

## Imputação de genótipos

1

## Imputação

- **Por que imputar?**
- **Conceitos básicos**
- **Abordagens para a imputação**
- **Fatores que afetam a acurácia da imputação**

2

## Processo de Imputação

- O que é imputação?
  - ato de imputar (ou determinar)
  - atribuição
  - estimação de dados perdidos (ex.: genótipos)
- Exemplo: Jogo da Forca
  - Melhor amigo do homem ? Ã O

3



### Como funciona a imputação?

**Identifica** os haplótipos em uma população usando muitos marcadores

**Rastreia** os haplótipos a partir de poucos marcadores  
e.g., rastreia um bloco de 20 SNP a partir de 4 SNP

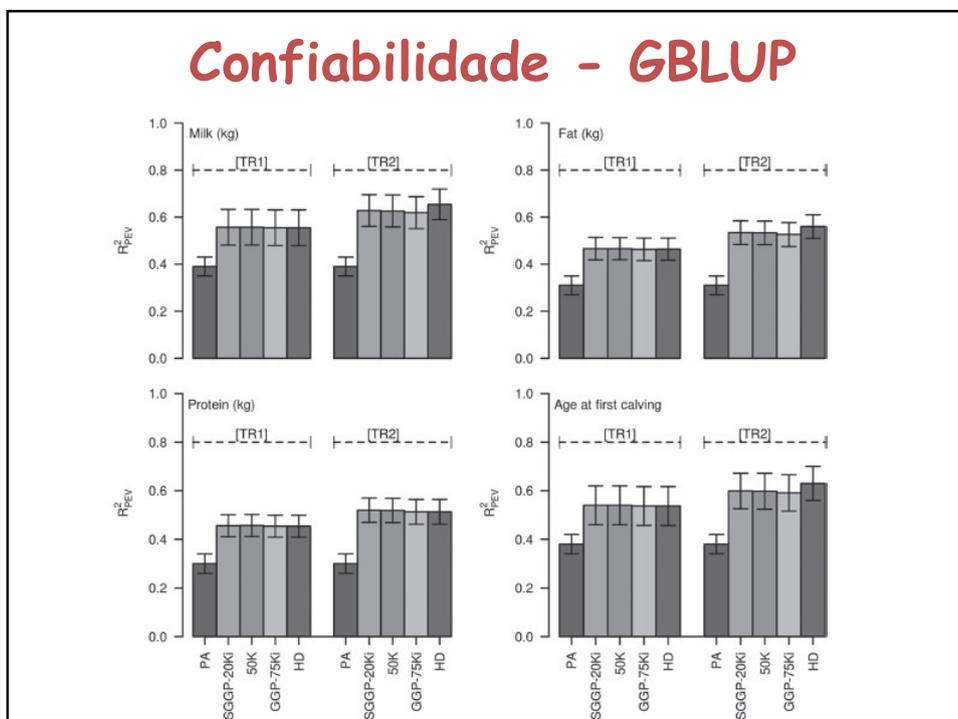
4 SNP: 2202

20 SNP: 20220200020020020002



## Por que se fazer imputação?

- Predizer genótipo da mãe a partir de sua progênie
- Predizer SNP conhecidos a partir de SNP desconhecidos  
 Genotipar **3,000**, predizer **50,000** SNPs  
 Genotipar **50,000**, predizer **800,000** SNPs
- Aumento da confiabilidade a menor custo (?)  
 \$25 ->\$125->\$300





**J. Dairy Sci. TBC:1–12**  
<https://doi.org/10.3168/jds.2016-11811>  
 © American Dairy Science Association®, TBC.

JDS11811

**Accuracy of genomic predictions in Gyr (*Bos indicus*) dairy cattle**

S. A. Boison,\* A. T. H. Utsunomiya,† D. J. A. Santos,† H. H. R. Neves,†‡ R. Carneiro,† G. Mészáros,\*  
 Y. T. Utsunomiya,† A. S. do Carmo,§ R. S. Verneque,§ M. A. Machado,§ J. C. C. Panetto,§ J. F. Garcia,#  
 J. Sölkner,\* and M. V. G. B. da Silva§<sup>1</sup>

\*Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences, 1180, Vienna, Austria  
 †Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista (UNESP), Jaboticabal, SP, 14884-900, Brazil  
 ‡GenSys Consultores Associados S/C Ltda, Porto Alegre, Brazil  
 §Empresa Brasileira de Pesquisa Agropecuária, Embrapa Gado de Leite, Juiz de Fora, MG, 360381330, Brazil  
 #Faculdade de Medicina Veterinária de Araçatuba, Universidade Estadual Paulista (UNESP), Araçatuba, SP, 16015-050, Brazil



**J. Dairy Sci. 98:4969–4989**  
<http://dx.doi.org/10.3168/jds.2014-9213>  
 © 2015, THE AUTHORS. Published by FASS and Elsevier Inc. on behalf  
 of the American Dairy Science Association®. Open access under [CC BY-NC-ND license](#)



**Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (*Bos indicus*) dairy cattle: Comparison of commercially available SNP chips**

S. A. Boison,\*<sup>1</sup> D. J. A. Santos,† A. H. T. Utsunomiya,† R. Carneiro,† H. H. R. Neves,† A. M. Perez O'Brien,\*  
 J. F. Garcia,‡ J. Sölkner,\* and M. V. G. B. da Silva§

\*University of Natural Resources and Life Sciences, Department of Sustainable Agricultural Systems, Gregor-Mendel 33, A-1180, Vienna, Austria  
 †Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista (UNESP), SP, 148841900, Brazil  
 ‡Faculdade de Medicina Veterinária de Araçatuba, Universidade Estadual Paulista (UNESP), Araçatuba, SP, 16015-050, Brazil  
 §Empresa Brasileira de Pesquisa Agropecuária, Embrapa Gado de Leite, Juiz de Fora, MG, 36038-330, Brazil



## Imputação de Genótipos

- Estimação de genótipos de marcadores usando dados de referência
- Amostra genotipada com low-density SNP chip

...A-----T-----C-----G-----...

- População de referência genotipada com high-density SNP chip

...AGCTTTAAGCCATACCTTAGGACATTACCTAGGAGCTTTAA-CCATAC...  
 ...AGCTTTAAGCCATACCTTAGGACATTACCTAGGAGCTTTAA-CCATAC...  
 ...

- Amostra imputada de low-density para high-density SNP chip

...AGCTTTAAGCCATACCTTAGGACATTACCTAGGAGCTTTAA-CCATAC...

FILLING THE GAPS!!!

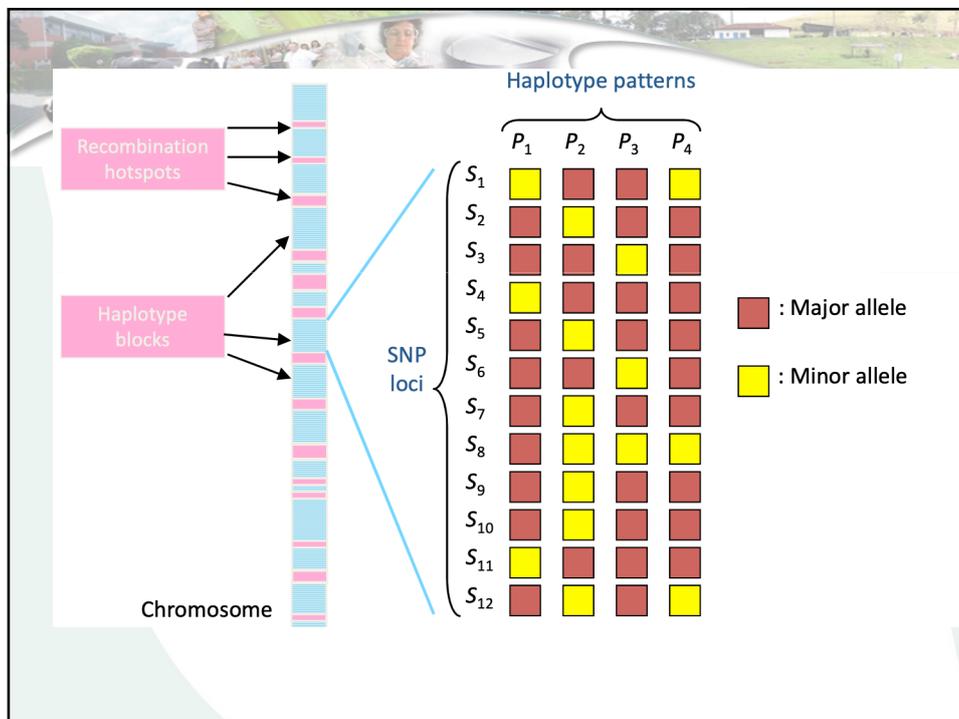
8

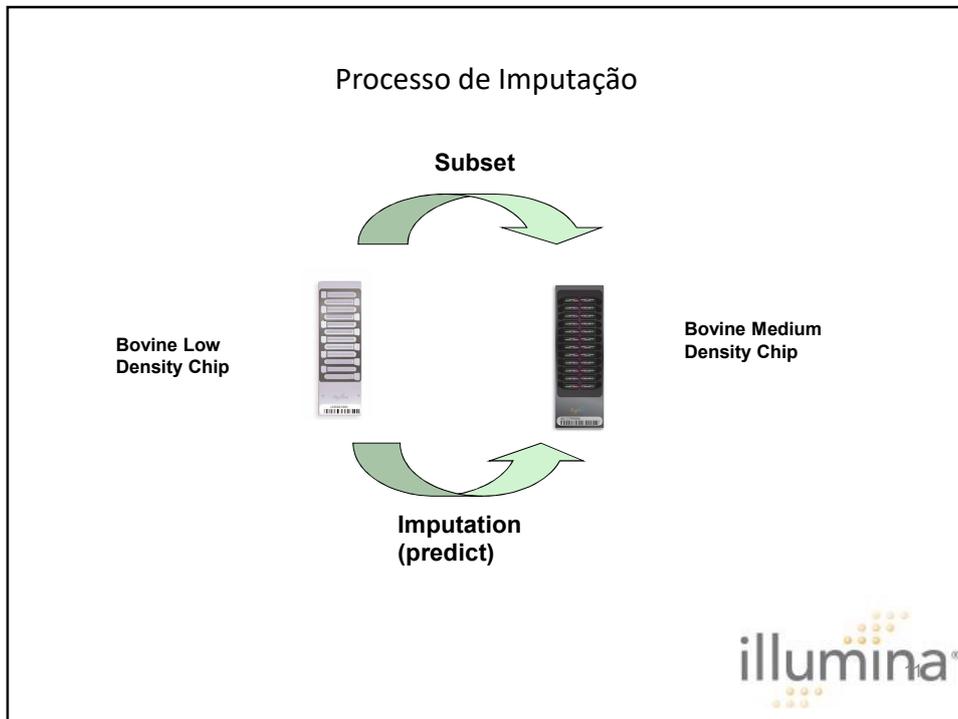
## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019

### Blocos de haplótipos e tagSNPs

- Blocos de haplótipos são regiões do genoma caracterizadas por apresentarem alto desequilíbrio de ligação e baixa diversidade;
- Nos blocos de haplótipos acontece pouca ou nenhuma recombinação;
- Dentro dessas regiões é possível capturar boa parte da variação utilizando apenas poucos marcadores, denominados tagSNPs (Wall & Pritchard, 2003);
- Zonas de recombinação são regiões do cromossomo da ordem de 1.000 a 2.000 pares de base, nas quais eventos de recombinação tendem a estar concentrados. Frequentemente, estão flanqueados por “coldspots”, que são regiões com menor frequência de recombinação;





**Imputação – de 20K (ZL2) para 50K**

**J. Dairy Sci. 100:9623–9634**  
<https://doi.org/10.3168/jds.2017-12732>  
 © American Dairy Science Association<sup>®</sup>, 2017.

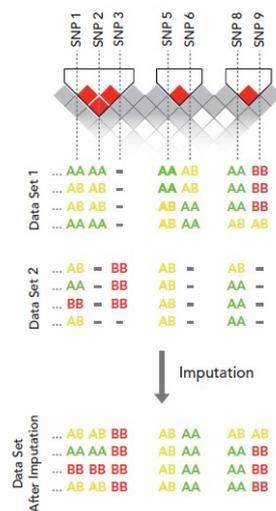
**Genotype imputation in a tropical crossbred dairy cattle population**

**Gerson A. Oliveira Júnior,\* Tatiane C. S. Chud,† Ricardo V. Ventura,‡§ Dorian J. Garrick,# John B. Cole,|| Danisio P. Munari,† José B. S. Ferraz,\* Erik Mullart,¶ Sue DeNise,\*\* Shannon Smith,\*\* and Marcos Vinicius G. B. da Silva††**

\*Departamento de Medicina Veterinária, Universidade de São Paulo (USP), Faculdade de Zootecnia e Engenharia de Alimentos, Pirassununga, SP, 13635-900, Brazil  
 †Departamento de Ciências Exatas, Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, SP, 14884-900, Brazil  
 ‡Beef Improvement Opportunities, Guelph, ON N1K1E5, Canada  
 §Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON N1G2W1, Canada  
 #Department of Animal Science, Iowa State University, Ames 50011-3150  
 ||Animal Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD, 20705-2350  
 ¶CRV Holding B.V., Arnhem, 454, the Netherlands  
 \*\*Zoetis, Kalamazoo, MI 49007  
 ††Embrapa Dairy Cattle, Brazilian Corporation of Agricultural Research, Juiz de Fora, MG, 36038-330, Brazil

CENÁRIOS	Correlação
300 GIROLANDO bulls	0.950
300 GIR bulls + 300 HOL bulls + 400 GIROL bulls and cows	0.957
300 GIR bulls + 600 HOL bulls + 400 GIROL bulls and cows	0.956
300 GIR bulls + 300 HOL_low* bulls + 400 GIROL bulls and cows	0.955
300 GIR bulls + 400 GIROL bulls and cows	0.955

## Combinação de dados de várias plataformas de genotipagem



SNPs 1-9 formam três blocos de alto LD. O conjunto de dados 1 e 2 representam 8 indivíduos genotipados com duas plataformas diferentes.

O conjunto de imputação possui as estimativas (imputação) dos SNP para o conjunto de dados 2.

Por exemplo, o SNP 2 é genotipado no conjunto 1 mas não no conjunto 2.

Devido ao alto LD entre os SNPs 1-3, o os alelos do SNP 2 podem ser inferidos no conjunto 2, a partir das informações do conjunto de dados 1.

13

## Métodos de imputação

- **Duas alternativas**
  - **Com base na família**
    - regras de segregação e de ligação mendeliana
    - mais preciso para animais com parentes genotipados
  - **Com base na população**
    - Usa as informações LD e os SNPs observado adjacentes
    - Animais não aparentados ou animais sem genótipo sem antepassados próximos
  - **Combinação das duas**

14

## Software para imputação

- Fimpute
- AlphaImpute
  - <https://sites.google.com/site/hickeyjohn/alphaimpute>
- Findhap
  - <http://aipl.arsusda.gov/software/findhap/>
- MACH 1.0
  - <http://www.sph.umich.edu/csg/yli/mach>
- IMPUTE2
  - [http://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
- Beagle
  - <http://faculty.washington.edu/browning/beagle/beagle.html>
- AlphaPhase
  - <https://sites.google.com/site/hickeyjohn/alphaphase>
- fastPHASE
  - <http://depts.washington.edu/uwc4c/express-licenses/assets/fastphase/>
- PLINK, SNPSTAT, UNPHASED and TUNA

15

## Considerações

- **Aumentando o tamanho da população de referência os resultados da imputação melhoram**
- **Diminuição do erro de imputação é decrescente com aumento na densidade de marcadores**
- **Avaliar o impacto da taxa de erro da imputação sobre as estimativas dos valores genômicos**
- **A densidade ótima vai depender do custo da genotipagem (baixa densidade) e a diminuição em acurácia dos valores genômicos**

16

Outra estratégia para diminuir os custos da genotipagem.....

## Fenotipagem e genotipagem seletiva dos animais mais informativos

17

### Estratégias de genotipagem para seleção genômica no gado leiteiro (Jiménez-Montero et al., 2010)

#### Estratégias de genotipagem seletiva:

2%, 5% e 10% indivíduos da população de referência foram selecionados como conjunto de treinamento com estratégias diferentes para características de 0,25 e 0,10 de herdabilidade

1. Aleatória (**RND**). As fêmeas escolhidas aleatoriamente da população de referência
2. Valores divergentes fenotípicos (**DPH**). Igual número de fêmeas no  $\alpha$  e  $(1-\alpha)$  percentis da distribuição "ajustada" fenotípica.
3. Valores divergentes EBV (**DBV**). As fêmeas com seus valores genéticos no  $\alpha$  e  $(1-\alpha)$  percentis.
4. Os maiores valores fenotípicos (**TopPH**). Vacas top no ranking de valores fenotípicos "ajustado"
5. Os maiores valores EBV (**TopBV**). Vacas top no ranking dos valores genéticos
6. Divergente família EBV (**DFM**). Fêmeas meio irmãos filhas dos touros melhores e piores em EBV.

#### Avaliação genômica do modelo

Bayesiano Lasso para estimar os coeficientes SNP na população analisada (6 estratégias)

Correlações de Pearson entre os valores genômicos estimados (GBV) e os valores genéticos verdadeiros (TBV), foram calculados na geração de 15

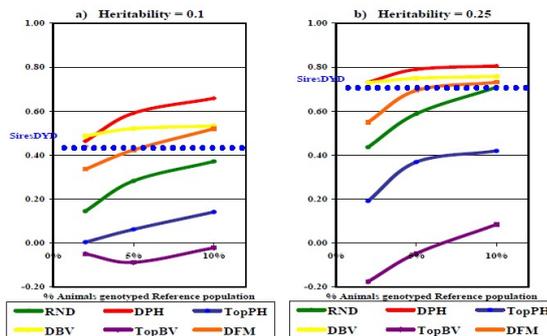
18

## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019

### Acurácia dos valores genéticos genômicos (corr (GBV, TBV))

**A acurácia preditiva do GBV depende da quantidade de animais genotipados e da estratégia de genotipagem seletiva usada.**



Acurácia dos GBV na geração 15, quando 2%, 5% e 10% das fêmeas na população de referência (G 11 - 14) foram genotipadas utilizando diferentes estratégias.

19

## Como aproveitar melhor seus recursos para genotipagem: Imputação de SNP genótipos

- Imputar / prever genótipos para:
  - Completar genótipos faltantes (limpeza de dados)
  - Marcadores não considerados no painel
- Utiliza a estrutura de haplótipos de amostras já existentes, tais como amostras do HapMap, para inferir os dados dos genótipos o marcadores faltantes na amostra analisada

20

## Imputação

- Identificar os haplotipos em uma população utilizando muitos marcadores (painel denso)
- Identificar os haplotipos com menos marcadores
- e.g., usar 5 SNP para identificar 25 SNP

- 5 SNP: 22020 População de validação
  - 25 SNP: 2022020002002002000202200 Haplótipo referência
- 

21

## Imputação de genótipos

- A imputação pode ser usado para inferir genótipos
  - Limpeza de dados
- Imputação pode ser utilizada para prever genótipos de alta densidade
  - 7k -> 50k ou 50k-> 700k
- Diminuir o custo da genotipagem (gado de corte e ovinos)
- Combinar conjuntos de dados
  - Aumento do tamanho da amostra

22

## Conceitos básicos

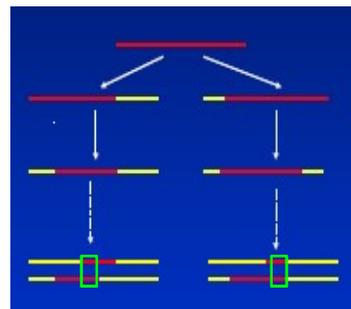
- **Genes ou segmentos Idênticos por estado (IBS)**
  - Um par de indivíduos têm o mesmo alelo num determinado locus
  
- **Idênticos por descendência (IBD)**
  - Um par de indivíduos têm os mesmos alelos em um locus e provem de um ancestral comum
  
- **Métodos de imputação determinam se um segmento de cromossomo é IBD**

23

• Pequenos segmentos do cromossomo na população atual são descendentes em comum do mesmo antepassado.

• Estes segmentos de cromossomos, que provem de um antepassado em comum sem intervenção da recombinação, carregam alelos ou haplótipos idênticos (IBD).

• Portanto estas regiões estão conservadas, ou seja dois indivíduos parentes vão carregar (compartilhar) os mesmos alelos (IBD)



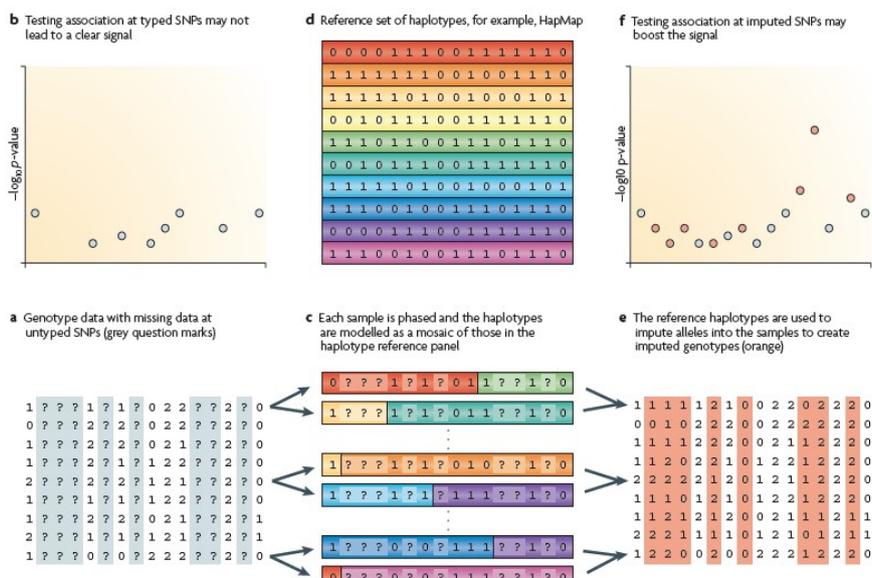
24

## Conceitos básicos

- Indivíduos na população podem compartilhar uma parte do seu genoma (IBD)
  - Segmentos IBD são iguais e têm origem em um ancestral comum
- Quanto mais próximo o relacionamento mais longos são os segmentos IBD (*Long range phase*)

25

## Como funciona a imputação?



26

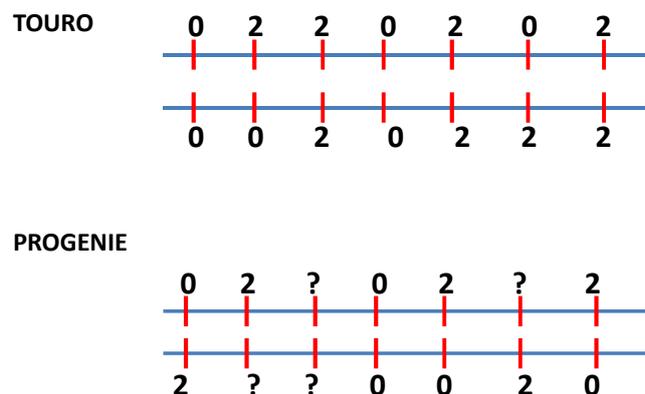
Marchini e Howie, 2010

## Métodos de imputação

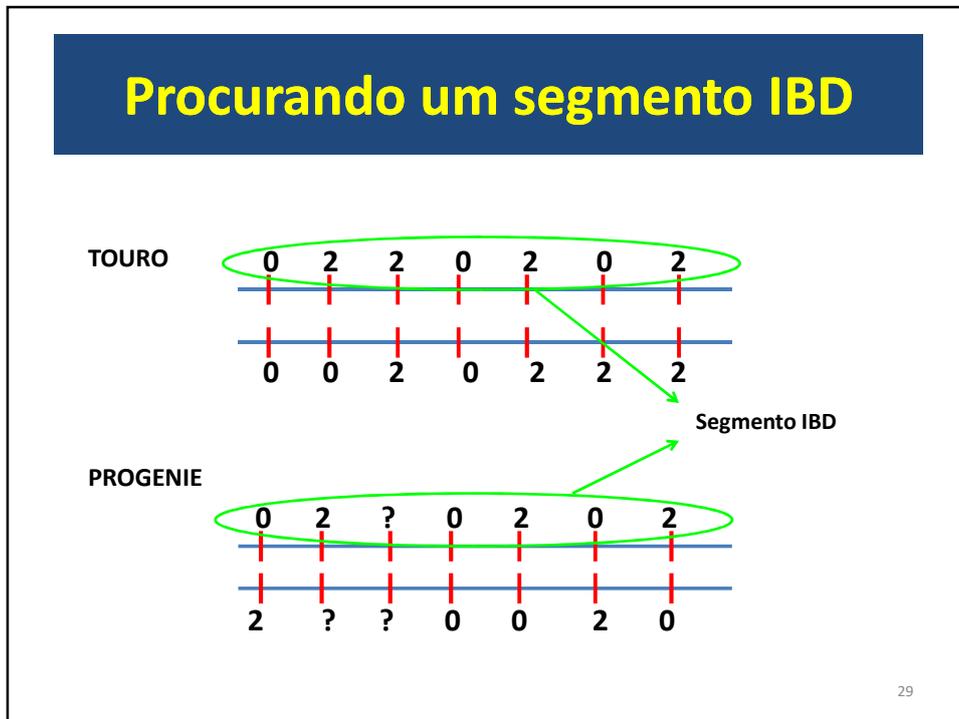
- **Duas alternativas**
  - **Com base na família**
    - regras de segregação e de ligação mendeliana
    - mais preciso para animais com parentes genotipados
  - **Com base na população**
    - Usa as informações LD e os SNPs observado adjacentes
    - Animais não aparentados ou animais sem genótipo sem antepassados próximos
  - **Combinação das duas**

27

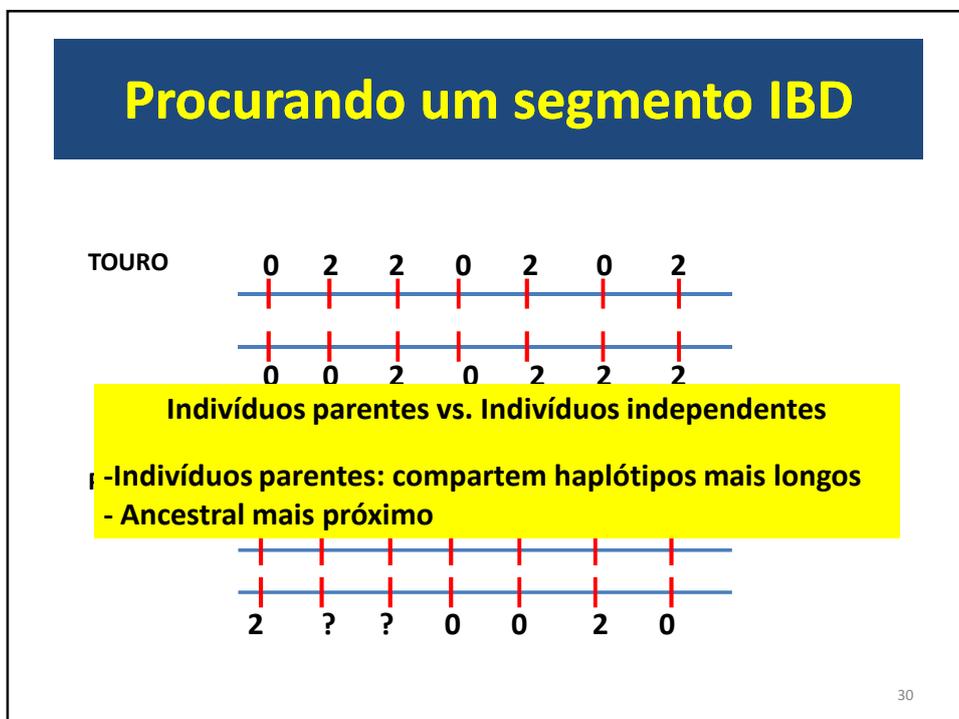
## Procurando um segmento IBD



28



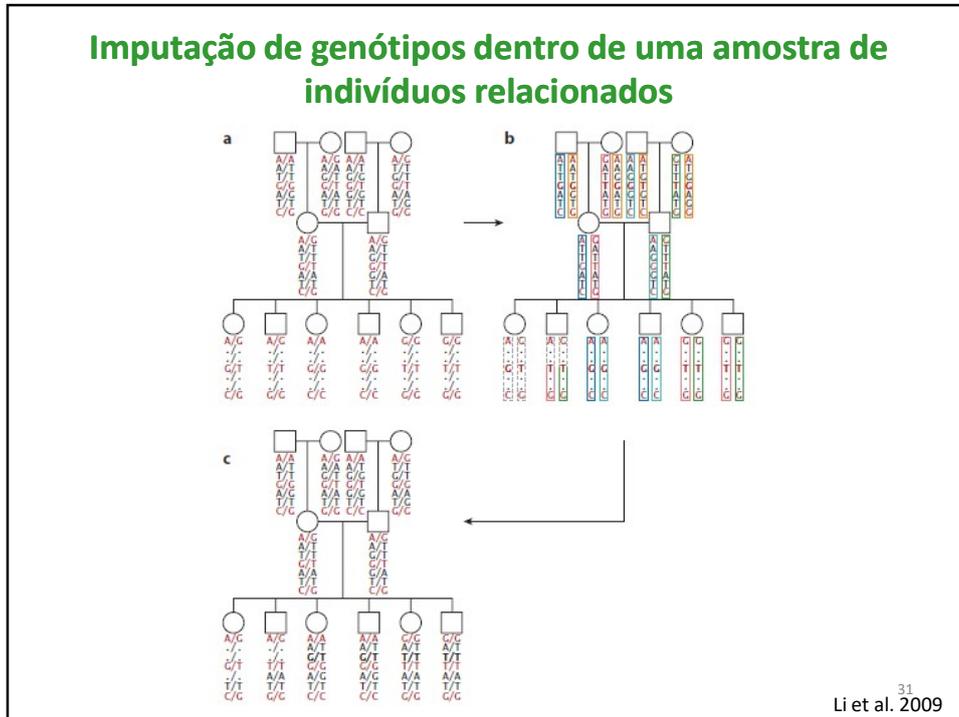
29



30

## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
 Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019



## Métodos de imputação com base na população

### Genótipos observados

. . . . . **A** . . . . . **A** . . . . . **A** . . . .  
 . . . . . **G** . . . . . **C** . . . . . **A** . . . .

### População de Referência

```

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
    
```

32

## Métodos de imputação com base na população

### Genótipos observados

. . . . . A . . . . . A . . . . . A . . . .  
 . . . . . G . . . . . C . . . . . A . . . .

### População de Referência

```

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
    
```

33

## Métodos de imputação com base na população

### Genótipos Observados

c g a g A t c t c c c g A c c t c A t g g  
 c g a a G c t c t t t t C t t t c A t g g

### População de Referência

```

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
    
```

34

## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019

### Ferramentas de imputação Genótipo se dividem em duas categorias principais:

- a. Ferramentas computacionalmente intensivas tais como Impute, MACH, e fastPHASE / BIMBAM, que levam em conta todos os genótipos observados quando imputam cada genótipo
- b. Ferramentas computacionalmente mais eficientes, tais como Plink, ATUM, WHAP, e Beagle, que geralmente se concentram em um pequeno número de marcadores adjacentes aos genótipos a serem imputados

35

### Hidden Markov model (HMM)

- ✓ Uma classe de modelo estatístico que pode ser utilizado para relacionar um processo observado no genoma para um processo não observado, subjacente de interesse.
- ✓ No contexto da imputação e inferência dos genótipo perdidos, os dados observados são os genótipos observados “*unphased*”, enquanto que o estado oculto (*hidden states*) representa os haplótipos e os genótipos verdadeiros.
- ✓ HMM têm sido amplamente utilizados para estimar a probabilidade de que um indivíduo carrega um genótipo particular (SNP particular), tendo em conta os dados genotípicos do indivíduo para outros SNPs e o resto da população.

36

## Imputação de base populacional: HMM

- **Hidden Markov Models**

- “hidden states” (fase de haplótipos e genótipos verdadeiros para todos os loci)
- Para os indivíduos de interesse estes fornecem mapas de referência dos haplótipos
- O problema da imputação é derivar probabilidades de determinado genótipo dado os estados ocultos (*hidden states*), genótipos esparsos, as taxas de recombinação, e parâmetros de outras populações

$$P(G|H, \theta, \rho) = \sum_s P(G|S, \theta) P(S|H, \rho)$$

37

## Imputação de base populacional: HMM

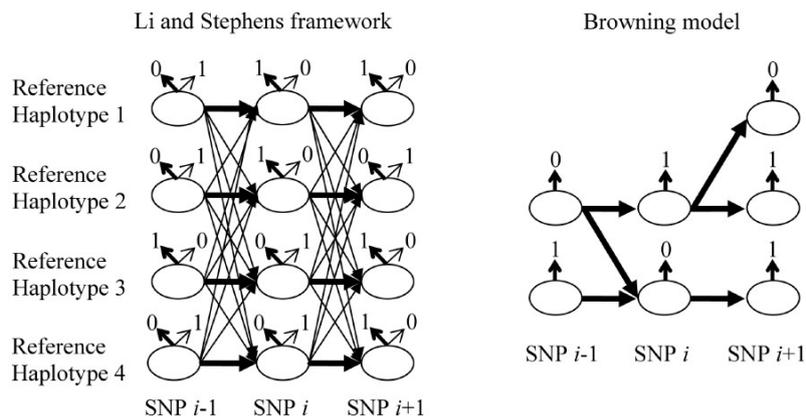


Ilustração destacando grandes diferenças entre os modelos com base no Li e Stephens (2003), a base para MACH, Impute e fastPHASE, e o modelo Browning (Browning 2006), a base de BEAGLE.

38

## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019

### Incluindo informações de pedigree para melhorar a acurácia de imputação

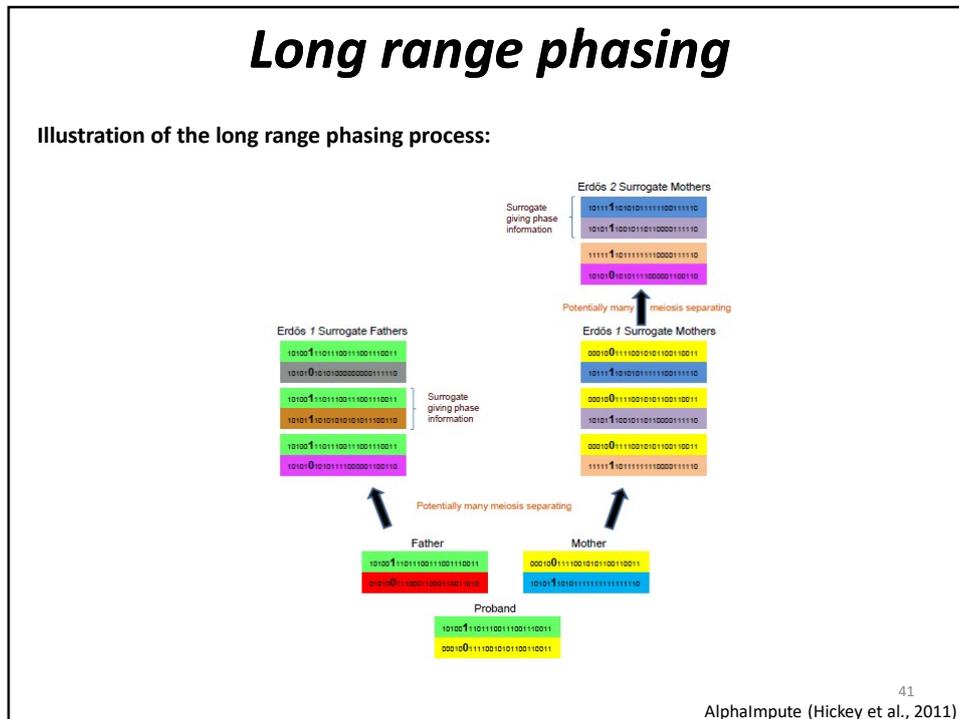
- Acurácia da imputação melhora quando o pedigree da população referencia e objetivo são conhecidos
- Quando o pedigree é conhecido, o número de estados ocultos (*hidden states*) que devem ser considerados pode ser reduzido a quatro, (2 alelos paternos 2 maternos).
- Importante quando a informação da população é pobre (baixo LD)

39

### Uma abordagem alternativa para *phasing* e imputação: *long range phasing* (Kong et al. (2008) and HICKEY et al. 2011)

- Alguns indivíduos compartilham um ancestral comum recente e, portanto, compartilhar segmentos de cromossomos longos (IBD)
- Isto é particularmente verdadeiro em populações de animais, onde alguns reprodutores têm número muito grande de descendentes
- Populações de animais têm  $N_e$  relativamente pequeno, grandes segmentos do cromossomo são compartilhados entre os indivíduos.

40



## Imputação de genótipos

- Requer recursos computacionais de grande escala
- Necessidade de avaliar a qualidade de imputação
  - Comparar os genótipos imputados com os genótipos reais
- Levar em conta a incerteza dos SNPs imputados na posterior análise

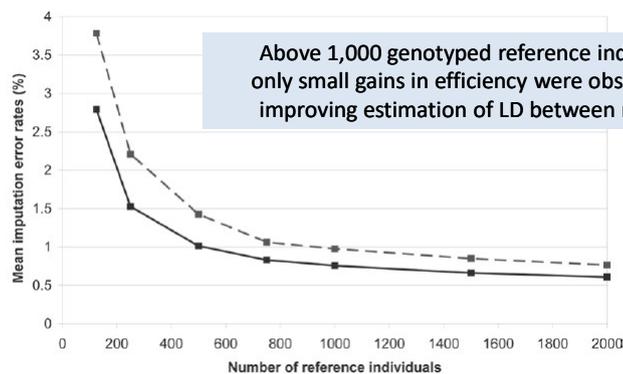
42

## Fatores que afetam a Imputação

- **Número de indivíduos genotipados com o painel de alta densidade (referência)**
- **Densidade de marcadores**
- **Frequência dos alelos do SNP**
- **Parentesco entre indivíduos genotipados com painel de alta e baixa**

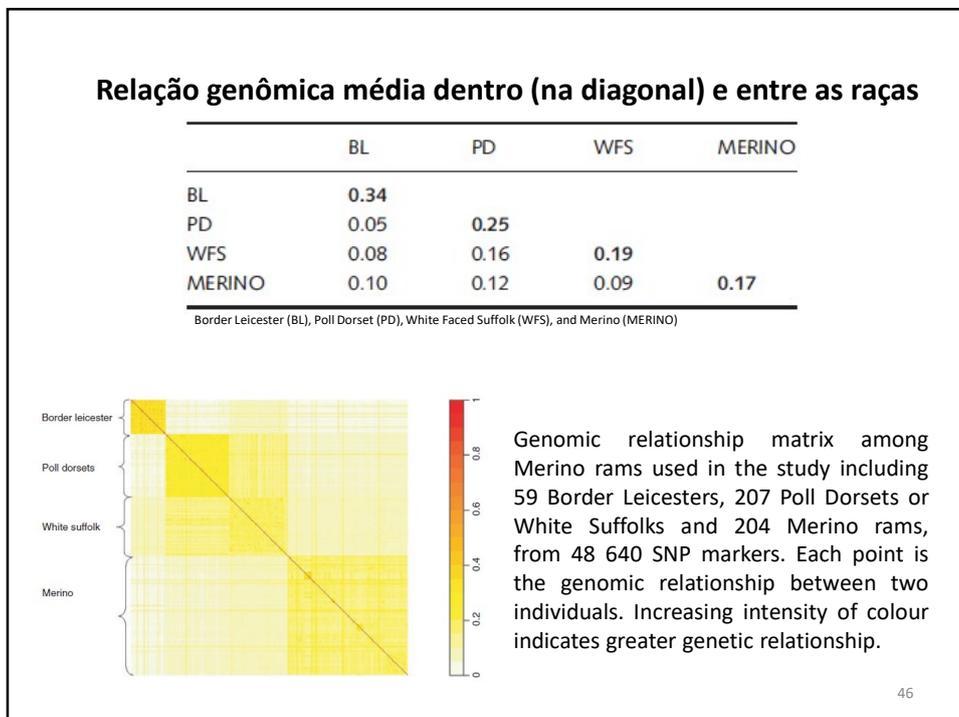
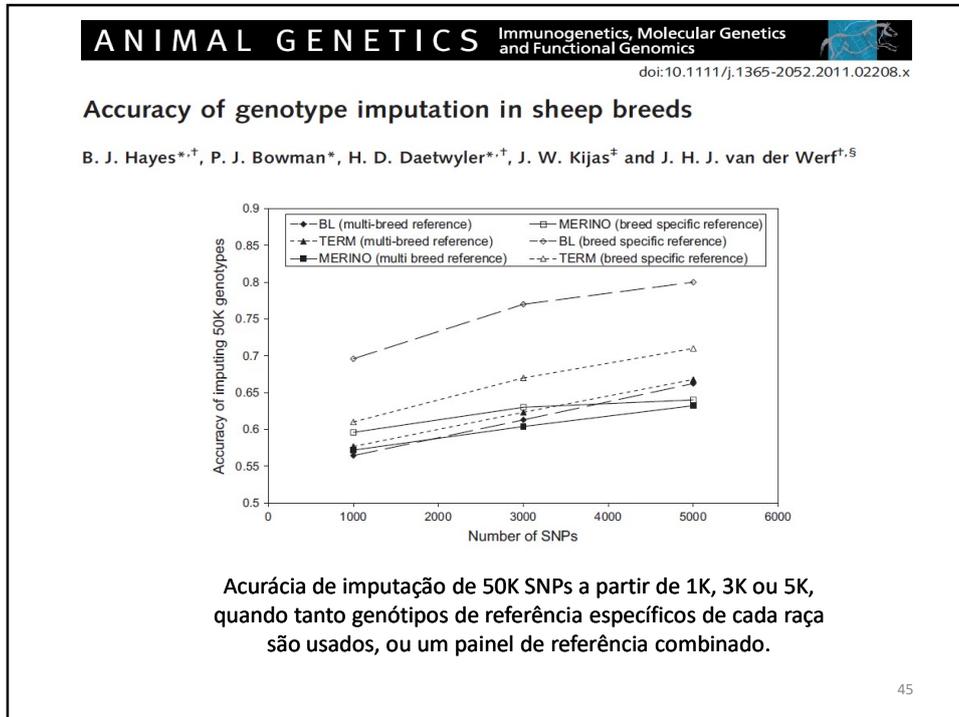
43

## Influência do número de indivíduos Genotipados no Painel de Referência



Relação entre a taxa média de erro de imputação e o número de indivíduos de referência (Druet et al., 2010)

44



## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019

### Acurácia de imputação de 50k para HD com Flmpute e BEAGLE usando uma única ou uma mistura de populações de referência.

		Reference	Imputed	Correct Call	Incorrect Call	Accuracy
Guernsey	Beagle	GU	100.000	95.367	4.633	0.954
		GU+AY+HO	100.000	96.671	3.329	0.967
	Flmpute	GU	99.919	97.179	2.740	0.973
		GU+AY+HO	100.000	97.423	2.577	0.974
Ayrshire	Beagle	AY	100.000	97.158	2.842	0.972
		GU+AY+HO	100.000	97.775	2.225	0.978
	Flmpute	AY	99.985	97.997	1.989	0.980
		GU+AY+HO	99.997	98.231	1.765	0.982
Holstein	Beagle	HO	100.000	99.296	0.704	0.993
		GU+AY+HO	100.000	99.286	0.714	0.993
	Flmpute	HO	100.000	99.234	0.764	0.992
		GU+AY+HO	100.000	99.225	0.774	0.992

Larmer et al.<sup>47</sup>

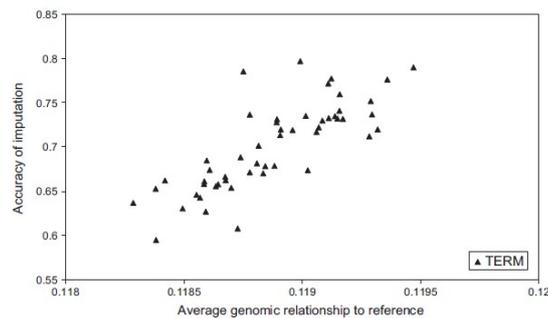
## ANIMAL GENETICS

Immunogenetics, Molecular Genetics  
and Functional Genomics

doi:10.1111/j.1365-2052.2011.02208.x

### Accuracy of genotype imputation in sheep breeds

B. J. Hayes<sup>\*,†</sup>, P. J. Bowman<sup>\*</sup>, H. D. Daetwyler<sup>\*,†</sup>, J. W. Kijas<sup>‡</sup> and J. H. J. van der Werf<sup>†,§</sup>



Acurácia de imputação (5-50K) para indivíduos da população imputada em função da relação média de parentesco com a população de referência

48

Carvalho et al. *Genetics Selection Evolution* 2014, **46**:69  
<http://www.gsejournal.org/content/46/1/69>



RESEARCH
Open Access

## Accuracy of genotype imputation in Nelore cattle

Roberto Carvalho<sup>1\*</sup>, Solomon A Boison<sup>2†</sup>, Haroldo H R Neves<sup>1,3†</sup>, Mehdi Sargolzaei<sup>4,5</sup>, Flavio S Schenkel<sup>4</sup>, Yuri T Utsunomiya<sup>1</sup>, Ana Maria Pérez O'Brien<sup>2</sup>, Johann Sölkner<sup>2</sup>, John C McEwan<sup>6</sup>, Curtis P Van Tassell<sup>7</sup>, Tad S Sonstegard<sup>7</sup> and José Fernando Garcia<sup>1,8</sup>

**Abstract**

**Background:** Genotype imputation from low-density (LD) to high-density single nucleotide polymorphism (SNP) chips is an important step before applying genomic selection, since denser chips tend to provide more reliable genomic predictions. Imputation methods rely partially on linkage disequilibrium between markers to infer unobserved genotypes. *Bos indicus* cattle (e.g. Nelore breed) are characterized, in general, by lower levels of linkage disequilibrium between genetic markers at short distances, compared to taurine breeds. Thus, it is important to evaluate the accuracy of imputation to better define which imputation method and chip are most appropriate for genomic applications in indicine breeds.

**Methods:** Accuracy of genotype imputation in Nelore cattle was evaluated using different LD chips, imputation software and sets of animals. Twelve commercial and customized LD chips with densities ranging from 7 K to 75 K were tested. Customized LD chips were virtually designed taking into account minor allele frequency, linkage disequilibrium and distance between markers. Software programs Fimpute and BEAGLE were applied to impute genotypes. From 995 bulls and 1247 cows that were genotyped with the Illumina<sup>®</sup> BovineHD chip (HD), 793 sires composed the reference set, and the remaining 202 younger sires and all the cows composed two separate validation sets for which genotypes were masked except for the SNPs of the LD chip that were to be tested.

**Results:** Imputation accuracy increased with the SNP density of the LD chip. However, the gain in accuracy with LD chips with more than 15 K SNPs was relatively small because accuracy was already high at this density. Commercial and customized LD chips with equivalent densities presented similar results. Fimpute outperformed BEAGLE for all LD chips and validation sets. Regardless of the imputation software used, accuracy tended to increase as the relatedness between imputed and reference animals increased, especially for the 7 K chip.

**Conclusions:** If the Illumina<sup>®</sup> BovineHD is considered as the target chip for genomic applications in the Nelore breed, cost-effectiveness can be improved by genotyping part of the animals with a chip containing around 15 K useful SNPs and imputing their high-density missing genotypes with Fimpute.

49

**Table 4 Average (standard deviation) imputation accuracy, for different imputation analyses using Fimpute**

Analysis <sup>1</sup>	SNP chip <sup>2</sup>	Nb (%) SNPs to be imputed	CORR <sup>3</sup>	PERC <sup>4</sup>
1	7 K	435509 (99.1)	0.9257 (0.0346)	90.56 (4.09)
2	50 K	418581 (95.2)	0.9783 (0.0136)	97.14 (1.76)

**Table 7 Summary statistics of imputation accuracy, using BEAGLE and Fimpute**

Anal. <sup>1</sup>	Validation set	SNP chip <sup>2</sup>	BEAGLE (Fimpute)			
			Minimum	Maximum	Mean	SD
24 (1)	Young sire	7 K	0.7525 (0.8003)	0.9717 (0.9845)	0.8982 (0.9257)	0.0392 (0.0346)
25 (3)	Young sire	GGP20Ki	0.8603 (0.8988)	0.9951 (0.9963)	0.9614 (0.9771)	0.0225 (0.0143)
26 (4)	Young sire	GGP75Ki	0.9142 (0.9568)	0.9986 (0.9990)	0.9842 (0.9922)	0.0120 (0.0056)
27 (8)	Young sire	15K_eml	0.8788 (0.9211)	0.9976 (0.9981)	0.9714 (0.9840)	0.0183 (0.0107)
28 (9)	Young sire	11a7 K	0.8773 (0.9163)	0.9979 (0.9975)	0.9697 (0.9823)	0.0190 (0.0117)
29 (12)	Young sire	48a7 K	0.9214 (0.9628)	0.9989 (0.9992)	0.9860 (0.9931)	0.0111 (0.0049)
30 (17)	Dam	7 K	0.6969 (0.7096)	0.9576 (0.9656)	0.8501 (0.8791)	0.0441 (0.0474)
31 (19)	Dam	GGP20Ki	0.8124 (0.8357)	0.9874 (0.9923)	0.9321 (0.9566)	0.0288 (0.0211)
32 (20)	Dam	GGP75Ki	0.8645 (0.9291)	0.9946 (0.9976)	0.9692 (0.9846)	0.0198 (0.0082)
33 (21)	Dam	15K_eml	0.8296 (0.8711)	0.9904 (0.9954)	0.9456 (0.9680)	0.0254 (0.0164)
34 (22)	Dam	11a7K	0.8249 (0.8640)	0.9893 (0.9951)	0.9430 (0.9658)	0.0260 (0.0173)
35 (23)	Dam	48a7K	0.8677 (0.9363)	0.9954 (0.9980)	0.9715 (0.9864)	0.0193 (0.0073)

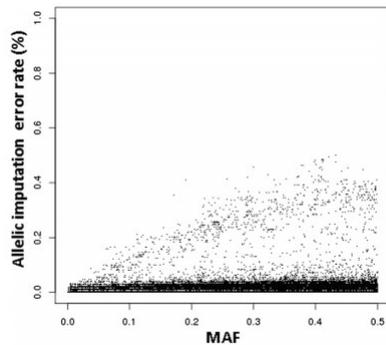
<sup>1</sup>Results of imputation analyses using BEAGLE or Fimpute (between brackets) and different validation sets (young sires and dams); the numbers of each analysis refer to those from Figure 1; <sup>2</sup>as described in the section "SNP chips" of "Methods"; SD = standard deviation.

<sup>3</sup>Imputation analyses using Fimpute software and 202 younger sires as the validation set; the numbers of each analysis refer to those in brackets from Figure 1; the first and the second numbers refer to analyses with and without pedigree information, respectively; <sup>4</sup>as described in the section "SNP chips" of "Methods".

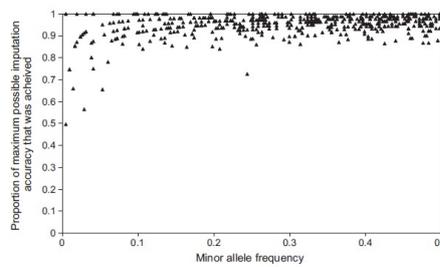
## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019

### Alelos raros e acurácia da imputação



Relationship between allelic imputation error rate and minor allele frequency in Montbéliarde breed



Proportion of maximum possible imputation accuracy that was achieved (50K to high density genotypes or full genome sequence) by minor allele frequency.

Hozé et al., Genetics Selection Evolution 2013, 45:33

Hayes et al., 2011

#### Imputation Accuracy from 6k to HD for 3 breeds using one and two step imputation procedures as well as single and multi-breed reference populations

	Reference	Imputed	Correct Call	Incorrect Call	Accuracy	
Guernsey	One-step	GU	99.989	91.891	8.096	0.919
		GU+AY+HO	100.000	88.920	11.080	0.889
	Two-step	GU	99.987	93.215	6.772	0.932
		GU+AY+HO	100.000	91.955	8.044	0.920
Ayrshire	One-step	AY	99.963	94.809	5.152	0.948
		GU+AY+HO	99.985	93.957	6.029	0.940
	Two-step	AY	99.967	94.711	5.256	0.947
		GU+AY+HO	99.979	94.417	5.562	0.944
Holstein	One-step	HO	100.000	97.110	2.888	0.971
		GU+AY+HO	100.000	96.923	3.075	0.969
	Two-step	HO	99.999	97.369	2.628	0.974
		GU+AY+HO	100.000	97.291	2.708	0.973

52

## Resultados de imputação (3k, 6k, 50k HD)

Software	50k -> HD	6k -> HD	3k -> HD
<i>Flmpute</i>	99.3	94.7	91.1
<i>findhap</i>	99.0	94.6	90.5

✓ Imputação mais acurada a partir de 50K para o HD do que de 3K e/ou 6K

✓ *Two step procedure*: Acurácia de imputação de 3K e 6K para HD melhorou cerca de 2% com Flmpute e 1% com findhap se primeiro imputa para 50K e depois para HD ao invés de imputar todos os genótipos em conjunto para HD.

53  
 Van Randen et al. Presented at Interbull annual meeting, Cork, Ireland, May 29, 2012

## Comparando os programas de imputação

Average accuracy of genotype imputation from 5K to 50K in target individuals using either fastPHASE or BEAGLE for genotype imputation.

	fastPHASE	BEAGLE
BL	0.80	0.81
TERM	0.70	0.80
MERINO	0.63	0.61

BL, Border Leicester.

Predicted accuracy of genotype imputation from high density genotypes or sequence from a 50K panel

	fastPHASE	BEAGLE
BL	0.96	0.93
TERM	0.94	0.93
MERINO	0.86	0.83

BL, Border Leicester.

- **Maior acurácia do BEAGLE para a imputação de genótipos esparsos para 50K**
- **fastPHASE mais acurado para a imputação de alta densidade (50k -> full sequence)**

54

## Software para imputação

- Fimpute
- AlphaImpute
  - <https://sites.google.com/site/hickeyjohn/alphaimpute>
- Findhap
  - <http://aipl.arsusda.gov/software/findhap/>
- MACH 1.0
  - <http://www.sph.umich.edu/csg/yli/mach>
- IMPUTE2
  - [http://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
- Beagle
  - <http://faculty.washington.edu/browning/beagle/beagle.html>
- AlphaPhase
  - <https://sites.google.com/site/hickeyjohn/alphaphase>
- fastPHASE
  - <http://depts.washington.edu/uwc4c/express-licenses/assets/fastphase/>
- PLINK, SNPMSTAT, UNPHASED and TUNA

55

## Considerações

- **Aumentando o tamanho da população de referência os resultados da imputação melhoram**
- **Diminuição do erro de imputação é decrescente com aumento na densidade de marcadores**
- **Avaliar o impacto da taxa de erro da imputação sobre as estimativas dos valores genômicos**
- **A densidade ótima vai depender do custo da genotipagem (baixa densidade) e a diminuição em acurácia dos valores genômicos**

56

## Várias opções de genotipagem no mercado

- **Existem diferenças em custo, serviço e assistência prestada.**
- **Qualquer genótipo é útil para realizar avaliação genômica?**
  - Em teoria, sim!
  - Na prática, depende!
- **Qual *chip* de genotipagem é recomendado no programa de melhoramento no qual participo?**
  - A acurácia é influenciada pela quantidade de marcadores em comum entre as plataformas de genotipagem
  - Necessidade de prever marcadores faltantes
- **Qual base de fenótipos foi utilizada na calibragem dos marcadores genéticos?**
  - Uma questão não menos importante!
  - Para um mesmo genótipo posso ter várias opções de DEP genômicas para a mesma característica.

57

## Acurácias Genômicas para diferentes densidades de marcadores **não imputados** ssGBLUP

Características	Genótipos sem imputar				
	N	ssGBLUP 777K	ssGBLUP Nelore Clarified v3.1	ssGBLUP GGP <i>indicus</i>	ssGBLUP <sup>1</sup> Nelore Clarified (imputado de GGP)
<b>PM120</b>	845	0,33	0,33	0,33	0,25
<b>PD120</b>		0,44	0,44	0,44	0,31
<b>P450</b>		0,49	0,49	0,49	0,37
<b>PAC</b>		0,29	0,29	0,28	0,21
<b>IPP</b>		0,30	0,31	0,30	0,22

<sup>1</sup>Animais genotipados no GGP *indicus* e imputado para Nelore Clarified 3.1

58

## Acurácias Genômicas para diferentes densidades de marcadores **imputados** ssGBLUP

Características	Genótipos imputados			
	N	ssGBLUP 777K	ssGBLUP Nelore Clarified v3.1	ssGBLUP GGP <i>indicus</i>
PM120	14.218	0,23	0,26	0,23
PD120		0,33	0,37	0,33
P450		0,38	0,42	0,39
PAC		0,21	0,23	0,21
IPP		0,22	0,25	0,22



59

## Perspectivas futuras sobre o uso de informações genômicas

- **Seleção genômica em raças compostas e cruzamentos**
  - Habilidade combinatória
- **Utilização da genômica para tomada de decisões**
  - Sincronizar o potencial genético com o manejo nutricional
- **Interação genótipo ambiente**
  - Equações de predição para cada ambiente ou normas de reação
- **Sequenciamento de próxima geração**
  - Novas variações genéticas
- **Melhoramento genético de precisão baseado em informações multiômicas e personalização da população**
  - Combinar informação da população de diferentes níveis ômicos

## Revolução multiômica e medicina de precisão

**Nova visão da medicina (P4):**

1. Previsível
2. Preventiva
3. Personalizada
4. Participativa

**P4 medicine: how systems medicine will transform the healthcare sector and society**

Ten years ago, the proposition that healthcare is evolving from reactive disease care to care that is predictive, preventive, personalized and participatory was regarded as highly speculative. Today, the core elements of that vision are widely accepted and have been articulated in a series of recent reports by the US Institute of Medicine. Systems approaches to biology and medicine are now beginning to provide patients, consumers and physicians with personalized information about each individual's unique health experience of both health and disease at the molecular, cellular and organ levels. This information will make disease care radically more cost effective by personalizing care to each person's unique biology and by treating the causes rather than the symptoms of disease. It will also provide the basis for concrete action by consumers to improve their health as they observe the impact of lifestyle decisions. Working together in digitally governed familial and affinity networks, consumers will be able to reduce the incidence of the complex chronic diseases that currently account for 75% of disease-care costs in the USA.

**KEYWORDS:** big data knowledge network learning healthcare new taxonomy of disease omics studies P4 medicine personal data clouds systems biology systems medicine wellness industry

Maurício Flores<sup>1,2</sup>, Gustavo Gusman<sup>1,2</sup>, Kristin Boonard<sup>1</sup>

Fonte: Giudice et al. (2017). Proteomics and phosphoproteomics in precision medicine: applications and challenges

## Medicina Personalizada e Aconselhamento Genético

**Sua Saúde**

**Tratamento contra câncer é personalizado com informações genéticas do paciente**

Fonte: Prof. Dr. Paulo Marcelo G. Hoff, diretor-geral do Centro de Oncologia do Hospital Sírio-Libanês. Publicado em 07/07/2016

A impossibilidade de existir no mundo duas pessoas exatamente iguais se deve ao fato de que cada um tem sua própria formação genética. No processo de desenvolvimento de um câncer, os tumores se desenvolvem a partir de alterações no material genético. Essas alterações acontecem pela ação de múltiplos estímulos e em diferentes combinações, de forma que um mesmo tipo de câncer pode ocorrer como resultado de certa variedade de alterações genéticas. A medicina vem evoluindo rapidamente para identificar essas alterações e reconhecer diferentes perfis de tumores, que até então eram tratados de forma indiferenciada.

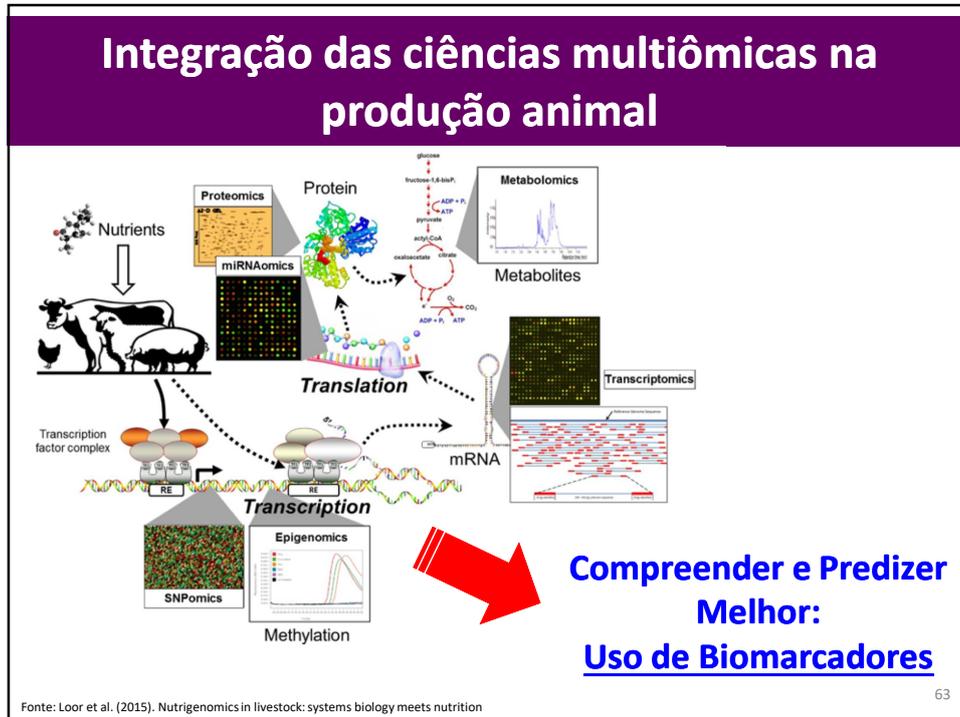
Essa abordagem do câncer baseada no perfil genético do tumor recebe o nome de **medicina de precisão**, ou terapia personalizada. Para explicar essa abordagem terapêutica, o diretor-geral do Centro de

**Especialidades**  
 Conheça nossos centros  
 Oncologia  
 Reprodução Humana

**Encontre um Médico**

**Mais sobre**  
 Câncer  
 Zika Vírus  
 Menopausa Precoce  
 Doenças Imunológicas  
 Criopreservação de Ovíulos  
 Congelamento de Ovíulos  
 Criopreservação de Oócitos  
 Ovuulação  
 Função Transvaginal  
 Sedação  
 Ultrassom  
 Dor Abdominal  
 Sangramento  
 Fertilização In Vitro  
 Gestação  
 Mulher

Fonte: <https://hospitalsiriolibanes.org.br/sua-saude/Paginas/tratamento-contracancer-personalizado-informacoes-geneticas-paciente.aspx>



**PLOS ONE**

RESEARCH ARTICLE

**Combining multi-OMICs information to identify key-regulator genes for pleiotropic effect on fertility and production traits in beef cattle**

Pablo Augusto de Souza Fonseca<sup>1,2</sup>, Samir Id-Lahoucine<sup>1</sup>, Antonio Reverter<sup>3</sup>, Juan F. Medrano<sup>4</sup>, Marina S. Fortes<sup>5</sup>, Joaquim Casellas<sup>6</sup>, Filippo Miglio<sup>1,7</sup>, Luiz Brito<sup>1</sup>, Maria Raquel S. Carvalho<sup>8</sup>, Flávio S. Schenkel<sup>1</sup>, Loan T. Nguyen<sup>5</sup>, Laercio R. Porto-Neto<sup>5</sup>, Milton G. Thomas<sup>8</sup>, Angela Cánovas<sup>1\*</sup>

**Abstract**

The identification of biological processes related to the regulation of complex traits is a difficult task. Commonly, complex traits are regulated through a multitude of genes contributing each to a small part of the total genetic variance. Additionally, some loci can simultaneously regulate several complex traits, a phenomenon defined as pleiotropy. The lack of understanding on the biological processes responsible for the regulation of these traits results in the decrease of selection efficiency and the selection of undesirable hitchhiking effects. The identification of pleiotropic key-regulator genes can assist in developing important tools for investigating biological processes underlying complex traits. A multi-breed and multi-OMICs approach was applied to study the pleiotropic effects of key-regulator genes using three independent beef cattle populations evaluated for fertility traits. A pleiotropic map for 32 traits related to growth, feed efficiency, carcass and meat quality, and reproduction was constructed. A set of SNP exhibited predictive correlations around 0.50 for several traits. The pleiotropic map identified a key-regulator gene, FCER1A, in the receptor signalling pathway ( $p = 0.0132$ ) and positive regulation of NFkB transcription factor activity ( $p = 0.00208$ ). We report FCER1A, a gene encoding a high-affinity receptor for the Fc region of immunoglobulin

**OPEN ACCESS**

**Citation:** Fonseca PAUS, Id-Lahoucine S, Reverter A, Medrano JF, Fortes MS, Casellas J, et al. (2018) Combining multi-OMICs information to identify key-regulator genes for pleiotropic effect on fertility and production traits in beef cattle. PLOS ONE 13(10): e0205295. <https://doi.org/10.1371/journal.pone.0205295>

**Editor:** Peter J. Hansen, University of Florida, UNITED STATES

**Received:** June 8, 2018

**Accepted:** September 21, 2018

**Published:** October 18, 2018

**Copyright:** © 2018 Fonseca et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution, and reproduction in any medium, provided the original author and source are credited.

\* [acanovas@uoguelph.ca](mailto:acanovas@uoguelph.ca)

**Introduction**

... effects through the use with the incorporation of a promising alternative to inclusion of gene set results es further research. prediction, gene set, kernel

**CONCLUSION**

ved genetic values or yet-for complex quantitative animal and plant breeding sonalized medicine. Given rolled by a large number environmental conditions, in cryptic ways, the acved or future values can he recent arrival of high-sequencing technologies f thousands of SNP sites revolutionized the genetic its. These whole-genome pic records allow the idenf causal mutations and the delts. Indeed, high-density

64

## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019

### Considerações Finais

- Existem **diferentes alternativas** para adoção da seleção genômica;
- Maior **resposta em acurácia** com uso da genômica em animais jovens e **características de difícil avaliação**;
- Informação genômica em populações com **estrutura de pedigree deficiente**;
- Importância das **variações estruturais e GWAS** para a identificação de QTLs associados com características adaptativas e produtivas;

### Considerações Finais

- Geração de conhecimentos dos **mecanismos biológicos** em nível genético;
- Implementação de abordagens **genômicas holísticas** “sistemas genéticos” para aprimorar o conhecimento sobre as **bases genéticas e fisiológicas** das características e melhorar a habilidade de **predição da informação genômica**.
- No futuro próximo, o **melhoramento genético de raças zebuínas de corte** deverá considerar de forma integrada conceitos de genética quantitativa, biologia de sistemas, genética funcional, bioinformática e engenharia genética.

**Acurácia dos valores genômicos diretos (DGV) para 5 características de leite com base em Bovine3K, BovineLD 7K, 50K, imputados até 50K, imputados até 800K e 800K-dosage**

Genotypes used	Mean allelic error rate (%) of imputation	Milk volume	Fat yield	Protein yield	Survival	Daughter fertility
50K	-	0.540	0.527	0.499	0.224	0.251
Subset Bovine3K	-	0.444	0.464	0.429	0.187	0.200
Subset Bovine LD 7K	-	0.481	0.516	0.443	0.186	0.232
50K-imputed (Test imputed <sup>A</sup> using Bovine3K)	3.86	0.533	0.523	0.496	0.200	0.244
50K-imputed (Test imputed <sup>A</sup> with BovineLD)	2.30	0.546	0.531	0.507	0.214	0.246
50K-imputed (Train & Test imputed <sup>B</sup> using Bovine3K)	5.52	0.505	0.515	0.481	0.207	0.245
50K-imputed (Train & Test imputed <sup>B</sup> using BovineLD)	3.06	0.530	0.524	0.492	0.209	0.248
800K-imputed <sup>C</sup>	-	0.558	0.530	0.526	0.232	0.256
800K-dosage <sup>C</sup>	-	0.554	0.525	0.520	0.229	0.253

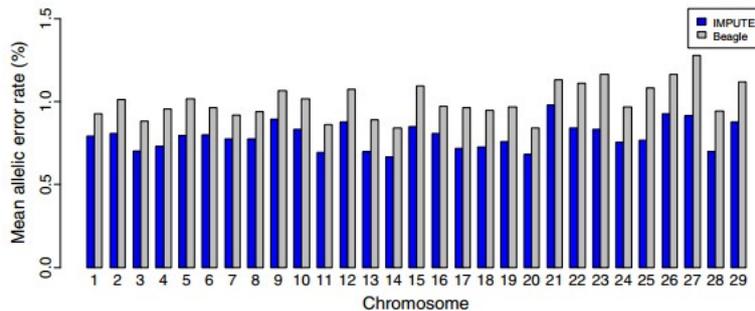
<sup>A</sup>Genotypes of 452 young bulls with subset of original SNPs were imputed (using IMPUTE2) up to 50K using 1753 bulls as reference set. Hence for DGV prediction entire test set (452 young bulls) had imputed genotypes and all the training bulls (1753) had actual 50K genotypes.

<sup>B</sup>Genotypes of 2055 bulls with subset of original SNPs were imputed (using IMPUTE2) up to 50K using 136 bulls as reference set. Hence for DGV prediction the entire test set (452 young bulls) and 1617 bulls out of the training set of 1753 bulls had imputed genotypes.

<sup>C</sup>Data on 2205 bulls genotyped for 50K were imputed using IMPUTE2 up to 800K using 845 cows genotyped on 800K as reference.

Khatkar et al. BMC Genomics 2012, 13:538

67



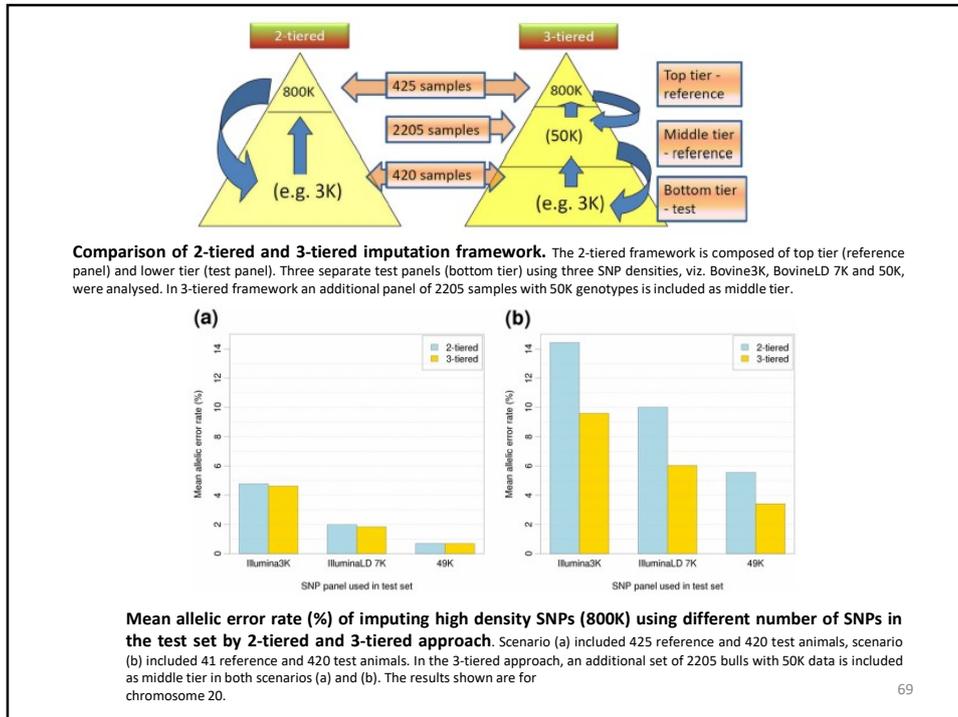
**Taxa de erro de alelos (%) de imputação de alta densidade de SNPs (800K) utilizando 49K SNPs no conjunto de teste comparando os dois métodos de imputação em todos os autossomas**

Khatkar et al. BMC Genomics 2012, 13:538

68

## Workshop de Melhoramento Genético Animal

Projeto ALT-Biotech<sup>RepGen</sup> - Recursos Genéticos Animais e Biotecnologias: projeção para o futuro  
Estação Zootécnica Nacional – Fonte Boa, 17 de Dezembro de 2019



Outra estratégia para diminuir os custos da genotipagem.....

Fenotipagem e genotipagem seletiva dos animais mais informativos

## Estratégias de genotipagem para seleção genômica no gado leiteiro (Jiménez-Montero et al., 2010)

### Estratégias de genotipagem seletiva:

2%, 5% e 10% indivíduos da população de referência foram selecionados como conjunto de treinamento com estratégias diferentes para características de 0,25 e 0,10 de herdabilidade

1. Aleatória (**RND**). As fêmeas escolhidas aleatoriamente da população de referência
2. Valores divergentes fenotípicos (**DPH**). Igual número de fêmeas no  $\alpha$  e  $(1-\alpha)$  percentis da distribuição "ajustada" fenotípica.
3. Valores divergentes EBV (**DBV**). As fêmeas com seus valores genéticos no  $\alpha$  e  $(1-\alpha)$  percentis.
4. Os maiores valores fenotípicos (**TopPH**). Vacas top no ranking de valores fenotípicos "ajustado"
5. Os maiores valores EBV (**TopBV**). Vacas top no ranking dos valores genéticos
6. Divergente família EBV (**DFM**). Fêmeas meio irmãos filhas dos touros melhores e piores em EBV.

### Avaliação genômica do modelo

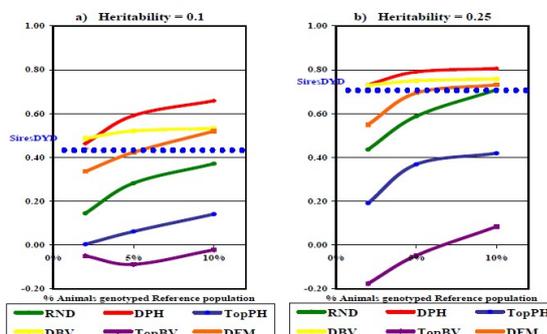
Bayesiano Lasso para estimar os coeficientes SNP na população analisada (6 estratégias)

Correlações de Pearson entre os valores genômicos estimados (GBV) e os valores genéticos verdadeiros (TBV), foram calculados na geração de 15

71

### Acurácia dos valores genéticos genômicos (corr (GBV, TBV))

**A acurácia preditiva do GBV depende da quantidade de animais genotipados e da estratégia de genotipagem seletiva usada.**



Acurácia dos GBV na geração 15, quando 2%, 5% e 10% das fêmeas na população de referência (G 11 - 14) foram genotipadas utilizando diferentes estratégias.

72