento

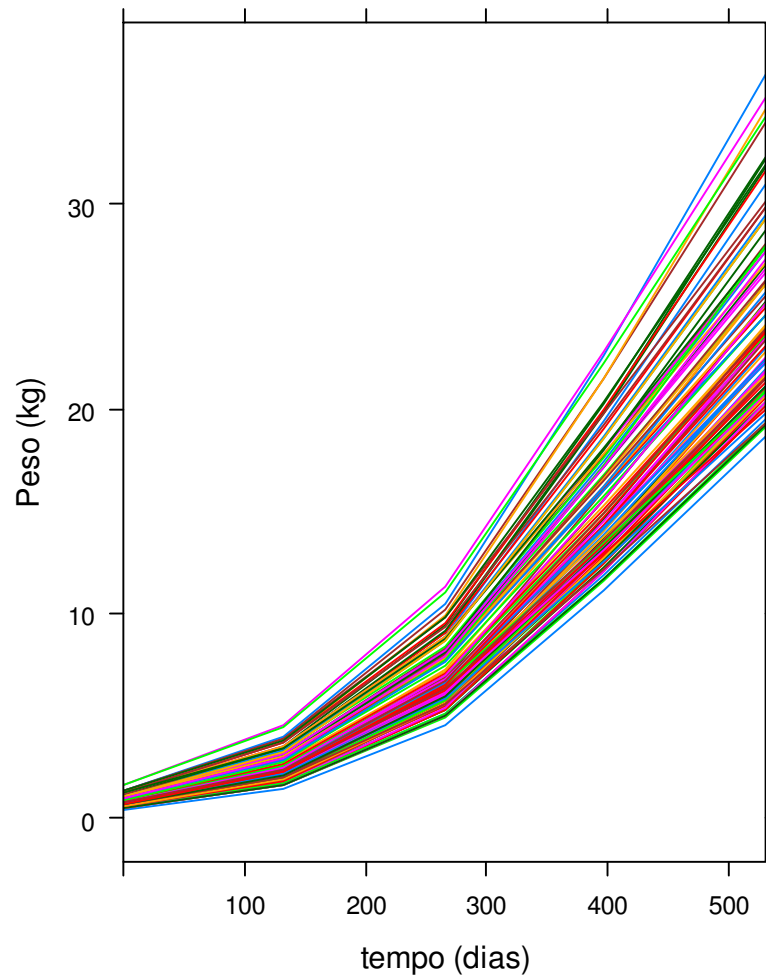
Modelos	Equações	Número de parâmetros
Brody	$W = A(1 - B \exp(-kt))$	3
Logístico	$W = A(1 + B \exp(-kt))^{-1}$	3
Gompertz	$W = A \exp(-B \exp(-kt))$	3
Richards	$W = A(1 - B \exp(-kt)^n)$	4



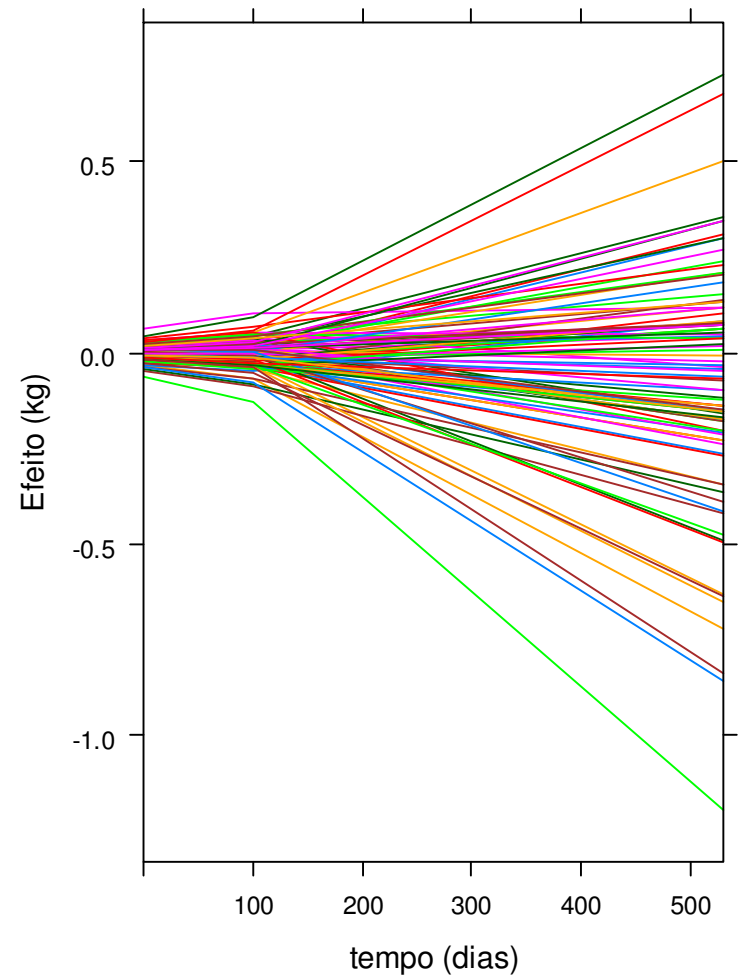
Research Article

Genomic growth curves of an outbred pig population

Fabyano Fonseca e Silva¹, Marcos Deon V. de Resende², Gilson Silvério Rocha¹, Darlene Ana S. Duarte^{1,3}, Paulo Sávio Lopes³, Otávio J.B. Brustolini⁴, Sander Thus⁵, José Marcelo S. Viana⁶ and Simone E.F. Guimarães³



“curva de crescimento genômica”



“Efeito de marcadores SNPs ao longo do tempo”

Modelos de regressão aleatória podem ser usados como alternativas aos modelos não lineares (vantagens e desvantagens???)



Journal of Animal Science



[HOME](#) | [CONTACT US](#) | [HELP](#) | [ARCHIVES](#) | [PAPERS IN PRESS](#) | [ASSOCIATION NEWS](#) | [MEETINGS](#)

Sire evaluation for total number born in pigs using a genomic reaction norms approach¹

F.F. Silva^{*,2}, H.A. Mulder[†], E.F. Knol[‡], M.S. Lopes[‡],
S.E.F. Guimarães^{*}, P.S. Lopes^{*}, P.K. Mathur[‡], J.M.S. Viana[§] and
J.W.M. Bastiaansen[†]

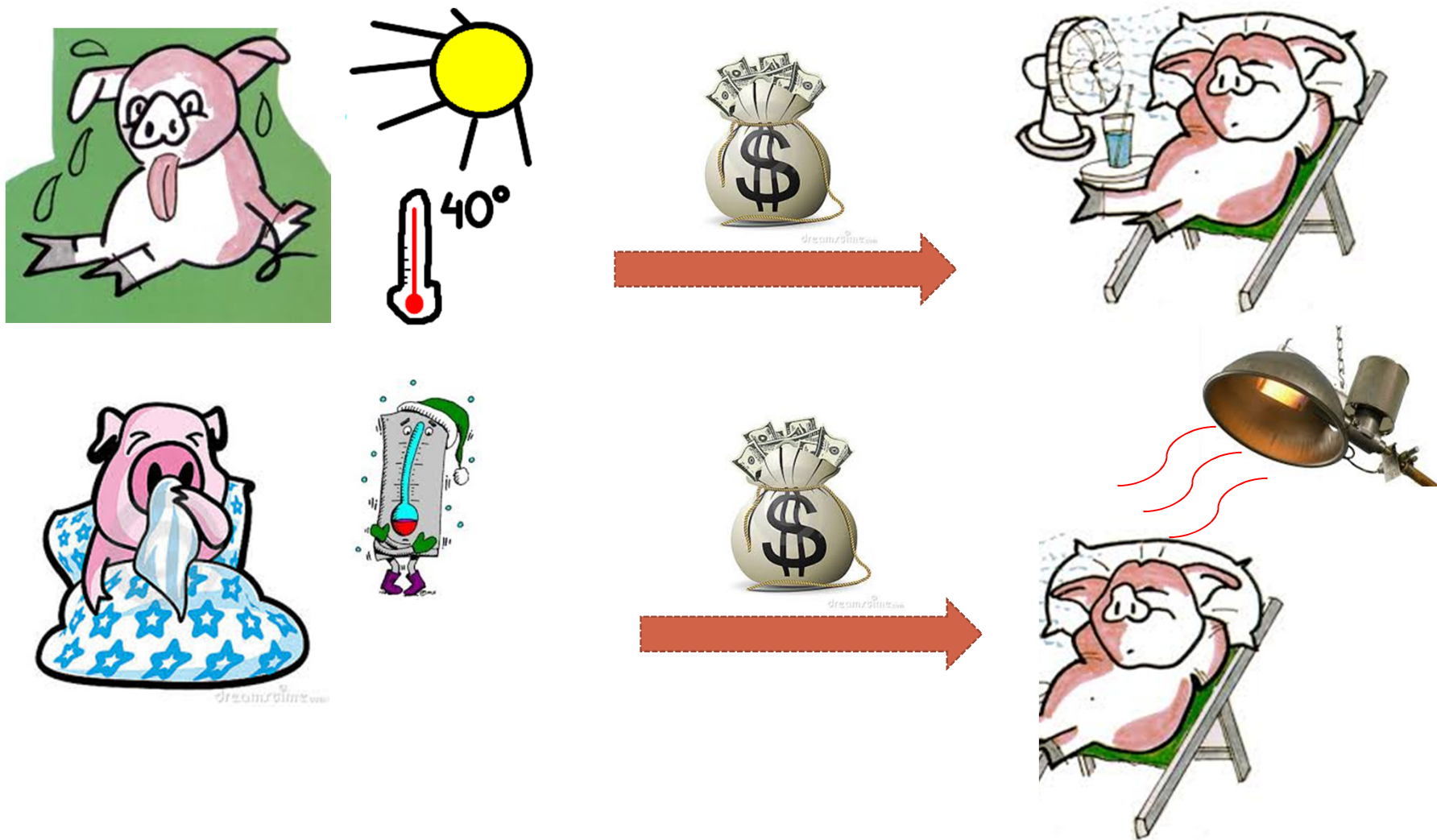
Melhoramento de suínos em rebanhos núcleos (empresas multinacionais)



**Mesmos reprodutores
com descendentes em
diferentes países
(técnicas de
biotecnologia)**

- **Inseminação artificial**
- **transferência de embrião)**
- **-clonagem**

Interação G x E em suínos



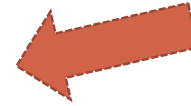
Statistical models

The two-step reaction norm approach (Calus et al., 2002; Kolmodin et al., 2002) was used to perform the G×E analysis.

From a practical GWS viewpoint, the two-step reaction norms may be preferred because this methodology is applied directly in traditional software for performing random regression, such as ASREML (Gilmour et al., 2002) and WOMBAT (Meyer, 2007), which allow easy replacement of the traditional relationship matrix (A) with a genomic relationship (Van Raden, 2008) matrix (G).

In the first step, a general sire model was fitted:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_{\text{line}} + \mathbf{X}_2\boldsymbol{\beta}_{\text{parity}} + \mathbf{X}_3\boldsymbol{\beta}_{\text{hys}} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$



O quer dizer
este modelo?

where \mathbf{y} is the vector of the TNB to the daughters of the considered sires; $\boldsymbol{\beta}_{\text{line}}$, $\boldsymbol{\beta}_{\text{parity}}$ and $\boldsymbol{\beta}_{\text{hys}}$ are the fixed effects of the line, parity and herd-year-season (HYS), respectively, with corresponding incidence matrices of \mathbf{X}_1 , \mathbf{X}_2 and \mathbf{X}_3 ; \mathbf{u} is the vector of the sires' additive genetic effects, $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{A})$; and \mathbf{e} is the residual random term, $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$.

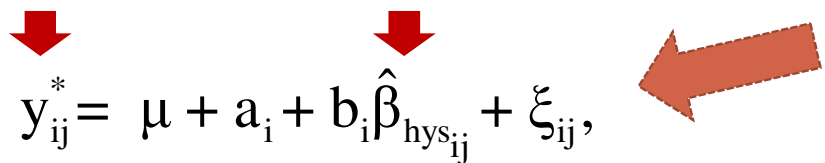
The aim of this first step was to provide a vector of the pre-corrected
phenotypes (\mathbf{y}^*), which were corrected for fixed effects,

$$\mathbf{y}^* = \mathbf{y} - (\mathbf{X}_1 \hat{\boldsymbol{\beta}}_{\text{line}} + \mathbf{X}_2 \hat{\boldsymbol{\beta}}_{\text{parity}} + \mathbf{X}_3 \hat{\boldsymbol{\beta}}_{\text{hys}}) = \mathbf{Z} \hat{\mathbf{u}} + \hat{\mathbf{e}}, \text{ and a vector of the}$$

HYS level estimates ($\hat{\boldsymbol{\beta}}_{\text{hys}}$). These variables were used in the
reaction norm model in the second step as dependent and
independent variables, respectively.

In the second step, the following random regression reaction norm

(RRRN) model was fitted:


$$y_{ij}^* = \mu + a_i + b_i \hat{\beta}_{hys_{ij}} + \xi_{ij},$$

Modelo que tenta explicar como o
componentes genético do primeiro modelo
varia em função dos níveis ambientais?

where y_{ij}^* represents pre-corrected TNB values from the daughters of sire i for level j of the estimated HYS effect ($\hat{\beta}_{hys}$); μ is the general mean; a_i and b_i are the random intercept and random slope, respectively, for the regression of the additive genetic value (u_i) over HYS levels; and ξ_{ij} is the residual term, $\xi_{ij} \sim N(0, \sigma_k^2)$.

Parentesco entre os animais

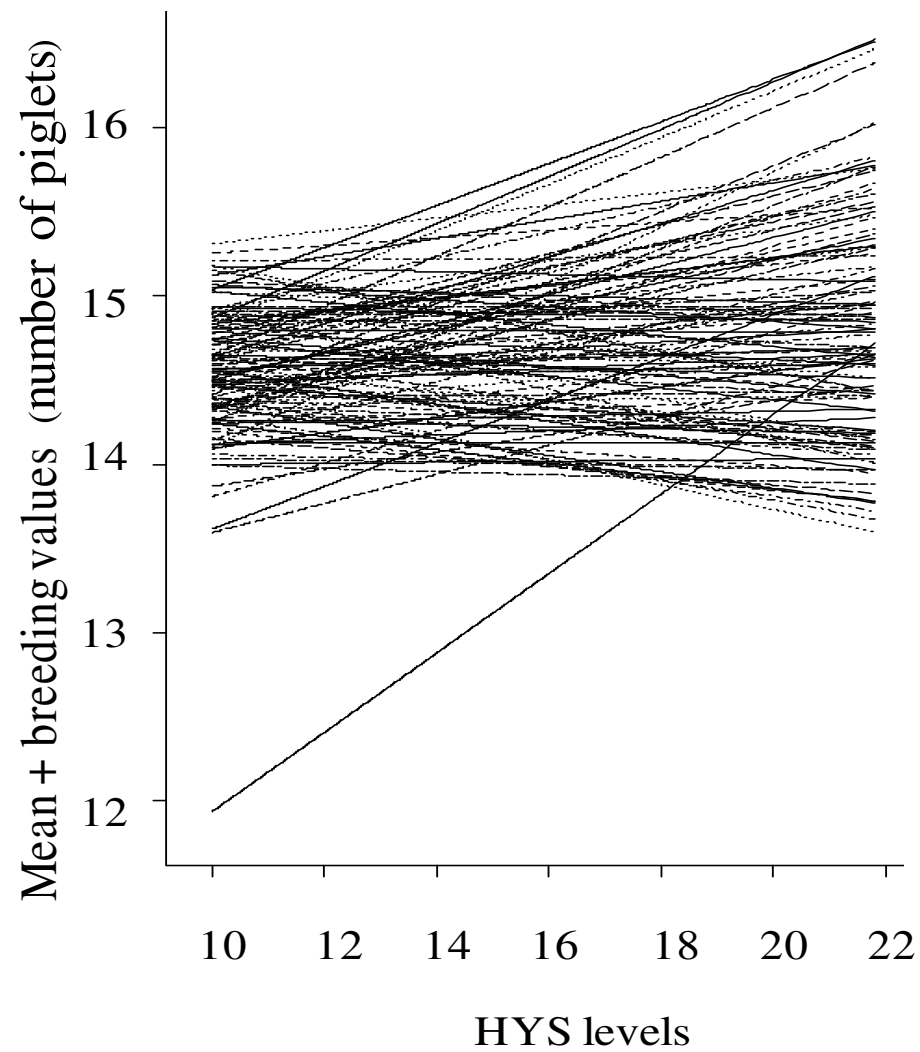
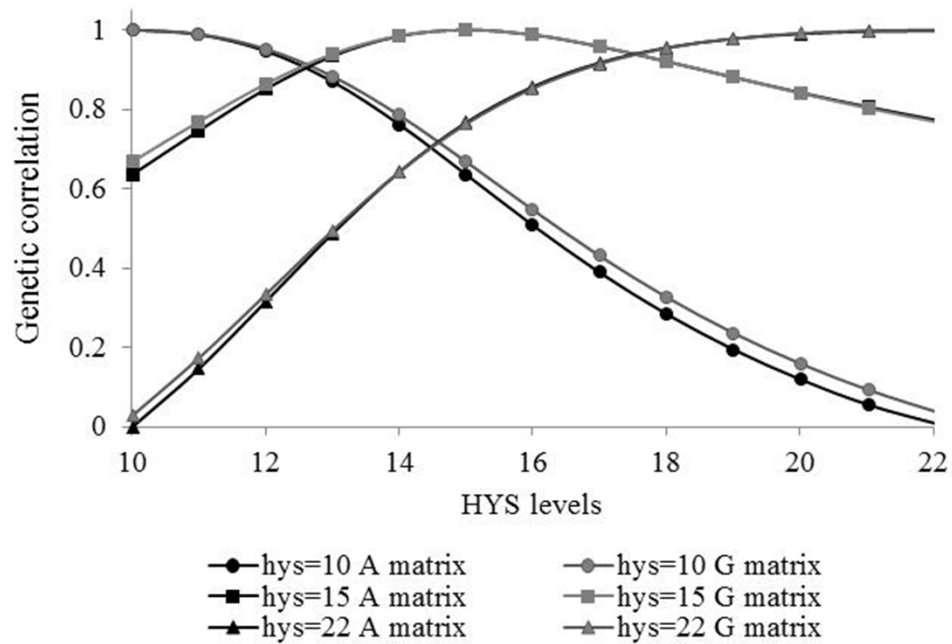
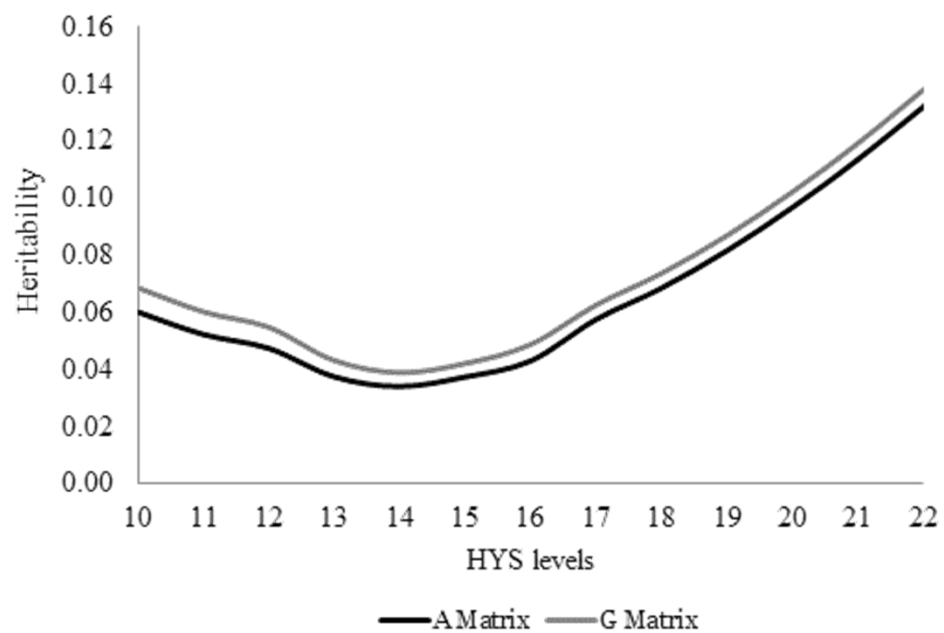
$$\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{ab} \otimes \mathbf{A}), \text{ where } \boldsymbol{\Sigma}_{ab} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}.$$

↓
Parentesco tradicional

Matriz A foi substituída por genômica (G-BLUP)

$$\mathbf{G} = \mathbf{M}\mathbf{M}' / \sum_{m=1}^M 2p_m(1-p_m) \text{ (Van Raden, 2008),}$$

Both matrices (**A** and **G**), as well as their inverses (**A**⁻¹ and **G**⁻¹), were obtained using *preGSf90* software (<http://nce.ads.uga.edu/~ignacy/>). Models [1] and [2] were fitted with *ASREML* (Gilmour et al., 2002) software using the *.giv* qualifier to enter **A**⁻¹ and **G**⁻¹ in the mixed model equations.



$$\begin{aligned}
\hat{\phi}_j &= (\mathbf{M}'\mathbf{M})^{-1}(\mathbf{M}'\hat{\mathbf{u}}_j) = (\mathbf{M}'\mathbf{M})^{-1}(\mathbf{M}'(\hat{\mathbf{a}} + \hat{\mathbf{b}}\hat{\beta}_{\text{hys}_j})) = (\mathbf{M}'\mathbf{M})^{-1}(\mathbf{M}'\hat{\mathbf{a}} + \mathbf{M}'\hat{\mathbf{b}}\hat{\beta}_{\text{hys}_j}) \\
&= \underbrace{(\mathbf{M}'\mathbf{M})^{-1}(\mathbf{M}'\hat{\mathbf{a}})}_{\text{SNP effect for a: } \hat{\mathbf{a}}_{\text{SNP}}} + \underbrace{(\mathbf{M}'\mathbf{M})^{-1}(\mathbf{M}'\hat{\mathbf{b}})}_{\text{SNP effect for b: } \hat{\mathbf{b}}_{\text{SNP}}} \hat{\beta}_{\text{hys}_j}.
\end{aligned}$$

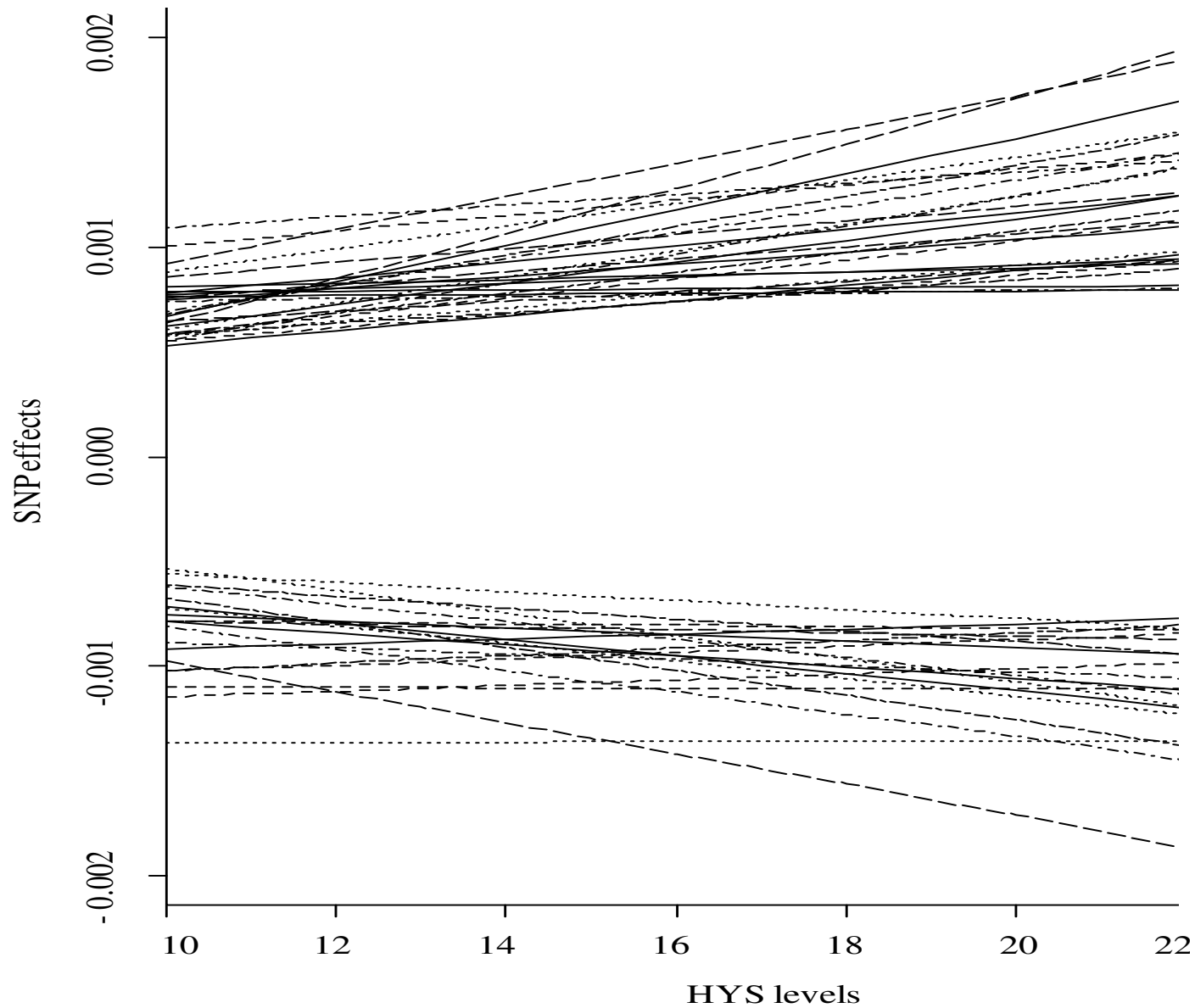
Thus, a general linear prediction equation can be obtained:

$$\hat{\phi}_j = \hat{\mathbf{a}}_{\text{SNP}} + \hat{\mathbf{b}}_{\text{SNP}}\hat{\beta}_{\text{hys}_j},$$

Primeira proposta de equação na literatura!

which allows the estimation of a vector of the SNP effects for each

HYS level of interest, within the observed range of $\hat{\beta}_{\text{hys}}$.



(b)

Number of common SNPs between 462 most relevant SNPs (a) in each level of HYS and effect estimates of the 47 most relevant SNPs for all HYS level (b).

Identificação dos genes pertencentes a um dado intervalo

- NCBI <http://www.ncbi.nlm.nih.gov/>
- Pacote NCBI2R: função *GetGenesInSNPs* (Returns a vector of genes that the provided SNPs are located within.)
- Map2NCBI (ainda não disponível)
Mapping genomic markers to closest feature using the R package Map2NCBI (livestock Science – in press)

Softwares para GO (gene ontologu)



GO**RILLA**



Gene Ontology enRIchment anaLysis and visuaLizAtion tool

<http://cbl-gorilla.cs.technion.ac.il/>



GO**stat**

by Tim Beißbarth
beissbarth@wehi.edu.au

Find statistically overrepresented GO terms within a group of genes

<http://gostat.wehi.edu.au/>

Aula prática 5

**Seleção genômica para
dados longitudinais**