

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



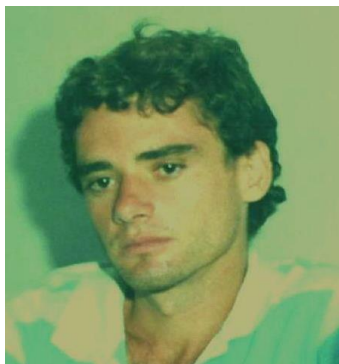
Seleção Genômica Ampla: aspectos teóricos e computacionais



Fabyano Fonseca e Silva

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

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



Colaboradores

Dr. Marcos Deon V. de Resende
Embrapa Florestas/UFV



Profa. Simone E. F. Guimarães – DZO/UFV

Prof. Paulo Sávio Lopes – DZO/UFV



Prof. Moysés Nascimento
DET-UFV

Profa. Camila Azevedo
DET-UFV





**Dr. Guilherme J.M. Rosa
(UW, Madison - USA)**



**Dr. Luis Varona –
(Universidad
Zaragoza , Espanha)**



**Dr. Egbert Frank Knol
(Topigs, Holanda)**



**Dr. John Bastiaansen
(Wageningen University,
Holanda)**



**Dr. Stephen Moore
(University of Queensland, Austrália)**



**Dr. Matt Kelly
(University of Queensland, Austrália)**

GWS x GWAS/MAS

Genome-wide association and genomic selection in animal breeding¹

Ben Hayes and Mike Goddard

Genome 53: 876–883 (2010)

A different approach is to use all the genome-wide markers simultaneously to predict breeding values in an approach known as genomic selection (Meuwissen et al. 2001). The difference between MAS and genomic selection is that MAS only utilizes the SNPs that are significant in a GWAS, whereas genomic selection uses a genome-wide panel of dense markers so that all QTL are expected to be in LD with at least one marker. There are two advantages

MOTIVAÇÃO

Journal of
Animal Breeding and Genetics



J. Anim. Breed. Genet. ISSN 0931-2668

J. Anim. Breed. Genet. **127** (2010) 336–337

EDITORIAL

More than a third of the WCGALP presentations on genomic selection

David Habier, Iowa State University, USA

“A relevância da Seleção Genômica (SG) no melhoramento animal já está comprovada”, ... mas ainda existem desafios a serem superados

- **GWS na prática: Incorporação de resultados GWS em programas de melhoramento (“blending”)**
- **Validação em populações diferentes daquelas em que os efeitos de marcadores SNPs foram estimados precisa ser melhor estudada**
- **Generalização dos métodos propostos para admitir fenótipos específicos (Ex. dados binários, de contagens, longitudinais e avaliações multicaracterísticas)**

1. Introdução à Seleção Genômica Ampla

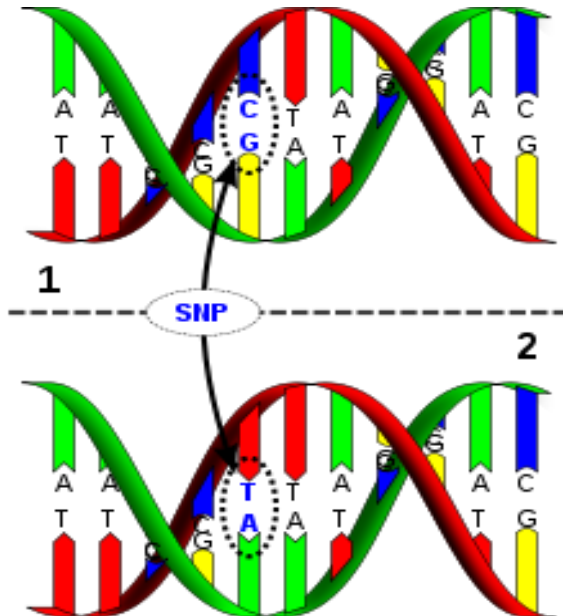
1.1 Marcadores SNPs

➤ *Single Nucleotide Polymorphism (SNP – Cho et al., 1999)*

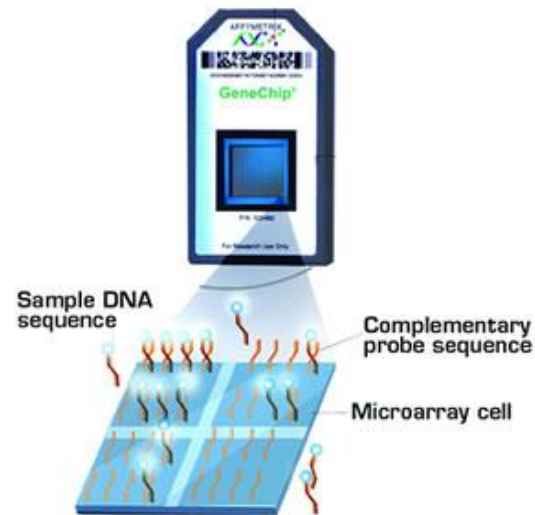
- se baseiam nas alterações mais elementares da molécula de DNA, isto é, mutações em apenas uma das bases nitrogenadas da cadeia (A, C, T, G).

- são a forma mais frequente de variação genética (90%)

- são extremamente abundantes nos genomas (amplamente distribuídos)



Possibilidade de genotipagem em massa (SNP chip)



Ind1

Ind2

DNA

DNA

Ind3

The diagram illustrates the Ind3-mediated DNA repair pathway. It shows a DNA double helix with a double-strand break. A circular inset shows a crossover event. The main diagram shows the DNA strands being repaired, with a sequence of bases (T, C, A, G) being added to the broken strand.

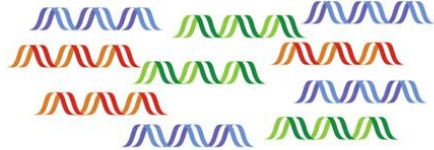
AT	T	G
AT	C	A
AT	T	G
AT	T	G
AT	C	A
AT	C	A

CCT TTT GAT
CCT TTT GAT

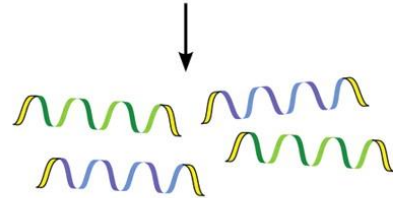
CCT TTT GAT
CCT TTT GAT

CCT TTT GAT
CCT TTT GAT

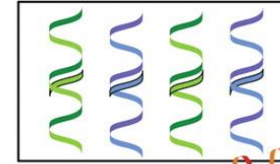
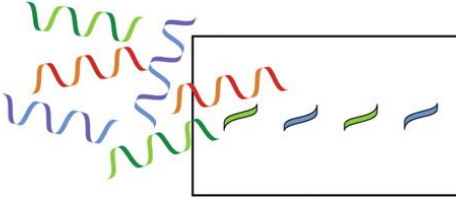
Como funciona o processo de genotipagem??



Genomic DNA is sheared into fragments and size selected, then separated into single strands.



Universal adaptor sequences are ligated to the target pool of DNA

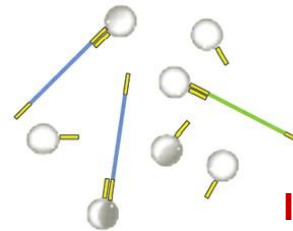
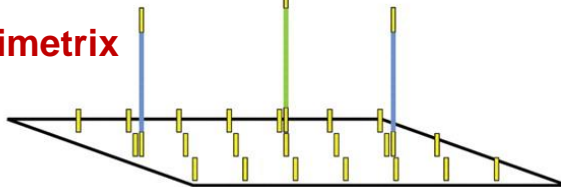


To select for particular areas of the genome, DNA is captured by complementary fragments of DNA or RNA on fixed arrays (shown) or on beads in solution.

SNPchip



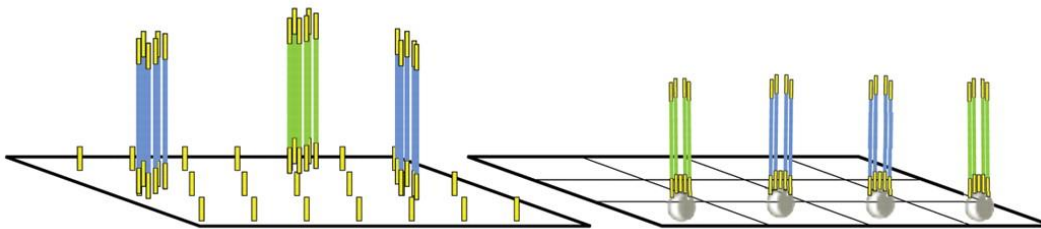
Affimetrix



Illumina



The DNA fragments are washed over an array or incubated with microscopic beads such that one DNA molecule is anchored by its adaptor on a single bead or away from other fragments on an array.

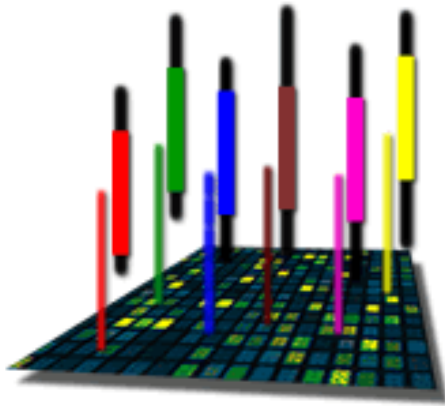


Raffan E , and Semple R K Br Med Bull 2011;bmb.lbr029

© The Author 2011. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

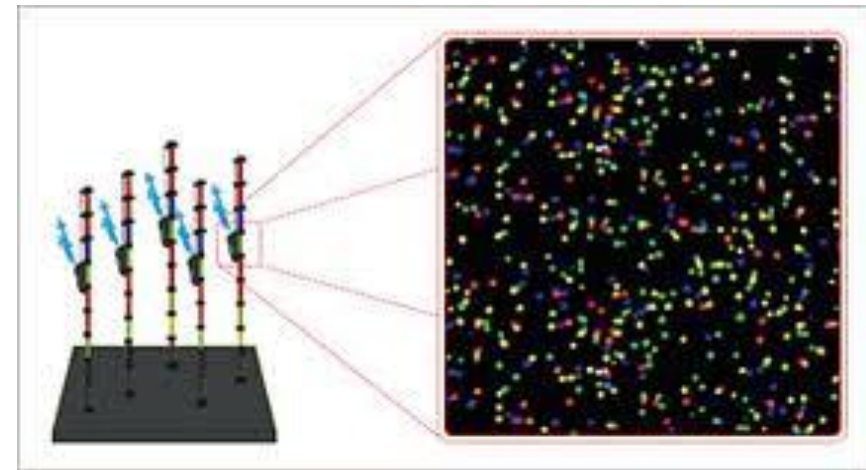
**BRITISH MEDICAL
BULLETIN**

DNA fixado em áreas distintas do array (SNPship)



Sequenciadores:

“basicamente” fazem uso de técnicas modernas de fluorescência para transformar bases (C, G, T, A) em cores distintas com diferentes comprimentos de ondas



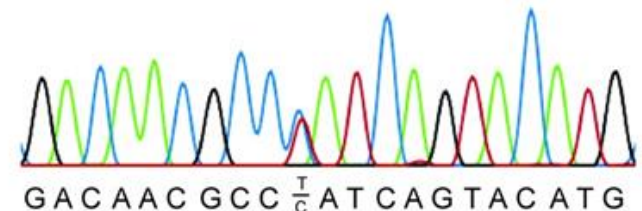
Identificação das sequências geradas

A ID3131: MYH7 (p.Y162H)

```

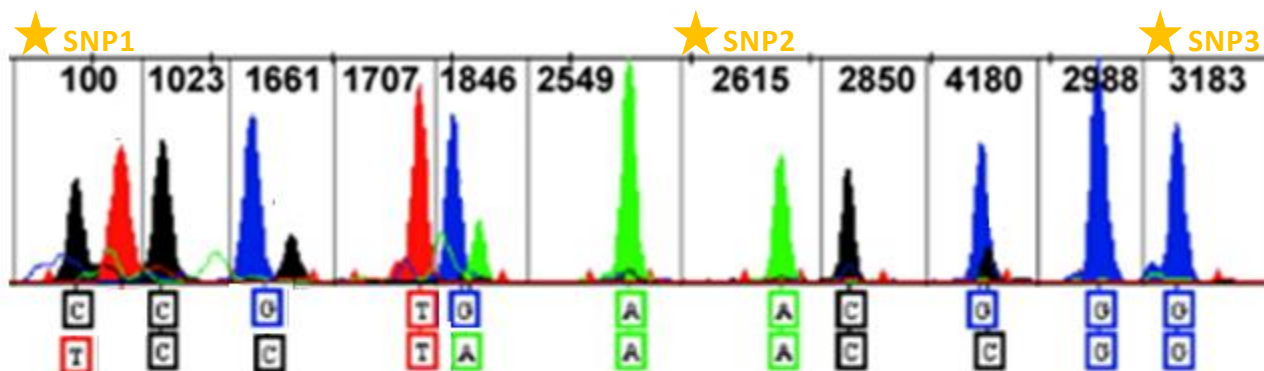
CCCACATCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAGGCCCTGGAA
ATCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAG
ATCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAG
ATCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAG
ATCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAG
TCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAGG
TCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAGG
TCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAGG
TCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAGG
TCTTCTCCATCTCCGACAACGCCATCAGTACATGCTGACAGGTGAGAGG

```

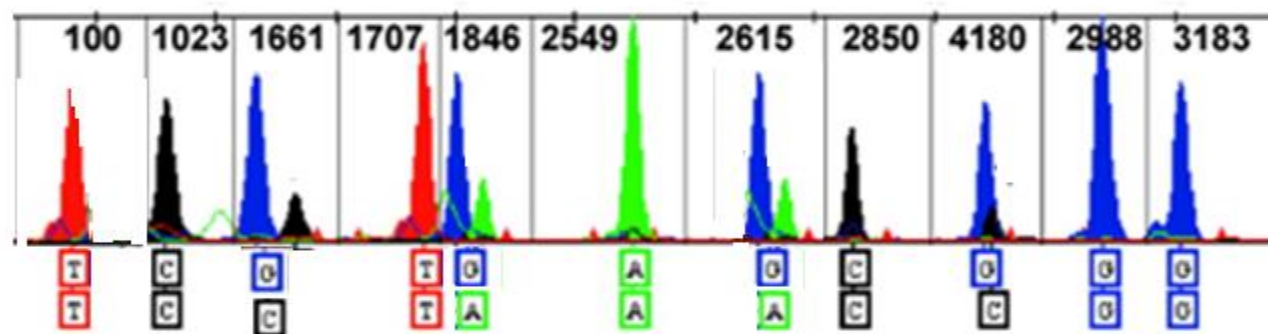


█ C
 █ T
 █ G
 █ A

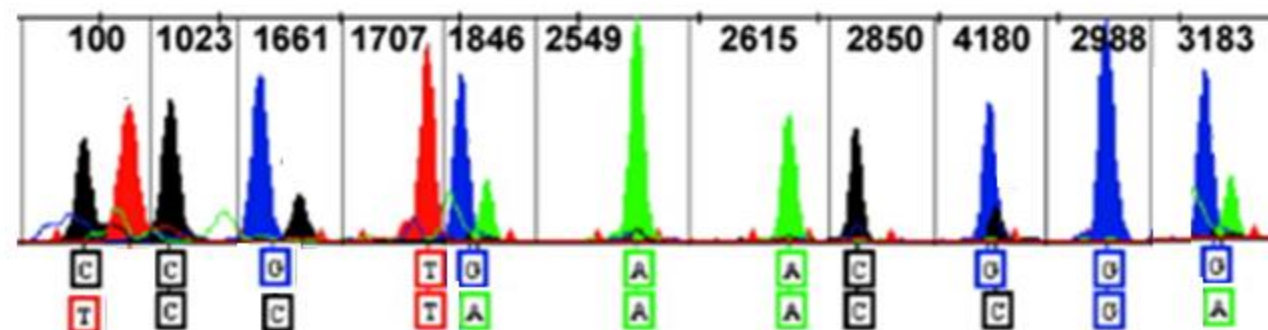
Position, bp



sample 1

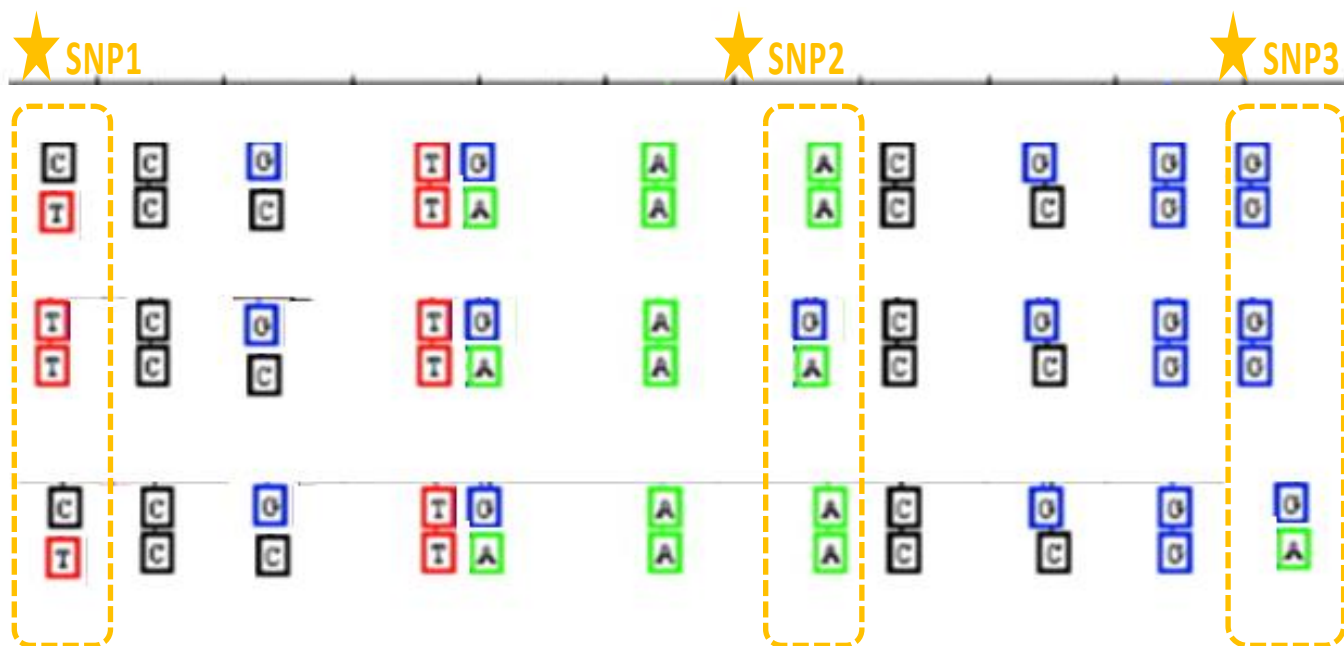


sample 2



sample 3

Identificação dos
SNPs via leitura
de diferentes
comprimentos de
ondas

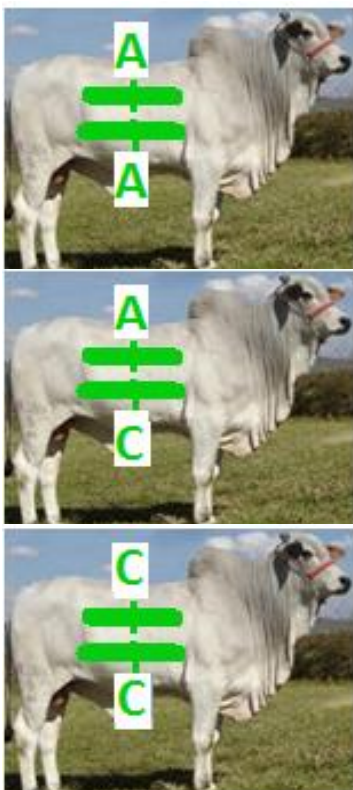


Output básico de uma genotipagem

SNPName	SampleID	Allele1-Top	Allele2-Top	GCscore	Allele1-AB	Allele2-AB
snp1	1	C	T	0.9374	A	B
snp2	1	A	A	0.9568	A	A
snp3	1	G	G	0.7996	B	B
snp1	2	T	T	0.9604	B	B
snp2	2	G	A	0.9296	A	B
snp3	2	G	G	0.8934	B	B
snp1	3	C	T	0.8967	A	B
snp2	3	A	A	0.7583	A	A
snp3	3	G	A	0.5377	A	B

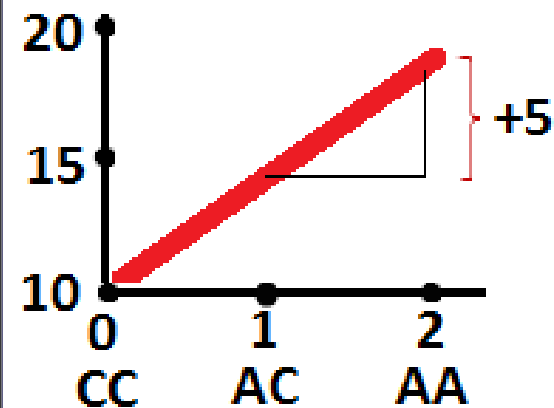
Qual a importância de um único SNP?

AAGCCTTGATAATT
AAGCCTTGCTAATT



Genótipo SNP	Média da característica
AA (2)	+20
AC (1)	+15
CC (0)	+10

Efeito estimado "alelo" A = + 5



Obs.:assumindo
mesma freq alélica

Regressão linear simples: a substituição do alelo C pelo A, aumenta em média, 5 unidades na característica de interesse

$$y = \beta_0 + \beta_1 x + e, \quad x = \{0, 1, 2\}$$

Resultado da genotipagem (vários SNPs):

Genotypes of chromosome pair

Individual 1

ATCGATCCTTAAATTTACTATT
ATCGATCCATAACAATTTACTTT

Individual 2

ATGGATCCATAAATTTACAAAT
ATCGATCCTTAAATTTACTATT

⋮

Individual N

ATGGATCCTTAAATTTACTATT
ATGGATCCTTAAATTTACTATT

snp_1 snp_2 ... snp_m

Qual a importância de vários SNPs?

	Phenotype	Genotype		
Individual 1	2.5	.. C	T	T
		.. C	A	T
Individual 2	4.8	.. G	A	A
		.. C	T	T
⋮				
Individual N	4.7	.. G	T	T
		.. G	T	T

	Phenotype	Genotype		
Individual 1	2.5	.. 0	1	2
Individual 2	4.8	.. 1	1	1
⋮				
Individual N	4.7	.. 2	2	2

$$y = \beta_0 + x_1\beta_1 + \dots + x_m\beta_m + e$$



Interpretação do efeito de substituição alélica para cada SNP



Um efeito para cada SNP



Regressão múltipla

y: fenótipo

X: 0,1 e 2 (aa, aA, AA)



β_1, \dots, β_m efeitos a serem estimados simultaneamente

1.2 Valor Genético Genômico (EGBV)

$$\hat{y} = \text{EGBV} = Z_1 \hat{b}_1 + Z_2 \hat{b}_2 + \dots + Z_m \hat{b}_m$$

Genotipo id 1 para snp 1 Efeito do snp 1 Genotipo id 1 para snp 2500 Efeito do snp 2500

$$\text{EGBV}_1 = z_{1,1} \hat{b}_1 + z_{2,1} \hat{b}_2 + \dots + z_{2500,1} \hat{b}_{2500}$$

Example: $\text{EGBV}_1 = 0 (0.25) + 1 (-0.10) + \dots + 2 (0.12) = 2.5$

aa 0 Aa -0.1 AA 0.24

$$\text{EGBV}_2 = z_{1,2} \hat{b}_1 + z_{2,2} \hat{b}_2 + \dots + z_{2500,2} \hat{b}_{2500}$$

Example: $\text{EGBV}_2 = 1 (0.25) + 0 (-0.10) + \dots + 2 (0.12) = 1.6$

Aa 0.25 aa 0 AA 0.24

⋮

$$\text{EGBV}_{634} = z_{1,634} \hat{b}_1 + z_{2,634} \hat{b}_2 + \dots + z_{2500,634} \hat{b}_{2500}$$

Example: $\text{EGBV}_{634} = 1 (0.25) + 2 (-0.10) + \dots + 2 (0.12) = 0.5$

Aa 0.25 AA -0.2 AA 0.24

2500 SNPs
634 indivíduos

$$\text{EGBV} = Z\hat{b}$$

Vector estimativas efeitos de SNPs

Vetor valor genômico estimado (EGBV)

Matriz de genotipos (SNPs)



1.3 Vantagens da Seleção Genômica Ampla (GWS)

- Não é necessário realizar análise de ligação (ordenamento de marcadores), uma vez que devido a grande saturação do genoma com marcadores SNPs, assume-se que estes estão diretamente em LD com o QTL. Além disso, as posições de todos os marcadores são conhecidas (evita a etapa de construção de mapas de ligação usada quando se tem marcadores microsatélites)
- Sendo o sucesso da seleção assistida por marcadores (MAS) dependente de um grande número de indivíduos genotipados, os SNPchips apresentam como uma solução viável, devido ao processo automatizado de genotipagem. Em relação aos marcadores microsatélites, para um grande número de indivíduos (>500) estes podem ser inviáveis, principalmente por logística laboratorial.
- Em relação ao BLUP, a vantagem é considerar o parentesco sem pressuposições fundamentadas em termos esperados (coef. parentesco de Wright), pois a matriz de parentesco tradicional é definida em termos médios (Ex. irmãos completos apresentam, EM MÉDIA, o mesmo coeficiente, mas sob o ponto de vista genômico é possível observar diferenças entre estes coeficientes)

1.3 Vantagens da Seleção Genômica Ampla (GWS)

➤ Possibilidade de “extrapolar” os resultados (efeitos de marcadores) obtidos de um grupo de indivíduos para outro grupo de indivíduos relacionados de alguma forma com o primeiro (Ex. gerações). Assim, diferentemente da análise clássica de QTL via microssatélites, a qual apresenta “validade” apenas dentro da população estudada (Ex. QTLs identificados em famílias F2 só podem ser usados nesta população), ao se usar SNPs informações genéticas podem ser exploradas para outras populações (gerações).



A Seleção Genômica Ampla (GWS) permite estimar o mérito genético (valores genômicos) de indivíduos que ainda não tiveram seus fenótipos coletados baseando-se apenas em seus genótipos e nos efeitos de marcadores estimados em análises prévias, desenvolvendo assim a teoria de populações de treinamento e validação.

OBS.: Curiosidades sobre a seleção genômica

- A GWS é proposta mais importante desde o BLUP (Dekkers, 2004)
- O potencial da GWS é demasiadamente grande para ser ignorado (Schaeffer, 2006)
- Incorporação de melhoristas quantitativos a genética molecular (Misztal, 2009)
- *More than a third of the WCGALP presentations on genomic Selection (Habier, 2010)*
J. Anim. Breed. Genet. 127 (2010) 336–337

Isto sim é tecnologia:

calcular EGBV para indivíduos que ainda não tiveram seu fenótipos coletados
(teoria de populações de treinamento e validação)

□ **população de treinamento: estimação dos efeitos de SNPs (\hat{a})**
Indivíduos genotipados e fenotipados

$$\text{Modelo geral: } y_i = \mu + \sum_{j=1}^J m_{ij} a_j + e_i \Leftrightarrow \mathbf{Y} = \mathbf{1}\mu + \mathbf{M}_t \mathbf{a} + \mathbf{e}$$

$$\mathbf{Y} = \begin{bmatrix} 19,53 \\ 37,53 \\ 31,30 \\ \vdots \\ 36,48 \end{bmatrix}_{634 \times 1}, \quad \mathbf{M}_t = \begin{bmatrix} 1 & 0 & 2 & \dots & 1 \\ 1 & 0 & 1 & \dots & 2 \\ 0 & 0 & 1 & \dots & 2 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 2 & 1 & 1 & 0 & 0 \end{bmatrix}_{634 \times 2500}, \quad \mathbf{e} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{2500} \end{bmatrix}_{2500 \times 1}$$

Objetivo:
 cor entre obs
 e pred
 (verificar
 possibilidade
 de selecionar
 com base
 apenas no
 genótipo)

□ **população de validação: estimação dos valores genômicos (EGBV)**
Apenas indivíduos genotipados (N indivíduos)

Matriz de marcadores
 pop. de validação

$$\text{EGBV} = \mathbf{1}\hat{\mu} + (\mathbf{M}_v \hat{\mathbf{a}})$$

Vetor de estimativas proveniente
 da pop. treinamento

$$\text{EGBV} = \begin{bmatrix} 30,4976 \\ 30,4976 \\ 30,4976 \\ \vdots \\ 30,4976 \end{bmatrix}_{N \times 1} + \begin{bmatrix} 0 & 2 & 1 & \dots & 0 \\ 0 & 2 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 2 & 2 & 1 & 1 \end{bmatrix}_{N \times 2500} \times \begin{bmatrix} -0,1871 \\ -0,0990 \\ 0,0887 \\ \vdots \\ 0,1773 \end{bmatrix}_{2500 \times 1} = \begin{bmatrix} 33,7189 \\ 38,4528 \\ 36,9261 \\ \vdots \\ 35,9370 \end{bmatrix}_{N \times 1}$$

EGBV

2. Análise de qualidade de SNPs

Genetic Epidemiology 34:591–602 (2010)

Quality Control and Quality Assurance in Genotypic Data for Genome-Wide Association Studies

Cathy C. Laurie,¹ Kimberly E. Doheny,² Daniel B. Mirel,³ Elizabeth W. Pugh,² Laura J. Bierut,⁴
Tushar Bhangale,¹ Frederick Boehm,¹ Neil E. Caporaso,⁵ Marilyn C. Cornelis,⁶ Howard J. Edenberg,⁷
Stacy B. Gabriel,³ Emily L. Harris,⁸ Frank B. Hu,⁶ Kevin B. Jacobs,⁵ Peter Kraft,⁹ Maria Teresa Landi,⁵
Thomas Lumley,¹ Teri A. Manolio,¹⁰ Caitlin McHugh,¹ Ian Painter,¹ Justin Paschall,¹¹ John P. Rice,⁴
Kenneth M. Rice,¹ Xiuwen Zheng,¹ and Bruce S. Weir^{1*} for the GENEVA Investigators

Em resumo, não se aproveita todos os SNPs, uma vez que a qualidade da genotipagem depende de vários fatores (Exemplos):

- **Relação entre a base genética genotipada e a sequencia referência (SNP)**
- **Qualidade do material biológico (amostras DNA)**
- **Variabilidade genética da população genotipada**
- **Sintonia (clareza) da análise de alinhamento (idoneidade da empresa)**

OBS. Marcadores microssatélites também demandavam análise de qualidade

2.1 Call rate

É uma medida de qualidade utilizada para eliminar SNPs com grande quantidade “valores perdidos” (*missing genotypes*). Esta medida é calculada proporcionalmente em relação ao número de observações válidas (*non-missing genotypes*), geralmente opta-se por trabalhar com SNPs cuja call rate seja maior que 95% (0.95).

$$CR = (\text{nº non-missing}) / (\text{nº total})$$

$$\text{Ex. } 437 / (437 + 63) = 0.874$$

2.2 MAF (Minor allele frequency)

É uma medida relacionada com a variação dos alelos na população, alelos pouco variáveis são pouco informativos e não apresentam relevância genética na população. Geralmente utiliza-se $MAF \geq 5\%$ (ou 0.05)

$$f(A) = \frac{(2 \times \text{número de AA}) + (\text{número de Aa})}{(2 \times \text{número total de indivíduos})} \quad f(a) = \frac{(2 \times \text{número de aa}) + (\text{número de Aa})}{(2 \times \text{número total de indivíduos})}$$

A MAF é o menor valor ($f(A)$ ou $f(a)$).

$$\text{Ex. } ((2 \times 53) + 196) / (2 \times 437) = \underline{0.3455378} \text{ (MAF)}$$

$$((2 \times 188) + 196) / (2 \times 437) = 0.654462$$

2.3 Equilíbrio de Hardy-Weinberg

É usado para verificar se as frequências genótípicas observadas estão de acordo com as esperadas conforme o EWH, caso não estejam pode haver problemas em exercer a seleção considerando os locos que desviam do EWH, pois estes podem estar altamente influenciados pelo tamanho da população, mutação, migração e seleção.

	<i>AA</i>	<i>aA</i>	<i>aa</i>
<i>Freq obs.</i>	53	196	188
<i>Freq esp.</i>	$Nf(A)^2$	$2Nf(A)f(a)$	$Nf(a)^2$
	$437*0.3455378^2 = 52.17$	$2*437*0.3455378*0.654462 = 197.64$	$437*0.654462^2 = 187.17$

$$\chi^2_c = \sum \frac{(O - E)^2}{E} = \frac{(53 - 52.17)^2}{52.17} + \frac{(196 - 197.64)^2}{197.64} + \frac{(188 - 187.17)^2}{187.17} = 0.0304$$

se $\chi^2_c \geq \chi^2_{\alpha}$ (1 gl) RHo, p-valor = $P(\chi^2 \geq \chi^2_c)$

p-valor = $1 - \text{pchisq}(0.0304, \text{df} = 1) = 0.8615857$

OBS. Na presença de m marcadores, são realizados m testes EHW, portanto deve-se corrigir para o nível global de significância. A forma mais prática é por meio da “proteção de Bonferroni”, a qual assume que para se acessar a significância p-valor < alfa/m. Recomenda-se que SNPs que se mostraram significativos para EHW sejam removidos da análise.

Aula prática 1

Controle de qualidade de marcadores SNPs