UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

METODOLOGIAS E ESTRATÉGIAS DE IMPUTAÇÃO DE MARCADORES GENÉTICOS EM BOVINOS DA RAÇA CANCHIM

Tatiane Cristina Seleguim Chud
Zootecnista

2014

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

METODOLOGIAS E ESTRATÉGIAS DE IMPUTAÇÃO DE MARCADORES GENÉTICOS EM BOVINOS DA RAÇA CANCHIM

Tatiane Cristina Seleguim Chud
Orientador: Prof. Dr. Danísio P. Munari
Coorientadores: Dr. Ricardo Vieira Ventura
Dr. Roberto Carvalheiro

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Mestre em Genética e Melhoramento Animal

2014

Chud, Tatiane Cristina Seleguim
C559m Metodologias e estratégias de imputação de marcadores genéticos em bovinos da raça Canchim / Tatiane Cristina Seleguim Chud. – – Jaboticabal, 2014
v, 47 p. : il. ; 28 cm

Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2014
Orientadora: Danísio Prado Munari
Co-orientador: Ricardo Vieira Ventura, Roberto Carvalheiro
Banca examinadora: Arione Augusti Boligon, Fernando Sebastián Baldi Rey
Bibliografia

1.SNP. 2. Genômica. 3. Bovinos de Corte. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 575:636.2

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação – Serviço Técnico de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal.

unesp **UNIVERSIDADE ESTADUAL PAULISTA**

CAMPUS DE JABOTICABAL

FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS DE JABOTICABAL

CERTIFICADO DE APROVAÇÃO

TÍTULO: METODOLOGIAS E ESTRATÉGIAS DE IMPUTAÇÃO DE MARCADORES GENÉTICOS EM BOVINOS DA RAÇA CANCHIM

AUTORA: TATIANE CRISTINA SELEGUIM CHUD

ORIENTADOR: Prof. Dr. DANÍSIO PRADO MUNARI
CO-ORIENTADOR: Prof. Dr. RICARDO VENTURA
CO-ORIENTADOR: Prof. Dr. ROBERTO CARVALHEIRO

Aprovada como parte das exigências para obtenção do Título de MESTRE EM GENÉTICA E MELHORAMENTO ANIMAL , pela Comissão Examinadora:

Prof. Dr. DANÍSIO PRADO MUNARI
Departamento de Ciências Exatas / Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal

Prof. Dr. FERNANDO SEBASTIÁN BALDI REY
Departamento de Zootecnia / Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal

Profa. Dra. ARIONE AUGUSTI BOLIGON
Universidade Federal de Pelotas / Pelotas/RS

Data da realização: 19 de fevereiro de 2014.

DADOS CURRICULARES DO AUTOR

Tatiane Cristina Seleguim Chud – nascida em Porto Ferreira - SP, no dia 25 de maio de 1988, ingressou no curso de Zootecnia em março de 2007 na Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de Jaboticabal, Jaboticabal-SP, foi bolsista de iniciação científica da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e obteve o título de bacharel em Zootecnia em março de 2012. Iniciou o curso de Mestrado pela mesma instituição de ensino, em março de 2012, sob orientação do Prof. Dr. Danísio Prado Munari e coorientação do Dr. Ricardo Vieira Ventura e do Dr. Roberto Carvalheiro. Foi bolsista de mestrado da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Realizou estágio de pesquisa de Mestrado, sob supervisão do Prof. Dr. Flavio Schramm Schenkel, no Centre for Genetic Improvement of Livestock, University of Guelph - Canadá, no período de abril a setembro de 2013.

Ando devagar
Porque já tive pressa
E levo esse sorriso
Porque já chorei demais

Hoje me sinto mais forte
Mais feliz, quem sabe
Só levo a certeza
De que muito pouco sei
Ou nada sei

Conhecer as manhas
E as manhãs
O sabor das massas
E das maçãs

É preciso amor
Pra poder pulsar
É preciso paz pra poder sorrir
É preciso a chuva para florir

Penso que cumprir a vida
Seja simplesmente
Compreender a marcha
E ir tocando em frente

Como um velho boiadeiro
Levando a boiada
Eu vou tocando os dias
Pela longa estrada, eu vou
Estrada eu sou

Todo mundo ama um dia
Todo mundo chora
Um dia a gente chega
E no outro vai embora

Cada um de nós compõe a sua história
Cada ser em si
Carrega o dom de ser capaz
E ser feliz

Almir Sater

AGRADECIMENTOS

À **Deus** pelas maravilhas que me tem proporcionado, pela força e proteção que recebo cada dia.

Aos meus pais Cidinha e Mauro pelo amor e pela confiança. Por terem me fornecido condições para me tornar o que sou hoje e por acreditarem nos meus sonhos.

À minha irmã-mãe **Patrícia**, à minha sobrinha **Victória** e meu cunhado-pai **Rodrigo** pelo carinho, respeito e amor. Obrigada por sempre estarem ao meu lado. Meus agradecimentos infinitos pelo apoio, dedicação e por me ajudarem nos momentos de dificuldade.

Ao meu amor e amigo **Hugo**. Pelo amor, pela paciência e compreensão nos momentos difíceis. Obrigada por tornar minha vida mais tranquila e trazer a paz quando preciso. Por sempre estar ao meu lado, me apoiando nas decisões. Por ser essa pessoa tão especial e fazer meus dias mais felizes. AMO VOCÊ.

Ao meu orientador **Prof. Dr. Danísio Prado Munari** pela amizade, dedicação e motivação . Pela confiança em mim depositada na execução desse trabalho e pelos ensinamentos que muito contribuíram para minha formação acadêmica e pessoal.

Aos meus coorientadores **Dr. Ricardo Vieira Ventura** e **Dr. Roberto Carvalheiro** pelo auxílio e discussões sobre o trabalho, pela atenção e dedicação.

Ao **Prof. Dr. Flavio Schramm Schenkel**. Pela oportunidade de realizar o estágio sanduíche em Guelph. Pela hospitalidade, atenção e amizade. Pelo auxílio que foram essenciais na realização deste trabalho.

Aos **membros da banca** do Exame Geral de Qualificação – Prof. Dr. Fernando Sebastián Baldi Rey e Dr. Marcos Eli Buzanskas. Obrigada pelas sugestões que enriqueceram meu trabalho.

Aos professores Dr. Fernando Sebastián Baldi Rey e Dra. Arione Augusti Boligon

pela disponibilidade em participar da banca examinadora de mestrado.

À todos meus queridos amigos que me apoiaram e me ajudaram: Em especial as minhas irmãs de coração: Laura Fantucci, Daniela Grossi, Jaqueline Rosa, Sabrina Caetano e Denise Ayres. Obrigada pela amizade que me dedicaram.

À toda família Fantucci Matheus. Em especial: Telma, Laércio, Laura, Márcia Venilton e Carol. Por me acolherem tão bem na família e compartilharem momentos maravilhosos.

À **Jaqueline Rosa** e **Emília Barreto** pela convivência no nosso Puxadinho e pelas conversas sobre tantas coisas.

À todas as moradoras e ex-moradoras da República Rep Hour: Juliana (Miss-Pórra), Josiane (Isposta), Laura (Kixana), Jaqueline (Só-k-Rolha), Mariana (Varanella), Gabriele (Tsunami), Ana Carolina (Forfé), Bianca (K-labok), Isabela (Engasga), Geórgia (Fitági), Carolina (Sokinão) e Ana Carolina (Topera). Pela amizade, pelo carinho e pelos grandes momentos compartilhados.

À todos os amigos do Departamento de Ciências Exatas e da Pós-Gradução da FCAV-Unesp. Em especial: Bruna Naressi, Diego Guidolin, Guilherme Nascimento, Guilherme Venturini, Ismael Urbinati, Jaqueline Rosa, Mariana Maciel, Marcos Buzanskas, Mirele Picinato, Natália Grupioni, Nedenia Stafuzza, Priscila Arrigucci, Rodrigo Savegnago, Salvador Ramos e Valdecy Cruz. Pela ajuda, pelos cafés, pelas risadas e pelas nossas conquistas.

À todos os funcionários e professores do Departamento de Ciências Exatas pela amizade, colaboração e incentivo.

To my dear friends in Canada. Daniela Grossi, Jane Kelly, Fabricia Braga, Claudia Bertoli, Luiz Brito, Daniel Gordo, Fabiana Mota, Ana Paula Terakado (Fuxis), Scott Gray, Carla Domingues (Bisko), Daniele Reis, Honghao Li, Narges Zare, Anna Neustaeter, Laila Schenkel, Diogo Magalhães, Kristen Rekker and family, Fabiana Rocha, Schenkel family, Eduardo e Leticia Figueiredo, Sheila Bittar, Raquel Reis, Yasmin Santos, Marina Cavalcante, Henrique Zanardo (Pirr), Thaiza Costa (Tiger), Livia Torres, Vania Zanello, Vinicius Deon and Mario Piccoli. Thank you for providing me such an excellent experience in Guelph.

To **my Canadian family**. Glenn, Gabe, Jessica and Olivia Urquhart. Thank you for welcoming into your home and taking care me.

À UNESP/FCAV pela formação profissional.

À **Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)** pela concessão da bolsa de estudos (Processo n. 2012/21891-8) e da Bolsa de Estágio no exterior (Processo n. 2013/02175-2). Obrigada pela oportunidade.

À **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior** (**CAPES)** pela bolsa de estudos concedida no início do curso de mestrado.

À **Embrapa** pela concessão dos dados utilizados na dissertação. Em especial as pesquisadoras Dr. Luciana Correia de Almeida Regitano e Dra. Cintia Righetti Marcondes.

À **todos os cachorros** que convivem comigo pela alegria e pelo afeto. Em especial ao meu querido filhote **Bentinho** por tornar meus dias mais doces e divertidos.

MUITO OBRIGADA!

SUMÁRIO Página

METODOLOGIAS E ESTRATÉGIAS DE IMPUTAÇÃO DE MARCADORES GENÉTICOS EM BOVINOS DA RAÇA CANCHIM

RESUMO - Painéis de marcadores genéticos de alta densidade (HD) possuem forte desequilíbrio de ligação, que permite melhores predições de valores genômicos. Entretanto, genotipar animais com estes painéis apresenta custo elevado, tornando-se uma limitação para a genotipagem de todos os candidatos à seleção. Uma alternativa para a redução desses custos é utilizar imputação de genótipos. A imputação é um método em que marcadores de uma população genotipada com painéis de baixa densidade (LD) são inferidos utilizando informações provenientes de uma população referência genotipada com painéis HD. O objetivo deste trabalho foi comparar em diferentes cenários metodologias de imputação de marcadores moleculares de polimorfismos de nucleotídeos únicos (SNP) em bovinos de corte da raça Canchim. Foram utilizadas informações de 285 animais da raça Canchim, 114 do grupo genético "MA" e 1 touro da raça Charolês genotipados com painel Illumina BovineHD BeadChip (786.799 SNP), nascidos entre 1999 e 2005 e provenientes da base de dados genômicos da Embrapa Pecuária Sudeste, São Carlos, SP. A edição dos dados foi realizada no *software* R e em linguagem C++. Para a frequência mínima de alelos (MAF) foram aplicados 3 diferentes critérios: sem remover MAF (QC1); SNP com MAF menor que 0,0025 (QC2) e menor que 0,10 (QC3) foram excluídos. O painel HD original foi reduzido para painéis de baixa densidade (LD) 3K, 6K, 9K, 50K, 20K, 80K e 90K, selecionando os marcadores em comum entre o painel HD original e os painéis comerciais Illumina Bovine3K (3K), BovineLD (6K), GeneSeek Genomic Profiler (GGP) Beef LD (9K), BovineSNP50 (50K), GGP Indicus LD (20K) ,GGP Beef HD (80K) e GGP Indicus HD (90K). Os animais foram divididos em diferentes cenários, denominados de população referência e imputação, sendo o cenário 1 (C1): População referência formada por animais nascidos de 1999 até 2004 e população de imputação composta por animais nascidos em 2005; Cenário 2; C2A : População referência formada pelos animais do grupo genético Canchim e população de imputação pelos animais do grupo genético MA; C2B: População referência composta pelos animais do grupo genético MA e a população de imputação por animais Canchim; C2C: População referência composta por animais Canchim e do grupo genético MA (nascidos até 2004) e população de imputação formada por animais nascidos em 2005 do grupo genético MA.; C2D: População referência constituída por animais MA e Canchim (nascidos até 2004) e a população de imputação composta por animais Canchim nascidos em 2005; Cenário 3; C3A: População referência composta pelos machos e a de imputação pelas 204 fêmeas (155 Canchim e 49 MA); C3B: População referência formada pelos machos e fêmeas nascidas até 2004 e a de imputação constituída somente pelas fêmeas nascidas em 2005. A imputação dos genótipos foi realizada pelo método baseado na população utilizando desequilíbrio de ligação entre marcadores pelo programa *FImpute* v2.2 e pelo programa *BEAGLE* v3.3.2. A acurácia de imputação foi estimada por meio da taxa de concordância e pelo quadrado da correlação alélica (R^2 alélico). A aplicação dos critérios para MAF (QC1, QC2 e QC3) no controle de qualidade do SNP não apresentou ganhos em acurácia de imputação. A acurácia de imputação de LD para HD pela taxa de concordância variou de 56 a 98% e o R^2 alélico de 0,25 a 0,96 utilizando o *FImpute*. Pelo *software BEAGLE* a variação foi de

50 a 96% pela taxa de concordância e de 0,17 a 0,94 pelo R^2 alélico. Para os cenários propostos, o mais eficiente em termos de acurácia foi o C3B. Considerando possível aplicação deste estudo na seleção genômica para esta população, o painéis LD mais adequados para a genotipagem seriam o de 80K e 90K. O tempo de execução para realização da imputação pelo *software FImpute* foi de 20 a 100 vezes menor em relação ao *software BEAGLE.* A aplicação de restrição de MAF no controle de qualidade dos dados avaliados não é necessária para análises de imputação. Os painéis 80K e 90K podem ser acuradamente imputados para o painel HD na raça Canchim. Para obtenção de melhores acurácias de imputação animais Canchim e do grupo genético MA devem ser considerados juntos na população referência. O algoritmo do *software FImpute* demonstrou maior eficiência na imputação dos marcadores.

Palavras-chave: bovinos de corte, genômica, painéis de baixa densidade, raça composta, SNP

METHODOLOGIES AND STRATEGIES FOR GENOTYPE IMPUTATION IN CANCHIM CATTLE

ABSTRACT- High-density panels (HD) have strong level linkage disequilibrium among genetic markers (i.e. single nucleotide polymorphism - SNP), which allows better predictions of genomic breeding values. However, HD genotyping still expensive and became a limitation for the quantity of candidate animals used in genomic studies. As an alternative to decrease costs, imputation methods are powerful tools to infer missing marker genotypes from low-density (LD) panels to HD. Imputation uses information from a reference population of animals genotyped with a HD panel to impute variants that are not directly genotyped in LD panels. The objective of this study was to compare different scenarios and methodologies of imputation for the Canchim cattle. Data set was provided by Embrapa Pecuária Sudeste and comprised 285 Canchim animals, 114 MA genetic group animals, and 1 ancestor Charolais bull. Animals born between 1999 and 2005 were genotyped with the Illumina BovineHD panel (786,799 SNP). Data editing was performed in the R software and in C ++ language. Multiple scenarios combining different minor allele frequencies (MAF) thresholds for SNPs were tested: no MAF filter (QC1), and exclusion of SNPs with MAF lower than 0.0025 (QC2) and MAF lower than 0.10 (QC3). LD panels were created by masking SNPs originally present in the HD panel, and then assigning markers into the Illumina Bovine3K (3K), Illumina BovineLD (6K), Beef LD GeneSeek Genomic Profiler (9K), Indicus LD GeneSeek Genomic Profiler (20K), Illumina BovineSNP50 (50K), GeneSeek Genomic Profiler Beef HD (80K) and GeneSeek Genomic Profiler Indicus HD (90K) panels. Reference and target populations were defined as scenario 1 (C1), reference animals were born up until 2004 and target animals were born in 2005; scenario 2A (C2A), reference animals from Canchim breed and target animals from MA genetic group; scenario 2B (C2B), reference animals from MA genetic group and target animals from Canchim breed; scenario 2C (C2C), reference population was composed by all animals from Canchim and animals from MA born up until 2004, and target population was composed by animals from MA were born in 2005; scenario 2D (C2D), reference population was composed by all animals from MA and animals from Canchim were born up until 2004, and target population was composed by animals from Canchim were born in 2005; scenario 3A (C3A), reference population was composed by bulls and target population was composed by cows; and scenario 3B (C3B), reference population was composed by all bulls and cows born until 2004, and target population was composed by cows born in 2005. Imputation analyses were carried out by means of FImpute and BEAGLE software. Imputation accuracy was obtained by genotype concordance rate and allelic R square (R2). The scenarios for MAF (QC1, QC2, and QC3) did not improve the imputation accuracy. The genotype concordance rate ranged from 56 to 98% and the R2 ranged from 0.25 to 0.96 using FImpute. Genotype concordance rate and R2 obtained from BEAGLE software varied from 50 to 96% and 0.17 to 0.94, respectively. The highest accuracy was observed for scenario C3B. The 80K and 90K panels were considered as the most adequate panels to imputate for HD. FImpute reducing run-time by 20 to 100 compared to BEAGLE. No advantages in genotype imputation were observed for SNP filtering according to MAF. Canchim and MA individuals should be considered in the reference population for imputation. The FImpute algorithm had better performance

than BEAGLE for genotype imputation in our data set.

Keywords: beef cattle, composite breed, genomic, low-density panel, SNP

CAPÍTULO 1 - CONSIDERAÇÕES GERAIS

1 INTRODUÇÃO

A seleção genética tradicional de animais tem sido realizada a partir da predição dos valores genéticos obtidos por meio de dados fenotípicos e informações de pedigree. Entretanto, a inclusão de informações genotípicas na avaliação genética dos animais de produção pode trazer grandes benefícios ao melhoramento genético animal, pois permite aumento da resposta à seleção. Meuwissen, Goddard e Hayes (2001) propuseram novo método de seleção, denominado de seleção genômica ampla, que consiste na predição de valores genéticos genômicos (GEBV-Genomic Estimated Breeding Values) por meio de marcadores moleculares de polimorfismos de nucleotídeos únicos ("Single Nucleotide Polymorphism" - SNP). Os SNP são alterações em um único nucleotídeo que ocorrem na sequência de DNA. Devido ampla distribuição ao longo do genoma é o marcador molecular mais comum utilizado em estudos genômicos. No genoma bovino foram identificados cerca de 13 milhões de SNP (National Center for Biotechnology Information – NCBI, 2014).

Segundo Goddard e Hayes (2007), painéis desenvolvidos para genotipagem de indivíduos em alta densidade, ou seja, contendo milhares de marcadores do tipo SNP, possibilitariam capturar variabilidade genética de determinada característica quantitativa, desde que os genes que a determinam estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores. Entretanto, genotipar animais com painéis de alta densidade apresenta custo elevado, tornando-se limitação para a genotipagem de todos os animais candidatos à seleção. Desta forma, alguns pesquisadores têm utilizado a metodologia de imputação de genótipos (DRUET; SCHROOTEN; ROSS, 2010; HAYES et al., 2011; SARGOLZAEI; CHESNAIS; SCHENKEL, 2011).

Imputação de genótipos é um termo utilizado para predição de genótipos não-observados nos painéis de genotipagem. A ausência dos marcadores pode ser devido às restrições aplicadas aos dados, SNP não considerados no painel ou erros de genotipagem (*no-call*) (DRUET;SCHROOTEN; ROSS, 2010; MARCHINI; HOWIE, 2010; WANG et al., 2012). A imputação permite inferir genótipos para um painel de

alta densidade (HD) a partir de um painel baixa densidade (LD), utilizando informações de outros indivíduos genotipados com painéis HD (população referência) (SARGOLZAEI; CHESNAIS; SCHENKEL, 2011). Deste modo é possível genotipar animais com painéis LD e predizer os genótipos do painel HD. De acordo com Ventura et al. (2011), a imputação tornou-se ferramenta poderosa para aumentar o número de animais com informações dos marcadores em painéis HD e, consequentemente, proporcionar avaliações genéticas mais confiáveis com menor custo.

Os métodos de imputação podem ser classificados em dois grupos: baseados em informações do pedigree ou da população. A imputação baseada no pedigree utiliza regras de ligação e segregação mendeliana para predizer os genótipos, sendo mais acurada para os indivíduos que possuem parentes genotipados. Métodos de imputação baseados em informações da população, beneficiam-se do desequilíbrio de ligação entre os SNP observados na população utilizada como referência. Este é viável para um conjunto de indivíduos que não são aparentados ou que não possuem ancestrais próximos genotipados. Pesquisadores têm desenvolvido diferentes técnicas para sua implementação, tanto baseado na metodologia de imputação pelo pedigree (HICKEY et al., 2011) como baseado na imputação pela população (BROWNING ; BROWNING, 2009) ou ainda pela combinação de ambas as metodologias (SARGOLZAEI; CHESNAIS; SCHENKEL, 2011; VANRADEN et al., 2011).

Considerando a eficiência dos métodos de imputação e os diversos desafios para implementação da seleção genômica nos rebanhos brasileiros de bovinos de corte, tais como a obtenção e manipulação das informações genômicas de alta densidade e o alto custo da genotipagem, é imprescindível que pesquisas com imputação sejam realizadas no Brasil.

2 OBJETIVOS

Comparar em diferentes cenários estratégias e metodologias de imputação de marcadores moleculares de polimorfismos de nucleotídeos únicos (SNP) em bovinos de corte da raça Canchim (62,5% de genes raça Charolês e 37,5% de Zebuíno) e do

grupo genético MA (65,6% de genes raça Charolês e 34,4% de Zebuíno), visando futura aplicação dos marcadores imputados para estudos de seleção genômica.

3 REVISÃO DE LITERATURA

3.1 Imputação de genótipos

Na seleção genômica (MEUWISSEN; GODDARD; HAYES, 2001), efeitos de marcadores moleculares (SNP- polimorfismo de nucleotídeo único) distribuídos ao longo do genoma são inferidos simultaneamente utilizando uma população referência com dados fenotípicos e genotípicos. Posteriormente, valores genéticos genômicos (GEBV) são calculados em outra população (validação) somando os efeitos dos marcadores ou haplótipos (conjunto de SNP que são encontrados no mesmo cromossomo) destes. Para gerações futuras, somente dados genotípicos são necessários, entretanto a população referência com informações de fenótipos deverá ser mantida para viabilizar posteriores predições. Nesta metodologia, é fundamental que regiões no genoma responsáveis pela expressão das características quantitativas (QTL - Quantitative trait loci) estejam em desequilíbrio de ligação (DL) com ao menos um marcador, sendo DL associações não-aleatórias entre a presença de alelos em dois ou mais loci de um mesmo cromossomo (BOHMANOVA; SARGOLZAEI; SCHENKEL, 2010). Para garantir este preceito a densidade dos marcadores deve ser suficientemente grande (GODDARD;HAYES, 2007).

Estudos de associações genômicas permitem identificar marcadores moleculares associados com QTL, possibilitando detectar regiões no genoma responsáveis por características de importância econômica para pecuária. Afim de melhorar a predição dos resultados obtidos pela seleção genômica e estudos de associação metodologias de imputação de genótipos tem sido aplicadas em dados genômicos (DRUET; SCHROOTEN; ROSS, 2010; HAYES et al., 2011; SARGOLZAEI; CHESNAIS; SCHENKEL, 2011). A imputação possibilita diversas aplicações como:

i. Gerar painéis de alta densidade (HD) para animais genotipados com os de

baixa densidade (LD) baseando-se em dados genotípicos de uma população referência de indivíduos genotipados com painéis HD (SARGOLZAEI; CHESNAIS; SCHENKEL, 2011; BOICHARD et al., 2012), visando redução dos custos com a genotipagem.

ii. Combinar dados provenientes de populações genotipadas com painéis de diferentes densidades ou raças distintas, permitindo painéis com de única densidade e assim aumentar o número de indivíduos na população, para promover resultados mais satisfatórios em estudos genômicos (HAYES et al., 2011; LARMER et al., 2012).

iii. Predizer genótipos faltantes (*missing genotype*) do painel de genotipagem provenientes do controle de qualidade do dados ou erros de leitura, aumentando o *call rate* (porcentagem de genótipos válidos) dos animais (MARCHINI ; HOWIE, 2010; MA et al., 2013). A inferência desses genótipos é importante para implementação da seleção genômica e estudos de associação, pois aumenta o número de SNP e animais disponíveis para estimar os efeitos dos marcadores (WENG et al., 2012).

Existem diversos *softwares* desenvolvidos para imputação com diferentes metodologias tais como *BEAGLE, IMPUTE2, FastPHASE, MACH, FImpute, AlphaImpute, findhap*. Entretanto, a maioria deles foram desenvolvidos para aplicação em genótipos humanos, em que o tamanho efetivo da população, aproximadamente 10.000 (KRUGLYAK, 1999), é relativamente maior em relação a de bovinos, aproximadamente 100 (RIQUET et al., 1999) e na maioria das vezes apresentam informações de parentesco desconhecidas, necessitando de algoritmos complexos para imputação dos genótipos, tornando-se inviável para aplicação em grandes conjuntos de dados, como de bovinos (JOHNSTON; KISTEMAKER; SULLIVAN, 2011). Os métodos de imputação podem ser efetuados considerando o desequilíbrio de ligação entre os marcadores na população (*BEAGLE, IMPUTE2, FastPHASE, MACH),* utilizando informações da família por meio do *pedigree (AlphaImpute) ou* ainda combinando ambas metodologias *(FImpute* e *findhap*).

Modelos ocultos de Markov (*Hidden Markov Models*- HMM) têm sido amplamente utilizados para predição de fases dos haplótipos em metodologias de imputação, principalmente em populações em que os indivíduos não são

aparentados. Este modelo permite identificar um estado não observado (*hidden states*) por meio de uma sequência de informações observadas. O modelo é definido por 5 elementos: i) estados ocultos: $S_1, S_2, \ldots, S_{N;}$; ii) valores observados: $v_1, v_2, \ldots, v_N$; iii) matriz de probabilidades de transição: a_{ij} é a probabilidade de ocorrer a transição do estado S_i para o estado S_j ; iv) matriz de probabilidades de emissão: $b_j(vk)$ é a probabilidade valor observado vk ser gerado no estado S_j; v) probabilidade inicial: π_i é a probabilidade do processo HMM iniciar no estado S_i (RABINER ; JUANG, 1986).

O *software BEAGLE* utiliza HMM para inferir as fases do haplótipos e imputar genótipos não-observados (BROWNING ; BROWNING, 2009). O algoritmo realiza um agrupamento ("cluster") de haplótipos de acordo com a posição de cada marcador ao longo do cromossomo. No modelo HMM, o estado oculto corresponde aos haplótipos que serão inferidos; os valores observados são os genótipos da população referência; a probabilidade de emissão é um determinado alelo ser igual a 1 (*major*) e o outro alelo ser igual a 0 (*minor*); probabilidade de transição é a proporção de haplótipos que sofreram recombinação ou mutação e a probabilidade inicial é a probabilidade do modelo iniciar-se em determinado haplótipo.

Métodos de imputação baseados na população assumem que os indivíduos não são aparentados, entretanto é possível identificar relação de parentesco entre os indivíduos por haplótipos compartilhados (BROWNING ; BROWNING, 2011). Pequenos segmentos do cromossomo na população sem intervenção da recombinação, carregam alelos ou haplótipos idênticos por descendência (provenientes de um ancestral comum - IBD). Portanto, estas regiões estão conservadas, ou seja, dois indivíduos aparentados irão compartilhar os mesmos alelos. Indivíduos mais próximos compartilham segmentos cromossômicos mais longos com menores frequências, pois não há quebra por recombinação dos haplótipos IBD. Entretanto parentes mais distantes apresentam haplótipos mais curtos com maior frequência pois ao longo das gerações os segmentos IBD são perdidos, principalmente por recombinação. Alguns autores (KONG et al., 2008; HICKEY et al., 2011) propuseram um método alternativo de imputação denominado *"Long range phase*" , em que genes idênticos por descendência (IBD) são identificados considerando que quanto mais próximo o relacionamento de

parentesco mais longos são os segmentos IBD.

A informação do pedigree representa importante fator para identificação da fase dos haplótipos e imputação (KONG et al., 2008). Os genótipos não-observados de um indivíduo podem ser inferidos comparando haplótipos IBD herdados com haplótipos presentes em outro indivíduo proveniente da mesma família (LI et al., 2009). De acordo com Sargolzaei, Chesnais e Schenkel (2011), este método é mais eficiente quando painéis LD apresentam poucos SNP (como o painel 3K SNP). O *software FImpute* descrito por estes autores, combina ambas metodologias de imputação (família e população), entretanto o método baseado na população difere da maioria dos softwares, pois assume-se que todos os animais estão relacionados entre si com algum grau de parentesco. Para encontrar os segmentos IBD, o algoritmo sobrepõe "janelas" formadas pelos segmentos cromossômicos dos indivíduos (*overlapping sliding windows*), possibilitando verificar a consistência da fase dos haplótipos. O procedimento para encontrar os segmentos é repetido várias vezes, com alteração do tamanho das janelas a serem comparadas, iniciando com grandes janelas. Os tamanhos são reduzidos gradualmente a cada varredura, até uma janela de tamanho bem pequeno e, após esse processo, as frequências dos haplótipos na população referência são utilizadas para predizer o genótipo faltante mais provável (VANRADEN et al., 2013).

3.2 Fatores que afetam a acurácia de imputação

Tamanho da população referência

A população referência deve ser grande o suficiente para que todos os haplótipos presentes na população imputada sejam representados na população referência. A média da taxa do erro de imputação diminui conforme aumenta o número de indivíduos na população referência (KHATKAR et al., 2012). O efeito do tamanho da amostra depende da estrutura populacional, uma vez que populações com baixa variabilidade genética requerem menos indivíduos na população referência, pois o desequilíbrio de ligação entre os marcadores é maior, além disso os indivíduos provém de número menor de antepassados (DRUET;SCHROOTEN; ROSS, 2010). Estes mesmos autores testaram tamanhos de população referência,

variando de 0 a 2000 indivíduos, e obtiveram menor taxa de erro de imputação considerando 500 ou mais indivíduos na população estudada.

Densidade dos marcadores

A densidade de marcadores é importante fator que afeta a acurácia de imputação, pois painéis com muitos SNP apresentam forte desequilíbrio de ligação entre os marcadores, reduzindo os erros de imputação (HABIER; FERNANDO; DEKKERS, 2009 ; MEUWISSEN, 2009; HICKEY et al., 2012). Segundo Wang et al. (2012), quanto maior a taxa de genótipos não observados menor é a acurácia de imputação. Para populações que apresentam desequilíbrio de ligação relativamente grande e tamanho efetivo pequeno, permite-se predizer genótipos provenientes de painéis com poucos marcadores para painéis de alta densidade (como 3K SNP para 50K SNP) com baixas taxas de erros de imputação. Dassonneville et al. (2011) relataram acurácia de 94,5% imputando de um painel de 3K SNP para 54K SNP em bovinos de leite. Hayes et al. (2011), estudando populações de ovinos relataram baixas acurácias de imputação do painel 5K SNP para 50K SNP (0,61 a 0,81) e sugeriram que a população de ovinos apresenta baixo desequilíbrio de ligação entre os marcadores e ampla diversidade genética comparados à população de bovinos de leite.

Para resultados mais acurados deve-se explorar imputação de marcadores LD utilizando a informação de *pedigree* (HABIER; FERNANDO; DEKKERS, 2009; SARGOLZAEI; CHESNAIS; SCHENKEL, 2011). O número de marcadores em comum entre painéis distintos (LD e HD) também afeta a acurácia de imputação, pois os SNP em comum vinculam ambos mapas e permitem estimar indiretamente o desequilíbrio de ligação entre os marcadores provenientes dos diferentes painéis (DRUET;SCHROOTEN; ROSS, 2010).

Distância genética da população de referência

Para a imputação de marcadores provenientes de painéis de baixa densidade (3K SNP para 50K SNP ou HD), acurácia de imputação depende da distância genética entre os indivíduos da população de imputação e os da população

referência (ZHANG ; DRUET, 2010). Se um indivíduo presente na população de imputação possui parentes na população referência, os marcadores serão facilmente identificados no painel HD. Khatkar et al. (2012) identificaram menores erros de imputação considerando pais dos indivíduos na população referência (média 2,61% versus 3,34%).

Frequência Alélica

A maioria dos indivíduos são homozigotos para o alelo mais comum da população, tornando a imputação desse alelo mais fácil do que a imputação do alelo raro (MA et al., 2013). Deste modo, imputação acurada de alelos raros (frequência do menor alelo - MAF ≤ 0,05) é fundamental, pois estes alelos podem explicar variações genéticas que não são explicadas pelos alelos comuns (CIRULLI ; GOLDSTEIN, 2010). Além disso, características definidas por alelos raros apresentam poucas observações fenotípicas, limitando estudos de associação. Todavia, se a imputação do alelo ocorrer erroneamente essa limitação será ainda maior. A taxa do erro de imputação para esse alelo também diminui de acordo com o tamanho da população referência e a densidade do painel.

Segundo Ma et al. (2013) a medida mais indicada para mensurar a acurácia de imputação de alelos raros é o coeficiente de correlação entre o genótipo imputado e o genótipo original, pois o coeficiente de correlação é influenciado por valores extremos, assim erros na predição de marcadores com baixo MAF reduz o valor da correlação, evitando acurácia de imputação superestimada. Além disso, esta medida desconsidera a imputação do genótipo por chance, ou seja, simplesmente imputar o alelo não-observado pelo alelo de maior efeito (*major alelle*).

3.3 Resultados de imputação aplicados na pecuária

Diferentes metodologias de imputação de marcadores tem sido testadas em populações de bovinos visando aplicação na seleção genômica e estudos de associação. Em bovinos de leite, Zhang e Druet (2010) verificaram que, com a inclusão de informação do pedigree, a metodologia de imputação baseada em família tem vantagens sobre o método baseado somente na população. Johnston,

Kistemaker e Sullivan (2011), relataram que quando os dois métodos são combinados, a acurácia da imputação aumenta. Entretanto, Larmer et al. (2012) ao conduzirem análises de imputação de um painel de 50K SNP para um painel de 777K SNP por meio do software *FImpute* (versão 2), obtiveram a mesma acurácia em ambos os métodos. Estes autores também observaram pequeno aumento na estimativa da acurácia de imputação considerando simultaneamente várias raças de bovinos de leite na população de referência.

Ventura et al. (2011) obtiveram acurácia de imputação do painel 6K SNP para 50K SNP variando de 71,3% a 97,8% utilizando o software *BEAGLE* em uma população mestiça de bovinos de corte. Comparando o software *BEAGLE* com o *FImpute* (versão2) também em bovinos de corte. Ventura et al. (2012) observaram aumento médio na acurácia de imputação de 2,21% utilizando o *FImpute* e também redução de 10 vezes no tempo de execução do software comparado ao *BEAGLE.* Sun et al. (2012) utilizaram os softwares BEAGLE, IMPUTE, fastPHASE, AlphaImpute, findhap e FImpute para imputar genótipos do painel de 5K SNP para 50K SNP em bovinos da raça Angus. Estes autores relataram acurácia de imputação variando de 0,89 a 0,98 e obtiveram melhores resultados utilizando os softwares BEAGLE e FImpute .

Khatkar et al. (2012) estudando bovinos de leite da Austrália, relataram valor da acurácia na predição do valor genômico de genótipos imputados variando de 0,50 a 0,56 para produção de leite, 0,51 a 0,53 para produção de gordura, 0,48 a 0,53 para proteína, 0,20 a 0,23 para fertilidade da progênie e 0,24 a 0,25 sobrevivência, sendo que a acurácia aumentou conforme reduziram os erros de imputação. Dassonneville et al. (2011) e Mulder et al. (2012) também observaram em bovinos leiteiros da raça Holandesa aumento na acurácia do valor genômico em relação ao menor erro de imputação para características reprodutivas e de produção de leite.

Boichard et al. (2012), imputando genótipos do painel comercial Illumina BovineLD BeadChip (6K SNP) para BovineSNP50 (50K SNP) por meio do software *BEAGLE* em populações constituídas de diferentes raças de bovinos provenientes da Austrália, França e América do Norte, encontraram acurácia de imputação variando 92,3 % a 98,8 % . Ma et al. (2013), predizendo marcadores dos painéis 3K

SNP e 54K SNP para HD em bovinos de leite, obtiveram taxa de alelos imputados corretamente variando de 94% a 97% (3K SNP para HD) e 99% (54K SNP para HD) e coeficientes de correlação entre marcadores imputados e verdadeiros de 0,74 a 0,90 (3K SNP para HD) e 0,96 a 0,98 (54K SNP para HD) utilizando os programas *FImpute* e *BEAGLE*. Estes programas alcançaram melhores resultados comparados a outros softwares utilizados no estudo (*findhap, AlphaImpute, IMPUTE2, AlphaBea*).

Hozé et al. (2013) estudaram dezesseis diferentes raças francesas de bovinos de corte e de leite e relataram erro de imputação do painel de 50K SNP para HD variando de 0,31% a 2,41% utilizando o software *BEAGLE.* Estes autores observaram que o erro de imputação em média foi maior em raças de bovinos de corte do que em bovinos de leite devido ao tamanho da população referência disponível, o número efetivo de ancestrais de cada raça e a relação de parentesco entre a população de imputação e referência.

Diante da eficiência da aplicação dos métodos de imputação na pecuária, estudos de imputação devem ser realizados nos rebanhos de bovinos de corte do Brasil. Raças compostas desenvolvidas por meio de cruzamentos de animais *Bos taurus taurus* e *Bos taurus indicus* se tornaram importantes para bovinocultura de corte brasileira (ALENCAR, 1988) com intuito de unir características de precocidade e rendimento dos animais taurinos com características de rusticidade dos zebuínos. A raça Canchim é exemplo de raça composta produzida no Brasil, o rebanho desses animais representa 3% do plantel brasileiro (ABCCAN, 2013). Estudos genômicos e de associação tem sido realizados nessa raça (BUZANSKAS et al., 2013; MOKRY et al., 2013) com a finalidade de identificar animais geneticamente superiores para características de interesse econômico como conformação da carcaça, espessura de gordura subcutânea e perímetro escrotal. Assim, estudos de imputação devem ser efetuados na população de animais Canchim visando reduzir custos de genotipagem e promover aumento das informações genotípicas no programa de melhoramento genético da raça.

4 REFERÊNCIAS

ABCCAN. Associação Brasileira de Criadores de Canchim. Disponível em: http://www.abccan.com.br/canchim/index.php/a-raca.html. Acesso em: 20 de outubro de 2013.

ALENCAR, M. M. Bovino – Raça Canchim: Origem e desenvolvimento. Documento, 4. Brasília, EMBRAPA – DPU, 1988, 102p.

BOHMANOVA, J.; SARGOLZAEI, M.; SCHENKEL, F. S. Characteristics of linkage disequilibrium in North American Holsteins. **BMC Genomics**, v. 11, n.421, 2010.

BOICHARD, D.; CHUNG, H.; DASSONNEVILLE, R.; DAVID, XEGGEN, A.; FRITZ, S.; GIETZEN, K. J.; HAYES, B. J.; LAWLEY, C. T.; SONSTEGARD, T. S.; VAN TASSELL, C. P.; VANRADEN, P. M.; VIAUD-MARTINEZ, K. A.; WIGGANS, G. R. Design of a bovine low-density SNP array optimized for imputation. **PLoS ONE,** v.7, e34130, 2012.

BROWNING, S. R., BROWNING, B. L. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. **American Journal of Human Genetics,** v.84*,* p.210-223, 2009.

BROWNING, S. R., BROWNING, B. L. Haplotype phasing: existing methods and new developments. **Nature Reviews Genetics,** v.12*,* p.703-714, 2011.

BUZANSKAS, M. E. ; GROSSI, D. A. ; SCHENKEL, F.S. ; VENTURA, R. V. ; REGITANO, L.C.A. ; ALENCAR, M. M. ; MUNARI, D. P. Linkage disequilibrium analysis in Canchim beef cattle. In: 50ª Reunião Anual da Sociedade Brasileira de Zootecnia, 2013, Campinas/SP. 50ª Reunião Anual da Sociedade Brasileira de Zootecnia, 2013.

CIRULLI, E. T.; GOLDSTEIN, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics, v. 11, p.415-425, 2010.

DASSONNEVILLE, R.; BRØNDUM, R. F.; DRUET, T.; FRITZ, S.; GUILLAUME, F.; GULDBRANDTSEN, B.; LUND, M. S.; DUCROCQ, V.; SU, G. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. **Journal of Dairy Science**, v. 94, p.3679–3686, 2011.

DRUET, T., SCHROOTEN, C., de ROOS, A.P.W. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. **Journal of Dairy Science,** v. 93, p. 5443-5454, 2010.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, p. 323-330, 2007.

HABIER, D. R.; FERNANDO, R. L.; DEKKERS, J. C. Genomic Selection using low-density markers panels. **Genetics,** v. 182, p.343-353, 2009.

HAYES, B. J.; BOWMAN, P. J.; DAETWYLER, H. D.; KIJAS, J. W.; VAN DER WERF, J. H. J. Accuracy of genotype imputation in sheep breeds. **Animal Genetics,** v 43, p. 72-80, 2011.

HICKEY, J.M., KINGHORN, B.P., TIER, B., WILSON, J.F., DUNSTAN, N.; VAN DER WERF, J.H.J. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. **Genetics Selection Evolution,** v. 43, n. 12, 2011.

HICKEY, J. M; CROSSA, J.; BABU, R.; DE LOS CAMPOS, G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. **Crop Science**, v.52, p.654–663, 2012.

HOZÉ, C.; FOUILLOUX, M.; VENOT, E.; GUILLAUME, F.; DASSONNEVILLE, R.; FRITZ, S.; DUCROCQ, V; PHOCAS, F.; BOICHARD, D; CROISEAU, P. High-density marker imputation accuracy in sixteen French cattle breeds. **Genetics Selection Evolution**, v.45, p.33, 2013.

JOHNSTON, J., KISTEMAKER, G. SULLIVAN, P.G. Comparison of different imputation methods**. Interbull Open Meeting**, Stanvager, Norway, 2011.

KHATKAR, M. S.; MOSER, G.; HAYES, B. J.; RAADSMA, W. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle, **BMC Genomics**, v.13, p.538, 2012.

KONG, A.; MASSON, G.; FRIGGE, M. L.; GYLFASON, A.; ZUSMANOVICH, P.; THORLEIFSSON, G.; OLASON, P. I.; INGASON, A.; STEINBERG, S.; RAFNAR, T; SULEM, P.; MOUY, M.; JONSSON, F.; THORSTEINSDOTTIR, U.; GUDBJARTSSON, D. F.; STEFANSSON, H.; STEFANSSON, K. Detection of sharing by descent, long-range phasing and haplotype imputation. **Nature Genetics**, v. 40, n.9, p.1068-1075, 2008.

KRUGLYAK, L. Prospects for whole-genome linkage disequilibrium mapping of commom disease genes. **Nature Genetics**, v.22, p.139-144, 1999.

LARMER, L., SARGOLZAEI, M., VENTURA, R., SCHENKEL, F. Imputation accuracy from low to high density using within and across breed reference populations in Holstein, Guernsey and Ayrshire cattle. Technical report to the Dairy Cattle Breeding and Genetics Committee on February 28, 2012. University of Guelph, Guelph, ON, Canada, 2012.

LI ,Y.; WILLER, C. J.; SANNA, S.; ABECASIS, G. R. Genotype Imputation. **Annual Review Genomics Human Genetics**, v.10, p.387-406, 2009.

MA, P.; BRØDUM, R. F.; ZHANG, Q.; LUND, M. S.; SU, G. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle, **Journal of Dairy Science,** v. 96, p.4666-4677, 2013.

MARCHINI, J.; HOWIE B. Genotype imputation for genome-wide association studies. **Nature Review Genetics**, v.11, p.499-511, 2010.

MEUWISSEN, T. H. E.; GODDARD, M. E.; HAYES, B. J. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2001.

MEUWISSEN, T. H. E. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. **Genetics Selection Evolution**, v.41, p.35, 2009.

MOKRY, F. B.; HIGA, R. H.; MUDADU, M. A.; LIMA, A. O.; MEIRELLES, S. L. C.; SILVA, M. V. B.; CARDOSO, F. F.; OLIVEIRA, M. M.; URBINATI, I.; NICIURA, S. C. M.; TULLIO, R. R.; ALENCAR, M. M.; REGITANO, L. A. C. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics**, v14, p.47-57, 2013.

MULDER, H. A.; CALUS, M. P. L.; DRUET, T.; SCHROOTENT, C. Imputation of genotypes with low –density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. **Journal of Dairy Science**, v.95, p.876-889, 2012.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION – NCBI. Disponível em: http://www.ncbi.nlm.nih.gov/snp/. Acesso em : 29 fev. 2014.

RABINER, L. R.; AND JUANG, B. H. An Introduction to Hidden Markov Models IEEE Acousfics. **Speech & Signal Processing Magazine**, v.3, p.1-16,1986.

RIQUET, J.; COPPIETERS, W.; CAMBISANO, N.; ARRANZ, J. J.; BERZI, P.; DAVIS, S.; GRISART, B.; FARNIR, F.; KARIM, L.; MNI, M.; SIMON, P.; TAYLOR, J. F.; VANMANSHOVEN, D.; WOMACK, J. E.; GEORGES, M. Fine-mapping of quantitative trait loci by identity-by-descent fine-mapping of QTL in outbred populations: Application to milk production in dairy cattle. **Proceedings of the National Academy of Sciences of the United States of America**, v. 96, p.9252–9257,1999.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F.S. FImpute. An efficient imputation algorithm for dairy cattle populations. **Journal of Animal Science**, v. 89(E-Suppl. 1)/**Journal of Dairy Science**, v. 94(E-Suppl. 1), p. 421 (abstr. 333). 2011.

SUN, C., WU, X., WEIGEL, K.; ROSA, G.J.M.; BAUCK, S.; WOODWARD, B.W.; SCHNABEL, R.D.; TAYLOR, J.F.; GIANOLA, G. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. **Genetics Research**, Cambridge, v. 94, p.133-150, 2012.

VANRADEN, P. M.; O'CONNELL, J. R.; WIGGANS, G. R.; WEIGEL, K. A. Genomic evaluations with many more genotypes. **Genetics, Selection, Evolution**, v.43, n.1, p. 10-20. 2011.

VANRADEN, P. M.; NULL, D. J., SARGOLZAEI, M., WIGGANS, G. R.; TOOKER, M. E.; COLE, J. B.; SONSTEGARD, T. S.; CONOOR, E. E.; WINTERS, M.; VAN KAAM, J. B. C. H. M.; VALENTINI, A.; VAN DOORMAAL, B. J. II.; FAUST, M. A.; DOAK, G. A. Genomic imputation and evaluation using high-density Holstein genotypes. **Journal of dairy science**, v.96, p.668-678, 2013.

VENTURA, R.V., SCHENKEL, F.S., WANG, Z., MILLER, S.P. Accuracy of imputation using 6K and 50K SNP chips in beef cattle. Livestock Gentec's 2nd Annual Conference , Edmonton, Alberta, Canada, 2011.

VENTURA, R.V., SCHENKEL, F.S, SARGOLZAEI, M., WANG, Z., CHANGXI, L., MILLER, S.P. Accuracy of imputation using 6K and 50K SNP chips in multi-breed and crossbred beef cattle populations . In: Proceeding of the 33rd ISAG Conference, July 15-20, Cairns, Australia, 2012.

WANG, Y.; ZHIPENG, C.; STOTHARD, P.; MOORE, S.; GOEBEL, R.; WANG, L.; LIN, G. Fast accurate missing SNP genotype local imputation. **BMC Research Notes**, v.5, p.404, 2012.

WENG , Z.; ZHANG, Z.; XIANGDONG, D.; WEIXUAN, F.; MA, P.; WANG, C.; ZHANG, Q. Application of imputation methods to genomic selection in Chinese Holstein cattle. **Journal of Animal Science and Biotechnology**, v.3, p.6, 2012.

ZHANG, Z., DRUET, T. Marker imputation with low-density marker panels in Dutch Holstein cattle. **Journal of Dairy Science**, v. 93, n.11, p. 5487-5494, 2010.

CAPÍTULO 2 – METODOLOGIAS E ESTRATÉGIAS DE IMPUTAÇÃO DE MARCADORES GENÉTICOS EM BOVINOS DA RAÇA CANCHIM

1 INTRODUÇÃO

Informações genotípicas de animais de produção têm sido utilizadas nos programas de melhoramento genético, principalmente em bovinos, objetivando o aumento da resposta à seleção. Seleção genômica, em que se permite predizer valores genéticos genômicos por meio de marcadores moleculares e estudos de associação entre esses marcadores com determinada característica (GWAS) são exemplos de análises realizadas com dados genômicos. Todavia, estas metodologias exigem grande número de indivíduos genotipados com painéis de alta densidade contendo milhares de marcadores de polimorfismos de nucleotídeo único (SNP) espalhados ao longo de todo genoma (MEUWISSEN; GODDARD; HAYES, 2001).

Há uma diversidade de painéis disponíveis no mercado com diferentes densidades de marcadores para genotipagem de bovinos, variando de 3K até painéis com 3 milhões de marcadores, além de técnicas de sequenciamento de todo genoma (KHATKAR et al., 2012). Um dos fatores que afetam a acurácia das análises genômicas é a densidade de marcadores presentes nos painéis de genotipagem (MEUWISSEN, 2009). Painéis de alta densidade (HD) possuem maior desequilíbrio de ligação entre seus marcadores, permitindo assim melhores predições de valores genômicos e acurado mapeamento de QTL (MA et al., 2013). Entretanto, genotipar animais com painéis HD apresenta custo elevado, limitando o número de animais genotipados. Uma alternativa para a redução desses custos é utilizar imputação de genótipos (HAYES et al., 2011; JOHNSTON; KISTEMAKER; SULLIVAN, 2011).

A imputação é um método em que marcadores de uma população genotipada com painéis de baixa densidade (LD) são inferidos utilizando informações provenientes de uma população referência genotipada com painéis HD (HOWIE; DONNELLY; MARCHINI, 2009; HAYES et al., 2011). Metodologias de imputação

tornaram-se importantes ferramentas para aumentar o número de informações genotípicas, pois permitem a genotipagem de mais indivíduos com painéis LD, intensificando a seleção com baixos custos (VENTURA et al., 2011; BOICHARD et al., 2012). Além disso, estas permitem a combinação de dados provenientes de animais genotipados com painéis de diferentes densidades e/ou raças distintas, aumentando o número de indivíduos na população referência (HAYES et al., 2011; LARMER et al., 2012). Também é possível utilizar dessas técnicas para predizer genótipos não-observados no mesmo painel, aumentando o *call rate* (porcentagem de genótipos válidos) dos animais genotipados (MARCHINI; HOWIE, 2010; MA et al., 2013).

Muitos métodos de imputação e programas computacionais estão disponíveis para imputação de genótipos, tais como *BEAGLE, IMPUTE2, findhap, FImpute e AlphaImpute.* Entretanto, nem todas as metodologias conseguem obter resultados acurados na imputação dos dados genômicos de bovinos, pois alguns programas foram desenvolvidos para imputação de genótipos humanos, necessitando assim de algoritmos mais complexos (JOHNSTON; KISTEMAKER; SULLIVAN, 2011). Os métodos podem ser classificados em dois grupos: (1) baseado em informações de família (por uso de pedigree), que utiliza regras de ligação e segregação mendeliana para predizer os genótipos; (2) baseado na população, o método prediz genótipos por meio do desequilíbrio de ligação entre os SNP observados na população utilizada como referência (LI et al., 2009).

Vários fatores podem influenciar a acurácia de imputação, tais como a estrutura da população, tamanho da população referência, densidade do painel LD, frequências dos marcadores e relação de parentesco entre a população referência e a de imputação (HOWIE; DONNELLY; MARCHINI. 2009; DASSONNEVILLE et al., 2011; HICKEY et al., 2012). Considerando a necessidade de avaliar metodologias e estratégias de imputação que maximizem a acurácia na população de interesse, o objetivo deste trabalho foi comparar em diferentes cenários metodologias de imputação de marcadores moleculares SNP em bovinos de corte da raça Canchim utilizando os softwares *FImpute* e *BEAGLE* para futura aplicação em estudos genômicos .

2 MATERIAL E MÉTODOS

2.1 A raça Canchim

A raça Canchim foi formada por meio de cruzamentos alternados entre animais da Charolês *(Bos taurus taurus*) e Zebuínos (*Bos taurus indicus*), principalmente das raças Indubrasil, Guzerá e Nelore, com a finalidade de explorar os efeitos da heterose e a complementariedade entre as características favoráveis do Charolês (conformação) e a do Zebu (rusticidade). Os esquemas de cruzamentos iniciaram-se na década de 40. Os melhores resultados obtidos na época para as características de interesse econômico como precocidade, conformação e resistência ao calor e parasitas foram para animais com proporção de genes 62,5% Charolês e 37,5% Zebuíno (ALENCAR,1988). Visando alcançar maior variabilidade genética na população, outros sistemas de acasalamentos com diferentes proporções de genes Charolês-Zebu, principalmente Nelore nos dias atuais, foram realizados para obtenção da raça, formando assim diversos grupos genéticos, como o grupo genético “MA”, constituído por filhos de touros Charolês acasalados com fêmeas cruzadas Canchim X Zebu (grupo genético “A”), sendo a proporção esperada de genes para MA de aproximadamente 65,6% Charolês e 34,4% Zebu. O acasalamento entre animais “MA” ou entre animais “MA” e Canchim irão gerar animais da raça Canchim (PEREIRA et al., 2005; ANDRADE et al., 2008).

2.2 Descrição dos dados e genotipagem

Foram utilizadas informações de 400 animais genotipados com painel BovineHD BeadChip (Illumina Inc., San Diego, CA), consistindo 786.799 SNP distribuídos ao longo do genoma, nascidos entre 1999 e 2005, provenientes da base de dados genômicos da Embrapa Pecuária Sudeste, São Carlos, SP. O conjunto de dados é constituído por 205 fêmeas e 195 machos, em que 194 animais são oriundos da fazenda da Embrapa Pecuária Sudeste com origem em 17 touros, sendo 186 animais Canchim e 8 animais pertencentes ao grupo genético MA (oriundos de acasalamento entre machos da raça Charolês e fêmeas cruzadas Canchim X Zebu). O restante das amostras são pertencentes à fazendas no Estado

de São Paulo (38 animais Canchim e 9 animais MA) e de Goiás (60 animais Canchim e 97 animais MA), e 2 touros (1 Canchim e 1 Charolês) que são pais de alguns dos indivíduos genotipados. O critério para escolha dos animais que foram genotipados foi baseado em animais (machos e fêmeas) com valores genéticos extremos (alto e baixo) para a característica área de olho de lombo. A matriz de parentesco destes animais constituiu de 4.095 animais. A endogamia média da população foi igual a 0,02, calculada pelo programa CFC (SARGOLZAEI; IWAISAKI ;COLLEAU, 2006) .

2.3 Controle da qualidade dos dados

A edição dos dados foi realizada no *software* R versão 3.0.1 (R Core Team , 2013) e em linguagem C++. Foram considerados para as análises somente os cromossomos autossômicos e SNP com posição conhecida no mapa UMD_3.1 bovine assembly (ZIMIN et al., 2009), totalizando 742.906 SNP. Genótipos foram identificados de acordo com o número de alelos, sendo AA = “0”, AB = “1” e BB = “2”. Para o controle de qualidade dos genótipos foram excluídos do arquivo de dados SNP com escore de leitura (“genotype calling score”) inferiores a 0,60, desvios significativos ($p<0,000001$) do equilíbrio de Hard-Weinberg, proporção de heterozigotos desviando ±0,15 da proporção esperada e taxa de leitura (*call rate*) menor que 0,90. Para a frequência mínima de alelos (MAF) foram aplicados 3 diferentes critérios :

i. Sem remover MAF (QC1);
ii. SNP com MAF menor que 0,0025 foram excluídos (QC2);
iii. SNP com MAF menor que 0,10 foram excluídos (QC3);

Para o controle das amostras, animais que apresentaram *call rate* menor que 0,90 foram excluídos. Permaneceram no arquivo final de dados 396 animais, 621.837 SNP para o QC1, 616.565 para o QC2 e 533.023 SNP para QC3.

2.4 Delineamento do estudo para a imputação dos dados

2.4.1 Painéis de baixa densidade

Foram criados painéis de baixa densidade (LD) 3K, 6K, 9K, 20K, 50K, 80K e 90 K SNP, selecionando os marcadores em comum entre o painel HD original e os painéis comerciais Illumina Bovine3K (3K), Illumina BovineLD (6K), GeneSeek Genomic Profiler (GGP) Beef LD (9K), GGP Indicus LD (20K), Illumina BovineSNP50 (50K), ,GGP Beef HD (80K) e GGP Indicus HD (90K) (Tabela 1). Os painéis de 9K e 80K, 20K e 90K foram customizados pela empresa GeneSeek/Neogen, visando respectivamente o *Bos taurus taurus* e *Bos taurus indicus*.

Tabela 1. Número de SNP em comum entre os painéis de baixa densidade (LD) e alta densidade (HD) para cada critério de frequência mínima de alelos (MAF) utlizado no controle de qualidade.

Painel LD original	QC1	QC2	QC3
Illumina Bovine3K (2900 SNP)	2342	2341	2280
Illumina BovineLD (6909 SNP)	6283	6280	6132
GGP Beef LD (8762 SNP)	7561	7548	7266
GGP Indicus LD (19721 SNP)	14330	14305	13905
Illumina BovineSNP50 (54609 SNP)	40234	38802	30838
GGP Beef HD (76992SNP)	67601	67143	61174
GGP Indicus HD (74085 SNP)	50062	50038	49006

QC1 = sem remover MAF; QC2 = SNP com MAF menor que 0,0025 foram excluídos; QC3 = SNP com MAF menor que 0,10 foram excluídos.

2.4.2 População referência e de imputação

Os animais foram divididos em sete grupos distintos utilizando diferentes cenários, denominados de população referência e imputação (Tabela 2).

Tabela 2. Número de animais por grupos genéticos (Charolês, Canchim e MA) em diferentes cenários. Cenário 1 (C1) indivíduos agrupados por ano de nascimento; Cenário 2 (C2A, C2B, C2C e C2D) indivíduos agrupados pelo grupo genético; Cenário 3 (C3A e C3B) indivíduos agrupados por sexo.

C1	Charolês	Canchim	MA	Total
População referência (nascidos de 1999 a 2004)	1	184	68	253
População de imputação (nascidos em 2005)	-	99	44	143
C2A				
População referência	-	283	-	283
População de imputação	-	-	112	112
C2B				
População referência	-	-	112	112
População de imputação	-	283	-	283
C2C				
População referência (nascidos até 2004)	-	283	68	351
População de imputação (nascidos em 2005)	-	-	44	44
C2D				
População referência (nascidos até 2004)	-	184	112	296
População de imputação (nascidos em 2005)	-	99	-	99
C3A				
População referência (Machos)	1	128	63	192
População de imputação (Fêmeas)	-	155	49	204
C3B				
População referência (Machos + Fêmeas)	1	228	86	315
População de imputação (Fêmeas nascidas em 2005)	-	55	26	81

2.5 Cálculo do parentesco médio entre população referência e de imputação

Foi calculado o parentesco máximo, mínimo e médio genômico de acordo com VANRADEN (2008) entre a população referência e a de imputação para todos os cenários de população (Apêndice 1).

2.6 Cálculo do desequilíbrio de ligação entre os marcadores

Foi calculado o desequilíbrio de ligação médio entre os marcadores adjacentes pelo software SNPPLD (SARGOLZAEI et al., 2008), utilizando o arquivo de dados do controle de qualidade sem remover o MAF (QC1) e o painel HD (Apêndice 2). A medida de desequilíbrio de ligação utilizada foi r^2 (HILL; ROBERTSON, 1968).

2.7 Imputação dos genótipos

A imputação dos genótipos foi realizada pelo método baseado na população utilizando desequilíbrio de ligação entre marcadores. As metodologias foram implementadas pelo programa *FImpute* v2.2 (SARGOLZAEI; CHESNAIS; SCHENKEL, 2011) e pelo programa *BEAGLE* v3.3.2 (BROWNING; BROWNING, 2009).

2.7.1 *FImpute*

O algoritmo de imputação do *FImpute* utiliza como procedimentos a identificação de regiões que possam ser inferidas pelas informações dos pais ou da progênie com alto grau de confiabilidade, na reconstrução dos haplótipos (SARGOLZAEI et al., 2008) de forma que os haplótipos são reconstruídos de forma iterativa. Esta iteração é repetida até que a soma dos quadrados das probabilidades dos haplótipos seja suficientemente pequena. Os haplótipos das progênies são então combinados com os haplótipos dos pais ou descendentes e assim os genótipos não observados são inferidos. Na imputação com base na população, ao contrário da maioria dos *softwares*, o FImpute assume que todos os animais estão relacionados com algum grau de parentesco e utiliza sobreposição de “janelas” (*overlapping sliding windows*) formadas por segmentos cromossômicos, para encontrar os segmentos que são compartilhados entre os indivíduos, provenientes de um ancestral comum (IBD). Por meio da sobreposição dos segmentos cromossômicos é possível verificar a consistência da fase dos haplótipos. O procedimento para encontrar os segmentos é repetido várias vezes, com alteração do tamanho das janelas a serem comparadas, iniciando com grandes janelas. Os tamanhos são reduzidos em cada varredura gradualmente, até uma janela de

tamanho bem pequeno e, após esse processo, as frequências dos haplótipos na população referência são utilizadas para predizer o genótipo faltante mais provável.

2.7.2 *BEAGLE*

A metodologia utilizada pelo *BEAGLE* foi descrita por Browning e Browning (2009). Este *software* realiza a imputação baseada na população, agrupando os haplótipos de acordo com a posição de cada marcador ao longo do cromossomo (“ *Local cluster*”). O algoritmo utiliza os grupos formados para capturar os marcadores que estão em desequilíbrio de ligação e predizer os genótipos. O programa utiliza o modelo estatístico “*Hidden Markov Model*”, em que estima a probabilidade de que cada indivíduo carrega um genótipo particular (SNP particular), considerando os dados genotípicos do indivíduo para outros SNP e o resto da população. O *BEAGLE* utiliza um algoritmo de eliminação para determinar a fase do haplótipo para cada indivíduo e amostras dos haplótipos são utilizadas para reconstrução do "*cluster* do haplótipo local". Os haplótipos são agrupados de forma que aqueles em um mesmo *cluster* tendem a ter probabilidades semelhantes para os alelos subjacentes. Isso é repetido ao longo de 10 iterações para atingir elevado valor de acurácia, conservando a eficiência computacional. O *BEAGLE* calcula um número mínimo de recombinações que podem ocorrer no *crossing-over* e também gera relatório com possíveis recombinações dos ancestrais.

2.8 Determinação da acurácia de imputação

Para estimar a acurácia de imputação foram utilizados 2 critérios: a taxa de concordância e o quadrado da correlação alélica (R^2 alélico).

2.8.1 Taxa de Concordância

Os marcadores imputados foram comparados com aqueles verdadeiros, presentes no painel HD original e assim foi calculada a proporção de genótipos imputados corretamente, erroneamente e os não imputados. A taxa de concordância corresponde a proporção de genótipos imputados corretamente.

2.8.2 R^2 alélico

O R^2 alélico é determinado pelo quadrado da correlação entre a contagem de alelos (alelo de efeito menor, ou seja, *minor alelic*) imputados e a contagem de alelos do genótipo original. Esta metodologia foi descrita por Browning e Browning (2009). Segundo estes autores, o R^2 alélico é uma medida que não depende da frequência alélica do marcador para avaliar a eficiência da imputação, dado que o tamanho da amostra seja suficientemente grande para conter um número suficiente de alelos menores a serem imputados, caso contrário a correlação alélica pode ser estimada com erros padrões muito elevados.

3 RESULTADOS E DISCUSSÃO

3.1 Acurácia de imputação

A aplicação dos critérios para MAF (QC1, QC2 e QC3) no controle de qualidade do SNP não apresentou ganhos em acurácia de imputação (Tabela 3, 4 e 5). Quando o critério mais rigoroso para MAF (QC3) foi aplicado (MAF ≤ 0,10), a acurácia de imputação diminuiu, principalmente nos painéis LD com menores densidades (3K, 6K, 9K e 20K). A aplicação de QC3 reduziu o número de marcadores em comum com o painel HD (Tabela 1), no entanto estes SNP conectam ambos os mapas auxiliando na predição dos haplótipos utilizados na inferência do genótipos não-observados por meio do desequilíbrio de ligação existentes entre estes. Além disso, o número de SNP presentes no painel referência quanto a proporção de genótipos não-observados no painel LD influencia a acurácia de imputação (HOWIE et al., 2009; MULDER et al., 2012).

A acurácia de imputação de LD para HD pela taxa de concordância variou de 56 a 98% e o R^2 alélico de 0,25 a 0,96 utilizando o *FImpute* (Tabela 3 e 4). A taxa de alelos imputados erroneamente não diminuiu com a inclusão da informação de *pedigree*. Larmer et al. (2012), estudando uma população de bovinos de leite, não encontraram ganhos em acurácia adicionando informações de pedigree no *FImpute.* Pelo *software BEAGLE* a variação foi de 50 a 96% pela taxa de concordância e de 0,17 a 0,94 pelo R^2 alélico (Tabela 5). Observou-se que quando a taxa de

concordância é alta, o valor do R^2 alélico se aproxima dessa taxa. Entretanto, na imputação de painéis menos densos (3K, 6K, 9K e 20K SNP) para HD , em que a acurácia é inferior, o R^2 alélico foi menor.

Tabela 3. Acurácia de Imputação de painéis de baixa densidade (LD) para alta densidade (HD) pela taxa de concordância e pelo R^2 alélico, utilizando o *software FImpute* para os diferentes critérios de MAF aplicados* e cenários de população** .

		QC1		QC2		QC3	
Cenários	Tamanho do painel LD	Taxa de Concordância	R^2 alélico	Taxa de Concordância	R^2 alélico	Taxa de Concordância	R^2 alélico
C1	3K	75,84	0,59	75,70	0,59	73,27	0,52
	6K	87,82	0,79	87,72	0,79	86,46	0,76
	9K	88,74	0,81	88,64	0,81	87,42	0,77
	20K	92,50	0,87	92,43	0,87	94,54	0,85
	50K	95,24	0,92	95,20	0,92	91,66	0,90
	80K	96,99	0,95	96,96	0,95	96,61	0,94
	90K	96,72	0,95	96,68	0,94	96,39	0,93
C2A	3K	63,31	0,38	62,86	0,37	59,26	0,29
	6K	76,35	0,59	76,17	0,58	73,70	0,52
	9K	77,22	0,61	77,54	0,61	75,17	0,78
	20K	83,72	0,72	83,61	0,71	81,97	0,66
	50K	89,63	0,82	89,55	0,82	88,13	0,85
	80K	93,30	0,88	93,24	0,88	92,46	0,54
	90K	92,54	0,87	92,48	0,87	91,79	0,86
C2B	3K	60,57	0,34	60,21	0,33	56,20	0,25
	6K	71,78	0,52	71,46	0,51	68,62	0,44
	9K	73,20	0,54	72,93	0,54	70,21	0,47
	20K	79,35	0,65	79,19	0,65	77,13	0,59
	50K	86,03	0,76	85,92	0,76	84,16	0,71
	80K	90,67	0,84	90,60	0,84	89,56	0,81
	90K	89,64	0,82	89,54	0,82	88,63	0,79
C2C	3K	73,09	0,54	72,75	0,53	70,01	0,46
	6K	85,29	0,75	85,17	0,74	83,73	0,70
	9K	86,35	0,77	86,12	0,76	84,75	0,72
	20K	90,70	0,84	90,60	0,84	89,69	0,81
	50K	94,15	0,90	94,12	0,90	93,29	0,87
	80K	96,31	0,94	96,28	0,93	95,83	0,92
	90K	95,97	0,93	95,94	0,93	95,56	0,92
C2D	3K	78,02	0,62	77,74	0,62	75,46	0,75
	6K	89,96	0,83	89,84	0,83	88,81	0,89
	9K	90,76	0,84	90,67	0,84	89,70	0,82
	20K	94,21	0,90	94,15	0,94	95,85	0,93
	50K	96,40	0,94	96,36	0,90	93,54	0,96
	80K	97,77	0,96	97,74	0,96	97,48	0,95
	90K	97,58	0,96	97,55	0,96	97,33	0,97

Tabela 3. Continuação...

		QC1	QC2	QC3			
Cenários	Tamanho do painel LD	Taxa de Concordância	R^2 alélico	Taxa de Concordância	R^2 alélico	Taxa de Concordância	R^2 alélico
C3A	3K	76,6	0,60	76,52	0,60	73,77	0,53
	6K	88,82	0,81	88,71	0,81	87,43	0,78
	9K	89,71	0,82	89,56	0,82	88,39	0,79
	20K	93,22	0,88	93,13	0,88	92,44	0,86
	50K	95,65	0,92	95,60	0,93	94,99	0,91
	80K	97,22	0,95	97,19	0,95	96,86	0,94
	90K	97,01	0,95	96,98	0,95	96,69	0,94
C3B	3K	78,73	0,64	78,69	0,64	76,62	0,58
	6K	90,05	0,83	89,98	0,83	88,96	0,80
	9K	90,81	0,84	90,76	0,85	89,75	0,82
	20K	94,12	0,90	94,06	0,90	93,5	0,88
	50K	96,31	0,94	96,27	0,94	95,75	0,92
	80K	97,68	0,96	97,66	0,96	97,39	0,95
	90K	97,49	0,96	97,47	0,96	97,22	0,95

* QC1: Sem remover MAF; QC2: SNP com MAF menor que 0,0025 foram excluídos; QC3: SNP com MAF menor que 0,10 foram excluídos.

**C1: População Referência: animais nascidos até 2004; População de Imputação: nascidos em 2005; C2A: População referência: animais Canchim; População de imputação: animais MA; C2B: População referência: animais MA; População de imputação: animais Canchim; C2C: População referência: animais Canchim e MA (nascidos até 2004); População de imputação animais MA nascidos em 2005; C2D: População referência: animais MA e Canchim (nascidos até 2004); População de imputação: animais Canchim nascidos em 2005; C3A: População referência: machos; imputação fêmeas C3B: População referência: machos e fêmeas nascidas até 2004; imputação : fêmeas nascidas em 2005.

Tabela 4. Acurácia de Imputação de painéis de baixa densidade (LD) para alta densidade (HD) pela taxa de concordância e pelo R^2 alélico, utilizando o *software FImpute* com informações de *pedigree* para os diferentes critérios de MAF aplicados* e cenários de população**

		QC1		QC2		QC3	
Cenários	Tamanho do painel LD	Taxa de Concordância	R^2	Taxa de Concordância	R^2	Taxa de Concordância	R^2
C1	3K	75,81	0,59	75,66	0,58	73,07	0,52
	6K	87,63	0,79	87,51	0,79	86,21	0,75
	9K	95,08	0,92	95,04	0,91	94,36	0,90
	20K	88,54	0,81	88,41	0,80	87,16	0,77
	50K	92,30	0,87	92,24	0,87	91,44	0,84
	80K	96,58	0,94	96,55	0,94	96,23	0,93
	90K	96,72	0,94	96,83	0,95	96,46	0,94
C2A	3K	63,28	0,38	62,96	0,37	59,18	0,28
	6K	76,36	0,59	76,18	0,58	73,72	0,52
	9K	77,69	0,61	77,54	0,61	75,20	0,54
	20K	83,74	0,72	83,61	0,71	81,95	0,66
	50K	89,62	0,82	89,54	0,82	88,13	0,78
	80K	92,54	0,87	92,47	0,87	91,78	0,84
	90K	93,29	0,88	93,23	0,88	92,46	0,86
C2B	3K	60,60	0,33	60,25	0,33	56,25	0,25
	6K	71,81	0,52	71,49	0,51	68,65	0,44
	9K	73,22	0,54	72,95	0,54	70,22	0,47
	20K	79,35	0,65	79,20	0,65	77,15	0,59
	50K	86,03	0,76	85,92	0,76	84,15	0,71
	80K	89,64	0,82	89,53	0,82	88,63	0,79
	90K	90,68	0,84	90,60	0,84	89,56	0,81
C2C	3K	73,13	0,54	72,88	0,53	69,95	0,46
	6K	85,32	0,75	85,15	0,74	83,65	0,70
	9K	86,32	0,76	86,15	0,7611	84,69	0,72
	20K	90,72	0,84	90,61	0,83	89,69	0,81
	50K	94,16	0,90	94,11	0,89	93,29	0,87
	80K	95,97	0,93	95,93	0,9295	95,57	0,92
	90K	96,32	0,94	96,28	0,9359	83,65	0,71
C2D	3K	78,09	0,62	77,83	0,6216	75,56	0,56
	6K	89,98	0,83	89,86	0,8294	88,83	0,80
	9K	90,77	0,84	90,68	0,8435	89,72	0,82
	20K	94,22	0,90	94,16	0,9025	93,55	0,88
	50K	96,40	0,94	96,36	0,9394	95,85	0,93
	80K	97,57	0,96	97,55	0,9596	97,32	0,95
	90K	97,77	0,96	97,74	0,9627	97,48	0,95

Tabela 4. Continuação...

		QC1	QC2	QC3			
Cenários	Tamanho do painel LD	Taxa de Concordância	R^2 alélico	Taxa de Concordância	R^2 alélico	Taxa de Concordância	R^2 alélico
C3A	3K	76,60	0,60	76,33	0,60	73,60	0,53
	6K	88,82	0,81	88,49	0,81	87,43	0,77
	9K	89,50	0,82	89,32	0,82	88,14	0,79
	20K	93,05	0,88	92,96	0,88	92,26	0,86
	50K	95,65	0,93	95,48	0,92	94,86	0,90
	80K	96,92	0,95	96,88	0,94	96,6	0,93
	90K	97,13	0,95	97,12	0,95	96,77	0,94
C3B	3K	78,87	0,64	78,69	0,637	76,51	0,58
	6K	89,95	0,83	89,87	0,83	88,83	0,80
	9K	90,71	0,84	90,62	0,84	89,64	0,81
	20K	94,01	0,90	93,97	0,90	93,39	0,88
	50K	96,25	0,94	96,21	0,94	95,69	0,92
	80K	97,44	0,96	97,42	0,96	97,17	0,95
	90K	97,65	0,96	97,63	0,96	97,35	0,95

* QC1: Sem remover MAF; QC2: SNP com MAF menor que 0,0025 foram excluídos; QC3: SNP com MAF menor que 0,10 foram excluídos.

**C1: População Referência: animais nascidos até 2004; População de Imputação: nascidos em 2005; C2A: População referência: animais Canchim; População de imputação: animais MA; C2B: População referência: animais MA; População de imputação: animais Canchim; C2C: População referência: animais Canchim e MA (nascidos até 2004); População de imputação animais MA nascidos em 2005; C2D: População referência: animais MA e Canchim (nascidos até 2004); População de imputação: animais Canchim nascidos em 2005; C3A: População referência: machos; imputação fêmeas C3B: População referência: machos e fêmeas nascidas até 2004; imputação : fêmeas nascidas em 2005.

Tabela 5. Acurácia de Imputação de painéis de baixa densidade (LD) para alta densidade (HD) pela taxa de concordância e pelo R^2 utilizando o *software BEAGLE* para os diferentes filtros de MAF aplicados* e cenários de população**.

		QC1		QC2		QC3	
Cenários	Tamanho do painel LD	Taxa de Concordância	R^2	Taxa de Concordância	R^2	Taxa de Concordância	R^2
C1	3K	66,58	0,44	66,27	0,44	63,34	0,37
	6K	80,99	0,68	80,79	0,68	79,01	0,63
	9K	82,35	0,71	82,19	0,70	80,48	0,66
	20K	87,66	0,80	87,50	0,71	86,24	0,76
	50K	92,22	0,87	92,14	0,87	91,04	0,84
	80K	95,29	0,92	95,26	0,92	94,68	0,90
	90K	95,06	0,92	95,03	0,92	94,52	0,90
C2A	3K	59,90	0,33	59,73	0,33	55,96	0,25
	6K	72,46	0,53	72,23	0,58	69,72	0,46
	9K	73,96	0,56	73,78	0,55	71,18	0,48
	20K	79,93	0,66	79,75	0,65	77,70	0,60
	50K	86,73	0,77	86,66	0,77	84,88	0,72
	80K	91,57	0,76	91,51	0,85	90,47	0,82
	90K	90,93	0,85	90,85	0,84	89,97	0,81
C2B	3K	55,15	0,25	54,83	0,25	50,35	0,16
	6K	63,26	0,38	63,00	0,38	59,39	0,30
	9K	64,41	0,40	64,15	0,40	60,70	0,32
	20K	70,18	0,50	69,91	0,49	67,19	0,42
	50K	80,09	0,67	79,95	0,66	77,44	0,59
	80K	87,46	0,79	87,35	0,79	85,96	0,75
	90K	85,91	0,76	85,79	0,76	84,55	0,72
C2C	3K	65,05	0,41	64,55	0,40	61,82	0,33
	6K	79,70	0,66	79,32	0,65	77,57	060
	9K	81,06	0,68	80,85	0,67	79,02	0,62
	20K	86,64	0,77	86,55	0,77	85,05	0,73
	50K	91,33	0,86	91,24	0,8517	90,01	0,82
	80K	94,59	0,91	94,53	0,9081	93,83	0,89
	90K	94,44	0,91	94,36	0,9050	93,76	0,89
C2D	3K	68,92	0,48	68,57	0,4729	65,95	0,41
	6K	84,00	0,73	83,86	0,7306	82,46	0,69
	9K	85,38	0,76	85,23	0,7547	83,84	0,72
	20K	90,26	0,84	90,23	0,8403	89,18	0,81
	50K	93,92	0,90	93,90	0,9005	92,98	0,88
	80K	96,35	0,94	96,30	0,9402	95,85	0,93
	90K	96,12	0,94	96,10	0,9371	95,69	0,92

Tabela 5. Continuação...

		QC1	QC2	QC3			
Cenários	Tamanho do painel LD	Taxa de Concordância	R^2	Taxa de Concordância	R^2	Taxa de Concordância	R^2
C3A	3K	66,17	0,44	65,80	0,43	62,80	0,36
	6K	80,42	0,68	80,35	0,67	78,46	0,62
	9K	81,89	0,70	81,71	0,70	79,94	0,65
	20K	87,39	0,79	87,33	0,80	86,06	0,76
	50K	92,28	0,87	92,23	0,87	91,09	0,84
	80K	95,46	0,93	95,40	0,92	94,85	0,91
	90K	95,29	0,92	95,25	0,92	94,79	0,91
C3B	3K	69,24	0,48	69,06	0,4808	66,51	0,42
	6K	84,30	0,74	84,16	0.7351	82,71	0,69
	9K	85,59	0,76	85,42	0,7566	84,09	0,72
	20K	90,32	0,84	90,20	0,8381	89,24	0,81
	50K	93.84	0,90	93,82	0,8980	92,85	0,87
	80K	96,23	0,94	96,20	0,9377	95,71	0,92
	90K	96,10	0,94	96,04	0,9352	95,65	0,92

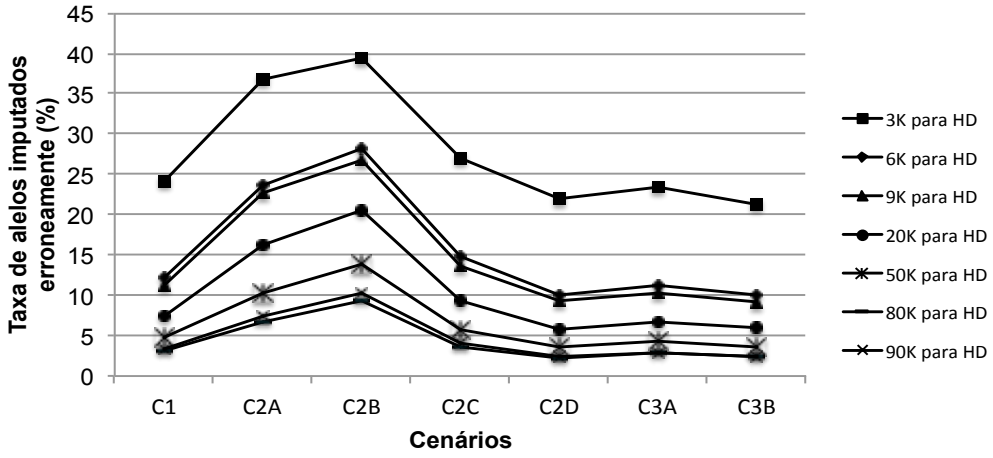
* QC1: Sem remover MAF; QC2: SNP com MAF menor que 0,0025 foram excluídos; QC3: SNP com MAF menor que 0,10 foram excluídos.

**C1: População Referência: animais nascidos até 2004; População de Imputação: nascidos em 2005; C2A: População referência: animais Canchim; População de imputação: animais MA; C2B: População referência: animais MA; População de imputação: animais Canchim; C2C: População referência: animais Canchim e MA (nascidos até 2004); População de imputação animais MA nascidos em 2005; C2D: População referência: animais MA e Canchim (nascidos até 2004); População de imputação: animais Canchim nascidos em 2005; C3A: População referência: machos; imputação fêmeas C3B: População referência: machos e fêmeas nascidas até 2004; imputação : fêmeas nascidas em 2005.

A taxa de alelos imputados erroneamente diminuiu conforme a quantidade de marcadores SNP presentes no painel LD (Figura1). O ganho médio em acurácia pela taxa de concordância do painel com maior densidade de marcadores (80K e 90K) comparado ao de 3K, este de menor densidade, para HD foi de 24% (*FImpute*) e 30% (*BEAGLE*). O R^2 alélico foi, em média, 1,82 e 2,26 vezes maior para 90K em relação ao 3K quando foi feita a imputação para HD utilizando o *FImpute* e o *BEAGLE*, respectivamente. A densidade do painel LD testado é importante fator que afeta a eficiência da imputação (HOWIE et al., 2009; HICKEY et al., 2012) quanto maior o número de SNP presentes no painel LD menor a taxa de erro da imputação. Estudos demonstraram que baixas acurácias de imputação não são satisfatórias para aplicação na seleção genômica, pois o valor genômico predito decresce conforme aumenta a taxa do erro de imputação (KHATKAR et al., 2012; MULDER et al., 2012).

Indivíduos da população de imputação com maior parentesco médio com indivíduos da população referência apresentaram maior taxa de concordância de alelos imputados (Figura 2). Para imputação do painel de 3K SNP para HD observou-se aumento linear na taxa de concordância de 0,55% ($p<0,001$) e 0,25% ($p<0,001$) para cada unidade de variação no parentesco entre os indivíduos da população referência e de imputação pelos softwares *FImpute* e *BEAGLE,* respectivamente. Entretanto, quando a imputação foi realizada do painel 80K SNP ambos softwares obtiveram 0,11% de aumento linear na taxa de concordância para cada unidade de variação no parentesco entre os indivíduos. Para imputação de painéis de baixa densidade (3K) os animais devem apresentar parentesco com os indivíduos da população referência para obtenção de melhores resultados nas análises de imputação (ZHANG; DRUET , 2010). Sugere-se que para imputação utilizando animais genotipados com painéis com poucos marcadores e aparentados com animais da população referência o algoritmo de imputação utilizado pelo *FImpute* é mais eficiente do que o algoritmo empregado pelo *BEAGLE.*

A) *FImpute software*



B) *BEAGLE software*

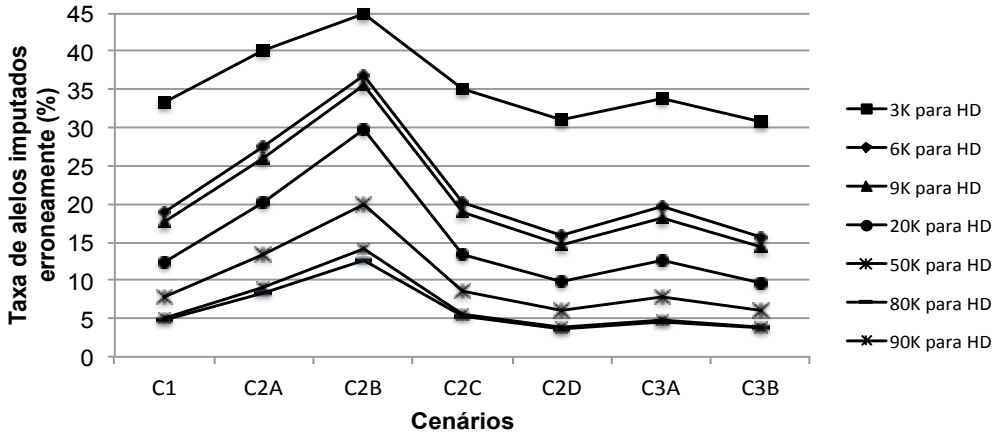


Figura1. Taxa de alelos imputados erroneamente considerando o controle de qualidade sem remover MAF (QC1), utilizando o software *FImpute* (sem o pedigree) (A) e *BEAGLE* para os cenários Cenário 1 (C1) indivíduos agrupados por ano de nascimento; cenário 2 (C2A, C2B, C2C, C2D) indivíduos agrupados por grupo genético e cenário 3 (C3A e C3B) indivíduos agrupados por sexo.

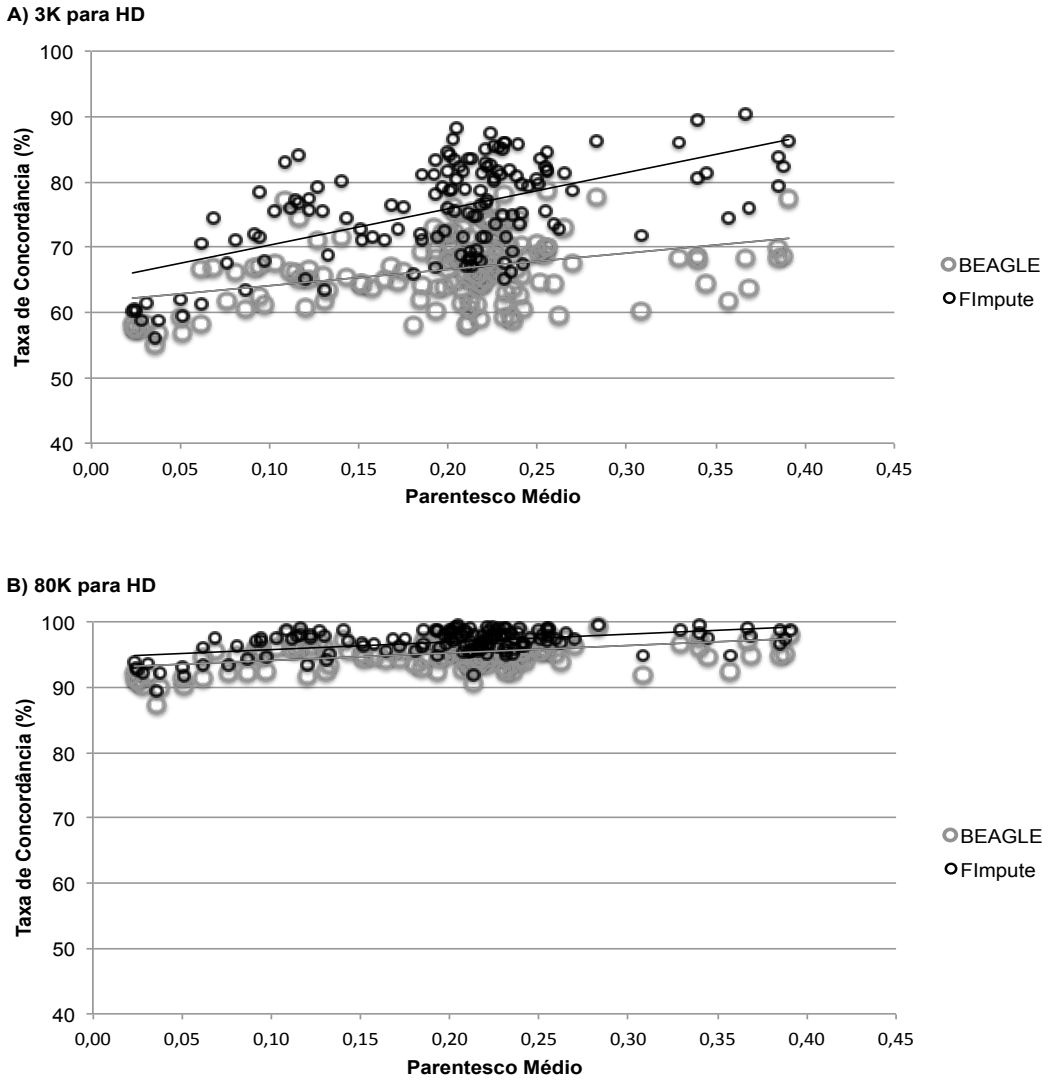


Figura 2. Parentesco médio entre os animais da população referência e imputação considerando os animais nascidos entre 1999 e 2004 na população referência e os animais nascidos em 2005 na população de imputação (C1) para os painéis 3K (A) e 80K (B).

Considerando possível aplicação deste estudo na seleção genômica para esta população, os painéis LD mais adequados para a genotipagem seriam o GGP Beef HD e GGP Indicus HD, pois estes apresentaram menor taxa de erro de imputação. Embora esses painéis tenham sido desenvolvidos respectivamente para *Bos taurus taurus* e *Bos taurus indicus*, apresentando diferentes marcadores no painel, não foram observadas diferenças na acurácia média de imputação entre esses painéis. O Canchim é uma raça composta, apresentando genes provenientes

de animais taurinos e zebuínos justificando esse resultado, além disso o número de SNP em ambos painéis são semelhantes.

No entanto, estes painéis apresentaram diferenças no erro de imputação por cromossomo (Figura 3) devido diferente distribuição de SNP por cromossomo. Para o painel de 80K SNP os cromossomos 27 e 28 apresentaram maior erro de imputação alélico, 3,9% e 3,8 % (*FImpute*) respectivamente e 5,9% em ambos cromossomos utilizando *BEAGLE,* enquanto que no painel de 90K o cromossomo 13 apresentou maior erro de imputação, 4,27% (*FImpute*) e 6,48% (*BEAGLE*). No painel de 80K SNP o cromossomo 13 apresenta maior número de SNP em relação ao painel de 90K promovendo maior informação para inferência dos haplótipos (Tabela 6). Além disso, o cromossomo 13 apresentou menor média de desequilíbrio de ligação (r^2) entre os marcadores adjacentes (Apêndice 2). Segundo Sun et al. (2012) a diferença de imputação por cromossomo ocorre devido ao tamanho do cromossomo, pois imputar alelos corretamente é mais difícil na região inicial e final do cromossomo, consequentemente cromossomos mais curtos são mais afetados do que cromossomos longos ocasionando menor taxa de alelos imputados erroneamente, além disso cromossomos mais longos possuem maior número de marcadores do que cromossomos mais curtos.

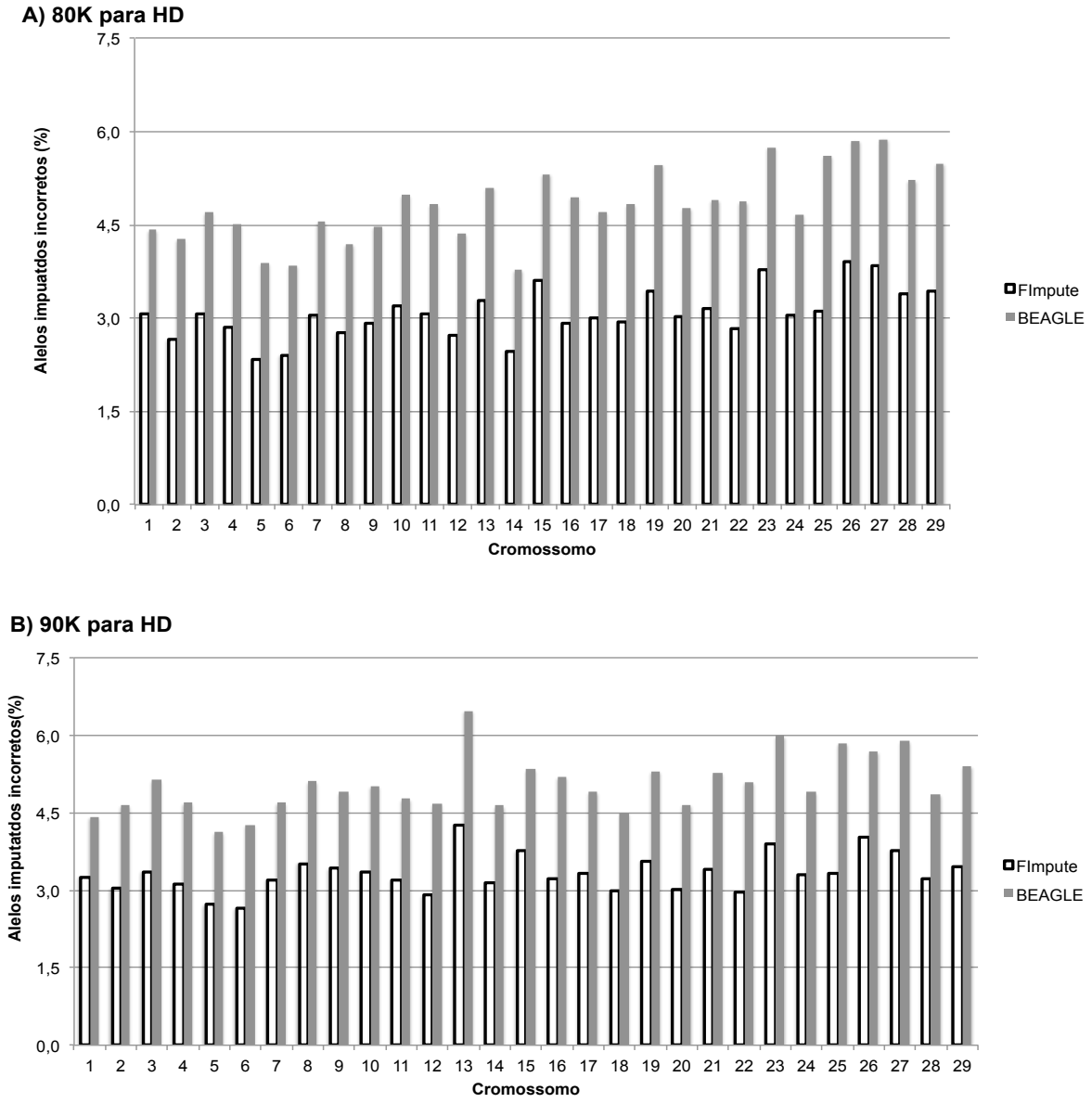


Figura 3. Taxa de alelos imputados erroneamente por cromossomo pelos softwares *FIMPUTE* e *BEAGLE* considerando o cenário agrupado por ano de nascimento (C1), sem remover MAF (QC1) para os painéis 80K para HD (A) e 90K para HD (B).

Tabela 6. Tamanho do cromossomo em mega pares de base (MB), número de SNP e taxa de alelos imputados corretamente (TC) e por cromossomo pelos *softwares FImpute* (sem o pedigree) e *BEAGLE* para os painéis imputados de 80K SNP para HD e 90K SNP para HD sem remover MAF (QC1).

Cromossomo	Tamanho (MB)	Painel HD	Painel 80K			Painel 90K		
		N SNP	N SNP	TC FImpute	TC BEAGLE	N SNP	TC FImpute	TC BEAGLE
BTA1	158,24	39363	4189	96,85	95,58	3050	96,74	96,74
BTA 2	136,67	33394	3557	97,28	95,73	2457	95,73	96,97
BTA 3	121,39	30200	3225	96,87	95,30	2386	95,97	96,64
BTA 4	120,62	29745	3121	97,01	95,49	2269	96,09	96,89
BTA 5	121,15	29085	3144	97,55	96,11	2239	96,22	97,28
BTA 6	119,41	30729	3149	97,45	96,15	2363	96,24	97,35
BTA 7	112,60	27791	2938	96,86	95,45	2172	96,43	96,79
BTA 8	113,35	24359	2956	97,12	95,81	1782	96,48	96,48
BTA 9	105,66	26080	2795	96,91	95,52	1991	96,55	96,58
BTA 10	104,28	26591	2732	96,73	95,02	2108	96,58	96,64
BTA 11	107,24	28023	2800	96,78	95,17	2156	96,59	96,80
BTA 12	91,10	21984	2387	97,13	95,64	1682	96,64	97,10
BTA 13	84,11	17640	2216	96,59	94,90	1274	96,64	95,73
BTA 14	83,36	18677	2281	97,41	96,22	1429	96,67	96,85
BTA 15	85,23	21121	2289	96,34	94,68	1805	96,67	96,24
BTA 16	81,64	20528	2211	96,94	95,05	1661	96,71	96,77
BTA 17	75,13	19454	2031	96,85	95,29	1566	96,77	96,67
BTA 18	65,84	17111	1860	96,86	95,17	1537	96,78	97,02
BTA 19	63,95	16655	1820	96,40	94,53	1491	96,79	96,43
BTA 20	71,93	18991	2043	96,72	95,23	1610	96,80	96,98
BTA 21	71,55	17768	1976	96,68	95,10	1378	96,85	96,59
BTA 22	61,23	16201	1737	96,99	95,13	1436	96,89	97,03
BTA 23	52,46	13453	1577	96,04	94,25	1298	96,97	96,09
BTA 24	62,54	15755	1779	96,90	95,33	1289	96,98	96,71
BTA 25	42,77	11420	1274	96,77	94,39	1029	97,02	96,67
BTA 26	51,55	13562	1474	95,85	94,15	1205	97,03	95,97
BTA 27	45,40	11809	1312	96,03	94,12	1102	97,10	96,22
BTA 28	46,24	11574	1332	96,51	94,78	1138	97,28	96,78
BTA 29	51,48	12774	1396	96,54	94,51	1159	97,35	96,55

3.2 Diferentes cenários de população referência e imputação

Visando futura aplicação de imputação em estudos genômicos, sugere-se o cenário 1 (C1) como mais apropriado, em que os animais mais velhos sejam considerados na população referência e os mais jovens na de imputação. Resultados satisfatórios foram observados para este cenário. Assim, novos

candidatos à seleção poderão ser genotipados com painéis de baixa densidade, possibilitando o aumento do número de informações genotípicas na população para posteriores avaliações genéticas na raça Canchim utilizando dados genômicos.

Para os cenários propostos, o mais eficiente em termos de acurácia, com menor erro de imputação foi o C3B (referência: machos e fêmeas; imputação: fêmeas jovens). Este resultado justifica-se pelo número de animais na população referência, quanto maior o número de indivíduos na população referência maior a acurácia de imputação (HOWIE; DONNELLY; MARCHINI, 2009; KHATKAR et al., 2012; PAUSCH et al., 2013).

A divisão por sexo nos grupos (C3A e C3B) permitiu verificar que os genótipos das fêmeas podem ser imputados utilizando somente machos na população referência, pois a acurácia de imputação entre estes cenários não apresentou diferenças nesta população. Recentemente, painéis de baixa densidade tem sido utilizados para genotipagem de fêmeas (VANRADEN et al., 2013). Esta estratégia poderá ser uma ferramenta indicada para seleção de fêmeas em larga-escala.

Quando a população referência foi composta apenas por animais Canchim ou MA, os resultados obtidos foram insatisfatórios para aplicação da imputação. O parentesco médio observado entre a população referência e de imputação (Apêndice 1) nos cenários C2A e C2B foi relativamente baixo (0,005 e 0,003, respectivamente), afetando a acurácia de imputação. Hozé et al. (2013) relataram que quanto maior o parentesco entre indivíduos da população referência e de imputação menor a taxa de alelos imputados incorretos, sugerindo que a relação de parentesco entre esses indivíduos deve ser sempre mantida para análises de imputação. Assim, sugere-se que os dois grupos sejam considerados juntos, pois isto possibilita o aumento do parentesco médio e da população referência, além de auxiliar na construção dos haplótipos necessários para a imputação, uma vez que Buzanskas et al. (2013), estudando esta mesma população, relataram que os dois grupos estão na mesma fase de ligação, recomendando que avaliações para esta população devem ser multirraciais.

3.3 *FImpute* versus *BEAGLE*

O *FImpute* demonstrou melhor desempenho para a imputação, principalmente em painéis de baixa densidade e a diferença do ganho em acurácia foi menor no painéis de 50K, 80K e 90K para HD (Figura 4). O algoritmo utilizado pelo *FImpute* identifica segmentos cromossômicos longos de IBD pela sobreposição das janelas, facilitando identificação dos haplótipos nos painéis com poucos marcadores, reduzindo o erro de imputação. O tempo de execução para realização da imputação pelo *software FImpute* foi de 20 a 100 vezes menor em relação ao *software BEAGLE* (Tabela 7). O software *BEAGLE* foi desenvolvido para população humana, em que algoritmos mais complexos são necessários para construção dos haplótipos, necessitando de maior demanda computacional. Este fator é muito importante, devido ao aumento do número de animais a serem genotipados nas populações de bovinos sob avaliação genética.

Comparando o *software BEAGLE* com o *FImpute* (versão2), Ventura et al. (2012) observaram aumento médio na acurácia de imputação de 2,21%, quando utilizaram o *FImpute* e também redução de 10 vezes no tempo de execução do *software* comparado ao *BEAGLE.* Embora a imputação tenha grandes vantagens, são necessários recursos computacionais de grande escala, além da necessidade de avaliar a qualidade da imputação. Segundo Ma et al. (2013), o *FImpute* poderá ser o mais indicado entre os programas computacionais para imputação devido a exatidão na acurácia de imputação e a eficiência para atender a demanda computacional.

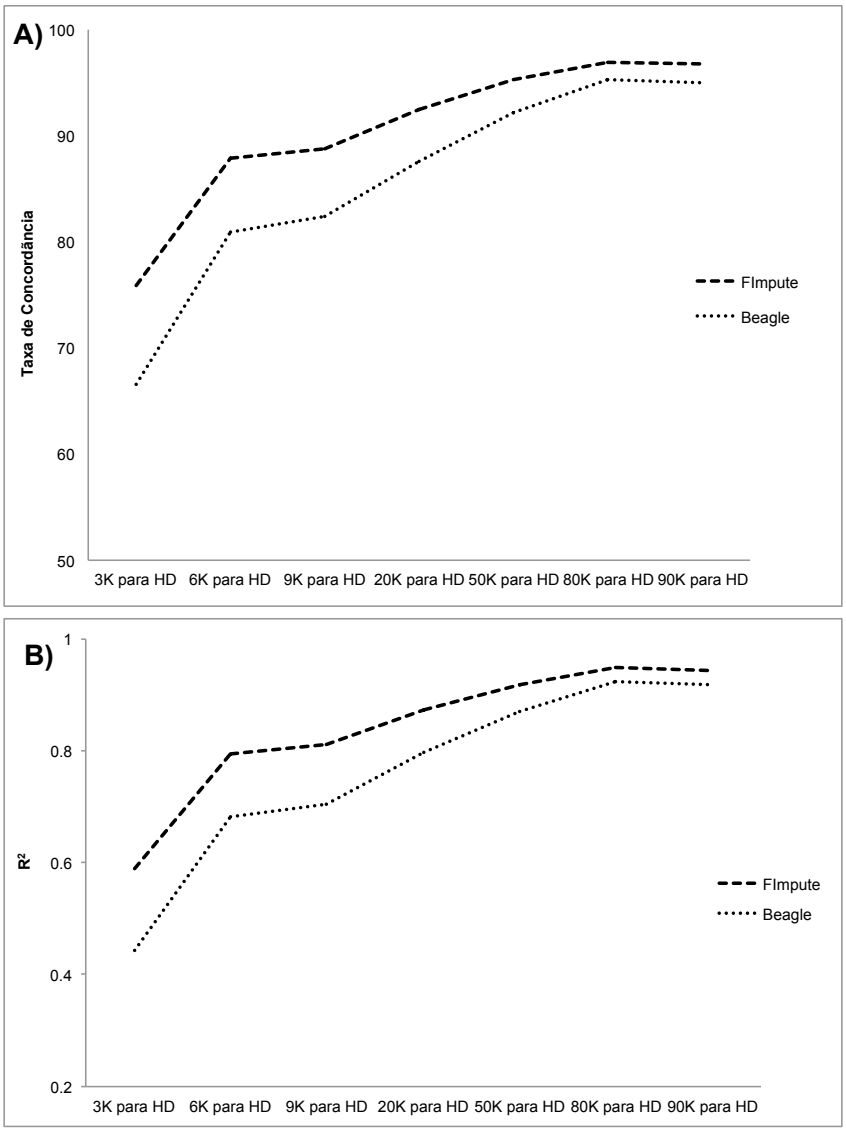


Figura 4. Taxa de Concordância (A) e R^2 alélico (B) utilizando os *softwares* *FImpute* (sem o pedigree) e *BEAGLE*, sem empregar o critério de MAF (QC1) no cenário em que as populações foram divididas pela data de nascimento (C1).

Tabela 7. Tempo (horas:minutos:segundos) geral de imputação (TI) de painéis de baixa densidade (LD) para alta densidade (HD) para os *softwares* *FImpute* e *BEAGLE*.

Cenários	Tamanho do Painel LD	TI- *FImpute*	TI- *BEAGLE*
C1	3K	00:03:45	10:00:00
	6K	00:04:10	07:27:23
	9K	00:04:16	07:04:33
	20K	00:04:34	04:55:32
	50K	00:05:08	03:12:35
	80K	00:05:36	02:24:48
	90K	00:05:23	02:24:48
C2A	3K	00:04:01	08:28:34
	6K	00:04:22	06:24:36
	9K	00:04:27	06:00:43
	20K	00:04:43	04:08:04
	50K	00:05:19	02:40:00
	80K	00:06:00	02:00:20
	90K	00:05:30	02:03:59
C2B	3K	00:02:11	08:19:23
	6K	00:02:31	07:20:07
	9K	00:02:36	06:11:13
	20K	00:02:53	05:31:25
	50K	00:03:47	03:42:42
	80K	00:04:26	03:20:35
	90K	00:04:02	02:53:52
C2C	3K	00:04:41	05:21:36
	6K	00:05:06	03:52:15
	9K	00:06:08	03:14:52
	20K	00:05:28	02:32:57
	50K	00:05:51	02:16:13
	80K	00:06:12	01:34:37
	90K	00:06:06	01:39:15
C2D	3K	00:04:22	08:36:28
	6K	00:04:39	06:15:00
	9K	00:04:41	05:56:21
	20K	00:05:08	04:30:45
	50K	00:05:32	03:12:29
	80K	00:06:03	01:56:57
	90K	00:05:50	02:37:34
C3A	3K	00:03:20	11:55:58
	6K	00:03:37	09:20:50
	9K	00:03:47	08:35:04
	20K	00:04:36	06:09:46
	50K	00:04:50	03:54:38
	80K	00:05:53	02:40:03
	90K	00:05:47	02:59:13
C3B	3K	00:05:01	07:30:27
	6K	00:05:33	05:24:50
	9K	00:05:19	06:15:24
	20K	00:06:12	03:38:16
	50K	00:06:35	02:27:22
	80K	00:07:02	01:50:22
	90K	00:06:38	02:00:27

**C1: População Referência: animais nascidos até 2004; População de Imputação: nascidos em 2005; C2A: População referência: animais Canchim; População de imputação: animais MA; C2B: População referência: animais MA; População de imputação: animais Canchim; C2C: População referência: animais Canchim e MA (nascidos até 2004); População de imputação animais MA nascidos em 2005; C2D: População referência: animais MA e Canchim (nascidos até 2004); População de imputação: animais Canchim nascidos em 2005; C3A: População referência: machos; imputação fêmeas C3B: População referência: machos e fêmeas nascidas até 2004; imputação : fêmeas nascidas em 2005.

4 CONCLUSÃO

Para análises de imputação a aplicação de restrição de MAF no controle de qualidade dos dados avaliados neste estudo não é necessária. Os painéis comerciais GeneSeek Genomic Profiler Beef HD (80K) e GeneSeek Genomic Profiler Indicus HD (90K) podem ser acuradamente imputados para o painel HD na raça Canchim. Animais Canchim e do grupo genético MA devem ser considerados juntos na população referência pois aumenta o número de informação utilizada na imputação, facilitando a construção dos haplótipos. O algoritmo utilizado pelo *software FImpute* demonstrou mais eficiente na imputação dos marcadores principalmente em painéis de menores densidade. Para validação dos resultados e possível genotipagem de animais jovens ou fêmeas da raça Canchim com esses painéis LD, visando redução de custos e aumento de informações genotípicas, futuras análises com maior número de animais na população devem ser realizadas.

5 REFERÊNCIAS

ALENCAR, M. M. Bovino – Raça Canchim: origem e desenvolvimento. Documento, 4. Brasília, EMBRAPA – DPU, 1988, 102p.

ANDRADE, P. C.; GROSSI, D. A; PAZ, C. C. P.; ALENCAR, M. M.;REGITANO, L. C. A; MUNARI, D. P. Association of an insulin-like growth factor1 gene microsatellite with phenotypic variation and estimated breeding values of growth traits in Canchim cattle. Animal Genetics, v. 39, n. 5, p. 480–485, 2008.

BOICHARD, D.; CHUNG, H.; DASSONNEVILLE, R.; DAVID, XEGGEN, A.; FRITZ, S.; GIETZEN, K. J.; HAYES, B. J.; LAWLEY, C. T.; SONSTEGARD, T. S.; VAN TASSELL, C. P.; VANRADEN, P. M.; VIAUD-MARTINEZ, K. A.; WIGGANS, G. R. Design of a bovine low-density SNP array optimized for imputation. **PLoS ONE,** v.7, e34130, 2012.

BROWNING, S.R., BROWNING, B.L. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. **American Journal of Human Genetics,** v.84*,* p.210-223, 2009.

BUZANSKAS, M. E. ; GROSSI, D. A. ; SCHENKEL, F.S. ; VENTURA, R. V. ; REGITANO, L.C.A. ; ALENCAR, M. M. ; MUNARI, D. P. Linkage disequilibrium analysis in Canchim beef cattle. In: 50ª Reunião Anual da Sociedade Brasileira de Zootecnia, 2013, Campinas/SP. 50ª Reunião Anual da Sociedade Brasileira de Zootecnia, 2013.

DASSONNEVILLE, R.; BRØNDUM, R. F.; DRUET, T.; FRITZ, S.; GUILLAUME, F.; GULDBRANDTSEN, B.; LUND, M. S.; DUCROCQ, V.; SU, G. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. **Journal of Dairy Science**, v. 94, p.3679–3686, 2011.

HAYES, B. J.; BOWMAN, P. J.; DAETWYLER, H. D.; KIJAS, J. W.; VAN DER WERF, J. H. J. Accuracy of genotype imputation in sheep breeds. **Animal Genetics,** v 43, p. 72-80, 2011.

HICKEY, J. M; CROSSA, J.; BABU, R.; DE LOS CAMPOS, G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. **Crop Science**, v.52, p.654–663, 2012.

HILL, W. G. G.; ROBERTSON, A. Linkage Disequilibrium in Finite Populations. **Theoretical and Applied Genetics**, v. 38, p. 226–231, 1968.

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:e1000529, 2009.

HOZÉ, C.; FOUILLOUX, M.; VENOT, E.; GUILLAUME, F.; DASSONNEVILLE, R.; FRITZ, S.; DUCROCQ, V; PHOCAS, F.; BOICHARD, D; CROISEAU, P. High-density marker imputation accuracy in sixteen French cattle breeds. **Genetics Selection Evolution**, v.45, p.33, 2013.

JOHNSTON, J., KISTEMAKER, G. SULLIVAN, P.G. Comparison of different imputation methods**. Interbull Open Meeting**, Stanvager, Norway, 2011.

KHATKAR, M. S.; MOSER, G.; HAYES, B. J.; RAADSMA, W. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle, **BMC Genomics**, v.13, p.538, 2012.

LARMER, L., SARGOLZAEI, M., VENTURA, R., SCHENKEL, F. Imputation accuracy from low to high density using within and across breed reference populations in Holstein, Guernsey and Ayrshire cattle. Technical report to the Dairy Cattle Breeding and Genetics Committee on February 28, 2012. University of Guelph, Guelph, ON, Canada, 2012.

LI ,Y.; WILLER, C. J.; SANNA, S.; ABECASIS, G. R. Genotype Imputation. **Annu. Rev. Genomics Human Genetics**, v.10, p.387-406, 2009.

MA, P.;BRØDUM, R. F.; ZHANG, Q.; LUND, M. S.; SU, G. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle, **Journal of Dairy Science,** v. 96, p.4666-4677, 2013.

MARCHINI J; HOWIE B. Genotype imputation for genome-wide association studies. **Nature Review Genetics**, v.11, p.499-511, 2010.

MEUWISSEN, T. H. E.; GODDARD, M. E.; HAYES, B. J. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2001.

MEUWISSEN, T. H. E. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. **Genetics Selection Evolution**, v.41, p.35, 2009.

MULDER, H. A.;CALUS, M. P. L.; DRUET, T.; SCHROOTENT, C. Imputation of genotypes with low –density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science, v.95, p.876-889, 2012.

PAUSCH, H.; BERNHARD, A.; EMMERLING, R.; EDEL, C.; GOTZ, K.; FRIES, R. Imputation of high-density genotypes in the Fleckvieh cattle population. **Genetics Selection Evolution**,v.45, p.3., 2013.

PEREIRA, A. P.; ALENCAR, M. M.; OLIVEIRA, H. N.; REGITANO, L. C. A. Association of GH and IGF-1 polymorphisms with growth traits in a synthetic beef cattle breed. Genetics and Molecular Biology, v. 28, n. 2, p. 230–236, 2005.

SARGOLZAEI, M.; IWAISAKI H.; COLLEAU, J. J. CFC: A tool for monitoring genetic diversity. In: World Congress on Genetics Applied to Livestock Production, 8., 2006, Belo Horizonte. **Proceedings...** Belo Horizonte, 2006. 1 CD-ROM.

SARGOLZAEI, M.; SCHENKEL, F. S.; JANSEN, G. B.; SCHAEFFER, L. R. Extent of linkage disequilibrium in Holstein cattle in North America. **Journal of Dairy Science**, v. 91, n. 5, p. 2106–2017, 2008.

SARGOLZAEI, M., CHESNAIS, J.P., SCHENKEL, F.S. FImpute. An efficient imputation algorithm for dairy cattle populations. **Journal of Animal Science**, v. 89(E-Suppl. 1)/**Journal of Dairy Science**, v. 94(E-Suppl. 1), p. 421 (abstr. 333) ,2011.

VANRADEN, P. M. Efficient methods to Compute Genomic predictions. **Journal of Dairy Science**, v 91, p.4414–4423, 2008.

VANRADEN, P. M.; NULL, D. J., SARGOLZAEI, M., WIGGANS, G. R.; TOOKER, M. E.; COLE, J. B.; SONSTEGARD, T. S.; CONOOR, E. E.; WINTERS, M.; VAN KAAM, J. B. C. H. M.; VALENTINI, A.; VAN DOORMAAL, B. J. II.; FAUST, M. A.; DOAK, G. A. Genomic imputation and evaluation using high-density Holstein genotypes. **Journal of dairy science**, v.96, p.668-678, 2013.

VENTURA, R.V., SCHENKEL, F.S. ,WANG, Z. MILLER, S.P. Accuracy of imputation using 6K and 50K SNP chips in beef cattle . Livestock Gentec's 2nd Annual Conference , Edmonton, Alberta, Canada, 2011.

VENTURA, R.V., SCHENKEL, F.S, SARGOLZAEI, M., WANG, Z., CHANGXI, L., MILLER, S.P. Accuracy of imputation using 6K and 50K SNP chips in multi-breed and crossbred beef cattle populations . In: Proceeding of the 33rd ISAG Conference, July 15-20, Cairns, Australia, 2012.

ZHANG, Z., DRUET, T. Marker imputation with low-density marker panels in Dutch Holstein cattle. **Journal of Dairy Science**, v. 93, n.11, p. 5487-5494, 2010.

ZIMIN, A. V; DELCHER, A. L.; FLOREA, L.; KELLEY, D. R.; SCHATZ, M. C.; PUIU, D.; HANRAHAN, F.; PERTEA, G.; VAN TASSELL, C. P.; SONSTEGARD, T. S.; MARÇAIS, G.; ROBERTS, M.; SUBRAMANIAN, P.;YORKE, J. A; SALZBERG, S. L. A whole-genome assembly of the domestic cow, Bos taurus. **Genome Biology**, v. 10, n. 4, p. r42, 2009.

Apêndice 1. Parentesco máximo, mínimo e médio entre população de imputação e referência para os diferentes cenários.

Cenários	Mínimo	Médio	Máximo
C1	0,023	0,198	0,390
C2A	0,010	0,050	0,220
C2B	0,003	0,003	0,225
C2C	0,028	0,193	0,330
C2D	0,050	0,198	0,390
C3A	0,090	0,210	0,409
C3B	0,108	0,228	0,390

**C1: População Referência: animais nascidos até 2004; População de Imputação: nascidos em 2005; C2A: População referência: animais Canchim; População de imputação: animais MA; C2B: População referência: animais MA; População de imputação: animais Canchim; C2C: População referência: animais Canchim e MA (nascidos até 2004); População de imputação animais MA nascidos em 2005; C2D: População referência: animais MA e Canchim (nascidos até 2004); População de imputação: animais Canchim nascidos em 2005; C3A: População referência: machos; imputação fêmeas C3B: População referência: machos e fêmeas nascidas até 2004; imputação : fêmeas nascidas em 2005.

Apêndice 2. Desequilíbrio de ligação médio (r^2) por cromossomo entre os marcadores adjacentes do painel de alta densidade considerando o arquivo de dados sem remover os alelos de menor frequência - MAF (QC1).

Cromossomo	Distância média (MB*)	r2 médio	Número de marcadores
1	4.043,72	0,40	39.121
2	4.121,77	0,40	33.155
3	4.046.88	0,40	29.995
4	4.080,57	0,41	29.556
5	4.204,40	0,43	28.815
6	3.870,82	0,42	30.552
7	4.037,92	0,41	27.617
8	4.680,89	0,38	24.207
9	4.072,88	0,42	25.942
10	3.933,11	0,39	26.508
11	3.852,74	0,41	27.833
12	4.180,11	0,40	21.791
13	4.803,52	0,36	17.510
14	4.498,18	0,39	18.531
15	4.056,76	0.39	21.009
16	3.965,15	0,42	20.394
17	3.852,67	0,41	19.370
18	3.877,15	0,41	16.982
19	3.857,93	0,38	16.577
20	3.805,86	0,40	18.898
21	4.053,16	0,40	17.653
22	3.797,84	0,41	16.123
23	3.914,24	0,37	13.402
24	3.999,31	0,41	15.638
25	3.770,64	0,37	11.344
26	3.818,05	0,39	13.501
27	3.879,77	0,38	11.729
28	4.008,62	0,39	11.536
29	4.030,92	0,38	12.694

*Mega pares de base.