Documentos

ISSN 1677-9274
Dezembro, 2016
143

Imputação de Genótipos em bovinos: um guia passo a passo

Embrapa

ISSN 1677-9274
Dezembro, 2016

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 143

Imputação de Genótipos em bovinos: um guia passo a passo

Maurício de Alvarenga Mudadu
Henry Gomes de Carvalho
Marcos Jun-Iti Yokoo
Fernando Flores Cardoso

Embrapa Informática Agropecuária
Campinas, SP
2016

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700
www.embrapa.br/informatica-agropecuaria
SAC: www.embrapa.br/fale-conosco/sac/

1ª edição
publicação digitalizada 2016

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Imputação de Genótipos em bovinos: um guia passo a passo / Maurício de Alvarenga Mudadu… [et al.].- Campinas : Embrapa Informática Agropecuária, 2016.

31 p. : il. ; cm. - (Documentos / Embrapa Informática Agropecuária, ISSN 1677-9274; 143).

1. SNP chip. 2. Marcadores genéticos.3. Fimpute. 4. Beagle. 5. Minimac. I. Mudadu, Mauricio de Alvarenga. II. Embrapa Informática Agropecuária. III. Título. IV. Série.

CDD 005.15

Autores

Maurício de Alvarenga Mudadu

Biólogo, doutor em Bioinformática
Pesquisador da Embrapa Informática Agropecuária, Campinas, SP

Henry Gomes de Carvalho

Informata, mestre em Ciência da Computação
Analista da Embrapa Pecuária Sul, Bagé, RS

Marcos Jun-Iti Yokoo

Zooetecnista, doutor em Genética e Melhoramento Animal - ênfase em Genética Quantitativa
Pesquisador da Embrapa Pecuária Sul, Bagé, RS

Fernando Flores Cardoso

Médico veterinário, doutor em Melhoramento Genético Animal
Pesquisador da Embrapa Pecuária Sul, Bagé, RS

Apresentação

O programa de melhoramento genético no Brasil está implementando o uso de marcadores genéticos, um procedimento chamado seleção genômica (SG). SG consiste em genotipar uma população de referência com fenótipos conhecidos e na descoberta de marcadores genéticos associados. O efeito dos marcadores são estimados e validados de forma a tornar possível predizer os valores genéticos dos cadidatos à seleção baseado nos seus genótipos.

A genotipagem de alta densidade tem custo elevado, de forma que usa-se genotipar a população de referência usando chips de alta densidade de marcadores e genotipar com menor custo os candidatos a seleção usando chips de baixa densidade de marcadores. A imputação de genótipos é então aplicada para expandir os dados de genotipagem dos candidatos, melhorando a intensidade de seleção e reduzindo custos.

Nesse trabalho, três softwares de imputação, Beagle v4.1, Minimac v3 e Fimpute v2.2 foram usados para imputar genótipos de chips de baixa densidade para alta densidade, usando dados de genotipagem de 233 touros da raça Hereford e Bradford provindos da região sul do Brasil. Dados de alta densidade de genotipagem (777 mil marcadores) foram disponibilizados para todas as amostras, de forma que os dados de um chip de baixa densidade (50 mil marcadores) puderam ser obtidos e as acurácias dos softwares puderam ser medidas.

Os resultados mostraram que os softwares permitiram imputar mais de 94% de todos os marcadores passíveis de imputação. A taxa de marcadores imputados corretamente variou de 86% a 94%. A performance dos softwares variou entre 26,9 a 378,1 marcadores imputados por segundo, usando uma amostra dos dados de genotipagem do cromossomo 1. De uma forma geral, todos os três softwares apresentaram boa performance e se mostraram como boas opções para a imputação de genótipos para se usar em SG.

Silvia Maria Fonseca Silveira Massruhá
Chefe-geral
Embrapa Informática Agropecuária

Sumário

Imputação de genótipos em bovinos: um guia passo a passo

Maurício de Alvarenga Mudadu
Henry Gomes de Carvalho
Marcos Jun-Iti Yokoo
Fernando Flores Cardoso

Introdução

A imputação de genótipos é a ação de predizer ou inferir genótipos que não estão presentes em um dado painel de marcadores genéticos. Genótipos podem ser imputados quando se possui uma amostra genotipada com um dado painel inicial de marcadores genéticos (por exemplo 700 mil marcadores do tipo SNP ou “Single Nucleotide Polymorphism”) e outra amostra genotipada com outro painel que contém um subconjunto dos marcadores contidos no painel inicial (por exemplo 50 mil marcadores SNP). Uma das vantagens da imputação é aumentar a densidade de marcadores do painel de menor densidade (MARCHINI; HOWIE, 2010),maximizando, por exemplo, os recursos financeiros disponíveis para seleção genômica. Dessa forma, a imputação permite a inferência dos genótipos não determinados com painéis de menor densidade (30 ou 50 mil marcadores) e a recuperação quase total da informação genotípica para o painel de alta densidade (700 mil marcadores ou 700k) com um custo de genotipagem consideravelmente menor (GODDARD; HAYES, 2009).

Uma das premissas para se realizar a imputação de genótipos é a inferência de fase de ligação ou “haplotype phasing”. Haplótipo é uma sequência

de marcadores ou alelos que estão em um mesmo cromossomo e têm a mesma origem parental. Os genótipos obtidos dos painéis de marcadores são pares de alelos desordenados sem a indicação da origem paterna ou materna desses alelos; por exemplo, em um genótipo AB não podemos afirmar se A veio do pai e B da mãe e vice-versa. Para que isso seja possível, é necessário inferir a fase haplotípica ou reconstruir os haplótipos (BROWNING, 2008). A inferência de fase de ligação invariavelmente usa os dados de desequilíbrio de ligação entre marcadores vizinhos e é realizada por meio de softwares específicos que usam metodologias diversas. Em se tratando de inferências de haplótipos entre indivíduos não relacionados existem diversos tipos de algoritmos como “Clark’s”, “EM”, método coalescente e cadeias de Markov. Esse último é usado em diversos softwares de imputação bastante conhecidos como “PHASE” (STEPHENS et al., 2001), “fastPHASE” (SCHEET; STEPHENS, 2006), MACH (LI et al., 2010) “minimac” (HOWIE et al., 2012), “IMPUTE2” (HOWIE et al., 2009) e “BEAGLE” (BROWNING; BROWNING, 2007);(BROWNING; BROWNING, 2016). Para inferir haplótipos de indivíduos relacionados, os softwares podem classificar haplótipos como obtidos por descendência Identity By Descent(IBD) assim como utilizar a frequência desses haplótipos na população BROWNING; BROWNING, 2011). Programas como “BEAGLE”, “SHAPEIT” (DELANEAU et al., 2011) e “FIMPUTE” (SARGOLZAEI et al., 2014) podem utilizar essa informação.

Atualmente, no Brasil, os programas de melhoramento genético estão buscando adotar a seleção assistida por marcadores em escala genômica, a qual é denominada seleção genômica (MEUWISSEN et al., 2001). Esse procedimento já vem sendo utilizado nos Estados Unidos para as avaliações do gado Holandês desde 2008 (http://www.aipl.arsusda.gov/eval.htm), dado os ganhos em acurácia e redução do intervalo de gerações que os valores genéticos genômicos proporcionam, entre outras vantagens (HAYES et al., 2009; RADEN et al., 2009). A seleção genômica consiste na genotipagem e conhecimento de marcadores de interesse em uma dada população de referência cujo fenótipo é conhecido. Posteriormente estima-se os efeitos dos marcadores ou haplótipos, para depois validar os efeitos estimados em um grupo de animais que não pertence à população referência e, finalmente predizer os valores genéticos de indivíduos candidatos à seleção, baseados nos genótipos dos marcadores e dos efeitos estimados (HAYES et al., 2009). O usual é realizar uma genotipagem de alta

densidade na população de referência usando painéis comerciais de 770 mil marcadores SNP, por exemplo com o chip "Illumina® BovineHD Chip" e uma genotipagem com painéis de menor densidade e custo nos candidatos à seleção, por exemplo com os chips de 50 mil marcadores SNP "Illumina® BovineSNP50 v2", ou 30 mil marcadores SNP "Low Density GeneSeek Genomic Profiler GGP-LD". A imputação de genótipos pode ser então realizada para expandir os dados de marcadores dos candidatos, reduzir os custos de genotipagem e aumentar a intensidade de seleção. No entanto, é necessário verificar a acurácia da imputação no intuito de se inferir a eficiência da seleção genômica a ser realizada (CALUS et al., 2014). Diversos trabalhos com esse objetivo já foram publicados em populações de bovinos no mundo todo e inclusive no Brasil (PICCOLI et al., 2014; CHUD et al., 2015; CARVALHEIRO et al., 2014).

Esse documento tem o intuito de explicar passo a passo a metodologia para se realizar a imputação de genótipos de um chip de menor densidade (50 mil marcadores) para alta densidade (770 mil marcadores) em uma população de bovinos. Para verificar a acurácia da imputação, serão utilizados dados de genotipagem em alta densidade porém com dados mascarados simulando um chip de 50 mil marcadores, permitindo-se dessa forma comparar os genótipos imputados com os genótipos reais.

Imputação de genótipos passo a passo

Os softwares e sistemas utilizados

Sistema Operacional e Sistema Computacional

O sistema operacional utilizado foi o "Red Hat Enterprise Linux Server release 6.8" de 64 bits. A máquina utilizada possuía 4 processadores "AMD Opteron(tm) Processor 6378" com 16 núcleos cada (possibilidade de 64 processos simultâneos), 512 Gb de memória RAM e espaço em disco suficiente.

Lista de softwares de terceiros

i. “Beagle 4.1” (Imputação).

http://faculty.washington.edu/browning/beagle/beagle.html

ii. “Fimpute v2.2” (Imputação).

http://www.aps.uoguelph.ca/~msargol/fimpute/

iii. “minimac v3” (Imputação).

http://genome.sph.umich.edu/wiki/Minimac3

iv. “Shapeit” (haplotype phasing).

https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html

v. “Plink v1.9” (manipulação e controle de qualidade de dados de genotipagem).

https://www.cog-genomics.org/plink2

vi. “GenGen” (manipulação de dados de genotipagem, script “convert_bim_allele.pl”).

http://gengen.openbioinformatics.org/en/latest/

vii. “Perl v5.12.4” (linguagem de programação).

http://www.perl.org/

viii. “R version 3.3.0, 2016-05-03” (linguagem de programação)

https://cloud.r-project.org/

“Bioconductor”, pacote “SnpStats”.

http://bioconductor.org/packages/release/bioc/html/snpStats.html

Scripts desenvolvidos “in house”

Os scripts desenvolvidos para análise de dados estão listados a seguir:

I. Pacote “fQC.R” na linguagem R, para realizar o controle de qualidade dos marcadores, desenvolvido pelo pesquisador Roberto H. Higa da Embrapa Informática Agropecuária (https://www.svnserver.cnptia.embrapa.br/svn/rga/fQC/)

II. Scripts na linguagem Perl com finalidades diversas e disponibilizados no sistema SVN da Embrapa Informática Agropecuária (https://www.svnserver.cnptia.embrapa.br/svn/rga/Imputacao):

i. “finalreport2ped.pl” – destinado a transformar os dados de genotipagem do formato Final Report para .ped.

ii. “mask2impute_higher2lower_ped.pl” – destinado a mascarar SNP do chip de maior densidade no formato .ped .

iii. “ped2fimpute.pl” – destinado a transformar os arquivos de dados .ped para o formato do Fimpute v 2.2.

iv. “fimpute2ped.pl” - destinado a transformar os arquivos de dados do Fimpute v 2.2 para formato .ped.

v. “fix_vcf_minimac.pl” – destinado a adequar o formato do campo SNP ID no arquivo de resultados do minimac v3.0.

vi. “fix_ref_alt_vcf.pl” – destinado a adequar o formato dos campo de alelos referência e alternativos do formato .vcf para imputação com o Beagle.

vii. “check_imputation_2.0_vcf.pl” – destinado a avaliar a imputação usando arquivos de resultados no formato .vcf (saidas dos softwares Beagle 4.1 e minimac v3).

viii. “check_imputation_2.0_ped.pl” – destinado a avaliar a imputação usando arquivos de resultados no formato .ped.

ix. “get_groups.pl” – destinado a separar as amostras em três grupos aleatórios para comparar a eficiência de imputação.

O conjunto de dados utilizado

Um conjunto de 233 touros das raças Hereford e Braford foram genotipados com o chip "Illumina® Bovine HD Chip" que possui 777692 marcadores SNP (http://support.illumina.com/array/array_kits/bovinehd_dna_analysis_kit.html).

Foram utilizados os dados do chip "Illumina® BovineSNP50 v3" que possui 53218 marcadores SNP (http://support.illumina.com/array/array_kits/bovinesnp50-beadchip-kit/downloads.html) para mascarar dados de genotipagem do chip HD e simular a genotipagem do chip "Illumina® BovineSNP50 v3". Isso só foi possível pelo fato de a maioria dos marcadores do chip "Illumina® BovineSNP50 v3" também pertencerem ao chip "Illumina® Bovine HD Chip".

Os dados de genotipagem no formato tipo texto em arquivos do tipo "Final Report" gerados pelo software "Genome Studio" da empresa Illumina foram transformados para o formato ".ped" (*pedigree*, ou *linkage format*, para mais detalhes do formato ver: https://www.broadinstitute.org/science/programs/medical-and-population-genetics/haploview/input-file-formats-0).

Para gerar arquivos no formato ".ped", foi criado o script "finalreport2ped.pl", disponível em : https://www.svnserver.cnptia.embrapa.br/svn/rga/Imputacao

O arquivo "FinalReport.txt" é um arquivo texto gerado pela ferramenta "Report Wizard" existente no software "GenomeStudio" da empresa Illumina (http://support.illumina.com/array/array_software/genomestudio.html). O separador de colunas frequentemente utilizado no "FinalReport.txt" é o caractere de tabulação. O arquivo geralmente possui um cabeçalho de 10 linhas e as colunas poderão estar em qualquer ordem. Usualmente o identificador de SNP é colocado na coluna 1, o identificador de amostra na coluna 2, os dados de genotipagem nas colunas 6 e 7 ("Allele1 – AB" e "Allele2 - AB") e "GC score" na coluna 8. Dependendo da disposição das colunas do arquivo "FinalReport.txt", poderá ser necessário alterar o script "finalreport2ped.pl" para que ele acesse as colunas na ordem correta. O script também necessita de um arquivo do tipo "SNP Map" ("SNP_Map.txt") tabular que contenha na segunda coluna a lista de identificadores SNP na mesma ordem em que aparecem no chip. O script "finalreport2ped.pl" tam-

bém poderá ser alterado para acessar corretamente as colunas do arquivo do tipo “SNP Map”.

Para executar o script foram necessários três passos (Figura 1), ou linhas de comando no “shell” do linux: obter a lista de amostras e gerar o arquivo “samples.txt”(Figura 1.i), obter a lista dos identificadores de famílias e amostras e gerar o arquivo “family.samples.txt”(Figura 1.ii), onde foi usada apenas a família de identificador “1” para todas as amostras e finalmente executar o script (Figura 1.iii). O resultado é um arquivo do tipo “.ped” chamado “FinalReport.ped”.

Controle de qualidade (QC)

Para executar o controle de qualidade (QC) nos dados de genotipagem, foram usados o script “fQC.R” e o software “Plink v1.9”.

O script “fQC.R” é um pacote bastante completo para execução do controle de qualidade em dados de genotipagem. Ele foi utilizado para realizar filtros menos comuns de controle de qualidade como remoção de amostras duplicadas, remoção de marcadores designados para uma mesma posição em um dado cromossomo e remoção de marcadores com correlação maior que 98% (dados não mostrados). O software “Plink v1.9” foi utilizado para realizar os filtros de QC mais comumente usados para esse tipo de dados, como call rate por amostra (<90%), call rate por marcadores (<98%), MAF (< 3%) e Hardy Weinberg (<1e-7). Respectivamente, as flags de filtro de qualidade para o “Plink v1.9” foram: --mind 0.1 --geno 0.02 --maf 0.03 --hwe 1e-7.

Após o QC, 229 amostras foram mantidas, 603304 marcadores SNP foram mantidos no chip de alta densidade (Bovine HD chip) e destes, 30741 SNPs pertencem ao chip de menor densidade (BovineSNP50 v3). O arquivo “HD.all.QC.ped” foi gerado com os dados de genotipagem para o chip de maior densidade.

```
i)
$ awk 'NR>10{print$2}' FinalReport.txt | sort -u > samples.txt

ii)
$ awk 'NR>10{print"1\t"$2}' FinalReport.txt | sort -u > family.samples.txt

iii)
$ perl finalreport2ped.pl FinalReport.txt family.samples.txt
SNP_Map.txt samples.txt HD.all.ped
```

Figura 1. Comandos usados em um "shell" do Linux para gerar o arquivo "HD.all.ped", no formato pedigree partindo de um arquivo no formato "Final Report.

Painéis de referência

Os softwares "Beagle v4.1" e "minimac v3" necessitam de painéis de referência de haplótipos como entrada para realizar a imputação. Já o software "Fimpute" infere os painéis de referência de haplótipos e não necessita dessa entrada. No intuito de se comparar a eficácia de imputação, a população resultante de 229 animais, foi separada em três grupos, 1 (76 animais), 2 (77 animais) e 3 (76 animais). O objetivo é comparar a imputação dos grupos 1, 2 e 3 entre si, usando sempre painéis de referência de haplótipos dos grupos restantes. Por exemplo, comparar a eficiência de imputação do grupo 1 usando painéis de referência de 2 e 3, com a eficiência de imputação do grupo 2, usando painéis de referência de 1 e 3, com a eficiência de imputação do grupo 3, usando painéis de referência de 1 e 2 (Tabela 1).

Tabela 1. Grupos de animais referentes aos candidatos à imputação e painéis de referência.

Grupos	
Candidatos a Imputação	Painéis de Referência
1	2 e 3
2	1 e 3
3	1 e 2

O software “Shapeit” foi usado em pares de grupos para inferir a fase de ligação e gerar os painéis de referência de haplótipos. No caso do software “minimac v3”, mesmo os candidatos a imputação devem ter a fase de ligação inferida de forma que o “Shapeit” também foi executado nesses grupos.

Para se formar os grupos 1, 2 e 3 e os grupos de referência, foi utilizado o script “get_groups.pl” e o software “Plink v1.9” (Figura 2). Os arquivos “HD.all.QC.ped” e “HD.all.QC.map” gerados após o QC foram usados. Inicialmente foi gerado um arquivo com todos os identificadores de amostras (Figura 2.i) e depois três arquivos contendo grupos de amostras aleatórias foram gerados (Figura 2.ii), os arquivos foram adaptados para entrada no “Plink v1.9” (Figura 2.iii) que separou o arquivo de genotipagem inicial em três de acordo com os grupos gerados (Figura 2.iv). Os grupos de referência (12, 13 e 23) foram então gerados por concatenação na linha de comando (Figura 2.v). Esses grupos serão usados na construção dos haplótipos de referência.

O próximo passo para se gerar os painéis de referência pelo software “Shapeit” foi transformar os arquivos “.ped” para o formato “VCF” (Figura 3). Para isso foi o usado o script “convert_bim_allele.pl” do pacote de programas “GenGen”, sendo necessário gerar arquivos “.bed” e “.bim”, pelo software “Plink v1.9”, que foram usados como entrada para o script “convert_bim_allele.pl” (Figura 3.i). Além disso esse script necessita de um arquivo de informação de codificação de alelos (“SNP_table.txt”), para a conversão do formato “AB” da Illumina para a codificação de nucleotídeos, ou alelos “ACTG” (ver mais informações em http://gengen.openbioinformatics.org/en/latest/tutorial/coding/). Essa informação está presente no arquivo “SNP Map” gerado pelo software “Genome Studio” em conjunto com o arquivo “Final Report” de genotipagem.

O arquivo “SNP_table.txt” contém quatro colunas do arquivo “SNP Map”: “Name” (nome do marcador SNP, ou “SNP ID”), “SNP” (o polimorfismo propriamente dito, por ex.: “[A/T]”), “ILMN Strand” e “Customer Strand” (correspondência quanto à fita de DNA usada, por ex:. “TOP” e “BOT”). O script “convert_bim_allele.pl” gerou arquivos do tipo .bim modificados para a codificação de alelos “ACTG” (Figura 3.ii). Esse arquivo foi usado para substituir o arquivo.bim (Figura 3.iii) gerado pelo software “Plink v1.9” anteriormente. Foi preciso converter os dados para o formato “VCF” e depois

separar os dados de genotipagem por cromossomo (29 autossomos bovinos) usando o “Plink v1.9” (Figura 3.iv). O software “Shapeit” foi usado pra inferir as fases de ligação e gerar os haplótipos de referência por cromossomo (Figura 3.v). Por fim, os haplótipos de referência foram convertidos para o formato “VCF” (Figura 3.vi).

Imputação

Beagle v4.1

```
i)
awk '{print$2}' HD.all.QC.ped > samples.txt

ii)
perl get_groups.pl samples.txt 3

iii)
awk '{print"1\t",$1}' samples.txt.1 > samples.txt.1.keep
awk '{print"1\t",$1}' samples.txt.2 > samples.txt.2.keep
awk '{print"1\t",$1}' samples.txt.3 > samples.txt.3.keep

iv)
plink --file HD.all.QC --keep samples.txt.1.keep --chr-set 29 --out HD.all.QC.1 --recode --make-bed
--missing-genotype 0
plink --file HD.all.QC --keep samples.txt.2.keep --chr-set 29 --out HD.all.QC.2 --recode --make-bed
--missing-genotype 0
plink --file HD.all.QC --keep samples.txt.3.keep --chr-set 29 --out HD.all.QC.3 --recode --make-bed
--missing-genotype 0

v)
cat HD.all.QC.1.ped HD.all.QC.2.ped > HD.all.QC.12.ped
cat HD.all.QC.1.ped HD.all.QC.3.ped > HD.all.QC.13.ped
cat HD.all.QC.2.ped HD.all.QC.3.ped > HD.all.QC.23.ped
```

Figura 2. Comandos para gerar os grupos de referência e de comparação, para serem executados em um “shell” do Linux.

O software Beagle v4.1 permite imputar dados de genótipos perdidos (“missing”), assim como dados de marcadores não existentes. Neste trabalho optamos pelo primeiro caso e mascaramos (isto é, tornamos

```
i)
plink --ped HD.all.QC.12.ped --map HD.all.QC.map --recode --make-bed --chr-set 29 --out HD.all.QC.12
plink --ped HD.all.QC.13.ped --map HD.all.QC.map --recode --make-bed --chr-set 29 --out HD.all.QC.13
plink --ped HD.all.QC.23.ped --map HD.all.QC.map --recode --make-bed --chr-set 29 --out HD.all.QC.23

ii)
perl convert_bim_allele.pl HD.all.QC.12.bim SNP_table.txt -outfile HD.all.QC.12.ACTG.bim -intype
illmn12 -outtype dbsnp
perl convert_bim_allele.pl HD.all.QC.13.bim SNP_table.txt -outfile HD.all.QC.13.ACTG.bim -intype
illmn12 -outtype dbsnp
perl convert_bim_allele.pl HD.all.QC.23.bim SNP_table.txt -outfile HD.all.QC.23.ACTG.bim -intype
illmn12 -outtype dbsnp

iii)
mv HD.all.QC.12.ACTG.bim HD.all.QC.12.bim
mv HD.all.QC.13.ACTG.bim HD.all.QC.13.bim
mv HD.all.QC.23.ACTG.bim HD.all.QC.23.bim

iv)
for i in {1..29}; do plink --chr-set 29 --chr $i --vcf HD.all.QC.12.vcf --recode vcf --out HD.all.QC.12.$i;
done
for i in {1..29}; do plink --chr-set 29 --chr $i --vcf HD.all.QC.13.vcf --recode vcf --out HD.all.QC.13.$i;
done
for i in {1..29}; do plink --chr-set 29 --chr $i --vcf HD.all.QC.23.vcf --recode vcf --out HD.all.QC.23.$i;
done

v)
for i in {1..29}; do shapeit --input-vcf HD.all.QC.12.$i.vcf -O HD.all.QC.12.$i.phased ; done
for i in {1..29}; do shapeit --input-vcf HD.all.QC.13.$i.vcf -O HD.all.QC.13.$i.phased ; done
for i in {1..29}; do shapeit --input-vcf HD.all.QC.23.$i.vcf -O HD.all.QC.23.$i.phased ; done

vi)
for i in {1..29}; do shapeit -convert --input-haps HD.all.QC.12.$i.phased --output-vcf
HD.all.QC.12.$i.phased.vcf; done
for i in {1..29}; do shapeit -convert --input-haps HD.all.QC.13.$i.phased --output-vcf
HD.all.QC.13.$i.phased.vcf; done
for i in {1..29}; do shapeit -convert --input-haps HD.all.QC.23.$i.phased --output-vcf
HD.all.QC.23.$i.phased.vcf; done
```

Figura 3. Comandos para se gerar os painéis de haplótipos de referência, para serem executados em um “shell” do Linux.

"missing" apagando a informação original) os marcadores do chip de alta densidade ("Bovine HD chip") de forma a deixar apenas os genótipos do chip de menor densidade ("BovineSNP50 v3") . Para isso foi usado o script "mask2impute_higher2lower_ped.pl" (Figura 4.i). Depois os grupos de candidatos a imputação (1, 2 e 3) foram separados usando o "Plink v1.9" (Figura 4.ii). O próximo passo foi converter os dados do formato "pedigree" para "VCF". Para isso foi usado o script "convert_bim_allele.pl" (Figura 4.iii e iv) e depois novamente o "Plink v1.9" (Figura 4.v).

Para realizar a imputação com o "Beagle" foi necessário separar os dados de genotipagem por cromossomo (Figura 4.vi). A conversão de ".ped" para ".vcf" usando o "Plink v1.9" remove os alelos referência e alternativos (colunas "REF" e "ALT") do arquivo de dados de genotipagem resultante. Porém, o Beagle necessita desses dados. Para recuperá-los, foi usado o script "fix_ref_alt_vcf.pl" (Figura 4.vii). Para finalizar, o "Beagle v4.1" foi executado (Figura 4.viii) seguindo o esquema da Tabela 1.

Minimac v3

Para a imputação com o software "minimac v3" foi usada a estratégia de imputação de genótipos não existentes (diferentemente do "Beagle" onde os genótipos foram mascarados (tornados "missing"). Dessa forma foram usados arquivos de entrada de menor densidade de genótipos além dos painéis de referência de haplótipos. Para gerar os arquivos de menor densidade o "Plink v1.9" foi usado com uma lista de identificadores de SNP do chip "BovineSNP50 v3" (arquivo "snps_LD_keep.txt", Figura 5.i). Os arquivos foram convertidos para formato "VCF" e separados por cromossomo usando o script "convert_bim_allelle.pl" e o "Plink v1.9" (Figura 5.ii,iii, iv e v). O "minimac v3" somente aceita como entradas dados com fase de ligação construída, de forma que foi necessário executar o "Shapeit" também para os dados em LD dos candidatos a imputação (Figura 5.vi e vii). O "minimac v3" foi executado usando os painéis de referência de haplótipos obtidos anteriormente seguindo o esquema da Tabela 1 (Figura 5.viii).

Fimpute v2.2

O software "Fimpute" não necessita de painéis de haplótipos de referência

```
i)
perl mask2impute_higher2lower_ped.pl HD.all.QC.ped BovineSNP50_v3_A1.map HD.all.QC.map

ii)
plink –file HD.all.QC.ped.masked --keep samples.txt.1.keep --chr-set 29 --out HD.all.QC.masked.1
--recode –make-bed --missing-genotype 0
plink –file HD.all.QC.ped.masked --keep samples.txt.2.keep --chr-set 29 --out HD.all.QC.masked.2
--recode –make-bed --missing-genotype 0
plink –file HD.all.QC.ped.masked --keep samples.txt.3.keep --chr-set 29 --out HD.all.QC.masked.3
--recode –make-bed --missing-genotype 0

iii)
for i in {1..3}; do perl convert_bim_allele.pl HD.all.QC.masked.$i.bim SNP_table.txt -outfile
HD.all.QC.masked.$i.ACTG.bim -intype ilmn12 -outtype dbsnp; done

iv)
for i in {1..3};do mv HD.all.QC.masked.$i.ACTG.bim HD.all.QC.masked.$i.bim; done

v)
for i in {1..3}; do plink –chr-set 29 --bfile HD.all.QC.masked.$i --out HD.all.QC.masked.$i --recode vcf;
done

vi)
for i in {1..29}; do plink --chr-set 29 –chr $i --vcf HD.all.QC.masked.3.vcf --recode vcf –out
HD.all.QC.masked.3.$i; done
for i in {1..29}; do plink --chr-set 29 –chr $i --vcf HD.all.QC.masked.2.vcf --recode vcf –out
HD.all.QC.masked.2.$i; done
for i in {1..29}; do plink --chr-set 29 –chr $i --vcf HD.all.QC.masked.1.vcf --recode vcf –out
HD.all.QC.masked.1.$i; done

vii)
for i in {1..29}; do perl fix_ref_alt_vcf.pl HD.all.QC.12.$i.phased.vcf HD.all.QC.masked.3.$i.vcf; done
for i in {1..29}; do perl fix_ref_alt_vcf.pl HD.all.QC.13.$i.phased.vcf HD.all.QC.masked.2.$i.vcf; done
for i in {1..29}; do perl scripts/fix_ref_alt_vcf.pl HD.all.QC.23.$i.phased.vcf HD.all.QC.masked.1.$i.vcf;
done

viii)
for i in {1..29}; do java -jar beagle.03May16.862.jar gt=HD.all.QC.masked.3.$i.vcf.ok
ref=HD.all.QC.12.$i.phased.vcf out=HD.all.QC.3.$i.ok12.impute; done
for i in {1..29}; do java -jar beagle.03May16.862.jar gt=HD.all.QC.masked.2.$i.vcf.ok
ref=HD.all.QC.13.$i.phased.vcf out=HD.all.QC.2.$i.ok13.impute; done
for i in {1..29}; do java -jar beagle.03May16.862.jar gt=HD.all.QC.masked.1.$i.vcf.ok
ref=HD.all.QC.23.$i.phased.vcf out=HD.all.QC.1.$i.ok23.impute; done
```

Figura 4. Passos para imputação usando o software “Beagle v4.1”. Comandos para serem executados em um “shell” do Linux.

como entrada e seu formato de dados é bem específico. Só é necessário especificar no arquivo de configuração do software os mapas de marcadores de alta e baixa densidade assim como entrar com os genótipos em alta e baixa densidade. Também não há necessidade de "mascarar" genótipos. Inicialmente foi necessário extrair os dados do chip de baixa densidade com o "Plink v1.9", utilizando o arquivo "snps_LD_keep.txt" (Figura 5.i) e os dados de genotipagem de alta densidade dos grupos 1, 2 e 3 obtidos anteriormente (Figura 6.i). Foi utilizado o script "ped2fimpute.pl" para converter os dados no formato ".ped" para o formato do "Fimpute" (Figura 6.ii). Foi necessário criar arquivos de configuração para cada rodada de imputação de acordo com o esquema da Tabela 1. Por exemplo, para imputar o grupo 1 em LD usando os grupos 2 e 3 em HD, foi criado o arquivo "HD_to_LD_1_23.conf" que contem os caminhos para os arquivos de genótipos e mapa de SNP gerados no passo anterior. O arquivo foi criado usando um editor de textos (conteúdo do arquivo mostrado no Figura 6.iii). O "Fimpute" foi executado carregando os arquivos de configurações respectivos para os grupos candidatos a imputação (Figura 6.iv). Por fim foi necessário transformar os arquivos de resultados do "Fimpute" para o formato ".ped" para serem analisados em um passo posterior usando o script "fimpute-2ped.pl" (Figura 6.v).

Análises e Resultados

Para realizar a análise dos dados e apurar a eficácia de imputação de cada software foram utilizados dois scripts, o "check_imputation_2.0_vcf.pl" e o "check_imputation_2.0_ped.pl", para resultados no formato "VCF" ("Beagle v4.1" e "minimac v3", Figura 7.i e ii respectivamente) e pedigree ("Fimpute", Figura 7.iii). Esses scripts contam o número de marcadores potencialmente imputáveis, o número de marcadores imputados corretamente e incorretamente e o número total de marcadores utilizados.

Os scripts geraram para cada cromossomo um arquivo texto tabular de resultado com extensão ".accuracy.result.txt". Esses arquivos foram concatenados (exemplo para o grupo 1 no Figura 7.iv) em um único arquivo com um cabeçalho e depois foram analisados usando a linguagem R (exemplo de código em R para análise dos resultados do grupo 1 na Figura 7.v), para cálculos estatísticos. Os resultados finais para os softwares "minimac

```
i)
awk '{print$2}' BovineSNP50_v3_A1.map > snps_keep_LD.txt
plink --file HD.all.QC --keep samples.txt.1.keep --chr-set 29 --out LD.all.QC.1 --extract snps_LD_keep.txt --recode
--make-bed --missing-genotype 0
plink --file HD.all.QC --keep samples.txt.2.keep --chr-set 29 --out LD.all.QC.2 --extract snps_LD_keep.txt --recode
--make-bed --missing-genotype 0
plink --file HD.all.QC --keep samples.txt.3.keep --chr-set 29 --out LD.all.QC.3 --extract snps_LD_keep.txt --recode
--make-bed --missing-genotype 0

ii)
perl convert_bim_allele.pl LD.all.QC.1.bim SNP_table.txt -outfile LD.all.QC.1.ACTG.bim -intype ilmn12 -outtype dbsnp
perl convert_bim_allele.pl LD.all.QC.2.bim SNP_table.txt -outfile LD.all.QC.2.ACTG.bim -intype ilmn12 -outtype dbsnp
perl convert_bim_allele.pl LD.all.QC.3.bim SNP_table.txt -outfile LD.all.QC.3.ACTG.bim -intype ilmn12 -outtype dbsnp

iii)
mv LD.all.QC.1.ACTG.bim LD.all.QC.1.bim
mv LD.all.QC.2.ACTG.bim LD.all.QC.2.bim
mv LD.all.QC.3.ACTG.bim LD.all.QC.3.bim

iv)
plink --chr-set 29 --bfile LD.all.QC.1 --out LD.all.QC.1 --recode vcf
plink --chr-set 29 --bfile LD.all.QC.2 --out LD.all.QC.2 --recode vcf
plink --chr-set 29 --bfile LD.all.QC.3 --out LD.all.QC.3 --recode vcf

v)
for i in {1..29}; do plink --cow --chr $i --vcf LD.all.QC.1.vcf --recode vcf --out LD.all.QC.1.$i; done
for i in {1..29}; do plink --cow --chr $i --vcf LD.all.QC.2.vcf --recode vcf --out LD.all.QC.2.$i; done
for i in {1..29}; do plink --cow --chr $i --vcf LD.all.QC.3.vcf --recode vcf --out LD.all.QC.3.$i; done

vi)
for i in {1..29}; do shapeit --input-vcf LD.all.QC.1.$i.vcf -O LD.all.QC.1.$i.phased; done
for i in {1..29}; do shapeit --input-vcf LD.all.QC.2.$i.vcf -O LD.all.QC.2.$i.phased; done
for i in {1..29}; do shapeit --input-vcf LD.all.QC.3.$i.vcf -O LD.all.QC.3.$i.phased; done

vii)
for i in {1..29}; do shapeit -convert --input-haps LD.all.QC.1.$i.phased --output-vcf LD.all.QC.1.$i.phased.vcf; done
for i in {1..29}; do shapeit -convert --input-haps LD.all.QC.2.$i.phased --output-vcf LD.all.QC.2.$i.phased.vcf; done
for i in {1..29}; do shapeit -convert --input-haps LD.all.QC.3.$i.phased --output-vcf LD.all.QC.3.$i.phased.vcf; done

viii)
for i in {1..29}; do Minimac3 --refHaps HD.all.QC.23.$i.phased.vcf –haps LD.all.QC.1.$i.phased.vcf --prefix
LD_HD.all.QC.1_23.$i.minimac.imputed --myChromosome $i; done
for i in {1..29}; do Minimac3 --refHaps HD.all.QC.23.$i.phased.vcf --haps LD.all.QC.1.$i.phased.vcf --prefix
LD_HD.all.QC.1_23.$i.minimac.imputed --myChromosome $i; done
for i in {1..29}; do Minimac3 --refHaps HD.all.QC.23.$i.phased.vcf --haps LD.all.QC.1.$i.phased.vcf --prefix
LD_HD.all.QC.1_23.$i.minimac.imputed --myChromosome $i; done
```

Figura 5. Comandos em “shell” do Linux para realizar a imputação usando o software “minimac v3”.

v3", "Fimpute v2.2" e "Beagle v4.1" estão dispostos na Tabela 2, Tabela 3 e Tabela 4 respectivamente. Pode-se verificar que todos os três softwares apresentaram um bom desempenho e imputaram acima de 94% dos marcadores com acerto entre 86% e 94%.

Para se ter uma ideia da velocidade de imputação dos softwares Beagle v4.1, minimac v3 e FImpute v2.2, foram feitos testes de imputação de 36297 marcadores, usando as amostras do grupo 1 (76 indivíduos) com os dados de genotipagem do cromossomo 1 (2039 marcadores em LD imputáveis para 38336 marcadores em HD). O FImpute v2.2 foi o mais rápido entre os três, pois em uma média de três repetições de imputações, ele gastou aproximadamente um minuto e trinta e seis segundos para terminar a imputação dos 36297 marcadores (média de 378,1 marcadores imputados por segundo). O Beagle v4.1 e o minimac v3, em três repetições, gastaram em média 22 minutos e 27 segundos e 3 minutos e 33 segundos, respectivamente,5 para executar a imputação dos mesmos marcadores, média de 26,9 e 170,4 marcadores por segundo, nesta ordem.

```
i)
plink --file HD.all.QC.1 --extract snps_keep_LD.txt --chr-set 29 --make-bed --recode --out HD.all.QC.1.LD
plink --file HD.all.QC.2 --extract snps_keep_LD.txt --chr-set 29 --make-bed --recode --out HD.all.QC.2.LD
plink --file HD.all.QC.3 --extract snps_keep_LD.txt --chr-set 29 --make-bed --recode --out HD.all.QC.3.LD

ii)
perl ped2fimpute.pl HD.all.QC.1.LD.ped HD.all.QC.23.ped HD.all.QC.1.LD.map HD.all.QC.23.map
perl ped2fimpute.pl HD.all.QC.2.LD.ped HD.all.QC.13.ped HD.all.QC.2.LD.map HD.all.QC.13.map
perl ped2fimpute.pl HD.all.QC.3.LD.ped HD.all.QC.12.ped HD.all.QC.3.LD.map HD.all.QC.12.map

iii)
title="LDtoHD group1_vs_23 imputation";
genotype_file="genotypes.toFImpute.txt";
snp_info_file="snps_info.toFImpute.txt";
output_folder="output_fimpute";
njob=20;

iv)
FImpute HD_to_LD_1_23.conf
FImpute HD_to_LD_2_13.conf
FImpute HD_to_LD_3_12.conf

v)
perl fimpute2ped.pl output_fimpute/genotypes_imp.txt HD.all.QC.23.ped HD.all.QC.1.LD.ped
perl fimpute2ped.pl output_fimpute/genotypes_imp.txt HD.all.QC.13.ped HD.all.QC.2.LD.ped
perl fimpute2ped.pl output_fimpute/genotypes_imp.txt HD.all.QC.12.ped HD.all.QC.3.LD.ped
```

Figura 6. Passos para imputação usando o software "FImpute v2.2". Comandos para serem executados em um "shell" do Linux.

```
i)
for i in {1..29}; do perl check_imputation_2.0_vcf.pl HD.all.QC.1.$i.vcf.ok HD.all.QC.1.$i.ok.23.impute.vcf
HD.all.QC.12.$i.phased.vcf; done
for i in {1..29}; do perl check_imputation_2.0_vcf.pl HD.all.QC.2.$i.vcf.ok HD.all.QC.2.$i.ok.13.impute.vcf
HD.all.QC.23.$i.phased.vcf; done
for i in {1..29}; do perl check_imputation_2.0_vcf.pl HD.all.QC.3.$i.vcf.ok HD.all.QC.3.$i.ok.12.impute.vcf
HD.all.QC.13.$i.phased.vcf; done

ii)
for i in {1..29}; do perl check_imputation_2.0_vcf.pl LD.all.QC.1.$i.phased.vcf
LD_HD.all.QC.1_23.$i.minimac.imputed.dose.vcf.fixmap HD.all.QC.12.$i.phased.vcf; done
for i in {1..29}; do perl check_imputation_2.0_vcf.pl LD.all.QC.2.$i.phased.vcf
LD_HD.all.QC.2_13.$i.minimac.imputed.dose.vcf.fixmap HD.all.QC.12.$i.phased.vcf; done
for i in {1..29}; do perl check_imputation_2.0_vcf.pl LD.all.QC.3.$i.phased.vcf
LD_HD.all.QC.3_12.$i.beagle.imputed.vcf HD.all.QC.13.$i.phased.vcf; done

iii)
perl check_imputation_2.0_ped.pl HD.all.QC.1.LD.ped HD.all.QC.1.LD.map
output_fimpute/genotypes_imp.txt.ped HD.all.QC.1.ped HD.all.QC.1.map
perl check_imputation_2.0_ped.pl HD.all.QC.2.LD.ped HD.all.QC.2.LD.map
output_fimpute/genotypes_imp.txt.ped HD.all.QC.2.ped HD.all.QC.2.map
perl check_imputation_2.0_ped.pl HD.all.QC.3.LD.ped HD.all.QC.3.LD.map
output_fimpute/genotypes_imp.txt.ped HD.all.QC.3.ped HD.all.QC.3.map

iv)
echo "id_sample      imputable_markers      non_imputable_markers   correct_markers incorrect_markers
    total_markers" > results1_23.txt
cat *.accuracy.result.txt | sort -k1,1 | grep "#" -v  >> results1_23.txt

v)
library(sqldf)
result <-  read.table(file="results1_23.txt",head= TRUE)
result$percentcorrect <- result$correct_markers/result$imputable_markers
result$percentincorrect <- result$incorrect_markers/result$imputable_markers
result$imputed <- result$imputable_markers/result$total_markers
sd(result$percentcorrect)
sd(result$percentincorrect)
sd(result$imputed)
summary(result)
result_grouped <- sqldf("select id_sample, sum(imputable_markers) as imputable_markers,
sum(non_imputable_markers) as non_imputable_markers, sum(correct_markers) as correct_markers,
sum(incorrect_markers) as incorrect_markers, sum(total_markers) as total_markers from result group by
id_sample")
summary(result_grouped)
write.table(file="result_1_23.txt",summary(result_grouped),quote=F,row.name=F)
```

Figura 7. Passos para analisar os resultados de imputação. Comandos para serem executados em um "shell" do Linux (i a iv) e em um ambiente com linguagem "R" (v).

Tabela 2. Resultado da imputação usando o software "minimac v3".

	Grupo 1	Grupo 2	Grupo 3
Total de indivíduos processados	76	77	76
Total de marcadores HD depois do QC	603304	603304	603304
Total de marcadores LD depois do QC	30741	30741	30741
Media de marcadores imputados (% do total de marcadores HD ± SD)	572563 (**94,93%** ±0,42%)	572563 (**94,93%** ±0,42%)	572563 (**94,93%** ±0,42%)
Media de marcadores que foram imputados corretamente (% do total de imputados ± SD)	524284 (**91,32%** ±5,64%)	521818 (**90,87%** ±5,79%)	523637 (**91,19%** ±6,14%)
Media de marcadores que foram imputados errôneamente (% do total de imputados ± SD)	48279 (8,68% ±5,64%)	50746 (9,13% ±5,79%)	48926 (8,80% ±6,14%)
Menor proporção de imputações corretas	61,97%	65,17%	62,22%
Maior proporção de imputações corretas	98,98%	98,61%	98,88

Tabela 3. Resultado de Imputação com o software "Fimpute v2.2".

	Grupo 1	**Grupo 2**	**Grupo 3**
Total de indivíduos processados	76	77	76
Total de marcadores HD depois do QC	603304	603304	603304 desempenho
Total de marcadores LD depois do QC	30741	30741	30741
Media de marcadores imputados (% do total de marcadores HD ± SD)	572476 (**94,89%** ±0,01%)	572426 (**94,88%** ±0,01%)	572434 (**94,88%** ±0,01%)
Media de marcadores que foram imputados corretamente (% do total de imputados ± SD)	535236 (**93,50%** ±4,21%)	533030 (**93,12%** ±4,69%)	534735 (**93,41%** ±4,79%)
Media de marcadores que foram imputados errôneamente (% do total de imputados ± SD)	37220 (6,50% ±4,21%)	39396 (6,88% ±4,69%)	37699 (6,59% ±4,79%)
Menor proporção de imputações corretas	81,60%	76,87%	79,23%
Maior proporção de imputações corretas	99,59%	98,57%	98,87%

Tabela 4. Resultado de Imputação com o software "Beagle v4.1".

	Grupo 1	Grupo 2	Grupo 3
Total de indivíduos processados	76	77	76
Total de marcadores HD depois do QC	603304	603304	603304
Total de marcadores LD depois do QC	30741	30741	30741
Media de marcadores imputados (% do total de marcadores HD ± SD)	572618 (**94,94%**) (±0,00%)	572628 (**94,95%**) (0,00%)	572629 (**94,95%**) (±0,00)
Media de marcadores que foram imputados corretamente (% do total de imputados ± SD)	496909 (**86,61%** (±4,39%)	494799 (**86,09%** (±4,56%)	496295 (**86,32%** (±4,61%)
Media de marcadores que foram imputados errôneamente (% do total de imputados ± SD)	75710 (13,39% (±4,39%)	77829 (13,91% (±4,56%)	76759 (13,68% (±4,61%)
Menor proporção de imputações corretas	61,01%	63,14%	62,11%
Maior proporção de imputações corretas	97,02%	96,85%	97,21%

Conclusões

A imputação de genótipos é um passo importante para a viabilidade econômica da implementação de seleção genômica no melhoramento animal. No entanto, realizar a imputação pelos softwares analisados nesse trabalho ainda é um processo complexo que demanda algum conhecimento de bioinformática. Neste documento foi descrito com detalhes um passo a passo para se realizar a imputação de genótipos usando chips bovinos comerciais. Os comandos utilizados para realizar a imputação com os dados utilizados foram descritos, a maioria para serem utilizados em um "shell" do Linux. Foram desenvolvidos diversos "scripts" para facilitar o processo

de forma a permitir que pessoas interessadas possam realizar a imputação por conta própria. Os softwares possuem entradas de dados em formatos diferentes e realizam a imputação usando algoritmos diferentes. Por exemplo, o software "FImpute" realiza a construção dos haplóptipos de referência enquanto os softwares "Beagle v4.1" e "minimac v3" necessitam que esses dados sejam inseridos pelo usuário (foi necessário utilizar outro software, o "Shapeit", para realizar essa tarefa). O "Beagle v4.1" e "minimac v3" usam dados no formato ".vcf" que é um formato muito utilizado em dados de marcadores obtidos por sequenciamento de nova geração. Já o "FImpute" possui um formato próprio de entrada de dados. Como visto, ainda é necessária uma significante interação do usuário para adequar dados e formatos de entrada para executar os softwares. Mas mesmo apresentando peculiaridades e diferenças nas velocidades de execução e acurácias, os três softwares apresentaram um bom desempenho de imputação e se mostraram como boas opções para realizar a imputação de genótipos com os dados utilizados.

Referências

Browning, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. **Human Genetics**, v. 124, n. 5, p. 439-450, 2008. DOI: 10.1007/s00439-008-0568-7.

Browning, B. L.; Browning, S. R. Genotype Imputation with Millions of Reference Samples. **American Journal of Human Genetics**, v. 98, n. 1, p. 116-126, 2016. DOI: 10.1016/j.ajhg.2015.11.020.

Browning, S.; Browning, B. Haplotype phasing: existing methods and new developments. **Nature Reviews Genetics**, v. 12, n. 10, p. 703-714, 2011. DOI: http://doi.org/10.1038/nrg3054.

Browning, S. R.; Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. **American Journal of Human Genetics**, v. 81, n. 5, p. 1084-1097, 2007. DOI: 10.1086/521987.

Calus, M. P. L.; Bouwman, A. C.; Hickey, J. M.; Veerkamp, R. F.; Mulder, H. A. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. **Animal**, v. 8, n. 11, p. 1743-1753, 2014. DOI: 10.1017/S1751731114001803.

Carvalheiro, R.; Boison, S. A.; Neves, H. H. R.; Sargolzaei, M.; Schenkel, F. S.; Utsunomiya, Y. T.; o'brien, a. m. p.; sölkner, j.; mCewan, j. c.; TASSELL, J. P. van; SONSTEGARD, T. S.; Garcia, J. F. Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution**, v. 46, n. 1, p.1-11, 2014. DOI: 10.1186/s12711-014-0069-1.

Chud, T. C. S.; Ventura, R. V.; Schenkel, F. S.; Carvalheiro, R.; Buzanskas, M. E.; Rosa, J. O.; MUDADU, M. de A.; SILVA, M. V. G. B. Da; MOKRY, F. B.; MARCONDES, C. R.; REGITANO, L. C. A.;MUNARI, D. P. Strategies for genotype imputation in composite beef cattle. **BMC Genetics**, v. 16, n. 1, p. 1-10, 2015. DOI: 10.1186/s12863-015-0251-7.

Delaneau, O.; Marchini, J.; Zagury, J.-F. A linear complexity phasing method for thousands of genomes. **Nature Methods**, v. 9, n. 2, p. 179-181, 2011. DOI: 10.1038/nmeth.1785.

Goddard, M. E.; Hayes, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews. Genetics,** v. 10, n. 6, p. 381-391, 2009. DOI: 10.1038/nrg2575 .

Hayes, B. J.; Bowman, P. J.; Chamberlain, A. J.; Goddard, M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**, v. 92, n. 2, p. 433-443, 2009. DOI: 10.3168/jds.2008-1646.

Howie, B.; Fuchsberger, C.; Stephens, M.; Marchini, J.; Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. **Nature Genetics**, v. 44, n. 8, p. 955-959, 2012. DOI: 10.1038/ng.2354.

Howie, B. N.; Donnelly, P.; Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. **PLoS Genetics**, v. 5, n. 6, p. e1000529, 2009. DOI: 10.1371/journal.pgen.1000529.

Li, Y.; Willer, C. J.; Ding, J.; Scheet, P.; Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. **Genetic Epidemiology**, v. 34, n. 8, p. 816-834, 2010. DOI: 10.1002/gepi.20533.

Meuwissen, T. H. E.; Hayes, B. J.; Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, 2001.

Piccoli, M. L.; Braccini, J.; Cardoso, F. F.; Sargolzaei, M.; Larmer, S. G.; Schenkel, F. S. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. **BMC Genetics**, v. 15, n. 1, p. 1-15, 2014. DOI: 10.1186/s12863-014-0157-9.

Raden, P. M. van; Tassell, C. P. van; Wiggans, G. R.; Sonstegard, T. S.; Schnabel, R. D.; Taylor, J. F.; Schenkel, F. S. Invited review: reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, v. 92, n. 1,p. 16-24, 2009. DOI: 10.3168/jds.2008-1514.

Sargolzaei, M.; Chesnais, J. P.; Schenkel, F. S. A new approach for efficient genotype

imputation using information from relatives. **BMC Genomics**, v. 15, n. 1, p. 1-12, 2014. DOI: 10.1186/1471-2164-15-478.

Scheet, P.; Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. **American Journal of Human Genetics**, v. 78, n. 4, p. 629-644, 2006. DOI: 10.1086/502802.

Stephens, M.; Smith, N. J.; Donnelly, P. A new statistical method for haplotype reconstruction from population data. **American Journal of Human Genetics**, v. 68, n. 4, p. 978-989, 2001. DOI: 10.1086/319501

Embrapa

Informática Agropecuária

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

BRASIL
GOVERNO FEDERAL

CGPE 13448