

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CAMPUS JABOTICABAL**

**IDENTIFICAÇÃO DE REGIÕES CROMOSSÔMICAS, GENES
E POLIMORFISMOS DE DNA ASSOCIADOS AO
DESEMPENHO DE EQUINOS DE CORRIDA DA RAÇA
QUARTO DE MILHA**

Guilherme Luis Pereira

Zootecnista

2017

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CAMPUS JABOTICABAL**

**IDENTIFICAÇÃO DE REGIÕES CROMOSSÔMICAS, GENES
E POLIMORFISMOS DE DNA ASSOCIADOS AO
DESEMPENHO DE EQUINOS DE CORRIDA DA RAÇA
QUARTO DE MILHA**

Guilherme Luis Pereira

Orientador: Prof. Dr. Rogério Abdallah Curi

Co-orientadora: Luciana Correia de Almeida Regitano

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

P436i Pereira, Guilherme Luis
Identificações de regiões cromossômicas, genes e polimorfismos de DNA associados ao desempenho de equinos de corrida da raça Quarto de Milha / Guilherme Luis Pereira. -- Jaboticabal, 2017
xix, 98 p. : il. ; 29 cm

Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2017
Orientador: Rogério Abdallah Curi
Banca examinadora: Julio Cesar de Carvalho Balieiro, Guilherme Costa Venturini, Fernando Sebastian Baldi Rey, Guilherme de Camargo Ferraz
Bibliografia

1. GWAS. 2. Exoma. 3. Índice de Velocidade. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.082:636.1

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

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: IDENTIFICAÇÃO DE REGIÕES CROMOSSÔMICAS, GENES E
POLIMORFISMOS DE DNA ASSOCIADOS AO DESEMPENHO DE EQUINOS
DE CORRIDA DA RAÇA QUARTO DE MILHA

AUTOR: GUILHERME LUÍS PEREIRA

ORIENTADOR: ROGERIO ABDALLAH CURI

COORIENTADORA: LUCIANA CORREIA DE ALMEIDA REGITANO

Aprovado como parte das exigências para obtenção do Título de Doutor em GENÉTICA E
MELHORAMENTO ANIMAL, pela Comissão Examinadora:

Prof. Dr. ROGERIO ABDALLAH CURI
Departamento de Melhoramento e Nutrição Animal / FMVZ / UNESP - Botucatu

Prof. Dr. JULIO CESAR DE CARVALHO BALIEIRO
Departamento de Ciências Básicas / FMVZ / USP / Pirassununga/SP

Prof. Dr. GUILHERME COSTA VENTURINI
Departamento de Melhoramento e Nutrição Animal / FMVZ / UNESP - Botucatu

Prof. Dr. FERNANDO SEBASTIAN BALDI REY
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Prof. Dr. GUILHERME DE CAMARGO FERRAZ
Departamento de Morfologia e Fisiologia Animal / FCAV / UNESP - Jaboticabal

Jaboticabal, 28 de abril de 2017

DADOS CURRICULARES DO AUTOR

Guilherme Luis Pereira, Itaporanga, 19 de fevereiro de 1986. Possui graduação em zootecnia pela Universidade Estadual de Ponta Grossa (2012). Possui Mestrado em Genética e Melhoramento Animal pela Universidade Estadual Paulista Júlio de Mesquita Filho Campus Jaboticabal (FCAV). Atualmente é Doutorando em genética e Melhoramento Animal pela Universidade Estadual Paulista Júlio de Mesquita Filho Campus Jaboticabal (FCAV). Tem experiência na área de Genética, com ênfase em Genética Animal, atuando principalmente nos seguintes temas: Genética Molecular, Melhoramento Genético e Melhoramento de Equinos.

*“Dificuldades preparam pessoas comuns para
destinos extraordinários”.*

C. S. Lewis

Aos meus pais, José e Lucia.

Ao meu querido irmão Eduardo.

À minha noiva Ariane.

Ao saudoso professor Marcilio Dias Silveira da Mota.

DEDICO

AGRADECIMENTOS

Agradeço à Deus, pela sua bondade em permitir que essa etapa fosse concluída, por estar comigo todos os dias e pelo Seu cuidado.

Aos meus pais José e Lúcia e ao meu irmão Eduardo, os quais foram meus maiores incentivadores e apoio nessa etapa.

Ao meu tio Cesar e aos meus avós Fíica e Dito Veiga, por suas preocupações, incentivo e apoio.

Agradeço à minha noiva e amor Ariane, pela paciência e incentivo nas fases mais difíceis do projeto.

Ao meu orientador (e amigo) Prof. Rogério Curi, em qual trago profundo respeito, admiração e muita gratidão; sempre esteve por perto, se preocupando, ajudando e incentivando, enfim, sou grato por todo tempo que passamos juntos.

Aos professores Dr. Marcelo Vicari e Dra. Luciana Regitano, (e todo o pessoal de seus laboratórios) por crerem que eu poderia chegar aonde cheguei e pelo esforço de me ajudarem a começar esse sonho, desde a minha graduação, sou muito grato.

Ao professor Dr. Henrique Oliveira pela tão grande ajuda e direcionamento dos trabalhos.

A todos do departamento de Melhoramento e Nutrição Animal (FMVZ – Botucatu), os professores e funcionários que me acolheram, incentivaram e me ajudaram em todo esse tempo.

Aos colegas de pós graduação e laboratório que me ajudaram e estiveram comigo, em especial, Rafael Matteis, Cintia Marchiori, Adriana Somavilla, Tatiane Chud.

A CAPES pela concessão da bolsa e a FAPESP pelo financiamento do projeto de pesquisa.

Aos membros da banca examinadora, Dr. Fernando Baldi, Dr. Julio Balieiro, Dr. Guilherme Ferraz e Dr. Guilherme Venturini.

Aos meus amigos que me acolheram em Botucatu e se tornaram meus irmãos, Antone, Ananda, Barbara, Camila, Gabi, Jéssica, Rafinha, Mari, Sabrina, Leandro, Estefânia, Rebeca, Felipe, Manu, Ivânio, Guilherme Vicentini, Natália, David e Mateus.

SUMÁRIO

CAPÍTULO 1 – Considerações Gerais	1
1. INTRODUÇÃO	1
2. OBJETIVOS GERAIS	3
2.1. Objetivos específicos	3
3. REVISÃO DE LITERATURA	3
3.1. Os equinos no Brasil e no mundo	3
3.2. A raça Quarto de Milha	5
3.3. Melhoramento genético de equinos	7
3.4. Características quantitativas e <i>major genes</i>	8
3.5. Polimorfismos de nucleotídeos únicos (SNPs)	9
3.6. Painéis de SNPs e imputação de genótipos	10
3.9. Estudos associação ampla do genoma (GWAS)	12
3.10. Marcadores para desempenho em corridas de equino	13
3.11. Sequenciamento de DNA	14
4. REFERÊNCIAS	16
CAPITULO 2 – Imputação e avaliação da acurácia em cavalos de corrida da raça Quarto de Milha genotipados com diferentes painéis comerciais de SNPs	23
1. Introdução	25
2. Material e Métodos	26
2.1. Animais	26
2.2. Genotipagem dos SNPs e controle de qualidade	26
2.3. Análise populacional	27
2.4. Imputação de genótipos e avaliação da acurácia de imputação	28
3. Resultados e discussão	29
3.1. Estrutura da população e análise de parentesco	29
3.2. Acurácia de imputação por SNP	30
3.3. Acurácia de imputação por amostra	33
3.4. Acurácia de imputação por cromossomo	36
4. Conclusões	38
5. Referências	39

CAPITULO 3 – Estudo de associação ampla genoma em equinos de corrida da raça Quarto de Milha	42
1. Introdução	44
2. Material e Métodos	45
2.1. Animais, coleta de sangue e extração do DNA	45
2.2. Controle de qualidade de genotipagem e imputação de genótipos.....	46
2.3. Informações fenotípicas	48
2.4. Estudo de associação ampla do genoma	49
2.5. Anotação de genes e análise funcional	50
3. Resultados e Discussão	51
3.1. Dados utilizados e estratificação da população	51
3.2. Associação ampla do genoma	53
3.3. Anotação de genes e análise funcional	59
4. Conclusões	66
5. Referências	67
CAPITULO 4 – Sequenciamento de exomas de equinos em regiões alvo relacionadas ao desempenho em corridas na raça Quarto de Milha	73
1. Introdução	75
2. Material e métodos	76
2.1. Amostras, genotipagem e obtenção de GEBVs	76
2.2. Seleção de amostras com GEBVs contrastantes para desempenho em corridas	77
2.3. Preparo das bibliotecas, captura e sequenciamento dos exomas equino	78
2.4. Alinhamento das <i>reads</i> ao genoma equino de referência	79
2.5. Detecção e filtragem de polimorfismos	80
2.6. Prospecção e anotação dos polimorfismos e genes de interesse	81
3. Resultados e discussão	82
3.1. Amostras sequenciadas	82
3.2. Sequenciamento e alinhamento	84
3.3. Detecção e filtragem de SNPs e InDels	86
3.4. Anotação funcional de variantes genômicas	88
3.5. Genes prospectados	92

4. Conclusões	94
5. Referências	95
Capítulo 5 – Considerações finais	98

IDENTIFICAÇÃO DE REGIÕES CROMOSSÔMICAS, GENES E POLIMORFISMOS DE DNA ASSOCIADOS AO DESEMPENHO DE EQUINOS DE CORRIDA DA RAÇA QUARTO DE MILHA

RESUMO – Dentre os equinos selecionados para velocidade, a linhagem de corrida da raça Quarto de Milha se destaca pelo alto desempenho em provas de curtas distâncias, sendo considerados os mais velozes do mundo. Apesar de, no Brasil, o efetivo de animais ser relativamente menor na linhagem de corrida do que nas demais, sua importância econômica é substancial. Tendo em vista o interesse econômico e científico relacionado a esta característica atlética, poucos esforços têm sido realizados para a maior compreensão de seus mecanismos genéticos e fisiológicos. Este trabalho teve como objetivos: 1) realizar a imputação de genótipos em duas vias entre indivíduos de uma amostra populacional relativamente pequena de cavalos de corrida da raça Quarto de Milha genotipados com painéis de 54k ou de 65k, bem como avaliar a acurácia de imputação por meio de simulações; 2) realizar estudo de associação ampla do genoma (GWAS) em cavalos da linhagem de corrida da raça Quarto de Milha por meio da utilização de chips equinos de genotipagem de SNPs, visando a prospecção de regiões cromossômicas, genes e polimorfismos relacionados ao desempenho; 3) analisar exomas de equinos de corrida da raça Quarto de Milha contrastantes para Índice de Velocidade máximo (IV max) em regiões previamente associadas à característica por meio de GWAS, visando a prospecção de polimorfismos gênicos causais, ligados ou em forte desequilíbrio de ligação com o desempenho em corridas. A imputação foi realizada utilizando 116 cavalos genotipados com o arranjo de SNPs de 54k e 233 genotipados com arranjo de 65k. Nas simulações foram escolhidas amostras aleatórias para constituírem as populações imputadas e referências em dois cenários. O cenário A simulou a imputação genótipos na primeira via (65k para 54k) e o cenário B na segunda (54k para 65k). No cenário A foram considerados 113 indivíduos para a população referência e 236 para a imputada, dos quais 116 e 120 foram genotipados com os arranjos de 54k e 65k, respectivamente. No cenário B foram considerados 50 indivíduos para a população referência e 299 para a imputada, dos quais 66 e 233 foram genotipados com os arranjos de 54k e 65k, respectivamente. Com isso, após o controle de qualidade, os painéis de 54k e de 65k contaram com 7.048 e 16.940 marcadores exclusivos, respectivamente. As médias de taxa de concordância para os cenários A e B foram 0,9815 e 0,9751 e para r^2 alélico foram 0,9791 e 0,9740, respectivamente. O GWAS foi realizado com base no método *single step* GBLUP por meio de duas abordagens: ssGWAS1, em que somente efeitos de SNPs são reestimados a cada iteração, e ssGWAS2, em que a cada iteração são reestimados efeitos de SNPs a partir de valores genético genômico (GEBVs) reestimados. Vinte e uma regiões foram encontradas explicando mais que 1% da variância genética total (gVar) da característica índice de velocidade máximo (IV max) para ssGWAS1 e doze para

ssGWAS2. No total mais de 40% da gVar foi explicada por estas regiões para ssGWAS1 e cerca de 30% para ssGWAS2. Entre os cromossomos que explicaram mais de 1% da variância genética, cinco foram comuns aos dois métodos (ECA 3, 10, 15, 22, 25). Foram identificados 108 genes na primeira abordagem e 59 na segunda. A partir de informações de GEBVs de cada cavalo foram formados dois grupos de animais contrastantes para desempenho em corridas (20 animais de IV max superior e 20 IV max inferior), para ser sequenciados. Foram observadas leituras de boa qualidade para toda extensão das *reads* sequenciadas (até 100pb) e cobertura média de 43x. Foram identificadas 1.203 variantes (1.105 SNPs e 93 InDels) em 33 regiões de interesse obtidas, anteriormente, por meio de estudo de GWAS, das quais 61,3% não estavam registradas/depositadas no banco de dados de variantes equino. Do total de polimorfismos, 29 (24 SNPs e 5 InDels) foram considerados de importância com base em três abordagens distintas e independentes: escores SIFT classificado como deletério (<0,05), grau de impacto na região consenso de cada polimorfismo, e frequências alélicas diferentes, identificadas pelo teste de Fisher ($p < 0,01$), entre os grupos de cavalos contrastantes para IV max. Com isso, oito genes descritos como candidatos em trabalhos anteriores (*ABCG5*, *COL11A1*, *GEN1*, *SOCS3*, *MICAL1*, *SPTBN1*, *EPB41L3* e *SHQ1*), e oito genes candidatos novos (*AKNA*, *ARMC2*, *FKBP15*, *LHX1*, *NOL10*, *TMEM192*, *ZFP37*, *FIG4* e *HNRNPU*) foram relacionados ao desempenho em corridas de cavalos da raça Quarto de Milha. Assim, os resultados obtidos neste trabalho mostraram que o desempenho em corridas na raça Quarto de Milha, dado pelo IV max, é característica quantitativa e que não há ocorrência de *major* genes.

Palavras-chave: exoma, GWAS, Imputação de genótipos, InDels, índice de velocidade, SNPs

IDENTIFICATION OF CHROMOSOMAL REGIONS, GENES AND DNA POLYMORPHISMS ASSOCIATED WITH PERFORMANCE OF QUARTER HORSE RACE HORSES

ABSTRACT – Among horses selected for speed, the racing line of Quarter Horses is characterized by high performance in sprint races, with these animals being considered the fastest horses in the world. Although in Brazil the effective number of animals in the racing line is relatively smaller compared to the other lines, its economic importance is substantial. Despite economic and scientific interest in this athletic trait, few efforts have been made to better understand the genetic and physiological mechanisms underlying this trait. The objectives of this study were: 1) to perform two-step genotype imputation between individuals in a relatively small population sample of racing Quarter Horses genotyped with the 54k or 65k panel, and to evaluate the accuracy of imputation through simulations; 2) to perform genome-wide association studies (GWAS) in Quarter Horses of the racing line using equine SNP genotyping chips for prospecting chromosome regions, genes and polymorphisms related to performance; 3) analyze exomes and UTRs in regions previously associated with this trait by GWAS in Quarter Horse racehorses with contrasting maximum speed index (SI_{max}), prospecting causal gene polymorphisms that are related to or are in strong linkage disequilibrium with racing performance. Genotypes were imputed using 116 horses genotyped with the 54k SNP array and 233 animals genotyped with the 65k array. For the simulations, random samples were chosen to compose the imputed and reference populations in two scenarios. Scenario A simulated the genotype imputation in the first step (from 65k to 54k) and scenario B in the second step (from 54k to 65k). Thus, after quality control, the 54k and 65k panels contained 7,048 and 16,940 exclusive markers, respectively. The mean concordance rate was 0.9815 and 0.9751 for scenarios A and B, and the mean allelic r^2 was 0.9791 and 0.974, respectively. After imputation was performed by the single-step GBLUP method using two approaches: ssGWAS1 in which only SNP effects are recalculated at each iteration, and ssGWAS2 in which SNP effects are recalculated from genomic estimated breeding values (GEBVs) at each iteration. Twenty-one regions that explained more than 1% of the total genetic variance (gVar) in the maximum speed index were identified by ssGWAS1 and 12 by ssGWAS2. More than 40% of gVar was explained by these regions in ssGWAS1 and about 30% in ssGWAS2. Among chromosomes that explained more than 1% of genetic variance, five were common to both methods (ECA 3, 10, 15, 22, 25). A total of 108 genes were identified with the first approach and 59 with the second approach. To exome sequencing, GEBVs were used for the formation of two groups of animals with contrasting racing performance (20 animals with superior SI_{max} and 20 with inferior SI_{max}). Good quality data were obtained throughout the reads sequenced, with an average coverage of 43x. A total of 1,203 variants (1,105 SNPs and 93 InDels) were identified in 33 regions of interest obtained previously by GWAS; of these, 61.3%

were not registered/deposited in the horse genomic variant database. Twenty-nine of the polymorphisms (24 SNPs and 5 InDels) were considered to be important based on three different and independent approaches: SIFT scores classified as deleterious (<0.05), degree of impact on the consensus region of each polymorphism, and different allele frequencies identified by Fisher's exact test ($p < 0.01$) between the groups of horses with contrasting SI_{max}. Thus, eight genes described as functional and positional candidates in previous studies (*ABCG5*, *COL11A1*, *GEN1*, *SOCS3*, *MICAL1*, *SPTBN1*, *EPB41L3*, and *SHQ1*) and eight new candidate genes (*AKNA*, *ARMC2*, *FKBP15*, *LHX1*, *NOL10*, *TMEM192*, *ZFP37*, *FIG4*, and *HNRNPU*), some of them with known function, were related to racing performance in Quarter Horses. Taken together, the present results show that the racing performance of Quarter Horses, given by the maximum speed index, is a quantitative trait and that no major genes exist.

Keywords: Exome, GWAS, Imputation of genotypes, InDels, Speed Index, SNPs,

CAPÍTULO 1 – Considerações Gerais

1. INTRODUÇÃO

A raça Quarto de Milha foi a primeira a ser criada na América do Norte, a partir do século XVII, com a entrada de equinos de origem árabe e turca, trazidos por colonizadores europeus. O seu maior desenvolvimento ocorreu com a ocupação do oeste Norte Americano, devido à necessidade de cavalos robustos e versáteis (ABQM, 2016). No século passado a raça foi intensamente difundida pelo mundo, sobretudo nas Américas, chegando ao Brasil em 1955. Atualmente encontram-se dentro da raça três segmentos distintos, originados por diferentes objetivos de seleção: as linhagens de corrida, de trabalho e de conformação (ABQM, 2016). No Brasil destacam-se, em números, cavalos da linhagem de trabalho, que são utilizados tanto no campo quanto em provas funcionais. Apesar do efetivo de animais ser relativamente menor na linhagem de corrida, sua importância econômica é substancial.

Em 2015 o complexo da agroindústria do cavalo gerou renda de R\$ 16,15 bilhões de reais no Brasil, dos quais R\$ 6,63 bilhões foram gerados a partir de cavalos destinados a lazer e esporte (MAPA, 2015). Embora os equinos tenham grande impacto na economia e sejam de grande interesse científico, pesquisas genéticas dedicadas ao melhoramento de raças, envolvendo ou não a exploração de ferramentas moleculares, tem sido bastante restritas quando comparados à espécies domésticas de produção. Tal fato pode estar relacionado a alguns fatores, como: 1) ausência de programas consistentes de melhoramento genético na espécie, a exemplo do que ocorrem com bovinos, aves e suínos; 2) dificuldades em mensurar características de desempenho, as quais podem ser dadas por meio indireto ou pouco preciso; 3) altos custos de ensaios customizados destinados a estudos genômicos; e 4) tradicionalismo e, por conta disso, resistência de associações e criadores à biotecnologias.

Na última década, plataformas de sequenciamento de nova geração tornaram-se acessíveis, permitindo o sequenciamento e mapeamento de milhares de polimorfismos espalhados por todo o genoma equino (WADE et al., 2009; DOAN et al., 2012), o que levou ao desenvolvimento de painéis de genotipagem de SNPs em larga escala. Em sua primeira versão, a plataforma de genotipagem específica para espécie equina, o *Equine SNP50 BeadChip* da empresa Illumina (Illumina Inc., EUA), contava com cerca de 54.000 SNPs (54K). Atualmente, em sua segunda geração, o SNP chip equino (*Equine SNP70 BeadChip*; Illumina Inc., EUA) possui aproximadamente 65 mil SNPs (65K), dos quais 19

mil são novos marcadores e 45 mil foram validados no *Equine SNP50 BeadChip*. A imputação de genótipos com base em informações de desequilíbrio de ligação (DL) entre SNPs oferece potencial solução para a resolução deste problema (HAYES et al., 2011; SARGOLZAEI; CHESNAIS; SCHENKEL, 2014). A fim de melhorar a predição dos resultados obtidos pelos estudos de associação e pela seleção genômica, diferentes estratégias de imputação de genótipos têm sido aplicados em dados genômicos (HAYES et al., 2011; SARGOLZAEI; CHESNAIS; SCHENKEL, 2014).

A relação causal entre polimorfismos genéticos dentro de uma espécie e as diferenças fenotípicas observadas entre indivíduos é de fundamental interesse biológico (KORTE; FARLOW, 2013) e, em muitas situações, econômico. Segundo Sahana et al. (2010), o principal propósito de estudos de associação ampla do genoma (*Genome-Wide Associations Studies* – GWAS), os quais envolvem milhares de marcadores distribuídos ao longo de todo o genoma, é identificar regiões dos cromossomos que abrigam genes que contribuem para a variação fenotípica de uma característica, servindo, posteriormente, como regiões putativas de QTL (*quantitative trait loci*) para mais estudos. Estudos de associação entre polimorfismos de DNA e desempenho em corridas em cavalos têm sido realizados na raça Puro-Sangue Inglês, tanto de forma pontual (GU et al., 2010; HILL et al., 2010), quanto de forma ampla (SHIN et al., 2015). Entretanto, os resultados encontrados por Pereira et al. (2016a,b) demonstraram que marcadores apropriados para a raça inglesa não servem para a identificação de animais superiores para desempenho em corridas na raça Quarto de Milha.

Embora a utilização de painéis de SNPs permita o estudo de todo o genoma, e, com base no desequilíbrio de ligação entre milhares de marcadores, apresente grande utilidade na detecção de QTL, na seleção genômica, e em estudos evolutivos entre outros, não geram informações adicionais de sequências de genes em regiões associadas do genoma. Deste modo, para refinamento de resultados, podem ser utilizadas outras estratégias. Atualmente, os custos relacionados ao sequenciamento de regiões genômicas de tamanho significativo, ou mesmo todo genoma (*Whole Genome Sequencing* – WGS) tem se tornado baixos (LIU et al., 2012), possibilitando o uso de tecnologias de nova geração (*Next Generation Sequencing* – NGS) para, dentre inúmeras possibilidades, refinamento de experimentos que utilizam painéis de SNPs. O sequenciamento de todo o exoma (*Whole Exome Sequencing* – WES) pode ser, do mesmo modo, empregado para esse propósito. Com baixo custo essa abordagem permite a detecção de variantes genéticas (*Single Nucleotide Polymorphisms* – SNPs, *Insertions/Deletions* – InDels e CNVs – *Copy Number Variation*) já mapeadas, ou *de novo*, em regiões gênicas, incluindo regiões codificantes de

aminoácidos, além de 5' e 3' UTR de todo o genoma, o que pode ser útil por disponibilizar dados extras para suprir mais de uma frente de pesquisa.

2. OBJETIVOS GERAIS

Considerando a falta de informações em relação ao potencial de utilização da seleção assistida por marcadores em equinos, em especial no Quarto de Milha, raça de grande importância no Brasil, este trabalho teve como objetivos encontrar polimorfismos, genes e regiões genômicas, que possam estar associados a diferenças no desempenho em corridas de equinos da raça Quarto de Milha.

2.1. Objetivos específicos

- Realizar imputação de genótipos exclusivos entre amostras de Quarto de Milha genotipadas com painéis de duas gerações (54k e 65k) para obter o mesmo número de marcadores para cada indivíduo.
- Realizar estudo de associação ampla do genoma (GWAS) e característica de desempenho em cavalos Quarto de Milha de corrida utilizando painéis de arranjos de SNPs.
- Analisar o exoma equino de cavalos Quarto de Milha dentro de regiões de interesse, as quais foram significativas pelo GWAS, visando a prospecção de polimorfismos gênicos responsáveis diretos (causais), ligados ou em forte desequilíbrio de ligação com locos gênicos responsáveis por características determinantes para o melhor desempenho em corridas.

3. REVISÃO DE LITERATURA

3.1. Os equinos no Brasil e no mundo

Existem aproximadamente 58,9 milhões de cavalos no mundo, maior parte vivendo na América, Ásia e alguns países da Europa (FAO, 2014). Os Estados Unidos da América é o país com maior número de cabeças (10.260.000), seguido da China (6.027.400), México (6.355.000) e Brasil (5.450.601) (GLIPHA, 2014). Deve-se destacar a redução da população de equinos na Ásia, principalmente na China, de 8.914.000 de cabeças em 2000 para 6.027.400 em 2014 (GLIPHA, 2014), está associada à migração interna da

população humana, com menor utilização dos equídeos no transporte e agricultura e o maior consumo de carne equina (ALMEIDA & SILVA, 2010). Por outro lado, nos Estados Unidos houve aumento expressivo da população de equinos, de 5.240.000 cabeças em 2000 para 10.260.000 cabeças em 2014 (GLIPHA, 2014), em parte devido às restrições legais internas ao abate e à exportação de carne de equídeos (ALMEIDA & SILVA, 2010). No Brasil, no período de 2000 à 2014, o número de equinos permaneceu praticamente estável, com pequenas oscilações (GLIPHA, 2014).

O complexo do agronegócio do cavalo no Brasil é bastante expressivo, movimentando valor econômico superior a R\$ 16,15 bilhões ao ano e gerando aproximadamente 3,2 milhões de empregos diretos e indiretos (MAPA, 2015). A utilização de animais (aproximadamente 3,9 milhões) no manejo de rebanhos bovinos vem sendo o seguimento de maior destaque dentro do agronegócio do cavalo no Brasil. Somando os valores do custo de manutenção, do valor anual da tropa e a renda associada à mão-de-obra, verifica-se que o segmento de lida responde por R\$ 8,58 bilhões, cerca de 53% da renda gerada no complexo do agronegócio do cavalo (MAPA, 2015). Esta intensa relação tem feito os cavalos acompanharem os bovinos em seu deslocamento para as regiões Centro-Oeste e Norte do País.

Embora a utilização do cavalo no manejo de bovinos seja marcante no contexto da equinocultura nacional, o seguimento que envolve os esportes equestres tem crescido acentuadamente nos últimos anos. Estima-se a existência de 1,1 milhão de animais no segmento de esportes e lazer, com 125.700 empregos diretos e movimentação econômica de R\$ 5,84 bilhões (MAPA, 2015). No período de 1999 a 2004 o número de eventos envolvendo as diversas modalidades equestres (salto, adestramento, concurso completo de equitação, enduro, competições de marcha, etc.) cresceu 315%, ou seja, aumento médio de 15,3% ao ano (LIMA *et al.*, 2006). O crescimento verificado nessa categoria desde o estudo realizado por Lima *et al.* (2006), em que o plantel foi estimado em 800 mil animais, foi de 37% (MAPA, 2015).

Além da utilização do cavalo no campo e nas provas equestres, a tendência de sua utilização para lazer vem aumentando significativamente (MAPA, 2015). O cavalo também é utilizado na equoterapia, modalidade disponível há milhares de anos e, agora, reconhecida como de grande eficácia para o tratamento de inúmeros males físicos, psíquicos e comportamentais.

De forma geral, de 2006 a 2015 o crescimento da renda total gerada pelo complexo do agronegócio do cavalo foi de 115% (LIMA *et al.*, 2006; MAPA, 2015). Esse resultado positivo pode ser explicado por algo que tem sido tendência mundial na equinocultura, o

forte crescimento da criação voltada para o público urbano, tanto para o lazer quanto para o esporte. São segmentos que movimentam com maior intensidade desde a indústria de medicamentos e ferragens até cosméticos e acessórios. Junto com este consumidor, cresce também o número e tamanho dos eventos, como provas de tambor e baliza, vaquejadas e tantos outros (MAPA, 2015).

3.2. A raça Quarto de Milha

A raça Quarto de Milha foi a primeira a ser formada na América do Norte, a partir do século XVII, com a introdução de equinos de origem árabe e turca, trazidos por colonizadores europeus. O seu maior desenvolvimento ocorreu com a ocupação do oeste Norte Americano, devido à necessidade de cavalos robustos e versáteis, com aptidão à sela e tração, visto a dificuldade de se manter plantel variado de animais para atender às diversas necessidades (ABQM, 2016). Em 1940, fundou-se, nos Estados Unidos da América, a *American Quarter Horse Association* (AQHA), primeira envolvendo a raça. Atualmente, a AQHA é considerada a maior associação de criadores do mundo, com cerca de 400 mil sócios e mais de cinco milhões de cavalos registrados, divididos em 43 países, representando 52% dos equinos em todo o mundo (dados de até 31 de dezembro de 2011) (ABQM, 2016).

Em 1969 foi fundada a Associação Brasileira de Criadores de Cavalo Quarto de Milha (ABQM), a qual conta atualmente com plantel composto por 474.862 animais registrados, 95.792 criadores, proprietários e associados cadastrados distribuídos por todos os estados brasileiros. No Brasil, a criação da raça apresenta impacto relevante no agronegócio nacional do cavalo, visto que seus haras ocupam área de aproximadamente um milhão de hectares, estimados em mais de R\$ 19,8 bilhões (dados de até 13 de agosto de 2015) (ABQM, 2016).

A seleção para diferentes propostas na raça Quarto de Milha (EVANS, 1996), levou à formação de linhagens, entre as quais as de: corrida, conformação e trabalho. A linhagem de trabalho destina-se às provas de caráter funcional, explorando habilidades como agilidade e obediência, características consideradas de grande importância no manejo do gado a campo. Dentro desta linhagem sempre houve grande interesse na produção de cavalos com *cow sense* superior (ELLERSIECK et al., 1985). *Cow sense*, ou habilidade de trabalhar com bovinos, pode ser medida pela capacidade do cavalo em cercar o gado e apartar do rebanho um animal escolhido (HINTZ, 1980), com pouca assistência do cavaleiro (ELLERSIECK et al., 1985). A linhagem de conformação enfatiza

a morfologia do padrão racial. A linhagem de corrida explora a aptidão dos animais quanto à velocidade em pistas retas e de curta distância.

Destacadamente equinos desta raça tem melhor desempenho em corridas de curtas distâncias do que qualquer outra raça (ABQM, 2016), podendo alcançar velocidades de até 88 km/h e percorrer, a partir de uma posição estática, $\frac{1}{4}$ de milha (402 metros, aproximadamente) em menos de 21 segundos (AMERICA'S HORSE DAILY, 2008). De acordo com Meira et al. (2013), existem diferenças significativas entre as linhagens de corrida e de trabalho da raça Quarto de Milha com relação às características morfológicas de peso, altura à cernelha, comprimentos corporal, da canela, da quartela, da garupa, da cabeça, e do pescoço e perímetros torácico, da canela e do casco. Os autores observaram que animais de corrida apresentaram maiores pesos, alturas, comprimentos e perímetros corporais em relação aos de trabalho.

Apesar do efetivo de animais ser relativamente menor na linhagem de corrida do que nas demais linhagens, sua importância econômica é substancial, não somente por gerar renda por meio de premiações e apostas (LIMA et al., 2006), mas também pelo elevado custo gerado na manutenção destes animais dentro desta modalidade esportiva.

O principal atributo de seleção utilizado pelos criadores do cavalo Quarto de Milha de corrida é a pontuação conhecida como Índice de Velocidade (IV). Este índice é obtido durante a campanha de um animal com o intuito de classificar o seu desempenho em diferentes condições (distâncias, hipódromo, clima, país) (EVANS, 1996). Cada hipódromo tem sua própria tabela de IV, que é elaborada a partir da média das três vitórias mais rápidas (três melhores tempos) para cada um dos três últimos anos consecutivos, em cada distância, sendo que o valor da média desses nove tempos equivalerá ao IV igual a 100 (JCS, 2002). Os pontos de IV são inteiros e variam de acordo com o tempo, ao nível de centésimos de segundo, seguindo ajustes em acordo com a distância percorrida.

A tabela de IV (Tabela 1) faz a conversão do tempo em pontos do IV com ajustes pelas distâncias. Como exemplo, nas distâncias de 365 metros (m), 402m e 503m, a cada quatro centésimos de segundo, a mais ou a menos, que um animal obtém, em relação ao tempo que representa o índice de velocidade igual a 100, diminui-se ou acresce-se um ponto neste índice. Assim, ao se considerar que a média das nove vitórias (IV =100) foi de 22 segundos para os 402m, o animal cujo tempo se situe entre 22,01 e 22,04 terá IV igual a 99, se o tempo estiver entre 22,05 e 22,08 o IV será 98, e assim por diante. Por outro lado, se o tempo estiver entre 21,96 e 21,99 será acrescido em um ponto, obtendo o IV 101, e os que tiverem tempo entre 21,92 e 21,95 seus IV serão de 102 pontos, e assim sucessivamente a cada quatro centésimos de segundo a menos. Para distâncias menores

a variação do IV ocorrerá em uma menor variação do tempo, em 320m se alternará a cada três e quatro centésimos de segundo, iniciando com três centésimos de segundo, enquanto que aos 301m e 275m será de três centésimos de segundo e aos 228m será a cada dois centésimos de segundo. Essa tabela é válida para animais que correm carregando um peso mínimo de 53 Kg. Para aqueles com peso inferior, deve-se acrescentar cinco centésimos de segundo ao seu tempo, para cada quilo a menos, antes de consultar a tabela.

Tabela 1. Esquema da variação para pontuação do Índice de Velocidade (IV), de acordo com a distância (m), tendo como ponto de partida os tempos referentes ao IV igual a 100.

	Centésimos de segundo			
	4	3 e 4*	3	2
	365	320	275	228
Distância (m)	402		301	
	503			

*alternados, iniciando-se com 3 centésimos (fonte: adaptado de Corrêa; Mota, 2007)

3.3. Melhoramento genético de equinos

Em relação a outras espécies de exploração zootécnica, as pesquisas na área de melhoramento genético são proporcionalmente menores em equinos, no mundo todo. No Brasil, em particular, dada a grandeza de sua tropa, esta distância é ainda mais evidente. Embora algumas pesquisas publicadas em cavalos envolvam a área de melhoramento genético (SANTOS, 2006; CORRÊA; MOTA, 2007; PRADO; MOTA, 2008), ainda não existem, efetivamente, programas consistentes de seleção nas diferentes raças criadas no Brasil. Neste sentido, pequenos grupos de pesquisadores ligados às universidades e institutos de pesquisa têm trabalhado na área, especialmente em relação a aspectos conservacionistas em raças nacionais ou quantitativos de caracteres de interesse econômico (herdabilidades e correlações) em raças nacionais e importadas. Recentemente, pesquisas envolvendo Genética e Biologia Molecular têm sido realizadas

em algumas raças da espécie no Brasil (MEIRA et al., 2014a,b,c; PEREIRA et al., 2016a,b).

Trabalhos mais refinados em relação à seleção e ao melhoramento genético de equinos vêm sendo realizados na Europa, com a utilização de modelos estatísticos mais adequados para a obtenção de valores genéticos individuais, como o BLUP (*Best Linear Unbiased Prediction*) (ARNASON, 2013). Esta alternativa tem sido aplicada na criação de cavalos destinados a esportes olímpicos como o *cross country*, o adestramento e o salto.

A maior causa desta dificuldade é a baixa herdabilidade do desempenho atlético e os valores de correlações genéticas existentes entre as características mais almejadas nos cavalos, principalmente ligadas a reprodução e desempenho em provas funcionais e corridas (KOENEN et al., 1995; WALLIN et al., 2003; BOKOR et al., 2005). A baixa herdabilidade do desempenho atlético em equinos é argumento a ser considerado para a utilização de novas ferramentas disponíveis para a seleção e o melhoramento genético, tais como os marcadores moleculares.

3.4. Características quantitativas e *major genes*

Da mesma forma que para a maior parte das características de importância econômica nas espécies de interesse zootécnico, o desempenho em corridas dos equinos Quarto de Milha, muitas vezes representado pelo Índice de Velocidade, deve ser governado por um grande número de genes, localizados em regiões cromossômicas denominadas locos de características quantitativas (*Quantitative Trait Loci* – QTL). Estes genes podem ter grandes efeitos e serem chamados de “genes principais” ou *major genes*, ou podem ter pequenos efeitos individualmente e serem chamados de “genes menores”. A estrutura da variação genética que está por trás de traços fenotípicos tem consequências importantes para a compreensão da evolução de características quantitativas (AGRAWAL et al., 2001). A frequência e o papel dos genes de grande efeito em genética quantitativa têm sido alvo de intenso debate e investigação (ORR, 2005). Apesar do fascínio pelo modelo infinitesimal, há evidência acumulada de várias fontes (experimentos de seleção artificial, evolução experimental, e mapeamento de QTL) sugerindo que os genes de grande efeito, muitas vezes contribuem para caracteres quantitativos (STINCHCOMBE et al., 2009).

3.5. Polimorfismos de nucleotídeos únicos (SNPs)

Marcador molecular é toda e qualquer variação oriunda de um gene expresso ou de um segmento específico de DNA, correspondente a regiões expressas ou não do genoma. Ao se verificar que esses marcadores segregam de acordo com as leis mendelianas para características monogênicas, ou apresentam distribuições compatíveis com as esperadas para características poligênicas, um marcador molecular é também definido como marcador genético (FERREIRA; GRATTAPAGLIA, 1998). Do ponto de vista molecular ocorrem três tipos principais de variações na molécula de DNA, as regiões repetitivas (minissatélites e microssatélites), as inserções e deleções (InDels) e as alterações de uma base (polimorfismos de nucleotídeos únicos – SNPs). À medida que as sequências de nucleotídeos dos genomas foram sendo desvendado, um grande número de variações de uma base foram encontradas ao se comparar segmentos correspondentes do mesmo genoma, ocorrendo, aproximadamente, a cada 600 bases.

As substituições mais frequentes observadas no DNA envolvem bases nitrogenadas de mesma característica estrutural, ou seja, são trocas entre duas purinas (A/G ou G/A) ou duas pirimidinas (C/T ou T/C) e são denominadas transições. As transversões são substituições de uma purina por uma pirimidina ou o contrário. Essas alterações podem ser provocadas por erros de incorporação de bases durante a replicação do DNA ou em outros casos, são causadas por agentes ambientais (químicos e físicos). Caso essas alterações ocorram em células sexuais e sejam transmitidas às gerações seguintes passam a ser denominadas mutações germinativas. Caso fixem-se na população a uma frequência mínima de 0,01 ou mais (*minor allele frequency* – MAF $\geq 1\%$), passam a ser denominadas de polimorfismos ou SNPs, no caso de alterações de apenas um nucleotídeo (KWOK & GU, 1999).

Os SNPs podem ocorrer em regiões codificadoras ou com função regulatória, bem como em espaços intergênicos, sem função determinada. Em regiões codificadoras, quando resultam em uma substituição de aminoácido na sequência proteica, são denominados não sinônimos, podendo a substituição ser conservativa ou não conservativa em função das características dos aminoácidos envolvidos na troca. Nesses casos, pode haver modificações estruturais e funcionais na proteína. Embora SNPs sinônimos não alterem a sequência proteica, eles podem modificar a estrutura e a estabilidade do RNA mensageiro, e, conseqüentemente afetar a quantidade de proteína produzida. Esta também pode ser afetada quando ocorrem alterações nas regiões não traduzidas do RNA mensageiro (5' UTR e 3' UTR). Além disso, polimorfismos gênicos podem promover

processamentos alternativos, geração ou supressão de códons de terminação, alteração nos códons de iniciação da tradução e alterações no padrão de expressão de genes quando a troca de bases ocorre em sequências promotoras (GUIMARÃES & COSTA, 2002). Polimorfismos em regiões de intron ganharam importância pelo fato de não serem mais descartados como possíveis responsáveis diretos por alterações fenotípicas. RNAs não codificantes transcritos a partir de regiões de introns (micro-RNAs) podem estar envolvidos em diferentes processos biológicos tais como os controles transcricional e pós-transcricional da expressão gênica (NAKAYA et al., 2007). Embora a função da maior parte das regiões intergênicas ainda seja desconhecida, conhece-se cada vez com mais profundidade a sua importância no controle da regulação da expressão gênica. Neste sentido, polimorfismos localizados nestas regiões também podem estar associados à variações fenotípicas.

Estudos em humanos (VENTER et al., 2001; LANDER et al., 2011) e em espécies de interesse zootécnico (WADE et al., 2009; DOAN et al., 2012;) mostraram a ocorrência de milhões SNPs ao longo do genoma de um indivíduo (*Human Genome Project Information, The SNP Consortium LTD, Bovine Genome Sequencing and Analysis Consortium, EquCab2.0 SNP Collection*), ou seja, apenas pela comparação da sequência dos seus dois cromossomos homólogos. Além dos marcadores SNP serem abundantes, suas bases moleculares permitem sua distribuição homogênea pelo genoma (CAETANO, 2009).

3.6. Painéis de SNPs e imputação de genótipos

De acordo com Chowdhary e Raudsepp (2008), entre os maiores destaques provenientes da análise do genoma de equinos estão o seu sequenciamento completo (EquCab2.0) e, a partir deste, a identificação de 1.162.753 de SNPs que ocorrem entre diversas raças (WADE et al., 2009).

Projetado para permitir a identificação de regiões genômicas modificadas pela seleção e a identificação de regiões cromossômicas, genes e SNPs que contribuem para características de interesse nas principais raças de equinos criadas atualmente no mundo, o *Equine SNP50 BeadChip* da empresa Illumina (Illumina Inc., EUA) constituiu poderosa plataforma para a seleção e o melhoramento genético da espécie, habilitando pesquisadores da área a conduzir vasta gama de experimentos em que a aplicação da genotipagem de polimorfismos de DNA é necessária. Já em sua segunda geração, o novo *SNP chip* equino (*Equine SNP70 BeadChip*; Illumina Inc., EUA) possui aproximadamente

65 mil SNPs, dos quais 19 mil são novos marcadores e 45 mil foram validados no *Equine SNP50 BeadChip*.

A análise simultânea de milhares de polimorfismos espalhados ao longo dos genomas têm possibilitado o estudo da estrutura genética de diferentes populações em várias espécies de animais domésticos (KIJAS et al., 2009; McKAY et al., 2008; GIBBS et al., 2009), estimar o grau de diversidade dentro e divergência genética entre populações (ZENGER et al., 2006), determinar a relação entre a perda de alelos e o aumento da endogamia devido à seleção (MUIR et al., 2008), e identificar e localizar regiões do genoma sujeitas à seleção (HAYES et al., 2006; BARENDSE et al., 2009; MACEACHERN et al., 2009; PRASAD et al., 2008).

Boa parte destas aplicações requerem o conhecimento ou acesso ao desequilíbrio de ligação (DL) existente entre os marcadores moleculares em dada população da espécie de interesse. O DL entre marcadores moleculares descreve a correlação entre os genótipos de dois marcadores, expressando o grau de associação não aleatória entre seus alelos (PORTO-NETO et al., 2014). Em geral, esta associação ou ligação é consequência da proximidade física existente entre dois locos (ARDLIE et al., 2002; BOLORMAA et al., 2011) e é influenciado por eventos como mutação, deriva, seleção, recombinação e tamanho efetivo da população.

A genotipagem de SNPs com diferentes arranjos em um mesmo experimento pode levar a uma perda importante de dados. A imputação de genótipos com base em informações de desequilíbrio de ligação (DL) entre marcadores oferece potencial solução para a resolução deste problema (HAYES et al., 2011; SARGOLZAEI et al., 2014).

A fim de melhorar a predição dos resultados obtidos pelos estudos de associação entre genótipos e fenótipos e pela seleção genômica, diferentes estratégias de imputação de genótipos têm sido aplicadas em dados genômicos (HAYES et al., 2011; SARGOLZAEI et al., 2014). A imputação possibilita diversas aplicações como: 1) gerar painéis de alta densidade (HD) para animais genotipados com os de baixa densidade (LD) (SARGOLZAEI et al., 2014; BOICHARD et al., 2012); 2) combinar dados provenientes de amostras populacionais genotipadas com painéis de diferentes densidades ou raças distintas, permitindo painéis com densidade única e assim aumentar o número de indivíduos na amostra, para promover resultados mais satisfatórios em estudos genômicos (HAYES et al., 2011; LARMER et al., 2012); e 3) predizer genótipos faltantes do painel de genotipagem em decorrência do controle de qualidade dos dados ou erros de leitura, aumentando a porcentagem de genótipos válidos dos animais (WENG et al., 2012).

Entretanto, a imputação em situações reais entre amostras de populações genotipadas com painéis distintos pode trazer grandes desafios à inferência da acurácia. Desse modo, algumas análises e simulações podem ser conduzidas no conjunto amostral com o intuito de direcionar a escolha de estratégias mais eficientes para imputação, bem como estimar a confiabilidade das imputações realizadas sob essas circunstâncias.

3.9. Estudos associação ampla do genoma (GWAS)

A relação causal entre polimorfismos genéticos dentro de uma espécie e as diferenças fenotípicas observadas entre indivíduos é de fundamental interesse biológico (KORTE; FARLOW, 2013) e, em muitas situações, econômico.

Segundo Sahana et al. (2010) o principal propósito de estudos de associação ampla do genoma (GWAS), os quais envolvem milhares de marcadores distribuídos ao longo de todo o genoma, é identificar regiões dos cromossomos que abrigam genes que contribuem para a variação fenotípica de uma característica, servindo, posteriormente, como regiões putativas de QTL para estudos mais aprofundados. Os GWAS baseiam-se também no princípio do DL ao nível da população. Os estudos de associação ampla têm emergido como poderosa ferramenta para revelar as bases genéticas de doenças herdáveis e de características quantitativas (COLLINS et al., 1998).

Em equinos, e em um primeiro momento, estudos de associação ampla serviram com sucesso, principalmente, à identificação de regiões genômicas e genes relacionados às importantes doenças e síndromes que acometem determinadas raças, tais como: lordose (COOK et al., 2010), osteocondrose (LYKKJEN et al., 2010; TEYSSEDRE et al., 2012), neuropatia laringeal recorrente (DUPUIS et al., 2011), nanismo (EBERTH et al., 2009) e síndrome do potro lavanda (BROOKS et al., 2010).

Mais recentemente, características complexas relacionadas ao desempenho em provas esportivas e aptidões específicas, como a marcha e suas variações, têm sido avaliadas (HILL et al., 2010; BINNS et al., 2010; SCHRÖDER et al., 2011; PETERSEN et al., 2013, MEIRA et al., 2014c, FONSECA et al., 2016, STAIGER et al., 2016a). Com relação às provas esportivas, estudos de associação amplos têm sido realizados, principalmente, com características relacionadas à distância e ao desempenho em corridas na raça Puro-Sangue Inglês (HILL et al., 2010; BINNS et al., 2010; TOZAKI et al., 2010) e ao salto em diversas raças (BRARD; RICARD, 2015). Características morfométricas (SIGNER-HASLER et al., 2012; TETENS et al., 2013; MEIRA et al., 2014b; STAIGER et

al., 2016b), de fertilidade (GOTTSCALK et al., 2016) e temperamento (STAIGER et al., 2016c) também têm sido contempladas.

3.10. Marcadores para desempenho em corridas de equinos

Embora polimorfismos de DNA já sejam utilizados para predição de distâncias ótimas e desempenhos superiores em corridas para a raça Puro-Sangue Inglês (PSI), esses marcadores não têm os mesmos efeitos em diferentes raças de igual propósito (corrida). Esta afirmação pode ser mantida ainda que se considere a linhagem de corrida da raça Quarto de Milha, a qual possui influência genética de PSIs que correm distâncias curtas (PEREIRA et al., 2016a,b).

Nos trabalhos realizados por Pereira e colaboradores não foram encontrados efeitos significativos dos alelos dos SNPs g.38973231A>G do gene *PDK4*, g.22684390C>T do gene *COX4I2* e g.22999655C>A do gene *CKM* sobre o Índice de Velocidade (IV) máximo de equinos de corrida da raça Quarto de Milha. Estes resultados sugerem que os alelos dos genes *PDK4* e *COX4I2*, relacionados com melhor desempenho em corridas na raça PSI, estejam relacionados à adaptações benéficas do metabolismo aeróbico e, dessa forma, tenham papéis secundários no desempenho em corridas de curtas distâncias e de explosão física (predominantemente anaeróbicas) da raça Quarto de Milha. Esperava-se que variações no *CKM* apresentassem papel relevante sobre o desempenho em corridas na raça Quarto de Milha uma vez que o produto proteico expresso por esse gene é responsável por manter os níveis de ATP celular constantes devido à fosforilação de adenosina difosfato (ADP) via creatina-fosfato, sendo importante no metabolismo energético muscular em exercícios intensos e curtos. Todavia, algumas pesquisas em humanos e em equinos PSI relacionaram o *CKM* com a eficiência energética do metabolismo energético muscular aeróbio (ECHEGARAY & RIVERA, 2001; GU et al., 2010).

Com base no exposto, estudos de associação ampla do genoma em cavalos Quarto de Milha de corrida seriam de grande interesse e importância para a descoberta e utilização de marcadores genéticos para seleção de fenótipos superiores na linhagem. Em trabalho desenvolvido por Meira et al. (2014c), considerado preliminar em razão do pequeno número de animais utilizados ($n = 112$), foram apontados alguns SNPs em genes candidatos possivelmente associados ($P < 0,0001$; *Q-value* de 0,25 – após correção para testes múltiplos) à característica de desempenho (índice de velocidade) na linhagem de corrida da raça Quarto de Milha (MEIRA et al., 2014c).

Estes SNPs, encontrados pelo GWAS com a característica IV, estão localizados, com base no EquCab2.0, nos cromossomos (Chr) 2: 97634986 e 97642160; Chr 4: 82402787; Chr 10: 52188059; Chr 18: 33671391; 36917870; 39128343 e Chr 27: 35222223 (MEIRA et al., 2014c), próximos à regiões selecionadas de forma divergente na linhagem de corrida em relação à de trabalho da raça Quarto de Milha, identificadas pela aplicação simultânea das metodologias homozigose relativa do haplótipo estendido (REHH – “relative extended haplotype homozygosity”) e índice de fixação (F_{ST}), localizadas nos Chr 2: 87453500-89546885; Chr 4: 96192056-983156500; Chr 10: 65461613-67583618; Chr 18: 30519448-32573824 e Chr 27: 14932202-16989870 (MEIRA et al., 2014a).

Embora os resultados encontrados sejam interessantes, em função de mostrarem pela primeira vez regiões do genoma possivelmente associadas ao desempenho em Quartos de Milha de corrida, necessitam ser corroborados em amostra maior de animais. Além disso, as regiões alvo de interesse devem ser sequenciadas em busca de polimorfismos em forte desequilíbrio de ligação, ligados ou responsáveis diretos pela variação da característica de interesse.

3.11. Sequenciamento de DNA

Quinze anos após a publicação da sequência de nucleotídeos do genoma humano no ano de 2001 nas revistas “Nature” e “Science” (VENTER et al., 2001; LANDER et al., 2001), muitos avanços ocorreram em relação às técnicas de sequenciamento e aos estudos genômicos. Os métodos de sequenciamento de próxima geração (*Next Generation Sequencing* – NGS) tornaram-se realidade, com rápidos e constantes avanços tecnológicos na última década, o que os deixou extremamente eficientes e de custo relativamente baixo, processos estes em contínua evolução. Com o auxílio de ferramentas de bioinformática, estes métodos permitem o sequenciamento e análise de milhões ou até bilhões de pares de bases (pb) em rodada única de leitura.

A primeira geração de NGS teve início com o desenvolvimento da 454 *Life Science* em 2000 por Jonathan Rothberg (www.454.com). Esse método, diferentemente da técnica que adota capilares (necessidade de eletroforese), tem abordagem de sequenciamento em tempo real, por síntese, e se baseia na detecção luminométrica de pirofosfatos inorgânicos (PPi) liberados com a incorporação dos desoxiribonucleotídeos trifosfatados (RONAGHI et al., 1998). A segunda geração de NGS é representada principalmente por quatro plataformas: 454 FLX (Roche, EUA); Solexa (Illumina Inc., EUA); e SOLiD (Applied Biosystems, EUA), as quais possuem elevada taxa de sequenciamento, gerando de 10 a

600 Gb por semana e tamanho médio de fragmentos de 100 a 700 pb (SHENDURE; JI, 2008).

Recentemente vêm sendo introduzidas novas plataformas de sequenciamento, incluindo os *Personal Genome Machine* – PGM e os NGS de terceira geração. O *Ion Personal Genome Machine* e *MiSeq* foram lançados pelas empresas Ion Torrent e Illumina, respectivamente. Ambos os equipamentos são de tamanho físico reduzido e apresentam altas taxas de rotatividade de trabalhos, mas a transferência de dados é limitada. Eles são direcionados para aplicações clínicas e pequenos laboratórios. As plataformas de sequenciamento NGS de terceira geração, diferentemente das de segunda, não necessitam de amplificação do DNA, cujo objetivo é fortalecer o sinal luminoso para captação fiável baseado em câmeras CCD, o que pode gerar distorções pela abundância de fragmentos. Na terceira geração, pela possível miniaturização em nano escala e utilização mínima de reagentes, é possível o sequenciamento de uma única molécula de DNA (CHIN et al., 2013). O primeiro sequenciador dessa geração foi o Heliscope™ (Pacific Bioscience, EUA), seguidos de outros como o PacBio RS ou SMRT™ e Sequel System (Pacific Bioscience, EUA) (KOREN et al., 2012; CHIN et al., 2013).

A utilização de painéis de SNPs de baixas e médias densidades possibilitam o estudo de todo o genoma, e, dessa forma, apresentam grande utilidade em estudos evolutivos e na detecção de QTL, entre outros. Todavia, não geram informações adicionais de regiões do genoma associadas a características de interesse. Deste modo, para refinamento de resultados, podem ser utilizadas estratégias adicionais. Atualmente, os custos relacionados ao sequenciamento de regiões genômicas de tamanho significativo, ou mesmo todo genoma (*Whole Genome Sequencing* – WGS) tem se tornado baixos (LIU et al., 2012), possibilitando o uso de tecnologias de próxima geração para, dentre inúmeras possibilidades, refinamento de experimentos que utilizam arranjos de SNPs por meio de resequenciamento de regiões alvo.

O sequenciamento de todo o exoma (*Whole Exome Sequencing* – WES) pode ser, do mesmo modo, empregado para esse propósito. Com custo apenas ligeiramente superior em relação ao resequenciamento de regiões alvo por captura por sondas, esta abordagem permite a detecção de variantes genéticas (*Single Nucleotide Polymorphisms* – SNPs, *Insertions/Deletions* – InDels e CNVs – *Copy Number Variation*) já mapeadas, ou *de novo*, em regiões expressas (exons), 5' e 3' UTR ao longo todo o genoma. Isto permite que a análise de polimorfismos possa ficar restrita à regiões de interesse do genoma, além de trazer a vantagem de disponibilizar dados extras para suprir outras frentes de pesquisa.

4. REFERÊNCIAS

- ABQM. Associação brasileira dos criadores de cavalos Quarto de Milha. Disponível em: <<http://www.abqm.com.br/item1.htm>>. Acesso em: 15 mar. 2016.
- AGRAWAL, A. F.; BRODIE, E. D.; RIESEBERG, L. H. Possible consequences of genes of major effect: transient changes in the G-matrix. **Genetica**, v. 112, p. 33–43, 2001.
- ALMEIDA, F. Q. de; SILVA, V. P. Progresso científico em equideocultura na 1ª década do século XXI. **Revista Brasileira de Zootecnia**, v. 39, p. 119-129, 2010.
- AMERICA'S HORSE DAILY. All About the Racing American Quarter Horse. 26 de fevereiro de 2008. Disponível em: <http://americashorsedaily.com/all-about-the-racing-american-quarter-horse/#.Urx8N_RDvik> . Acesso em: 12 nov. 2013.
- ARDLIE, K. G.; KRUGLYAK, L.; SEIELSTAD, M. Patterns of linkage disequilibrium in the human genome. **Natural Review Genetics**, v. 3, p. 299-309, 2002.
- ARNASON, T. Bright future for research in horse breeding! **Journal of Animal Breeding and Genetics**, v. 130, p. 167-169, 2013.
- BARENDSE, W.; HARRISON, B. E.; BUNCH, R. J.; THOMAS, M. B.; TURNER, L. B. Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. **BMC Genomics**, v. 10, n. 178, 2009. Disponível em: < doi: 10.1186/1471-2164-10-178>.
- BINNS, M. M.; BOEHLER, D. A.; LAMBERT, D. H. Identification of the myostatin locus (MSTN) as having a major effect on optimum racing distance in the Thoroughbred horse in the USA. **Animal Genetics**, v. 41, supl. 2, p. 154–158, 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, 2012. Disponível em: < doi: 10.1371/journal.pone.0034130>
- BOKOR, A.; BLOUIN, C.; LANGLOIS, B.; STEFLER, J. Genetic parameters of racing merit of Thoroughbred horses in steeplechase races. **Italian Journal of Animal Science**, v. 4, p. 43-45, 2005.
- BOLORMAA, S.; HAYES, B. J.; SAVIN, K.; HAWKEN, R.; BARENDSE, W.; ARTHUR, P. F.; HERD, R. M.; GODDARD, M. E. Genome-wide association studies for feedlot and growth traits in cattle. **Journal of Animal Science**, v. 89, p. 1684-1697, 2011.
- BRARD, S.; RICARD, A. Genome-wide association study for jumping performances in French sport horses. **Animal genetics**, v. 46, n. 1, p. 78-81, 2015.
- BROOKS, S. A.; N. GABRESKI, D.; MILLER, A.; BRISBIN, H. E.; BROWN. Whole-genome SNP association in the horse: Identification of a deletion in Myosin Va responsible for Lavender Foal Syndrome. **PLoS Genetics**, v. 6, n. 4, 2010. Disponível em: < <http://dx.doi.org/10.1371/journal.pgen.1000909>>

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, p. 64-71, 2009.

CHIN CS, ALEXANDER DH, MARKS P, KLAMMER AA, DRAKE J, HEINER C, CLUM A, COPELAND A, HUDDLESTON J, EICHLER E.E., TURNER SW, KORLACH J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. **Nature Methods**, v. 10, p. 563-569, 2013.

CHOWDHARY, B.P., RAUDSEPP, T. The Horse Derby: racing from map to whole genome sequence. **Chromosome Research**, v. 16, p. 109-127, 2008.

COLLINS, F. S.; BROOKS, L. D.; CHAKRAVARTI, A. A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. **Genome Research**, v. 8, p. 1229–1231, 1998.

COOK, D.; GALLAGHER, P.; BAILEY, E. Genetics of swayback in American Saddlebred horses. **Animal Genetics**, v. 41, supl. 2, p. 64–71, 2010.

CORRÊA, M. J. M.; MOTA, M. D. S. Genetic evaluation of performance traits in Brazilian Quarter Horse. **Journal of Applied Genetics**, v. 48, p. 145-151, 2007.

DOAN, R.; COHEN, N.D.; SAWYER, J.; GHAFARI, N.; JOHNSON, C.D.; DINDOT, S.V. Whole-Genome Sequencing and Genetic Variant Analysis of a Quarter Horse Mare. **BMC Genomics**, v. 13, n. 78, p. 1471-2164, 2012.

DUPUIS, M.C.; ZHANG, Z., DRUET, T., DENOIX, J.M.; CHARLIER, C.; LEKEUX P.; GEORGES M. Results of a haplotype-based GWAS for recurrent laryngeal neuropathy in the horse. **Mammalian Genome**, v. 22, n. 9, p.613-20, outubro de 2011.

EBERTH, J.; SWERCZAK, T.; BAILEY, E. Investigation of Dwarfism Among Miniature Horses using the Illumina Horse SNP50 Bead Chip. *Journal of Equine Veterinary Science*, v.29 (5), p.315, 2009.

Echegaray M, Rivera MA. Role of creatine kinase isoenzymes on muscular and cardiorespiratory endurance: genetic and molecular evidence. **Sports Medicine**, v. 31, p. 919-34, 2001.

ELLERSIECK, M.R.; LOCK, W.E.; VOGT, D.W.; Aipperspach, R. Genetic evaluation of cutting scores in horses. **Equine Veterinary Science**, v. 5, p. 287-289, 1985.

EVANS, J. W. Horses: a guide to selection, care and enjoyment. Freeman and Company: New York, 4 ed., 1996. 797p.

FAO. FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS STATISTICS DIVISION, 2014. Disponível em: < <http://faostat3.fao.org/browse/Q/QA/E>>. Acesso em: 15 jul. 2016.

FERREIRA, M. E.; GRATAPAGLIA, D. Introdução ao uso de marcadores moleculares em análise genética. Embrapa-Cenargen: Brasília, 3 ed., 1998. 220p.

FONSECA, M. G.; FERRAZ, G. C.; PEREIRA, G. L.; CURI, R. A. A genome-wide association study reveals differences in the genetic mechanism of control of the two gait

patterns of the Brazilian Mangalarga Marchador breed. **Journal of Equine Veterinary Science**, v.37, p.1-98, 2016.

GIBBS, R. A.; TAYLOR, J. F.; VAN TASSELL, C. P. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. **Science**, v. 324, p. 528-532, 2009.

GLIPHA. Global Livestock Production and Health Atlas, 2014. Disponível em: <http://www.fao.org/ag/againfo/home/en/news_archive/AGA_in_action/glipha.html>. Acesso em: 15 jul. 2016.

GOTTSCHALK, M.; METZGER, J.; MARTINSSON, G.; SIEME, H.; DISTL O. Genome-wide association study for semen quality traits in German Warmblood stallions. **Animal Reproduction Science**, v. 171, p. 81-86, 2016.

GU, J.; MACHUGH, D.E.; MCGIVENY, B.A.; PARK, S.D.E.; KATZ, L.M.; HILL, E.M. Association of sequence variants in CKM (creatine kinase, muscle) and COX4I2 (cytochrome c oxidase, subunit 4, isoform 2) genes with racing performance in Thoroughbred horses. **Equine Veterinary Journal**, v. 42, p. 569-575, 2010.

GUIMARÃES, P. E. M.; COSTA, M. C. R. SNPs: Sutis diferenças de um código. **Biociência e Desenvolvimento**, n. 26, p. 24-27, 2002.

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.

HAYES, B. J.; LIEN, S.; NIELSEN, H.; OLSEN, H. G.; BERG, P.; MACEACHERN, S.; POTTER, S.; MEUWISWSEN, T. E. The origin of selection signatures on bovine chromosome 6. **Animal Genetics**, v. 39, p. 101-105, 2006.

HILL, E. W.; MCGIVNEY, B. A.; GU, J.; WHISTON, R.; MACHUGH, D. E. A genome-wide SNP-association study confirms a sequence variant (g.66493737C>T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. **BMC genomics**, v. 11, n 552, 2010. Disponível em < doi: 10.1186/1471-2164-11-552>.

HINTZ, R.L. Genetics performance in the horse. **Journal of Animal Science**, v.51, p.582-594, 1980.

JCS. Especial Quarto de Milha de Corrida. Jockey Club de Sorocaba, Sorocaba, SP, 2002.

KIJAS, J. W.; TOWNLEY, D.; DALRYMPLE, B. P.; HEATON, M. P.; MADDOX, J. F.; MCGRATH, A.; WILSON, P.; INGERSOLL, R. G.; MCCULLOCH, R.; MCWILLIAM, S.; TANG, D.; MCEWAN, J.; COCKETT, N. V.; ODDY, H.; NICHOLAS, F. W.; RAADSMA, H. A. Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds. **PLoSOne**, v. 4, n. 3, 2009. Disponível em:< doi: 10.1371/journal.pone.0004668>.

KOENEN, E.P.C.; VAN VELDHUIZEN, A.E.; BRASCAMPA, E.W. Genetic parameters of linear scored conformation traits and their relation to dressage and show-jumping

performance in the Dutch Warmblood Riding Horse population. *Livestock Production Science*, v. 43, p. 85-94, 1995.

KOREN, S.; SCHATZ, M.C.; WALENZ, B.P.; MARTIN, J.; HOWARD, J.T.; GANAPATHY, G.; WANG, Z.; RASKO, D.A.; MCCOMBIE, W.R.; JARVIS, E.D.; PHILLIPPY, A.M. Hybrid error correction and de novo assembly of single-molecule sequencing reads. **Nature Biotechnology**, v. 30, p. 693–700, 2012.

KORTE, A.; FARLOW, A. The advantages and limitations of trait analysis with GWAS: a review. **Plant Methods**, v. 9, p. 29, 2013.

KWOK, P. Y.; GU, Z. Single nucleotide polymorphism libraries: why and how are we building them? **Molecular Medicine Today**, v. 5, p. 538-543, 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.

LIMA, R. A. S.; SHIROTA, R.; BARROS, G. S. C. Estudo do Complexo do Agronegócio Cavalo no Brasil. (Relatório Final), CEPEA – Centro de Estudos Avançados em Economia Aplicada, Piracicaba: ESALQ/USP, SP, Brasil, 2006. 250p.

LIU, L.; LI, Y.; LI, S.; HU, N.; HE, Y.; PONG, R.; LIN, D.; LU, L.; LAW, M. Comparison of next-generation sequencing systems. **Journal of Biomedicine and Biotechnology**, v. 2012, 2012. Disponível em: < <http://dx.doi.org/10.1155/2012/251364> >

LYKKJEN, S.; DOLVIK, N.I.; MCCUE, M.E.; RENDAHL, A.K.; MICKELSON, J.R.; ROED, K. H. Genomewide association analysis of osteochondrosis of the tibiotarsal joint in Norwegian Standardbred trotters. **Animal Genetics**, v. 41, suppl. 2, p. 111–120, 2010.

MACEACHERN, S.; HAYES, B.; MC EWAN, J.; GODDARD, M. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. **BMC Genomics**, v.10, p.181, 2009.

MAPA. Revisão do estudo do complexo do agronegócio do cavalo. Câmara de Equideocultura do Ministério da Agricultura, Pecuária e Abastecimento, Brasília, 2015.

MCKAY, S. D.; SCHNABEL, R. D.; MURDOCH, B. M.; MATUKUMALLI, L.K.; AERTS, J.; COPPIETERS, W.; CREWS, D.; NETO, E. D.; GILL, C. A.; GAO, C.; MANNEN, H.; WANG, Z.; VAN TASSELL, C. P.; WILLIAMS, J. L.; TAYLOR, J. F.; MOORE, S. S. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. **BMC Genetics**, v. 9, n. 37, 2008. Disponível em: <doi: 10.1186/1471-2156-9-37>.

MEIRA, C. T.; CURI, R. A.; SILVA J. A .II. V.; CORRÊA, M. J. M.; OLIVEIRA, H. N.; MOTA, M. D. S. Morphological and genomic differences between cutting and racing lines of Quarter Horses. **Journal of Equine Veterinary Science**, v. 13, p. 244-249, 2013.

MEIRA, C.T.; CURI, R.A.; FARAH, M.M.; OLIVEIRA, H.N.; BELTRAN, N.A.R.; SILVA, J.A.IV.; MOTA, M.D.S. Prospection of genomic regions divergently selected in racing line of Quarter Horses in relation to cutting line. **Animal**, v. 8, n. 11, p. 1754-1764, 2014a.

MEIRA, C. T.; FARAH, M. M.; FORTES, M. R. S.; MOORE, S. S.; PEREIRA, G. L.; SILVA, J. A. V.; DA MOTA, M. D. S.; CURI, R. A. A Genome-Wide Association Study for morphometric traits in Quarter Horse. **Journal of Equine Veterinary Science**. v. 34, p.1028-1031, 2014b.

MEIRA, C. T.; FORTES, M. R. S.; FARAH, M. M.; PORTO-NETO, R. L.; KELLY, M.; MOORE, S. S.; PEREIRA, G. L.; CHARDULO, L. A. L.; CURI, R. Speed Index in the Racing Quarter Horse: A Genome-wide Association Study. **Journal of Equine Veterinary Science**, v. 34 , n. 11 , p. 1263 – 1268, 2014c.

MUIR, W. M.; WONG, G. K.; ZHANG, Y.; WANG, J.; GROENEND M. A. M.; CROOIJMANS, R. P. M. A.; MEGENS, H. J.; ZHANG, H.; OKIMOTO, R.; VEREIJEN, A.; JUNGERS, A.; ALBERS, G. A. A.; LAWLEY, C. T.; DELAN, M. E.; MACEACHERNE, S.; CHENG, H. H. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. **Proceedings of the National Academy of Sciences of the United States of America**, v. 105, p. 17312-17317, 2008.

NAKAYA, H. L.; AMARAL, P. P.; LOURO, R.; LOPES, A.; FACHEL, A. A.; MOREIRA, Y. B.; EL-JUNDI, T. A.; SILVA, A. M.; REIS, E. M.; VERJOVSKI-ALMEIDA, S. Comment reviews reports deposited research refereed research interactions information Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. **Genome Biology**, v. 8, p. 1-25, 2007.

ORR, H. A. The genetic theory of adaptation: a brief history. **Nat. Rev. Genet**, v.6, p. 119–127, 2005.

PEREIRA, G.L.; DE MATTEIS, R.; MEIRA, C.T.; REGITANO, L.C.A.; SILVA, J.A.IV.; CHARDULO, L.A.L.; CURI, R.A. Comparison of Sequence Variants in the PDK4 and COX4I2 Genes Between Racing and Cutting Lines of Quarter Horses and Associations With the Speed Index. **Journal of Equine Veterinary Science**, v. 39, p. 1–6, 2016a.

PEREIRA, G.L.; DE MATTEIS, R.; REGITANO, L.C.A.; CHARDULO, L.A.L.; CURI, R.A. MSTN, CKM, and DMRT3 Gene Variants in Different Lines of Quarter Horses. **Journal of Equine Veterinary Science**, v. 39, p. 33-37, 2016b.

PETERSEN, J.L.; MICKELSON, J.R.; RENDAHL, A.K.; VALBERG, S.J.; ANDERSSON, L.S. et al. Genome-Wide Analysis Reveals Selection for Important Traits in Domestic Horse Breeds. **PLoS Genetics**, v.9, n.1, p. 1-17, 2013.

PORTO-NETO, L.; KIJAS, J.; REVERTER, A. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. **Genetics Selection Evolution**, v.46(2), 2014.

PRADO, R. S. A.; MOTA, M. D. S. Correlações genéticas entre deslocamento e conformação em equinos Mangalarga. **Archivos de Zootecnia**, v. 57, p. 165-169, 2008.

PRASAD, A.; SCHNABEL, R. D.; MCKAY, S.D.; MURDOCH, B.; STOTHARD, P.; KOLBEHDARI, D.; WANG, Z.; TAYLOR, J. F.; MOORE, S. S. Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. **Animal Genetics**, v. 39, p.597-605, 2008.

RONAGHI M, UHLEN M, NYREN P. A sequencing method based on real-time pyrophosphate. **Science**. v. 281, p.363–365, 1998.

SAHANA, G.; GULDBRANDTSEN, B.; BENDIXEN, C.; LUND, M. S. Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. **Animal Genetics**, v. 41, p. 579–588, 2010.

SANTOS, L. M. Morfologia e genética do cavalo Campolina. Belo Horizonte: Universidade Federal de Minas Gerais, 2006. 48p. Dissertação (Mestrado em Genética) – Universidade Federal de Minas Gerais, 2006.

SARGOLZAEI, M., CHESNAIS, J.P., SCHENKEL, F.S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v.15, p.478, 2014.

SCHRÖDER, W.; KLOSTERMANN, A.; STOCK, K.F.; DISTL, O. A genome-wide association study for quantitative trait loci of show-jumping in Hanoverian Warmblood horses. **Animal Genetics**, p.1-9, 2011.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature Biotechnology**, v. 26, p. 1135–1145, 2008.

SHIN, D-H.; LEE, J. W.; PARK, J-E.; CHOI, I-Y.; OH, H-S.; KIM, H. J.; KIM, H. Multiple Genes Related to Muscle Identified through a Joint Analysis of a Two-stage Genome-wide Association Study for Racing Performance of 1,156 Thoroughbreds. **Asian-Australasian Journal of Animal Sciences**, v. 28, n. 6, p. 771-781, 2015.

SIGNER-HASLER H., FLURY, C., HAASE, B., BURGER, D., SIMIANER, H., LEEB, T., RIEDER, S.A. Genome-wide association study reveals loci influencing height and other conformation traits in horses. **PloS One** 7, e37282., 2012.

STAIGER, E. A. et al. Skeletal variation in Tennessee Walking Horses maps to the LCORL/NCAPG gene region. **Physiological genomics**, v. 48, n. 5, p. 325-335, 2016b.

STAIGER, E. A.; ABRI, M. A.; SILVA, C. A. S.; BROOKS, S. A. Loci impacting polymorphic gait in the Tennessee Walking Horse. **Journal of Animal Science**, v.94, n. 4, p.1377-1386, 2016a.

STAIGER, E. A.; ALBRIGHT, J. D.; BROOKS, S. A. Genome-wide association mapping of heritable temperament variation in the Tennessee Walking Horse. **Genes, Brain and Behavior**, v. 15, n. 5, p. 514-526, 2016c.

STINCHCOMBE, J. R.; WEINIG, C.; HEATH, K. D.; BROCK, M. T.; SCHMITT, J. Polymorphic Genes of Major Effect: Consequences for Variation, Selection and Evolution in Arabidopsis Thaliana. **Genetics**, v.182, p.911, 2009.

TETENS, J.; WIDMANN, P.; KÜHN, C.; THALLER, G. A genome-wide association study indicates LCORL/NCAPG as a candidate locus for withers height in German Warmblood horses. **Animal Genetics**, v.44, p.467–471, 2013.

TEYSSÈDRE, S.; DUPUIS, M.C.; GUÉRIN, G.; SCHIBLER, L.; DENOIX, J.M.; ELSEN, J.M. et al Genome-wide association studies for osteochondrosis in French Trotters. **J Animal Science**, v.90, p.45-53, 2012.

TOZAKI, T.; MIYAKE, T.; KAKOI, H.; GAWAHARA, H.; SUGITA, S.; HASEGAWA, T. et al. A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene. **Animal Genetics**, v.41, p.28-35, 2010.

VENTER, J.C.; ADAMS, M.D.; MYERS, E.; LI, P.W.; MURAL, R.J.; SUTTON, G.G.; SMITH, H.O.; et al. The sequence of the human genome. **Science**, v. 291, p. 1304–1351, 2001.

WADE, C. M.; GIULOTTO, E.; SIGURDSSON, S.; ZOLI, M.; GNERRE, S.; IMSLAND, F.; LEAR, T. L.; ADELSON, D. L.; BAILEY, E.; BELLONE, R. R.; BLOCKER, H.; DISTL, O.; EDGAR, R. C.; GARBER, M.; LEEB, T.; MAUCALI, E.; MACLEOD, J. N.; PENEDO, M. C. T.; RAISON, M.; SHARPE, T.; VOGEL, J.; ANDERSSON, L.; ANTCHAK, D. F.; BIAGI, T.; BINNS, M. M.; CHOWDHARY, B. P.; COLEMAN, S. J.; DELLA VALLE, G.; FRYC, S.; GUERIN, G.; HASEGAWA, T.; HILL, E. W.; JURKA, J.; KIIALAINEN, A.; LINDGREN, G.; LIU, J.; MAGNANI, E.; MICKELSON, J. R.; MURRAY, J.; NERGADZE, S. G.; ONOFRIO, R.; PEDRONI, S.; PIRAS, M. F.; RAUDSEPP, T.; ROCCHI, M.; ROED, K. H.; RYDER, O. A.; SEARLE, S.; SKOWN, L.; SWINBURNE, J. E.; SYVANEN, A. C.; TOZAKI, T.; VALBERG, S. J.; VAUDIN, M.; WHITE, J. R.; ZODY, M. C.; Broad Institute Genome Sequencing Platform, Broad Institute Whole Genome Assembly Team; LANDER, E. S.; LINDBLAD-TOH, K. Genome sequence, comparative analysis, and population genetics of the domestic horse. **Science**, v.326, p.865-867, 2009.

Wallin L, Strandberg E, Philipsson J. Genetic correlations between field test results of Swedish Warmblood Riding Horses as 4-year-olds and life time performance results in dressage and show jumping. **Livestock Production Science**, v. 82, p. 61-71, 2003.

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.

ZENGER, K. R.; KHATKAR, M. S.; CAVANAGH J. A. L.; HAWKEN, R. J.; RAADSMA, H. W. Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. **Animal Genetics**, v. 38, p. 7-14, 2006.

CAPITULO 2 – Imputação e avaliação da acurácia em cavalos de corrida da raça Quarto de Milha genotipados com diferentes painéis comerciais de SNPs.

RESUMO – A confecção de painéis de genotipagem de SNPs em larga escala para equinos trouxe novas possibilidades para estudos genéticos na espécie. Em sua primeira versão, contava com cerca de 54.000 SNPs (54k). Atualmente, em sua segunda geração possui aproximadamente 65 mil SNPs (65k), dos quais 19 mil são novos marcadores e 45 mil estavam presentes no primeiro painel. Este trabalho teve como objetivos realizar a imputação de genótipos em duas vias entre indivíduos de uma amostra populacional relativamente pequena de cavalos de corrida da raça Quarto de Milha genotipados com painéis de 54k ou de 65k, bem como avaliar a acurácia de imputação por meio de simulações. A imputação foi realizada utilizando 116 cavalos genotipados com o arranjo de SNPs de 54k e 233 genotipados com arranjo de 65k. Nas simulações foram escolhidas amostras aleatórias para constituírem as populações imputadas e referências em dois cenários. O cenário A simulou a imputação genótipos na primeira via (65k para 54k) e o cenário B na segunda (54k para 65k). No cenário A foram considerados 113 indivíduos para a população referência e 236 para a imputada, dos quais 116 e 120 foram genotipados com os arranjos de 54k e 65k, respectivamente. No cenário B foram considerados 50 indivíduos para a população referência e 299 para a imputada, dos quais 66 e 233 foram genotipados com os arranjos de 54k e 65k, respectivamente. Em cada cenário, os indivíduos genotipados com o mesmo arranjo da população referência, mas que foram incluídos no grupo a ser imputado, tiveram seus genótipos exclusivos mascarados para posterior averiguação da acurácia de imputação. Com isso, após o controle de qualidade, os painéis de 54k e de 65k contaram com 7.048 e 16.940 marcadores exclusivos, respectivamente. As médias de taxa de concordância para os cenários A e B foram 0,9815 e 0,9751 e para r^2 alélico foram 0,9791 e 0,9740, respectivamente. Não foram observadas influências importantes dos coeficientes de parentesco genômicos entre as amostras a serem imputadas e as de referência sobre a acurácia de imputação. Desta forma, a imputação em duas vias de genótipos exclusivos entre os painéis equinos de 54k e 65k proporcionou aumento considerável de SNPs para todas as amostras (12 a 26%), sem que houvesse perdas na qualidade das informações.

Palavras-chave: cromossomos, estrutura de população, IBD, MAF, taxa de concordância

GENOTYPE IMPUTATION AND ACCURACY EVALUATION IN RACING QUARTER HORSES GENOTYPED USING DIFFERENT COMMERCIAL SNP PANELS

Abstract – The creation of large-scale SNP genotyping panels for horses has opened up new possibilities for genetic studies in the species. The first version contained about 54,000 SNPs (54k). The current second generation panel possesses approximately 65,000 SNPs (65k), 19,000 of them are new markers and 45,000 were present in the first panel. The objectives of this study were to perform two-step genotype imputation between individuals in a relatively small population sample of racing Quarter Horses genotyped with the 54k or 65k panel, and to evaluate the accuracy of imputation through simulations. Genotypes were imputed using 116 horses genotyped with the 54k SNP array and 233 animals genotyped with the 65k array. For the simulations, random samples were chosen to compose the imputed and reference populations in two scenarios. Scenario A simulated the genotype imputation in the first step (from 65k to 54k) and scenario B in the second step (from 54k to 65k). In each scenario, the individuals genotyped with the same panel as the reference population, but that were included in the imputed group, had their exclusive genotypes masked for subsequent evaluation of imputation accuracy. Thus, after quality control, the 54k and 65k panels contained 7,048 and 16,940 exclusive markers, respectively. The mean concordance rate was 0.9815 and 0.9751 for scenarios A and B, and the mean allelic r^2 was 0.9791 and 0.974, respectively. No important influences of the genomic relationship coefficients on imputation accuracy were observed between the imputed and reference samples. Thus, two-step imputation of exclusive genotypes between the 54k and 65k equine panels resulted in a considerable increase of SNPs for all samples (12 to 26%) without losses in the quality of information.

Keywords: concordance rate, chromosomes, IBD, MAF, population structure

1. Introdução

O surgimento de tecnologias de sequenciamento de nova geração (*Next Generation Sequencing* – NGS) possibilitou o sequenciamento e mapeamento de milhares de polimorfismos espalhados por todo o genoma equino (WADE et al., 2009; DOAN et al., 2012), o que permitiu a criação de painéis de genotipagem de SNPs em larga escala, abrindo novas possibilidades para estudos genéticos na espécie.

Em sua primeira versão, a plataforma de genotipagem específica para espécie, o *Equine SNP50 BeadChip* da empresa Illumina (Illumina Inc., EUA), contava com cerca de 54.000 SNPs (54k). Atualmente, em sua segunda geração, o SNP chip equino (*Equine SNP70 BeadChip*; Illumina Inc., EUA) possui aproximadamente 65 mil SNPs (65k), dos quais 19 mil são novos marcadores e 45 mil estavam presentes no *Equine SNP50 BeadChip*. A genotipagem de SNPs com os dois diferentes arranjos equinos em um mesmo experimento ou em experimentos diferentes pode levar à perda importante de dados. Entretanto, a imputação de genótipos de SNPs com base em informações de desequilíbrio de ligação (LD) populacional oferece potencial solução para a resolução deste problema, ou seja, permite a compatibilidade e a continuidade de pesquisas existentes.

De forma geral a imputação de genótipos possibilita combinar dados provenientes de amostras populacionais genotipadas com painéis de diferentes densidades (HAYES et al., 2011; LARMER et al., 2012), gerar painéis de alta densidade (*High Density* – HD) para animais genotipados com os de baixa densidade (*Low Density* – LD) (SARGOLZAEI et al., 2014; BOICHARD et al., 2012) e predizer genótipos faltantes do painel de genotipagem em decorrência da aplicação do controle de qualidade dos dados ou erros de leitura. Neste sentido, com o intuito de se obter maior sucesso em imputações de genótipos, diversas pesquisas têm estudado fatores que influenciam diretamente a acurácia de imputação, como: tamanho da população de referência (SARGOLZAEI et al., 2011, VENTURA et al., 2016), grau de parentesco entre as populações referência e imputada (CHUD et al., 2015; VENTURA et al., 2016), densidade de painéis (ZHANG et al., 2010; SARGOLZAEI et al., 2011) e alelos de baixa frequência (*Minor Allele Frequency* – MAF) (ZHANG et al., 2010; SARGOLZAEI et al. 2014; VENTURA et al., 2016). Contudo, a imputação em situações reais entre amostras de populações genotipadas com painéis distintos, pode trazer desafios à inferência da acurácia. Deste modo, algumas análises e simulações podem ser realizadas em um conjunto amostral com o intuito de direcionar a escolha de estratégias

mais eficientes para imputação, bem como para obter parâmetros de confiabilidade de imputações realizadas sob essas circunstâncias.

Dentre as raças equinas, a Quarto de Milha é das mais criadas no mundo e se destaca por sua versatilidade, robustez e temperamento dócil, sendo muito utilizada no trabalho com o gado, em provas equestres funcionais e de velocidade (ABQM, 2016). Como os mais velozes do mundo, cavalos da linhagem de corrida da raça Quarto de Milha apresentam melhor desempenho em pistas de curtas distâncias do que qualquer outra raça de equinos (AMERICA'S HORSE DAILY, 2008).

Dada a importância da linhagem de corrida da raça Quarto de Milha na equideocultura mundial, este trabalho teve como objetivos realizar a imputação de genótipos em duas vias entre indivíduos de uma amostra populacional relativamente pequena genotipados com painéis de 54k ou de 65k, bem como avaliar a acurácia de imputação por meio de simulações.

2. Material e Métodos

2.1. Animais

Foram utilizados 360 equinos da linhagem de corrida da raça Quarto de Milha, de ambos os sexos, registrados na Associação Brasileira de Criadores de Cavalo Quarto de Milha (ABQM). Deste total, 120 cavalos tiveram o sangue coletado no ano de 2011. Os demais animais utilizados, n= 240, foram coletados no primeiro semestre do ano de 2015. Estes animais, 78 machos e 282 fêmeas, são filhos de 83 garanhões e 249 éguas, resultando em média de 4,3 progênies/garanhão e 1,4 progênies/égua. Estes animais eram pertencentes a 159 criadores e encontravam-se alojados no Jockey Club de Sorocaba (Sorocaba/SP – Brasil) e em 25 propriedades rurais localizadas em cidades do interior estado de São Paulo/Brasil. A coleta e, conseqüentemente, a presença de irmãos completos na amostra foi evitada. Todos os procedimentos envolvendo os animais foram realizados segundo as normas brasileiras de bem-estar animal (Protocolo nº 157/2014 – Comissão de Ética no Uso de Animais / CEUA, FMVZ, Unesp, Botucatu/SP).

2.2. Genotipagem dos SNPs e controle de qualidade

Dos 360 cavalos utilizados neste estudo, 120 foram genotipados no ano de 2011 utilizando-se o *Equine SNP50 BeadChip* (Illumina, Inc., EUA). As demais amostras de

DNA (n= 240) foram genotipadas no ano de 2015 com o *Equine SNP70 BeadChip* (Illumina, Inc., EUA). A leitura de ambos os arranjos foi realizada utilizando-se o sistema *HiScan* (Illumina, Inc., EUA). Os controles de qualidade (*quality control* – QC) das informações de genotipagem para os indivíduos analisados com os arranjos de 54k ou de 65k (n=120 e n= 240) foram realizados por meio do pacote *snpStats* (Clayton, 2015) do programa *R* (R Core Team, 2016). Animais com *call rate* abaixo de 0,90 foram excluídos do conjunto amostral (Tabela 1). Foram excluídos SNPs localizados no cromossomo X, os com *call rate* inferior a 0,9 e os com *p-value* inferior a 1×10^{-5} para o equilíbrio de *Hardy-Weinberg* (*Hardy-Weinberg* Equilibrium – HWE) (Tabelas 1). A MAF não foi utilizada como critério para exclusão de SNPs em um primeiro momento tendo em vista que diferentes faixas de MAF foram utilizadas no processo de verificação da eficiência de imputação de genótipos em duas vias entre os chips equinos SNP50 e SNP70.

Tabela 1 – Número de marcadores dos chips *Equine SNP50 BeadChip* e *Equine SNP70 BeadChip* e número de SNPs excluídos por cada critério de QC utilizados.

Critérios utilizados para exclusão de amostras	SNP50 (n)	SNP70 (n)
<i>Call rate</i> < 0,9	4	7
Amostras restantes	116	233
Critérios utilizados para exclusão de SNPs	SNP50 (n)	SNP70 (n)
SNPs genotipados	54.602	65.157
Localizados no cromossomo X	3.223	3.411
<i>Call rate</i> < 0,9	148	504
Equilíbrio de <i>Hardy-Weinberg</i> – $p < 1 \times 10^{-5}$	4.519	4.638
SNPs restantes	46.712	56.604

2.3. Análise populacional

Análise de estratificação da população foi conduzida com todos os SNPs comuns aos dois painéis que passaram pelo controle de qualidade foram submetidos à *pruning* para seleção de marcadores que não estavam em LD, considerando $r^2 < 0,2$. As probabilidades de compartilhar marcadores idênticos por estado (*identity-by-state* – IBS) entre todas as amostras par a par foram calculadas utilizando os SNPs autossômicos que permaneceram após o *prune*. Por fim, para análise visual, foi obtido o gráfico de

escalonamento multidimensional por meio da opção *MDS plot* do programa PLINK 1.07 (PURCELL et al., 2007).

Os coeficientes de parentesco, ou seja, as probabilidades de indivíduos compartilharem marcadores IBD foram calculadas para todos os pares de amostras considerando apenas os SNPs comuns entre os dois painéis. Marcadores idênticos por descendência (*identity-by-descent* – IBD) podem ser utilizados para identificar indivíduos mais próximos do que seria esperado em uma amostra homogênea, ou seja, aparentados (PURCELL et al., 2007). As estimativas IBD foram dadas por Z0, Z1 e Z2, que representam a probabilidade de duas amostras compartilharem nenhum marcador IBD, um marcador IBD e dois marcadores IBD, respectivamente. A proporção de IBD entre amostras foi dada por $PI_HAT = p(Z2) + 0,5 \times p(Z1)$.

Os procedimentos de *pruning* e a obtenção das estimativas IBS e IBD foram realizados por meio do programa PLINK 1.07 (PURCELL et al., 2007). As médias da proporção IBD de cada amostra, bem como as médias de parentesco entre as amostras do conjunto imputadas (as quais tiveram marcadores mascarados) e do conjunto referência foram obtidas por meio do programa R (R Core Team, 2016).

2.4. Imputação de genótipos e avaliação da acurácia de imputação

O procedimento de imputação dos genótipos foi realizado por meio do programa FImpute v.2.2 (SARGOLZAEI et al., 2014). Como descrito anteriormente, os indivíduos utilizados neste estudo foram genotipados de forma exclusiva com um ou outro painel. Dessa forma, a imputação de genótipos final foi realizada em duas vias (i.e. do chip de 65k para o chip de 54k e vice-versa).

Pelo fato de não haver no conjunto amostral animais genotipados com ambos os painéis, a avaliação da acurácia do processo de imputação de marcadores entre os painéis foi estimada por meio de simulações envolvendo diferentes cenários. Foram propostos dois cenários de imputação visando avaliar a capacidade de imputar corretamente os SNPs exclusivos em cada painel. Primeiramente, foi utilizado o painel de 65k como referência e o de 54k como imputado (cenário A). Para isso foram considerados todos os marcadores presentes no painel de 65k e somente os marcadores em comum do de 54k. Para acessar a eficiência de imputação dos SNPs exclusivos do SNP70 foram escolhidas de forma aleatória, aproximadamente 50% das amostras genotipadas com esse chip. As demais amostras genotipadas com o SNP70 tiveram seus marcadores exclusivos mascarados e foram incorporadas às amostras a ser imputadas genotipadas com painel

de 54k. O mesmo procedimento foi realizado na segunda via (cenário B), em que houve somente a inversão dos painéis utilizados como referência e imputado.

Para determinar a acurácia de imputação foi utilizada a taxa de concordância (*concordance rate* – CR), a qual corresponde à proporção de genótipos imputados corretamente. Também foi estimado o r^2 alélico (correlação alélica), o qual é determinado pelo quadrado da correlação entre a contagem de alelos (alelo de efeito menor) imputados e a contagem de alelos do genótipo original. Os valores de acurácia foram obtidos tanto por SNPs (médias por intervalos de MAFs e cromossomos) quanto por amostras.

3. Resultados e discussão

3.1. Estrutura da população e análise de parentesco

A análise de estratificação da população foi realizada para verificar a existência de *clusters* constituídos, exclusivamente, de indivíduos genotipados com o mesmo chip dentro da amostra total de animais. No gráfico MDS (Figura 1) foram observadas subestruturas distintas, contudo, nenhuma delas envolvendo apenas amostras genotipadas com um ou outro arranjo nas regiões mais densas do gráfico.

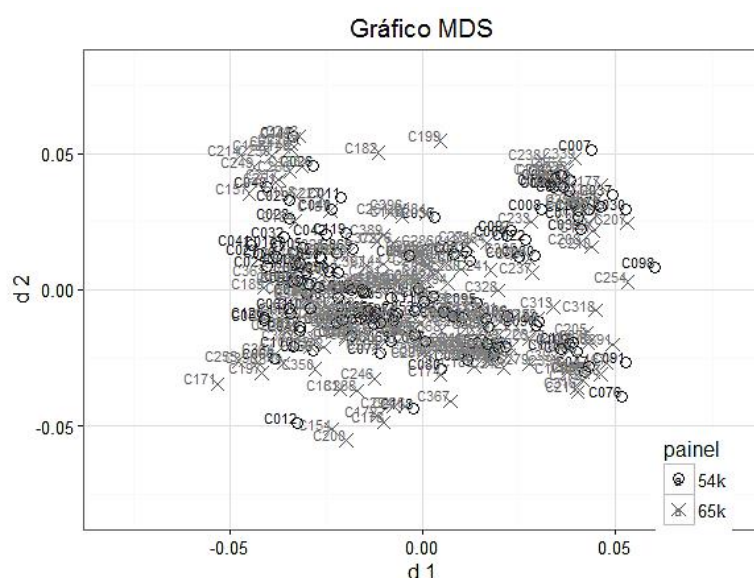


Figura 1 – MDS plot do conjunto total de amostras considerando os 39.664 SNPs comuns a ambos os arranjos equinos. Os círculos e os losangos correspondem a animais genotipados com os arranjos de 54k e 65k, respectivamente.

Este resultado indicou que não há formação de *clusters* exclusivos de animais genotipados com o painel de 54k ou com o de 65k, ou seja, que as amostras genotipadas com os painéis de menor e maior densidade não representam duas populações distintas. A ocorrência de populações distintas entre referências e imputados poderia acarretar em baixos valores de acurácia, ou seja, genótipos imputados não confiáveis.

As estimativas IBD para a determinação do coeficiente de parentesco médio com base em marcadores SNPs foi dada pela probabilidade média de conter um ou dois marcadores IBD (*IBD proportion*). De forma geral, as estimativas de parentesco genômico médio para cada par de indivíduos concordaram com as informações do pedigree. Animais com proporção IBD acima da média (0,0962) foram os que apresentaram maior grau de parentesco (1º e 2º grau) em relação a todos os indivíduos da população, sendo geralmente avós e pais. O parentesco médio e máximo da população imputada em relação à de referência foi maior no cenário B se comparado ao cenário A, contudo o parentesco mínimo foi maior em A (Tabela 2). Os coeficientes de parentesco médio e máximo encontrados para as populações a serem imputadas em cada cenário foram semelhantes aos encontrados para o conjunto amostral total (Tabela 2).

Tabela 2 – Coeficientes mínimo, médio e máximo de parentesco obtidos por meio de matriz IBD para todas as amostras e para amostras imputadas em cada via (54k para 65k e 65k para 54k)

Amostras	Mínimo	Médio	Máximo
Todas	0,00360	0,09620	0,21100
65k para 54k* (cenário A)	0,02072	0,09909	0,18300
54k para 65k* (cenário B)	0,01449	0,10030	0,21890

*Médias IBD obtidas para amostras imputadas em relação às de referência

3.2. Acurácia de imputação por SNP

A imputação de SNPs em duas vias trouxe aumento considerável de informação a cada painel. Na Tabela 3 estão descritos os números de SNPs compartilhados e exclusivos de cada painel, bem como o número total de SNPs após a imputação em duas vias. O ganho em número de marcadores em relação ao painel imputado final (64k) foi de 26% para o SNP50, ultrapassando o número original de SNPs, e 12% para o SNP70.

Tabela 3 – Número de marcadores dos chips *Equine SNP50 BeadChip* e *Equine SNP70 BeadChip* e comuns aos dois arranjos após o controle de qualidade e número de marcadores remanescentes após o processo de imputação em duas vias.

Painéis	SNPs após QC (n)
SNP50	46.712
*Exclusivos SNP50	7.048
SNP70	56.604
*Exclusivos SNP70	16.940
Painel comun (SNP50 e SNP70)	39.664
Painel após a imputação	63.652

*SNPs que foram imputados para o outro painel

Para o cenário A (do painel de 65k para o de 54k) foram considerados 113 indivíduos para a população referência e 236 para a imputada, dos quais 116 e 120 foram genotipados com os arranjos de 54k e 65k, respectivamente. Já para o cenário B (do painel de 54k para o de 65k) foram considerados 50 indivíduos para a população referência e 299 para a imputada, dos quais 66 e 233 foram genotipados com os arranjos de 54k e 65k, respectivamente. A utilização de aproximadamente 50% das amostras como referências e 50% como imputadas, permitiu manter número satisfatório de indivíduos tanto para o processo de imputação quanto para o procedimento de avaliação de acurácia. Para as populações imputadas, somente os genótipos comuns aos dois painéis foram considerados (39.664). Em cada cenário, os indivíduos genotipados com o mesmo arranjo da população referência, mas que foram incluídos no grupo a ser imputado, tiveram seus genótipos exclusivos mascarados para posterior averiguação da acurácia de imputação. Com isso, após o QC, os painéis de 54k e de 65k contaram com 7.048 e 16.940 marcadores exclusivos, respectivamente.

As médias de CR para os cenários A e B foram $0,9815 \pm 0,02882$ e $0,9751 \pm 0,03731$, e para r^2 alélico foram $0,9791 \pm 0,04209$ e $0,9740 \pm 0,04911$, respectivamente. As acurácias de imputação na via 65k para 54k foram maiores que na via oposta. Também foram analisadas as acurácias de imputação em cada cenário pela média de CR e r^2 alélico para cada intervalo de MAF.

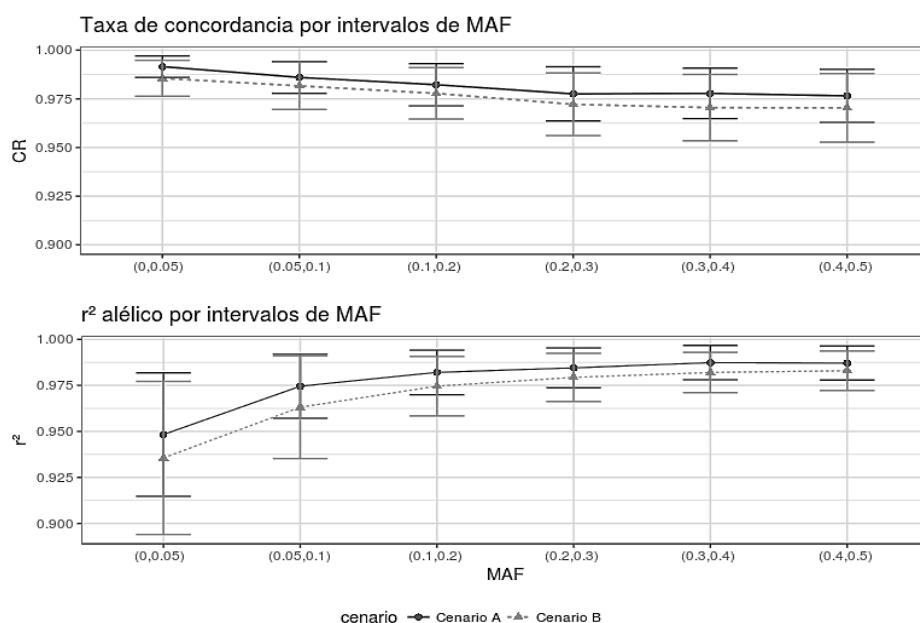


Figura 2 – Eficiência de imputação em diferentes cenários e em diferentes intervalos de MAF (eixo x) dado pelas médias de taxa concordância (CR) e r^2 alélico (eixo y). Cada tipo de linha e ponto representa um cenário diferente de imputação.

Como esperado para CR, as maiores acurácias foram obtidas para os menores intervalos de MAF (Figura 2). A correlação alélica (r^2) tem sido comumente utilizada para estimar a acurácia de imputação para alelos com MAF baixas, por minimizar a dependência da frequência alélica (SARGOLAZEI et al., 2014). Contudo, o cálculo de r^2 dado por SNPs (e não por amostras) requer grande quantidade de animais para compor o conjunto imputado, de modo a obter estimativas não viesadas (VENTURA et al., 2016). Ainda, segundo Browning e Browning (2009), é necessário que o tamanho da amostra seja suficientemente grande para conter número mínimo de alelos com baixa MAF a serem imputados, caso contrário valores de r^2 poderiam ser estimados com erros-padrão muito elevados. A partir disso, as estimativas de correlação alélica podem ser pouco confiáveis para cenários de imputação que utilizam poucos animais, como os apresentados neste trabalho.

Por sua vez, o CR não é afetado pelo tamanho da população. Contudo, como relatado por MA et al. (2013) a taxa de concordância tende a ser superestimada, principalmente quando se trata de alelos raros ou de MAF baixa. Outros fatores que podem afetar a taxa de concordância seriam a composição genética entre amostras referências e imputadas, bem como a densidade de marcadores dos arranjos a ser imputados. Em trabalho conduzido por Chud et al. (2015), utilizando bovinos de corte da mesma raça, a CR variou de 0,75 a 0,97 aproximadamente, entre imputações feitas de

arranjo de 3k para HD (777k) e de arranjo de 80k para HD, respectivamente. Por outro lado, números demasiadamente baixo de marcadores (1-2k) a serem imputados podem diminuir ligeiramente a acurácia de imputação como visto por Ogawa et al. (2015) em bovinos. Estimativas muito baixas de CR em intervalos de MAFs inferiores a 0,1 podem revelar baixa confiabilidade dos genótipos imputados tendo em vista que seriam, em geral, valores superestimados.

3.3. Acurácia de imputação por amostra

Com relação às amostras imputadas, as acurácias foram obtidas por meio de médias de CR e r^2 alélico (Tabela 4). Posteriormente, com o intuito de investigar o impacto do parentesco sobre a acurácia de imputação, foram obtidas acurácias para subconjuntos divididos de acordo com as médias IBD para as amostras imputadas obtidas em cada cenário, em que os indivíduos com coeficiente de parentesco acima da média foram classificados como de alto parentesco e os abaixo da média como de baixo parentesco (Tabela 4). Devido ao número reduzido de amostras no presente estudo, a criação de cenários extras com base no número de indivíduos que compõem cada população (referência ou imputada) não seriam relevantes.

Não foram observadas diferenças significativas entre as médias de CR e r^2 alélico para as amostras imputadas em ambos os cenários, bem como para as de alto e baixo parentesco. Nota-se que as médias de CR e de r^2 alélico para amostras de alto parentesco tiveram menores desvios-padrão, se comparado as de baixo parentesco, tanto para o cenário A, quanto para o B. No entanto, no cenário B foram observados desvios-padrão maiores, possivelmente relacionados à estrutura de parentesco entre as amostras imputadas e referências e o menor número de amostras presentes.

Tabela 4 – Acurácias de imputação dadas por taxa de concordância (CR) e r^2 alélico considerando todas as amostras imputadas, as amostras com parentesco acima da média IBD e abaixo da média IBD para os cenários A e B em equinos de corrida da raça Quarto de Milha.

Cenário	Via	Pop. Imputada	Médias		Min.	Max.	Mediana
			de CR	DP ¹			
A	65k para 54k	Todos (n=120)	0,9814	0,01173	0,9152	0,9989	0,9851
		Alto parentesco (n=56)	0,9841	0,00769	0,9671	0,9952	0,9864
		Baixo parentesco (n=58)	0,9794	0,01013	0,9534	0,9961	0,9813
B	54k para 65k	Todos (n=66)	0,9751	0,01362	0,9408	0,9949	0,9774
		Alto parentesco (n=30)	0,9801	0,01108	0,9536	0,9949	0,9825
		Baixo parentesco (n=34)	0,9704	0,01412	0,9408	0,9911	0,9738

Cenários	Via	Pop. Imputada	Médias		Min.	Max.	Mediana
			de r^2	DP			
A	65k para 54k	Todos (n=120)	0,9868	0,00779	0,9538	0,9997	0,9894
		Alto parentesco (n=56)	0,9884	0,00631	0,9733	0,9967	0,9904
		Baixo parentesco (n=58)	0,9848	0,00789	0,9643	0,9974	0,9872
B	54k para 65k	Todos (n=66)	0,9811	0,01098	0,9527	0,9968	0,9828
		Alto parentesco (n=30)	0,9849	0,00930	0,9630	0,9968	0,9870
		Baixo parentesco (n=34)	0,9775	0,01127	0,9527	0,9943	0,9792

¹Desvio-padrão

No cenário A (65k para 54k) não foi observada influência do parentesco sobre a acurácia de imputação ($p > 0,1$). Contudo, no cenário B animais com maiores coeficientes médios de parentesco em relação a população referência obtiveram maior acurácia de imputação ($p < 0,01$) (Figura 3).

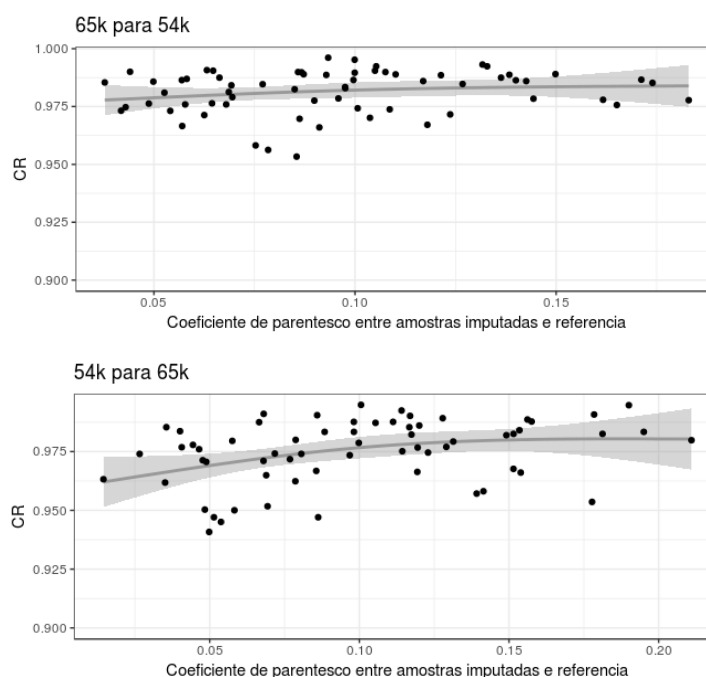


Figura 3 – Regressão da taxa de concordância (CR) (eixo y) e o coeficiente de parentesco dados pelas medias IBD entre amostras da população de imputação e referência (eixo x) em dois cenários em equinos de corrida da raça Quarto de Milha.

Os coeficientes de determinação (R^2) foram baixos para ambas as situações (0,029 e 0,113 para os cenários A e B, respectivamente), indicando que, neste estudo, a acurácia de imputação não foi fortemente influenciada pelo grau de parentesco entre os indivíduos das populações imputadas e referências. Diferentemente dos resultados encontrados aqui, estudos conduzidos por Chud et al. (2015) mostraram forte relação do grau de parentesco dos indivíduos da população com a acurácia de imputação. Embora o software FImpute combine metodologias de imputação (família e população), o método baseado em população difere da maioria dos outros softwares, pois assume que todos os animais estão relacionados entre si com algum grau de parentesco. Segundo Kong et al. (2008), a informação de pedigree é fator importante para a identificação da fase dos haplótipos e a imputação. Contudo, ganhos significativos de acurácia têm sido obtidos apenas em situações em que indivíduos genotipados com diferentes painéis são geneticamente próximos e as imputações ocorrem de painéis de baixa densidade (384, 3k ou 6k) para os de alta (HD) (SARGOLZAEI et al., 2014).

Pesquisas que avaliaram o impacto da informação de parentesco em imputações de genótipos em bovinos consideraram milhares de animais e varias gerações completas em seus pedigrees (SARGOLZAEI, et al., 2014; PICCOLI et al., 2014; VENTURA et al., 2016).

No caso aqui estudado, a inclusão da informação de pedigree seria irrelevante, levando em consideração o pequeno número de SNPs a ser imputados em relação aos que as amostras genotipadas possuem em comum e o conjunto amostral total ser relativamente pequeno. Além disso, a não disponibilização de forma sistematizada de arquivos de parentesco por parte das associações de criadores, bem como a ocorrência de erros de registro, poderiam acarretar perdas de acurácia no processo de imputação.

O fato de que os animais aqui estudados são relativamente próximos geneticamente, sobretudo se forem tomados o parentesco mínimo entre as amostras imputadas e as de referência de 0,02072 para o cenário A e de 0,01449 para o cenário B (Tabela 3), indicando grau importante de parentesco mesmo entre os indivíduos menos aparentados, contribuiu para a imputação de genótipos com acurácia elevada.

3.4. Acurácia de imputação por cromossomo

As médias de CR por cromossomo variaram de 0,9505 a 0,9845 e 0,9592 a 0,9877, e as de r^2 alélico de 0,9536 a 0,9840 e 0,9544 a 0,9862 para os cenários B e A, respectivamente (Figura 4). Foram observadas valores baixos de CR e de r^2 para ECA 3, 12, 13 e 25 no cenário B, e para ECA 6 e 12 no cenário A. Os valores médios de CR e de r^2 encontrados para cada cromossomo são semelhantes aos relatados por PICCOLI et al. (2014) e VENTURA et al. (2016). Entretanto, as acurácias médias para cada cromossomo não seguiram os padrões descritos por Piccoli et al. (2014) em estudo conduzido com bovinos de corte. Nesse estudo, o comprimento cromossômico foi associado com a precisão da imputação, em que a maior CR foi encontrada para o cromossomo 1 e a menor para o 28, respectivamente o maior e o menor cromossomo bovino.

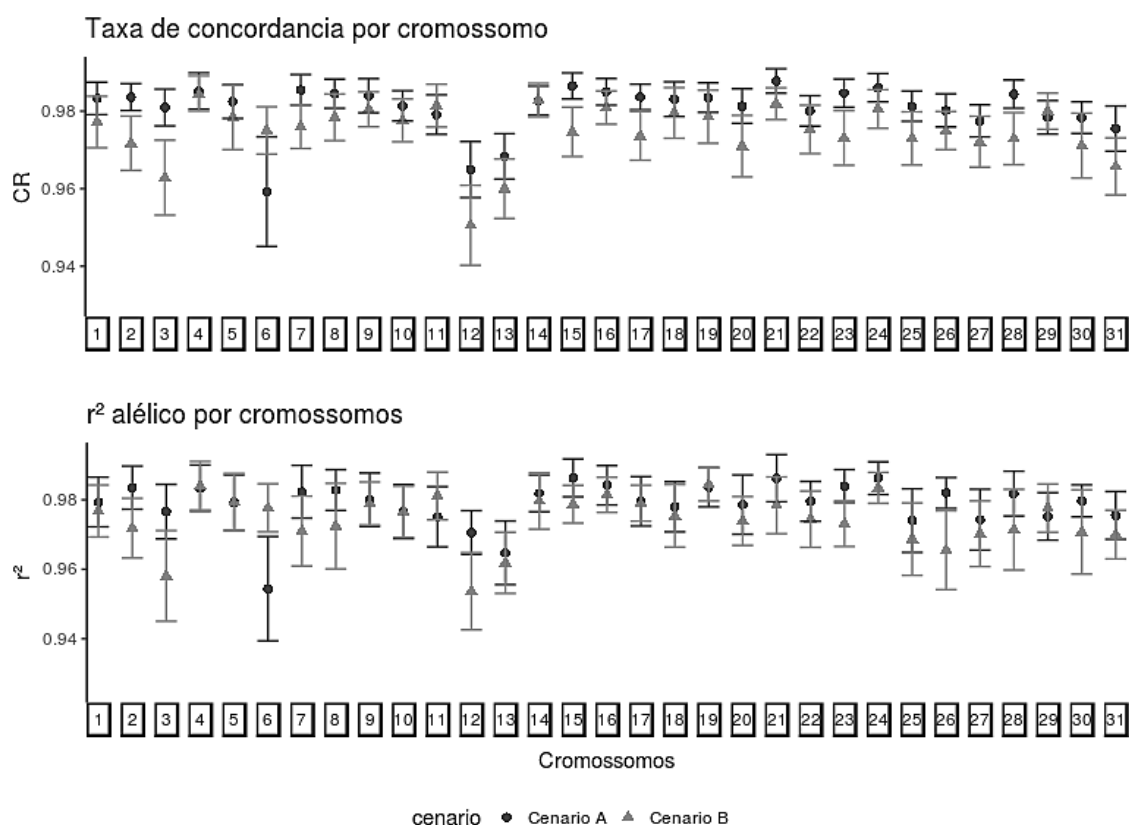


Figura 4 – Acurácia de imputação em diferentes cenários para cada cromossomo autossomo equino (31) (eixo x) dado pelas médias de concordância e r^2 alélico (eixo y) entre genótipos reais e imputados. Cada tipo de ponto representa um cenário de imputação diferente.

A figura 5 mostra a *plotagem* dos CR para cada SNP em cada cromossomo. No cenário A nota-se no ECA 6 (menor média de CR para esse cenário) baixa acurácia de imputação em uma grande porção na região terminal do cromossomo e para ECA 12 e 13 em região central. O cenário B mostrou, em geral, maior dispersão dos valores de acurácia de imputação em praticamente todos os cromossomos, principalmente a partir do ECA 17. Observou-se baixa acurácia de imputação em porções inicial e final do ECA 3 e central para ECA 12 (cromossomos que obtiveram menores médias de CR para esse cenário).

Ventura et al. (2016) analisaram regiões proximais e teloméricas (utilizando 100 SNPs para cada extremidade) em cada um dos 26 cromossomos autossômicos de ovinos. Esses autores constataram que a maioria dos cromossomos apresentaram problemas em pelo menos uma das extremidades, com exceção do cromossomo 26, que apresentou acurácia superior a 60% em ambas as extremidades. Tal fato não foi observado aqui, em que SNPs com baixa acurácia de imputação foram distribuídos por toda a extensão dos cromossomos.

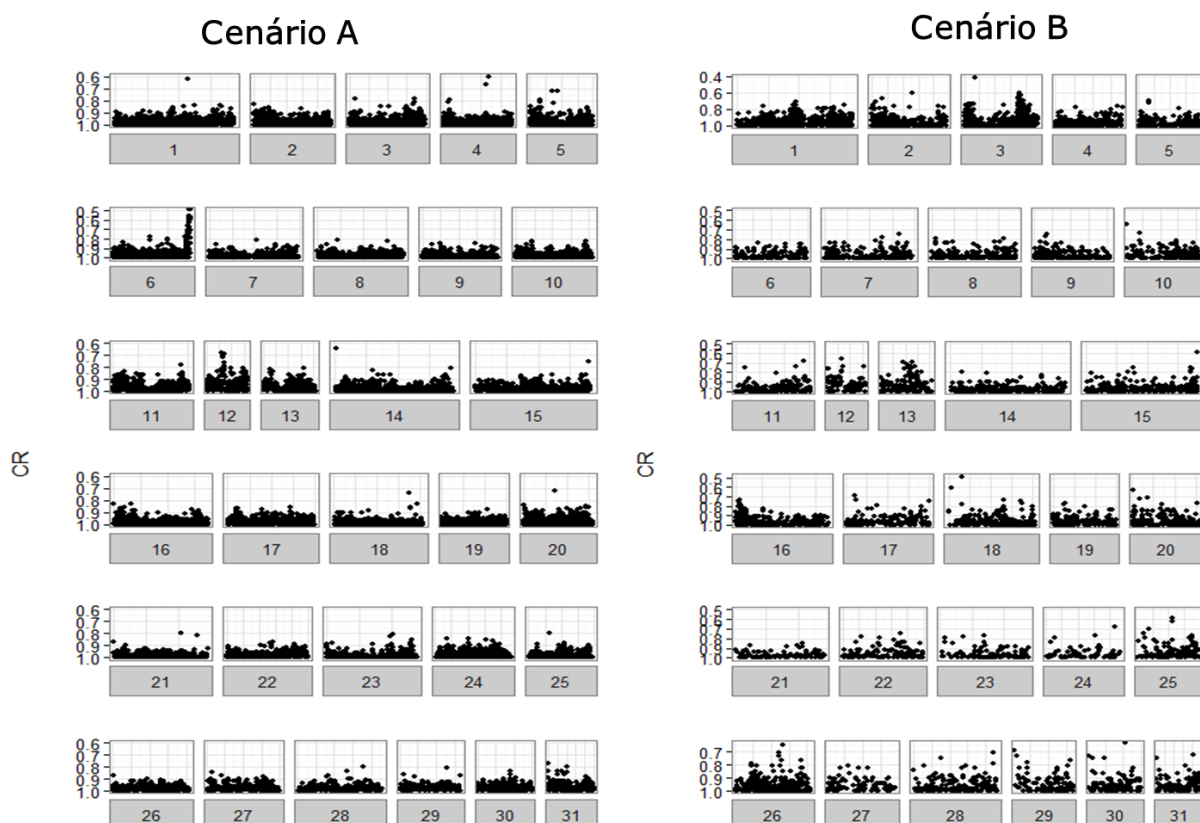


Figura 5 – Acurácia de imputação dado por CR (eixo y) pra cada SNP e cromossomo (eixo x) em dois cenários (A e B) em equinos de corrida da raça Quarto de Milha.

4. Conclusões

O ganho de informação de genótipos com qualidade para todas as amostras de cavalos de corrida da raça Quarto de Milha utilizadas neste estudo, proporcionado pela imputação em duas vias (do painel de 65k para o de 54k e do de 54k para o de 65k), foi considerável, sobretudo por se tratar de estudo com tamanho amostral reduzido, situação comum em pesquisas realizadas com equinos quando comparadas com pesquisas desenvolvidas com outras espécies de interesse zootécnico, como, por exemplo, a bovina.

As estratégias de imputação e avaliação de acurácia adotadas no presente trabalho foram eficazes e poderão ser úteis em situações semelhantes às aqui encontradas, ou seja: tamanho amostral relativamente pequeno; animais genotipados exclusivamente com um ou outro arranjo; inconsistências nas informações de genealogia, porém com indivíduos geneticamente próximos.

5. Referências

ABQM. Associação brasileira dos criadores de cavalos Quarto de Milha. Disponível em: <<http://www.abqm.com.br/item1.htm>>, acesso em: 15 mar. 2016

AMERICA'S HORSE DAILY. All About the Racing American Quarter Horse. 26 de fevereiro de 2008. Disponível em: <http://americashorsedaily.com/all-about-the-racing-american-quarter-horse/#.Urx8N_RDvik>, acesso em: 12 nov. 2013.

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, n. 3, 2012. Disponível em: <doi: 10.1371/journal.pone.0034130>

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.

CHUD, T.C.S; VENTURA, R.V; SCHENKEL, F.S; CARVALHEIRO, R.; BUZANSKAS, M.E; ROSA, J.O; MUDADU, M.A; DA SILVA, M.V.G.B.; MOKRY, F.B; MARCONDES, C.R; REGITANO, L.C.A; MUNARI, D.P. Strategies for genotype imputation in composite beef cattle. **BMC Genetics**, v. 16, p. 99, 2015.

DOAN, R.; COHEN, N.D.; SAWYER, J.; GHAFARI, N.; JOHNSON, C.D.; DINDOT, S.V. Whole-Genome Sequencing and Genetic Variant Analysis of a Quarter Horse Mare. **BMC Genomics**, v.13, n.78, p. 1471-2164, 2012.

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.

KONG, A.; MASSON, G.; FRIGGE, M. L.; GYLFASSON, A.; ZUSMANOVICH, P.; THORLEIFSSON, G. Detection of sharing by descent, long-range phasing and haplotype imputation. **Nature Genetics**, v. 40, p. 1068-1075, 2008.

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.

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.

MEIRA, C. T.; FORTES, M. R. S.; FARAH, M. M.; PORTO NETO, L.; KELLY, M.; MOORE, S. S.; PEREIRA, G. L.; CHARDULO, L. A. L.; CURI, R. A. Speed index in the racing Quarter Horse: a genome-wide association study. **Journal of Equine Veterinary Science**, v. 34, p. 1263-1268, 2014a.

MEIRA, C.T.; CURI, R.A.; FARAH, M.M.; OLIVEIRA, H.N.; BELTRAN, N.A.R.; SILVA, J.A.IV.; MOTA, M.D.S. Prospection of genomic regions divergently selected in racing line of Quarter Horses in relation to cutting line. **Animal**, v. 8, n. 11, p. 1754-1764, 2014b.

OGAWA, S.; MATSUDA, H.; TANIGUCHI, Y.; WATANABE, T.; SUGIMOTO, Y.; IWAISAKI, H. Estimation of variance and genomic prediction using genotypes imputed from low-density marker subsets for carcass traits in Japanese black cattle. **Animal Science Journal**, v. 87, n. 1, p. 3–12, 2016.

PEREIRA, G.L.; DE MATTEIS, R.; MEIRA, C.T.; REGITANO, L.C.A.; SILVA, J.A.IV.; CHARDULO, L.A.L.; CURI, R.A. Comparison of Sequence Variants in the PDK4 and COX4I2 Genes Between Racing and Cutting Lines of Quarter Horses and Associations With the Speed Index. **Journal of Equine Veterinary Science**, v. 39, p. 1-6, 2016a.

PEREIRA, G.L.; DE MATTEIS, R.; REGITANO, L.C.A.; CHARDULO, L.A.L.; CURI, R.A. MSTN, CKM, and DMRT3 Gene Variants in Different Lines of Quarter Horses. **Journal of Equine Veterinary Science**, v. 39, p. 33-37, 2016b.

PICCOLI, M.; BRACCINI, J.; CARDOSO, F.; SARGOLZAEI, M.; SCHENKEL, F. Impact of imputation in Braford and Hereford beef cattle. **BMC Genetics**, v. 15, p. 157, 2014.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M.A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P.I.; DALY, M.J.; SHAM, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics, Baltimore**, v. 81, p. 559-575, 2007.

R CORE TEAM. R: A language and Environment for Statistical Computing. R foundation for Statistical Computing 2016. <<http://www.R-project.org/>>.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F.S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v. 15, p. 478, 2014

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.

VENTURA, R.V.; MILLER, S.P.; DODDS, K.G.; AUVRAY, B.; LEE, M.; BIXLEY, M.; CLARKE, S.M.; MCEWAN, J.C. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. **Genetics, Selection, Evolution : GSE**, v. 48, p. 71, 2016.

WADE, C. M.; GIULOTTO, E.; SIGURDSSON, S.; ZOLI, M.; GNERRE, S.; IMSLAND, F.; LEAR, T. L.; ADELSON, D. L.; BAILEY, E.; BELLONE, R. R.; BLOCKER, H.; DISTL, O.; EDGAR, R. C.; GARBER, M.; LEEB, T.; MAUCELI, E.; MACLEOD, J. N.; PENEDO, M. C. T.; RAISON, M.; SHARPE, T.; VOGEL, J.; ANDERSSON, L.; ANTCZAK, D. F.; BIAGI, T.; BINNS, M. M.; CHOWDHARY, B. P.; COLEMAN, S. J.; DELLA VALLE, G.; FRYC, S.; GUERIN, G.; HASEGAWA, T.; HILL, E. W.; JURKA, J.; KIIALAINEN, A.; LINDGREN, G.; LIU, J.; MAGNANI, E.; MICKELSON, J. R.; MURRAY, J.; NERGADZE, S. G.; ONOFRIO, R.; PEDRONI, S.; PIRAS, M. F.; RAUDSEPP, T.; ROCCHI, M.; ROED, K. H.; RYDER, O. A.; SEARLE, S.; SKOWN, L.; SWINBURNE, J. E.; SYVANEN, A. C.; TOZAKI, T.; VALBERG, S. J.; VAUDIN, M.; WHITE, J. R.; ZODY, M. C.; Broad Institute Genome Sequencing Platform, Broad Institute Whole Genome Assembly Team; LANDER, E. S.

LINDBLAD-TOH, K. Genome sequence, comparative analysis, and population genetics of the domestic horse. **Science**, v. 326, p.865-867, 2009.

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.

CAPITULO 3 – Estudo de associação ampla genoma em equinos de corrida da raça Quarto de Milha

RESUMO – Equinos de corrida raça Quarto de Milha têm o melhor desempenho em corridas de curtas distâncias do que qualquer outra raça. Apesar de, no Brasil, o efetivo de animais ser relativamente menor na linhagem de corrida do que nas demais, sua importância econômica é substancial. Tendo em vista o interesse econômico e científico relacionado a esta característica atlética, poucos esforços têm sido realizados para a maior compreensão de seus mecanismos genéticos e fisiológicos. O objetivo deste trabalho foi realizar estudo de associação ampla do genoma (GWAS) em cavalos da linhagem de corrida da raça Quarto de Milha por meio da utilização de chips equinos de genotipagem de SNPs, visando a prospecção de regiões cromossômicas, genes e polimorfismos relacionados ao desempenho. Para isso, foram utilizadas inicialmente informações genotípicas e fenotípicas de 360 equinos genotipados com chips de 54k ou de 65k. O GWAS foi realizado com base no método *single step* GBLUP por meio de duas abordagens: ssGWAS1, em que somente efeitos de SNPs são reestimados a cada iteração, e ssGWAS2, em que a cada iteração são reestimados efeitos de SNPs a partir de valores genético genômico (GEBVs) reestimados. Foram encontradas 21 regiões explicando mais que 1% da variância genética total (gVar) da característica índice de velocidade máximo (IV max) para ssGWAS1 e 12 para ssGWAS2. No total mais de 40% da gVar foi explicada por estas regiões para ssGWAS1 e cerca de 30% para ssGWAS2. Entre os cromossomos que explicaram mais de 1% da variância genética, cinco foram comuns aos dois métodos (ECA 3, 10, 15, 22, 25). A capacidade de capturar maior quantidade de QTLs de menor efeito foi superior no ssGWAS1 se comparado ao ssGWAS2. Foram identificados 108 genes na primeira abordagem e 59 na segunda. Assim, os resultados obtidos neste trabalho mostraram que o desempenho em corridas na raça Quarto de Milha, dado pelo IV max, é característica quantitativa e que não há ocorrência de *major* genes. De forma geral, os genes *PDZRN3*, *PPP4R2*, *SLC31A1*, *SLC31A2*, *GEN1*, *SMC6*, *WHRN*, *SYT4*, *ABCG5*, *ABCG8*, *PLEKHH2*, *MICAL1*, *COL11A1*, *COL27A1* e *COX20*, os quais foram destacados por meio de análise de enriquecimento, estão relacionados ao sistema neuromotor, ao desenvolvimento ósseo e de tendões, à respostas hormonais, à contração muscular, e ao metabolismo energético e de esteróis. Estas funções podem ser consideradas de relevância para a característica estudada, corroborando estes genes como candidatos para o desempenho em corridas na raça Quarto de Milha.

Palavras-chave – arranjo de SNPs, InDels, polimorfismos, *single-step* GWAS, SNPs

GENOME-WIDE ASSOCIATION STUDY IN RACING QUARTER HORSES

ABSTRACT – Quarter Horses of racing are able to run short distances faster than any other breed. Although in Brazil the effective number of animals in the racing line is relatively smaller compared to the other lines, its economic importance is substantial. Despite economic and scientific interest in this athletic trait, few efforts have been made to better understand the genetic and physiological mechanisms underlying this trait. The objective of this study was to perform genome-wide association studies (GWAS) in Quarter Horses of the racing line using equine SNP genotyping chips for prospecting chromosome regions, genes and polymorphisms related to performance. Genotypic and phenotypic data from 360 horses genotyped with the 54k or 65k chip were used. GWAS was performed by the single-step GBLUP method using two approaches: ssGWAS1 in which only SNP effects are recalculated at each iteration, and ssGWAS2 in which SNP effects are recalculated from genomic estimated breeding values (GEBVs) at each iteration. Twenty-one regions that explained more than 1% of the total genetic variance (gVar) in the maximum speed index were identified by ssGWAS1 and 12 by ssGWAS2. More than 40% of gVar was explained by these regions in ssGWAS1 and about 30% in ssGWAS2. Among chromosomes that explained more than 1% of genetic variance, five were common to both methods (ECA 3, 10, 15, 22, 25). The capacity to capture the largest number of small-effect QTLs was greater for ssGWAS1 than ssGWAS2. A total of 108 genes were identified with the first approach and 59 with the second approach. Taken together, the present results show that the racing performance of Quarter Horses, given by the maximum speed index, is a quantitative trait and that no major genes exist. In general, the *PDZRN3*, *PPP4R2*, *SLC31A1*, *SLC31A2*, *GEN1*, *SMC6*, *WHRN*, *SYT4*, *ABCG5*, *ABCG8*, *PLEKHH2*, *MICAL1*, *COL11A1*, *COL27A1* and *COX20* genes, which were identified by enrichment analysis, are related to the neuromotor system, bone and tendon development, hormonal responses, muscle contraction, and energy and sterol metabolism. These functions are considered important for the trait studied, corroborating these genes as candidates for racing performance in the Quarter Horse breed.

Keywords: SNP array, InDels, polymorphisms, single-step GWAS, SNPs

1. Introdução

Originária da América do Norte, a raça Quarto de Milha é uma das mais criadas no mundo (ABQM, 2016). Introduzida no Brasil em 1969, tem sido amplamente utilizada no trabalho a campo, em provas funcionais e de velocidade (ABQM, 2016). Destacadamente, equinos de corrida desta raça têm o melhor desempenho em corridas de curtas distâncias do que qualquer outra (AMERICA'S HORSE DAILY, 2008). Apesar de, no Brasil, o efetivo de animais ser relativamente menor na linhagem de corrida do que nas demais, sua importância econômica é substancial (LIMA et al., 2006). Tendo em vista o interesse econômico e científico relacionado a esta linhagem e a sua aptidão para corridas, poucos esforços têm sido realizados para a maior compreensão dos mecanismos genéticos que governam esta característica (MEIRA et al., 2015a,b ; PEREIRA et al., 2016a,b), bem como para o melhoramento de equinos de corrida da raça Quarto de Milha (CORRÊA; MOTA, 2007).

Estudos de associação entre polimorfismos de DNA e desempenho em corridas de cavalos têm sido realizados, principalmente na raça Puro-Sangue Inglês (PSI), tanto de forma pontual (GU et al., 2010; HILL et al., 2010c), quanto de forma ampla (SHIN et al., 2015). Entretanto, os resultados encontrados por Pereira et al. (2016a;b) mostraram que alguns marcadores apropriados para a raça inglesa não são úteis para a identificação de animais superiores para desempenho em corridas no Quarto de Milha. Com o intuito de encontrar genes e polimorfismos de DNA específicos para a linhagem de corrida da raça Quarto de Milha e ligados à velocidade e ao desempenho em provas, Meira et al. (2015a;b) realizaram estudos de assinaturas de seleção e de associação ampla do genoma (*genome wide association study* – GWAS). Entretanto, devido ao pequeno número de animais utilizado, os resultados obtidos não foram terminantes em determinar variantes genéticas associadas ao desempenho em corridas em equinos da raça.

Tal como ocorre em humanos (SCHRÖDER et al., 2011), o desempenho atlético em equinos deve ser influenciado por grande número de regiões cromossômicas, ou seja, por muitos genes de pequeno efeito. Estudos genômicos a partir de painéis de SNPs podem ser empregados tanto para detecção de QTLs por meio de GWAS, quanto para predição de valores genéticos genômicos (*genomic estimated breeding value* – GEBVs) utilizados na seleção genômica (*genome selection* – GS) (GODDARD & HAYES, 2007; RESENDE et al., 2012). Convencionalmente, o GWAS procura encontrar associações entre locos e características fenotípicas por meio de testes de hipóteses, visando detectar efeitos com significância estatística. De modo a evitar a ocorrência de falsos positivos, o GWAS pode

ser penalizado para altas taxas de falsos negativos devido ao uso de pontos de corte muito rigorosos (RESENDE et al., 2012). Deste modo, apenas marcadores ligados a QTLs de grande efeito tendem a ser identificados. A GS, realizada por métodos bayesianos e gBLUP (*genomic BLUP*), de forma semelhante ao GWAS, é aplicada a todo o genoma, porém de forma simultânea e fundamentando-se em estimação e predição de efeitos ao invés de testes de hipóteses (RESENDE et al., 2012). O método ssGBLUP (*single-step GBLUP*) foi, inicialmente, descrito para realização de GWS. Contudo, Wang et al. (2012) descreveram adaptação do ssGBLUP para estudos de GWAS (*single-step GWAS*), em que são utilizados processos iterativos com ponderadores para os marcadores genéticos, e, a partir disso, são obtidos os GEBVs e os efeitos dos polimorfismos. O processo leva a ganhos na detecção de regiões genômicas de importância para a característica de interesse, por evidenciar a presença de QTLs de maior efeito ou de maior número de QTLs de pequeno efeito (WANG et al., 2012).

O objetivo deste trabalho foi realizar estudo de associação ampla do genoma (GWAS) em cavalos da linhagem de corrida da raça Quarto de Milha por meio da utilização de chips equinos de genotipagem de SNPs e do método ssGBLUP, visando a prospecção de QTLs e genes possivelmente relacionados ao desempenho.

2. Material e Métodos

2.1. Animais, coleta de sangue e extração do DNA

Foram utilizados 360 equinos da linhagem de corrida da raça Quarto de Milha, 78 machos e 282 fêmeas, registrados na associação brasileira de criadores (Associação Brasileira de Criadores de Cavalo Quarto de Milha – ABQM). Estes animais são filhos de 83 garanhões e 249 éguas, resultando em média de 4,3 progênies/garanhão e 1,4 progênies/égua, eram pertences de 159 criadores, e encontravam-se alojados no Jockey Club de Sorocaba (Sorocaba/SP – Brasil) e em 25 propriedades rurais localizadas em cidades do interior estado de São Paulo/Brasil. Deste total, 120 cavalos tiveram o sangue coletado no ano de 2011. Os demais, n= 240, foram coletados em 2015. A coleta e, consequentemente, a presença de irmãos completos na amostra foi evitada. Todos os procedimentos envolvendo os animais foram realizados segundo as normas brasileiras de bem-estar animal (Protocolo nº 157/2014 – Comissão de Ética no Uso de Animais/CEUA, FMVZ, Unesp, Botucatu/São Paulo/Brasil).

O processo de extração do DNA a partir de alíquotas de sangue foi realizado por meio do *Illustra Blood GenomicPrep Mini Spin Kit* (GE Healthcare, EUA), de acordo com as instruções do fabricante. Ao final do procedimento de extração, o DNA foi quantificado no espectrofotômetro de pequenos volumes *NanoDrop Lite* (Thermo Scientific, EUA) e analisado quanto a sua integridade em gel de agarose a 0,8%. Após o ajuste para a concentração de trabalho necessária à genotipagem dos SNPs (entre 40 e 60 ng/uL), as amostras foram armazenadas a -20 °C.

2.2. Controle de qualidade de genotipagem e imputação de genótipos

Os painéis utilizados neste trabalho foram o *Equine SNP50 BeadChip* (Illumina, Inc., EUA), o qual possui em sua matriz 54.602 marcadores (54k) e o *Equine SNP70 BeadChip* (Illumina, Inc., EUA), sucessor comercial do *Equine SNP50 BeadChip*, que possui 65.157 marcadores (65k). Entre o primeiro painel e o segundo, há 45.703 marcadores comuns, ou seja, os painéis de 54k e 65k possuem 8.899 e 19.454 marcadores não compartilhados, respectivamente. Dos 360 equinos utilizados neste estudo, 120 foram genotipados no ano de 2011 com o *Equine SNP50 BeadChip*. As demais amostras de DNA (n= 240) foram genotipadas no ano de 2015 com o *Equine SNP70 BeadChip*. A leitura de ambos os arranjos foi realizada por meio do sistema *HiScan* (Illumina, Inc., EUA). Os controles de qualidade (*quality control* – QC) das informações de genotipagem, para indivíduos analisados por um ou outro arranjo – 54k (n= 120) e 65k (n= 240), foram realizados por meio do pacote *snpStats* (CLAYTON, 2015) do programa R v.3.3.1 (R Core Team, 2016). Em relação aos indivíduos, animais com *call rate* abaixo de 0,90 foram excluídos do conjunto amostral. No que diz respeito aos SNPs, foram excluídos os localizados no cromossomo X, os com *call rate* inferior a 0,9, e os com *p-value* inferior a 1×10^{-5} para o equilíbrio de Hardy-Weinberg (*Hardy-Weinberg Equilibrium* – HWE) (Tabela 1).

Tabela 1 – Número de marcadores dos chips *Equine SNP50 BeadChip* e *Equine SNP70 BeadChip* e número de SNPs excluídos por cada critério de QC utilizado.

Critérios utilizados para exclusão de amostras	SNP50 (n)	SNP70 (n)
<i>Call rate</i> < 0,9	4	7
Amostras remanescentes	116	233
Critérios utilizados para exclusão de SNPs	SNP50 (n)	SNP70 (n)
SNPs genotipados	54.602	65.157
Localizados no cromossomo X	3.223	3.411
<i>Call rate</i> < 0,9	148	504
Equilíbrio de <i>Hardy-Weinberg</i> – $p < 1 \times 10^{-5}$	4.519	4.638
SNPs remanescentes	46.712	56.604

As informações utilizadas para a realização da imputação de genótipos em duas vias (i.e. do 54k para o de 65k e do 65k para o de 54k) foram relativas ao número de indivíduos e SNPs obtidos após o QC. A imputação dos genótipos foi realizada por meio do programa FImpute v.2.2 (SARGOLZAEI et al., 2014).

Na imputação em duas vias independentes foi utilizado primeiramente o painel de 65k como referência e o 54k como imputado. Para isso foram considerados todos os marcadores presentes no 65k e somente os marcadores em comum do SNP50 (45k). Os marcadores não compartilhados (9k) foram novamente acrescentados após a imputação. Deste modo, as amostras genotipadas com o painel de menor densidade (54k) passaram a ter também os genótipos exclusivos do de 65k. O mesmo procedimento foi realizado na segunda via, em que houve apenas a inversão dos painéis utilizados como referência e imputado. Assim, as amostras genotipadas com o arranjo de 65k passaram a ter também os genótipos exclusivos do painel de 54k. Posteriormente, ambos os painéis imputados foram ordenados e concatenados, formando um conjunto com indivíduos genotipados para todos os marcadores disponíveis nos painéis de 54k e 65k. O critério de QC *minor allele frequency* – MAF (MAF < 0,05) foi aplicado ao painel de genótipos já imputado em duas vias, o qual foi utilizado nas análises de GWAS.

Após o controle de qualidade de genotipagem e a imputação de genótipos exclusivos, foram conduzidas análises para investigação de estratificação populacional no conjunto amostral utilizado. Para isso foram obtidos os valores de inflação genômica por meio do fator de inflação da variância (VIF) e realizada a análise de componentes

principais (PCA) utilizando o software Plink 1.9 (PURCELL et al., 2007). Os valores de λ foram obtidos por meio de associação simples entre genótipos e os fenótipos não corrigidos para efeitos sistemáticos de ambiente e, posteriormente, entre os genótipos e fenótipos corrigidos (EBV) obtidos para cada animal. Para obtenção dos EBVs foi utilizado o programa blupf90 (MISZTAL, 2014), em que foram incluídas as informação de parentes (matriz de parentesco) e ajustes para efeitos fixos seguindo o modelo estatístico a ser utilizado para o ssGWAS. A visualização de subestruturas foi realizada por meio da plotagem dos dois primeiros componentes principais, PC1 e PC2, utilizando o pacote ggplot2 (WIKIHAM, 2009) do programa R v.3.3.1 (R Core Team, 2016).

2.3. Informações fenotípicas

Os dados de desempenho dos cavalos Quarto de Milha de corrida, atualizados até o mês de dezembro de 2015, foram obtidos junto ao Departamento de Estatística ou ao banco de dados online do Jockey Club de Sorocaba (<http://www.jcsorocaba.com.br/indice-de-velocidade/>) e junto ao banco de dados online Equibase (<http://www.equibase.com/homehorseplayer.cfm>).

Foi utilizado o registro de desempenho dado pelo índice de velocidade (IV) máximo (IV max) obtido ao longo do histórico de competição de cada animal. Para a maior parte dos indivíduos utilizados foram encontradas informação de IV, contudo para outros foram encontradas apenas informação de IV max. Como o IV médio apresenta alta correlação com o IV max ($r = 0,8762$; PEREIRA et al., 2016b) optou-se por utilizar o IV max a fim de que não houvesse perda de informações de desempenho. Ainda, com objetivo de garantir a confiabilidade da medida IV max, foram considerados para o GWAS apenas indivíduos com três ou mais corridas em seu histórico de competição.

Parâmetros genéticos e estatística descritiva para a característica IV max estão apresentados na Tabela 2. Os componentes de variância foram estimados por método bayesiano, por meio de amostrador de *Gibbs*, utilizando os programas gibbs1f90 e postgibbs1f90 (MISZTAL, 2). As amostras dos componentes de variância foram obtidas de cadeia de 500.000 ciclos. Os 50.000 ciclos iniciais foram descartados (*burn-in*) e as amostras foram retiradas a cada 50 ciclos (*thinning interval*), totalizando 9.000 amostras. A partir das amostras obtidas foram calculadas as médias posteriores para os componentes de variância, a herdabilidade e os desvios-padrão posteriores para cada parâmetro por meio do programa R v.3.3.1 (R Core Team, 2016). O modelo estatístico utilizado foi: $y = X\beta + Za + e$, onde: y é o vetor de observações; β é o vetor de efeitos fixos (sexo, idade e

ano da corrida referentes à observação do IV max), os quais foram testados por meio de análise de modelo linear; a é o vetor de efeitos genéticos aditivos diretos; X é a matriz de incidência; Z é a matriz de incidência dos efeitos aleatórios genéticos aditivos diretos; e é o vetor do efeito residual. A matriz de parentesco contou com a informação de 620 animais dos quais 492 tinham informação de IV max.

Tabela 2 – Estimativas de componentes de variância (σ_a^2 e σ_e^2), herdabilidade (h^2) e estatística descritiva para a característica IV max em cavalos de corrida da raça Quarto de Milha.

Parâmetros genéticos	Média (PSD*)
h^2	0,118 (0,062)
σ_a^2	6,809 (3,609)
σ_e^2	50,74 (4,848)
Estatística descritiva	
Média	97,3
Desvio-padrão	8,12
Mínimo	72,0
Máximo	120,0
Mediana	98,0

* desvio-padrão a posteriori.

2.4. Estudo de associação ampla do genoma

Com o objetivo de encontrar regiões genômicas que expliquem parte importante da variância genética total da característica IV max, foram estimados efeitos de SNPs e variâncias explicadas por janelas de SNPs por meio de *single-step* GWAS (ssGWAS) como descrito por Wang et al. (2012). As análises foram realizadas pelo pacote de programas computacionais blupf90 (MISZTAL et al., 2002; AGUILAR et al., 2011).

Para o ssGWAS foi aplicado modelo estatístico descrito no item 2.3. As variâncias de a e e foram dadas por:

$$\text{var} \begin{bmatrix} a \\ e \end{bmatrix} = \begin{bmatrix} H\sigma_a^2 & 0 \\ 0 & I\sigma_e^2 \end{bmatrix}$$

em que, σ_a^2 e σ_e^2 são a variância genética total e residual, respectivamente, e H é a matriz que combina as matrizes de parentesco baseado no pedigree e no genoma, como em Aguilar et al. (2010). A sua inversa é:

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

em que, A é o numerador da matriz de pedigree para todos os animais; A_{22} é o numerador da matriz de pedigree para animais genotipados; e G é a matriz de parentesco genômica. A matriz G , construída com base em VanRaden et al. (2008), assume frequências alélicas da população sob análise e ajusta para compatibilidade com A_{22} (CHEN et al., 2011; VITEZICA et al., 2011).

A equação para predição de efeito de SNP que utiliza matriz de parentesco genômica ponderada G^* , proposta por Wang et al. (2014), foi dada por: $\hat{u} = \lambda DZ'G^{*-1}\hat{a}_g = DZ'[ZDZ']^{-1}\hat{a}_g$ onde, λ é uma razão de variâncias ou uma constante de normalização e $\hat{a}_g$ é o GEBV. Em que $\lambda = \frac{\sigma_u^2}{\sigma_a^2} = \frac{1}{\sum_{i=1}^M 2p_i(1-p_i)}$, onde M é o numero de SNP e p é frequência do segundo alelo do i -th marcador; D é a diagonal da matriz de ponderação para as variâncias de SNP ($D = I$ para GBLUP); Z é a matriz que relaciona os genótipos de cada loco, e G^* é a matriz de parentesco genômica ponderada. Estimativas de efeito de SNP foram utilizadas para estimar a variância individual do efeito de cada marcador (ZHANG et al., 2010) conforme segue: $\sigma_{u,i}^2 = u_i^2 2p_i(1-p_i)$.

Para o GWAS, os efeitos de SNP foram obtidos a partir de GEBVs ou EBVs de indivíduos genotipados seguindo a metodologia e algoritmo descritos por Wang et al. 2014, como segue: 1. $t = 0$, $D_{(t)} = I$; $G_{(t)}^* = ZD_{(t)}Z' \lambda$; 2. Computar $\hat{a}_g$ por ssGBLUP; 3. Cálculo de: $\hat{u}_{(t)} = \lambda D_{(t)}Z'G_{(t)}^{*-1}\hat{a}_g$; 4. Cálculo de: $G_{(t+1)}^* = \hat{u}_{i(t)}^2 2p_i(1-p_i)$ para todos i como em Zhang et al. (2010); 5. Normalização por: $D_{(t+1)} = \frac{tr(D_{(0)})}{tr(D_{(t+1)}^*)} D_{(t+1)}^*$; 6. Cálculo de $G_{(t+1)}^* = ZD_{(t+1)}Z' \lambda$; 7. $t = t + 1$; 8. Parar, ou voltar ao passo 2 ou passo 3. Em que t representa o número de interações e i o i -th SNP, ao retornar ao passo 3 (ssGWAS1) é aplicado uma matriz G^* “re-ponderada” em que somente a predição dos efeitos de SNP são recalculados. Ao retornar para o passo 2 (ssGWAS2) ambos os efeitos, de animal e SNP, são recalculados. Foram utilizadas cinco iterações nos diferentes cenários (ssGWAS1 e ssGWAS2) e janelas de SNPs com tamanhos de 250 Kb, 500 Kb e 1 Mb.

A porcentagem da variância genética explicada pela i -th região foi calculada como segue:

$$\frac{Var(s_i)}{\sigma_a^2} \times 100\% = \frac{Var(\sum_{j=1} z_j \hat{u}_j)}{\sigma_a^2} \times 100\%$$

onde s_i é valor genético da i -th região, que consistem em janelas de 1 Mb com diferentes números de SNPs, σ_a^2 é a variância genética total, $\mathbf{Z}_j$ é vetor de conteúdo genético do j -th SNP para todos os indivíduos, e $\hat{u}_j$ é efeito marcador da j -th SNP na região i -th.

2.5. Anotação de genes e análise funcional

Para a prospecção de genes nas regiões cromossômicas consideradas de importância e para as análises de enriquecimento foram consideradas janelas de SNPs de 1 Mb de tamanho e que explicaram, pelo estudo de GWAS, proporção da variância genética (gVar) maior que 1%. As posições e os nomes de cada gene foram determinados com base na montagem mais recente da sequência do genoma equino (EquCab2.0) por meio do pacote *BioMart* (DURINCK et al., 2009) do programa R (R Core Team, 2016). A análise de enriquecimento foi realizada por meio do pacote *topGO* do programa R (ALEXA e RAHNENFUHRER, 2016), utilizando os genes que estavam contidos em janelas de SNPs com gVar maior que 1% (genes alvo) e um universo de genes localizados em janelas de SNPs com gVar maior que 0,1%. Para isso foi considerado o algoritmo padrão do pacote *topGO* (opção “weight01”) e aplicado o teste estatístico de Fisher com significância a 5%.

3. Resultados e Discussão

3.1. Dados utilizados e estratificação da população

Dos 349 indivíduos que passaram pelo controle de qualidade de genotipagem, 44 animais foram excluídos por não apresentarem informação de IV max ($n= 12$) ou por apresentarem em seu histórico menos de três corridas ($n= 32$). Assim sendo, restaram 305 cavalos genotipados e com informações confiáveis de desempenho. Adicionalmente, foram incorporadas às análises de ssGWAS informações de 187 cavalos da raça Quarto de Milha não genotipados, mas presentes na matriz de parentesco. Estes indivíduos possuíam registros de IV max e mais de três corridas em seu histórico. Dessa forma, foram utilizados para o GWAS os registros de IV máximo de 492 animais e informações de pedigree de 620 cavalos.

A imputação de genótipos em duas vias trouxe aumento considerável de informação a cada painel. Após o QC, os dois painéis passaram a contar com 39.664 SNPs compartilhados e 7.048 e 16.940 marcadores exclusivos para os arranjos de 54k e 65k,

respectivamente. A densidade final do painel imputado foi de 63.652 SNPs sem o filtro de MAF. O ganho em número de marcadores em relação ao painel imputado final foi de 26% para o SNP50, ultrapassando o número original de SNPs, e 12% para o SNP70. Após a aplicação do filtro de MAF ($MAF < 0,05$) foram excluídos 8.538 SNPs. Com isso, o painel imputado passou a contar com 55.114, os quais foram efetivamente utilizados na análise de associação.

A análise de componentes principais, que utilizou 305 indivíduos, mostrou subestruturas na população estudada (Figura 1). Contudo, estas não foram claramente distintas. A aplicação dos modelos mistos foi eficiente para correção do viés causado pela estratificação. Modelos mistos podem modelar a estrutura de população e de família, bem como a relação de parentesco de indivíduos não genotipados ou mortos (PRICE et al., 2010). Ao estudar métodos de correção de estratificação em nove características quantitativas em humanos, Kang et al. (2010) mostraram que a utilização de modelos mistos foi mais eficiente na correção da inflação genômica do que outros métodos. Contudo, faz-se necessário a incorporação de componentes principais no modelo quando observado alto valor de inflação genômica em conjunto amostral com grau significativo de parentesco (PRICE et al., 2010). As estimativas de VIF obtidas neste estudo foram de $\lambda=1,11$ com a utilização dos fenótipos não corrigidos e $\lambda=1,01$ com os fenótipos corrigidos (EBVs).

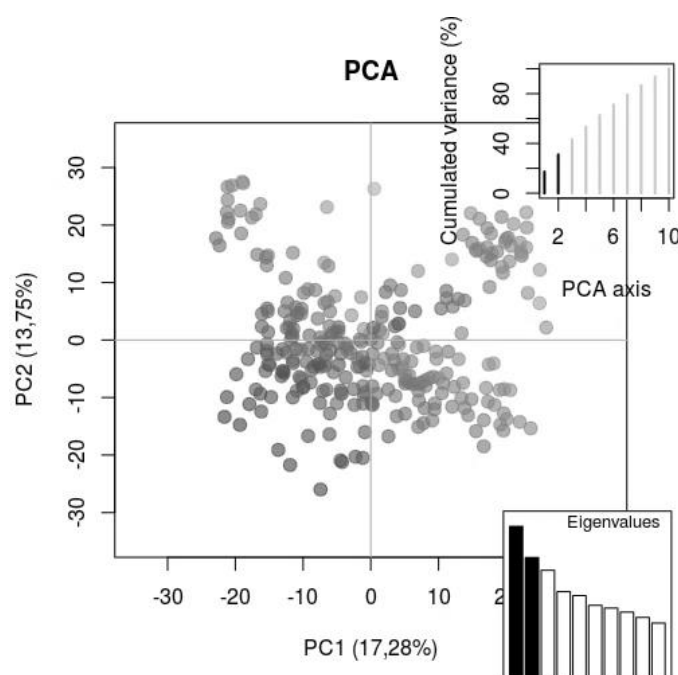


Figura 2 – Representação dos dois primeiros componentes principais (PC1 e PC2) e as porcentagens das variâncias explicadas por cada componente. Na parte superior direito é dada a % acumulada da variância explicada (*Cumulated variance*) dos CP1 e CP2 em relação aos dez primeiros PCs. Na parte inferior

direita estão representados os autovalores de 10 PCs, em que estão destacados os PC1 e PC2.

3.2. Associação ampla do genoma

Nos estudos de GWAS foram empregadas duas abordagens, ssGWAS1 e ssGWAS2. Isto se deveu ao fato de cada uma delas utilizar procedimentos diferentes para a obtenção dos efeitos de SNPs e consequentemente das variâncias explicadas, resultando em capacidades distintas de capturar QTLs de maior ou menor efeito. Com isso, a análise combinada de ambos os cenários pode levar ao maior ganho de informação em relação às regiões e genes ligados à característica de interesse. Os resultados foram obtidos para ambos os métodos levando em consideração a terceira iteração das cinco realizadas, visto que a partir da terceira iteração as regiões com pequeno efeito passaram a ser fortemente penalizadas pelos ponderadores (pesos) enquanto o oposto aconteceu com as de maiores efeitos. O número médio de SNPs observados no ssGWAS1 foi de 8,8; 13,1 e 25,9 e no ssGWAS2 foi de 9,6; 17 e 28,5 para as janelas de 250 Kb, 500 Kb e 1000 Kb, respectivamente.

No ssGWAS1, em que foram reestimados somente efeitos de SNPs a cada iteração, observou-se poucas alterações na quantidade de regiões associadas entre os diferentes tamanhos de janelas (250 Kb, 500 Kb e 1 Mb) (Figura 2). Ademais, aparentemente algumas regiões de efeito menor foram beneficiadas em janelas com maiores extensões (ECA 1, 8 e 29). Por outro lado, no ssGWAS2, em que a cada iteração os efeitos de SNPs foram obtidos a partir de GEBVs reestimados, observou-se alterações em quantidades de regiões associadas entre os diferentes tamanhos de janelas (250 Kb, 500 Kb e 1 Mb) (Figura 3). Nesse sentido, regiões de menor efeito foram mais fortemente penalizadas pelos ponderados em janelas de SNPs maiores, ocorrendo o inverso com regiões de maior efeito. A capacidade de capturar QTLs de menor efeito foi maior no ssGWAS1 se comparado ao ssGWAS2. Deste modo, como observado por Wang et al. (2012) a partir de dados simulados, o primeiro método (ssGWAS1) foi eficaz na detecção de maior número de regiões de efeitos menores quando comparado ao segundo (ssGWAS2), o qual, por sua vez, se mostrou mais eficiente em capturar menor número de QTLs de efeitos maiores.

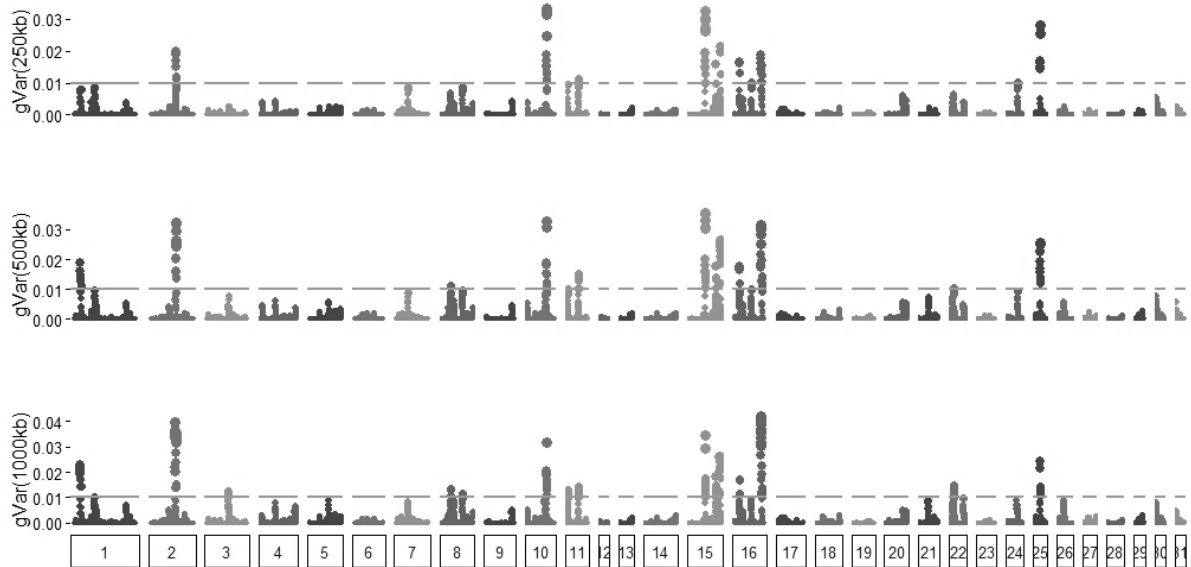


Figura 2 – *Manhattan plot* da proporção da variância genética total explicada por cada janela de SNP (250 Kb, 500 Kb e 1000 Kb) no ssGWAS1. A proporção da variância genética explicada por cada região está plotada no eixo y. Os cromossomos estão plotados no eixo x. A linha pontilhada indica o limiar de um 1%.

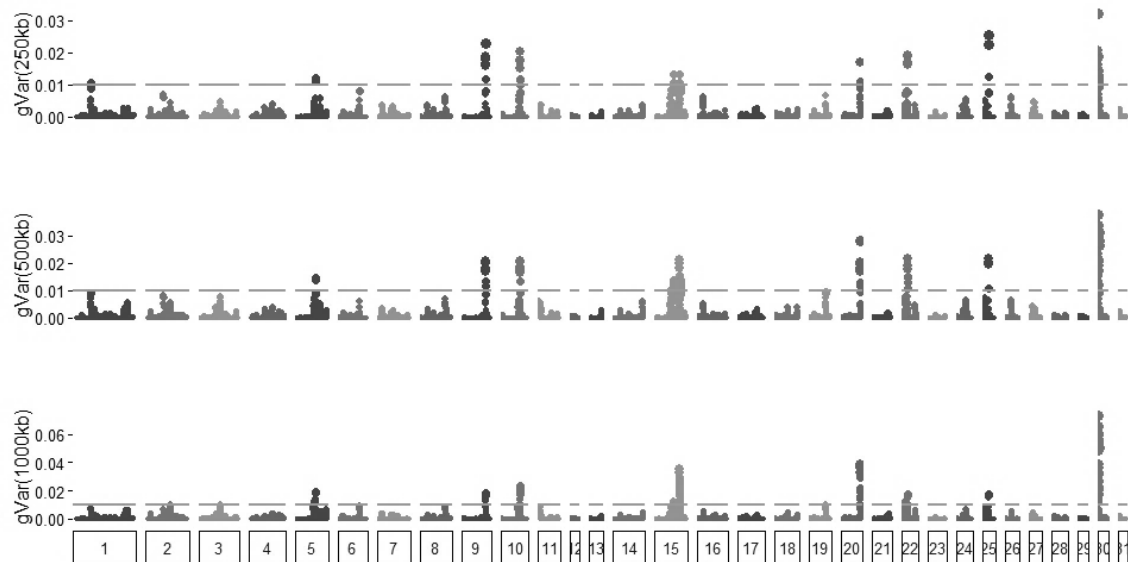


Figura 3 – *Manhattan plot* da variância genética total explicada por cada janela de SNP (250 Kb, 500 Kb e 1000 Kb) no ssGWAS2. A proporção da variância genética explicada por cada região está plotada no eixo y. Os cromossomos estão plotados no eixo x. A linha pontilhada indica o limiar de um 1%.

Considerando as janelas de SNPs com 1 Mb de extensão, foram encontradas 21 regiões explicando mais do que 1% da variância genética total da característica IV max para ssGWAS1 (Tabela 3) e 12 para ssGWAS2 (Tabela 4). No total, mais de 40% da gVar foi explicada por estas regiões para ssGWAS1 e cerca de 30% para ssGWAS2. Entre os cromossomos com regiões que explicaram mais de 1% da variância genética, cinco foram comum aos dois métodos (ECA 3, 10, 15, 22 e 25), cinco foram exclusivos ao ssGWAS1 (ECA 1, 2, 8, 11 e 16) e quatro ao ssGWAS2 (ECA 5, 9, 20, 30). Deste modo, tomando os dois métodos de forma conjunta, foram descritas regiões importantes em 14 cromossomos. Em relação às regiões associadas, porção única do ECA 16 (SS1_REG18) explicou 4,2% da variância genética (gVar) no ssGWAS1 (Tabela 3), enquanto que no ssGWAS2 a região SS2_REG12 do ECA 30 foi responsável por explicar 7,3% (Tabela 4). O ECA 15 foi o cromossomo com mais regiões de importância nos dois métodos, explicando no ssGWAS1 a maior proporção da gVar (10,67%), se somadas as cinco regiões mais relevantes (SS1_REG11 ao SS1_REG15) e no ssGWAS2 6,7% se somadas três regiões (SS2_REG05 ao SS2_REG07). Ao considerar os resultados obtidos por ambos os métodos, foram encontradas 33 regiões de importância, das quais quatro do ssGWAS1 coincidem com três do ssGWAS2 (SS1_REG08 e SS1_REG09 com SS2_REG04; SS1_REG19 com SS2_REG09; SS1_REG20 com SS2_REG11).

Tabela 3 – Localização cromossômica de regiões de importância para IV max (gVar maior que 1%) na amostra estudada de equinos da raça Quarto de Milha de corrida, número de SNPs e variância explicada por cada janela no ssGWAS1, considerando janelas de SNPs com 1 Mb de tamanho.

Região	Chr*	Posições de início e fim de cada janela em pb	SNPs	Variância genética (%)
SS1_REG01	1	20.721.039 - 21.689.628	17	2,30
SS1_REG02	1	61.826.124 - 62.818.065	36	1,01
SS1_REG03	2	69.370.047 - 70.369.774	35	3,98
SS1_REG04	3	65.380.573 - 66.369.357	20	1,23
SS1_REG05	8	30.526.194 - 31.504.074	21	1,32
SS1_REG06	8	62.566.792 - 63.559.486	34	1,18
SS1_REG07	10	57.990.860 - 58.961.242	25	3,19
SS1_REG08	10	58.997.563 - 59.976.633	29	1,68
SS1_REG09	11	03.345.863 - 04.341.072	24	1,31
SS1_REG10	11	35.523.097 - 36.481.549	29	1,43
SS1_REG11	15	45.572.062 - 46.563.225	26	3,46
SS1_REG12	15	46.601.694 - 47.575.761	23	1,50
SS1_REG13	15	76.296.526 - 77.289.510	20	1,80
SS1_REG14	15	82.855.154 - 83.838.250	21	1,29
SS1_REG15	15	87.766.362 - 88.753.736	29	2,62
SS1_REG16	16	16.542.440 - 17.474.810	23	1,71
SS1_REG17	16	17.542.693 - 18.503.480	30	1,17
SS1_REG18	16	77.697.272 - 78.681.644	31	4,19
SS1_REG19	22	10.868.934 - 11.861.413	24	1,47
SS1_REG20	25	17.757.118 - 18.713.444	23	2,41
SS1_REG21	25	18.768.518 - 19.722.839	24	1,20
Total	Total**	-	Média	Total
21	10	-	28,5	41,45

*Chr = cromossomos; ** contagem de cromossomo sem considerar os repetidos. Em negrito estão destacadas as cinco regiões que mais explicaram parte da variância genética total

Tabela 4 – Localização cromossômica de regiões de importância para IV max (gVar maior que 1%) na amostra estudada de equinos da raça Quarto de Milha de corrida, e número de SNPs e variância explicada por cada janela no ssGWAS2, considerando janelas de SNPs com 1 Mb de tamanho.

Região	Chr*	Posições de início e final de cada janela em pb	SNPs	Variância genética (%)
SS2_REG01	3	63.351.099 - 64.348.556	27	1,03
SS2_REG02	5	62.937.896 - 63.920.127	27	1,92
SS2_REG03	9	66.879.753 - 67.826.611	24	1,84
SS2_REG04	10	58.735.380 - 59.710.571	34	2,32
SS2_REG05	15	53.973.327 - 54.889.258	28	1,28
SS2_REG06	15	72.126.768 - 73.121.731	35	3,56
SS2_REG07	15	75.795.552 - 76.782.973	30	1,55
SS2_REG08	20	59.429.342 - 60.422.986	24	3,86
SS2_REG09	22	10.600.456 - 11.589.137	28	1,28
SS2_REG10	22	14.927.558 - 15.876.595	25	1,72
SS2_REG11	25	18.073.334 - 19.043.039	26	1,80
SS2_REG12	30	00138.095 - 01.097.585	34	7,34
Total	Total**	-	Média	Total
12	9	-	28,5	29,51

*Chr = cromossomos; ** contagem de cromossomo sem considerar os repetidos. Em negrito estão destacadas as cinco regiões que mais explicaram parte da variância genética total

O grande número de regiões associadas em ambos os métodos (ssGWAS1 e ssGWAS2) evidencia o caráter quantitativo da característica IV max e, conseqüentemente, do desempenho em corridas na raça Quarto de Milha, ou seja, que esta é influenciada por grande número de genes de pequeno efeito. A baixa herdabilidade encontrada para IV max aponta para uma importante influência do ambiente sobre a característica, como treinamento a que o animal foi submetido, entre outros. Ou ainda, a falta de informações de fatores sistemáticos de ambiente (páreo, *paddock*, peso do *jockey*, entre outros), o que acarretaria em aumento da estimativa do componente de variância residual (não explicada pelo modelo).

Em grande parte dos estudos realizados em diversas espécies utilizando o método ssGWAS, observou-se porcentagens de gVar menores do que as aqui encontradas, com valores máximos de aproximadamente 6%, considerando, na maioria dos casos, janelas de 10 a 20 SNPs (WANG et al., 2014; MAGALHÃES et al., 2016; SILVA et al., 2017). Contudo, em trabalho desenvolvido por Lemos et al. (2016), relatou-se janela de 10 SNPs explicando aproximadamente 12% da variância genética total da característica estudada na raça Nelore. Alguns autores chegaram a utilizar 150 SNPs por janela, que explicaram

no máximo de 5% da variância genética total da característica (SILVA et al., 2017). Ressalta-se que, nesses estudos com bovinos e frangos de corte, os painéis de genotipagem utilizados variaram de 60k a HD (~770k) e que as janelas foram escolhidas por número de SNPs e não por tamanho em pares de base, o que causa grande variação na sua extensão, ficando em sua maioria com tamanhos médios similares às aqui utilizadas. Por meio da utilização de dados simulados (dois cromossomos de 100 Mb cada e 3000 SNPs), Wang et al. (2012) estudaram a precisão do método ssGBLUP para estimar efeitos de QTLs, tomando para isso as correlações entre efeitos de QTL e a soma de m efeitos de marcador adjacentes de SNP, em que m variou de 1 a 40. Foram observados maiores correlações em janelas contendo de 8 a 16 SNPs (de 0,5 Mb a 1 Mb de tamanho em média, respectivamente), com forte queda a partir de janelas com 40 SNPs (> 2,5 Mb em média).

Grande parte das pesquisas genômicas realizadas no Quarto de Milha não consideraram, de forma separada, os segmentos com diferentes aptidões, ou seja, as linhagens existentes dentro da raça. Desta forma, trabalhos em que foram utilizados somente cavalos de corrida são escassos, principalmente os que consideraram características de desempenho. Neste sentido, ou seja, tomando a raça como um todo, estudos genômicos têm abordado principalmente aspectos populacionais (PETERSEN et al., 2013; MEIRA et al., 2015a; McCUE et al., 2012), morfométricos (MEIRA et al., 2015b) e doenças hereditárias (GO et al., 2012; McQUEEN et al., 2014). Contudo, alguns destes trabalhos, ainda que de forma indireta, apresentaram informações importantes que podem contribuir para o entendimento de mecanismos genéticos relacionados ao desempenho em corridas na raça Quarto de Milha. Meira et al. (2015) identificaram genes em assinaturas de seleção que apresentaram funções que podem levar à variações no desempenho em corridas na raça Quarto de Milha. Estes autores encontraram assinaturas de seleção em 19 cromossomos, dos quais sete coincidiram com os encontrados neste trabalho (ECA 1, 2, 3, 5, 10, 11, e 25). Contudo, não foram observadas regiões de assinatura sobrepostas com as aqui identificadas pelo GWAS. Entretanto, nos ECA 5, 10 e 25, observou-se quatro janelas de SNPs ($gVar > 1\%$) que estavam à, aproximadamente, 5 Mb distantes de regiões de assinatura de seleção descritas por estes autores.

No ECA 18, em região de aproximadamente 5,5 Mb em que o gene da miostatina (*MSTN*) se encontra, foi descrita assinatura de seleção nas raças Paint Horse e Quarto de milha (linhagens de corrida e trabalho) e, ainda, foram associados polimorfismos relacionados ao tipo de fibra muscular (PETERSEN et al., 2013). Os dados gerados por este trabalho mostraram que SNPs nesta região estavam fixados, e por isso foram

excluídos pelo QC, concordando com resultados obtidos por Pereira et al. (2016a), que demonstraram que o alelo C do SNP g.66493737C>T do gene *MSTN* estava fixado na linhagem de corrida da raça Quarto de Milha. De forma semelhante, esse alelo foi encontrado em alta frequência em animais PSI que disputam corridas de curta distância (*sprint*) (HILL et al., 2010a,b; TOZAKI et al., 2010). Portanto, nota-se, além de forte pressão de seleção sobre essa região do ECA 18 na linhagem de corrida da raça Quarto de Milha, a influência genética de cavalos PSI com a aptidão para corridas curtas sobre esse segmento.

De fato a raça PSI teve importante papel na formação da linhagem de corrida da raça Quarto de Milha (AQHA, 2016). Assim, resultados de estudos genéticos envolvendo a raça Inglesa podem ser semelhantes na linhagem de corrida da raça Norte Americana. Em estudo de associação genômica realizado na raça PSI, Shin et al. (2015) descreveram importantes genes associados à fisiologia muscular em sete cromossomos (ECA 5, 8, 16, 20, 21, 28 e 30), dos quais cinco (ECA 5, 8, 16, 20 e 30) coincidiram com os aqui encontrados. Esses autores descrevem cinco genes ligados no ECA 16, dos quais quatro estavam em uma região compreendida entre 14 e 18 Mb, mesma região em que estão localizados SS1_REG16 e SS1_REG17. No ECA 30 a posição de importância identificada por esses autores ficou 1,5 Mb a jusante da região SS2_REG12. Embora, os resultados obtidos por Shin et al. (2015) sejam relevantes, polimorfismos descritos em estudos genético moleculares e genômicos relacionados ao desempenho em corridas na raça PSI (TOZAKI et al., 2010, BINNS et al., 2010; HILL et al., 2010a,b,c; GU et al., 2010), em sua maior parte não tiveram o mesmo efeito na linhagem de corrida da raça Quarto de Milha (PEREIRA et al. 2016a,b). Segundo esses autores, este fato poderia ser explicado pelas diferenças nas provas disputadas entre as duas raças (mais curtas no Quarto de Milha e mais longas no PSI), em que alterações genéticas significativas na raça PSI teriam efeito na melhoria do metabolismo energético muscular anaeróbio, o qual é predominante em provas mais longas. Contudo, alterações em outros genes ligados à fisiologia do exercício poderiam contribuir para adaptações metabólicas importantes e ser compartilhados por ambas as raças, como foi demonstrado pela comparação dos resultados obtidos aqui e no PSI (SHIN et al., 2015).

3.3. Anotação de genes e análise funcional

Para ssGWAS1 foram encontrados 108 genes presentes em 19 das 21 regiões associadas (Tabela 5) e para ssGWAS2 foram descritos 59 genes em 10 das 12 janelas

de SNPs (Tabela 6). Foram encontrados 81 genes exclusivos da abordagem ssGWAS1, 32 da ssGWAS2 e 27 comuns a ambos os métodos. Estes 27 genes encontram-se localizados nos ECA 10, 22 e 25. Considerando os dois métodos simultaneamente, no total foram prospectados 140 genes. Destaca-se que aproximadamente 50% dos genes encontrados no ssGWAS2 também foram identificados no ssGWAS1.

Tabela 5 – Genes prospectados nas regiões cromossômicas de importância para IV max (gVar maior que 1%) na amostra estudada de equinos da raça Quarto de Milha de corrida utilizando o método ssGWAS1.

Região	Chr*	Posições de início e fim de cada janela em pb	Genes (<i>HGCN Symbol</i>)	Genes (n)
SS1_REG01	1	20.721.039 - 21.689.628	<i>DUSP5, ADD3, XPNPEP1</i>	3
SS1_REG02	1	61.826.124 - 62.818.065	<i>ADK, KAT6B, DUSP13, SAMD8, COMTD1, VDAC2</i>	6
SS1_REG03	2	69.370.047 - 70.369.774	<i>MSMO1, KLHL2, TMEM192</i>	3
SS1_REG04	3	65.380.573 - 66.369.357	-	
SS1_REG05	8	30.526.194 - 31.504.074	<i>EPB41L3</i>	1
SS1_REG06	8	62.566.792 - 63.559.486	<i>RIT2, SYT4</i>	2
SS1_REG07	10	57.990.860 - 58.961.242	<i>NR2E1, SNX3, LACE1, FOXO3, ARMC2, SESN1, CEP57L1</i>	7
SS1_REG08	10	58.997.563 - 59.976.633	<i>PPIL6, SMPD2, MICAL1, ZBTB24, FIG4, WASF1, CDC40, METTL24, DDO</i>	9
SS1_REG09	11	03.345.863 - 04.341.072	<i>RBFOX3, ENGASE, C1QTNF1, CANT1, LGALS3BP, TIMP2, USP36, CYTH1, DNAH17, PGS1, SOCS3, TMEM235, AFMID</i>	11
SS1_REG10	11	35.523.097 - 36.481.549	<i>CA4, ZNHIT3, MYO19, PIGW, GGNBP2, DHRS11, MRM1, LHX1, AATF, ACACA, SPTBN1, C2orf73, ACYP2</i>	10
SS1_REG11	15	45.572.062 - 46.563.225	<i>PSME4, ERLEC1, CHAC2</i>	3
SS1_REG12	15	46.601.694 - 47.575.761	<i>KCNS3, MSGN1, GEN1, SMC6, VSNL1</i>	3
SS1_REG13	15	76.296.526 - 77.289.510	<i>NOL10, ODC1, HPCAL1, RRM2, KLF11, GRHL1, TAF1B, YWHAQ</i>	5
SS1_REG14	15	82.855.154 - 83.838.250	<i>ALLC, COLEC11, RPS7, ADI1, TRAPPC12</i>	8
SS1_REG15	15	87.766.362 - 88.753.736	<i>PDZRN3, PPP4R2, GXYLT2,</i>	5
SS1_REG16	16	16.542.440 - 17.474.810	<i>SHQ1, RYBP</i>	3
SS1_REG17	16	17.542.693 - 18.503.480	-	2
SS1_REG18	16	77.697.272 - 78.681.644	-	0
SS1_REG19	22	10.868.934 - 11.861.413	<i>RYBP, TASP1, ISM1, SPTLC3</i>	4
SS1_REG20	25	17.757.118 - 18.713.444	<i>ZFP37, SLC31A2, FKBP15, SLC31A1, CDC26, PRPF4, RNF183, WDR31, BSPRY, HDHD3, ALAD, RGS3</i>	12
SS1_REG21	25	18.768.518 - 19.722.839	<i>ZNF618, AMBP, KIF12, COL27A1, AKNA, WHRN, ATP6V1G1, TMEM268, TNFSF15, TNFSF8, TNC</i>	11
Total				108

*Chr = cromossomos; ** contagem de cromossomo sem considerar os repetidos; as regiões e o genes coincidentes entres os métodos ssGWAS1 e ssGWAS2 estão destacados em negrito

Tabela 6 – Genes prospectados nas regiões cromossômicas de importância para IV max (gVar maior que 1%) na amostra estudada de equinos da raça Quarto de Milha de corrida utilizando o método ssGWAS2.

Região	Chr*	Posições de início e fim de cada janela em pb	Genes (HGCN Symbol)	Genes (n)
SS2_REG01	3	63.351.099 - 64.348.556	<i>GC, DCK, MOB1B, GRSF1, RUFY3, UTP3, ENAM</i>	7
SS2_REG02	5	62.937.896 - 63.920.127	<i>COL11A1</i>	1
SS2_REG03	9	66.879.753 - 67.826.611	<i>ZNF572, SQLE, KIAA0196, SMCE2, TRIB1</i>	5
SS2_REG04	10	58.735.380 - 59.710.571	<i>ARMC2, SESN1, CEP57L1, PPIL6, SMPD2, MICAL1, ZBTB24, FIG4, WASF1</i>	9
SS2_REG05	15	53.973.327 - 54.889.258	<i>SLC3A1, PREPL, PPM1B, LRPPRC, ABCG8, ABCG5, DYNC2LI1, PLEKHH2</i>	8
SS2_REG06	15	72.126.768 - 73.121.731	<i>KLHL29</i>	1
SS2_REG07	15	75.795.552 - 76.782.973	-	
SS2_REG08	20	59.429.342 - 60.422.986	-	
SS2_REG09	22	10.600.456 - 11.589.137	<i>TASP1, ISM1, SPTLC3</i>	3
SS2_REG10	22	14.927.558 - 15.876.595	<i>PLCB1, TMX4, HAO1</i>	3
SS2_REG11	25	18.073.334 - 19.043.039	<i>SLC31A2, FKBP15, SLC31A1, CDC26, PRPF4, RNF183, WDR31, BSPRY, HDHD3, ALAD, RGS3, ZNF618, AMBP, KIF12, COL27A1</i>	15
SS2_REG12	30	00138.095 - 01.097.585	<i>HNRNPU, COX20, DESI2, C1orf101, ADSS, C1orf100, ZBTB18</i>	7
Total				59

*Chr = cromossomos; ** contagem de cromossomo sem considerar os repetidos; as regiões e o genes coincidentes entres os métodos ssGWAS1 e ssGWAS2 estão destacados em negrito.

Dentre os genes encontrados em regiões concordantes, *PDZRN3*, *PPP4R2*, *GXYLT2* e *SHQ1* do ECA 16 foram identificados por SHIN et al. (2015) em estudo realizado com a raça PSI. Destacam-se os genes *PDZRN3* e *PPP4R2* que foram enriquecidos (Tabela 7) e estão relacionados a termos de formação, diferenciação e manutenção de funções neuromotoras. Em humanos foi destacado o papel importante do produto proteico do *PPP4R2* em uma disfunção motora que leva a perda progressiva de neurônios motores e causa paralisia muscular (atrofia muscular espinal – SMA) por meio de interação regulatória do complexo proteico SMN (*Survival of Motor Neuron*), o qual em níveis reduzidos pode levar a essa neuropatia (BOSIO et al., 2012). Por sua vez, o *PDZRN3* tem participação na morfogênese embrionária, atuando em uma via importante na orientação planar e organização de plexos vasculares altamente ramificados (SEWDUTH et al., 2014).

A partir da análise de enriquecimento foram observados quatro termos significativos no ssGWAS1 e 16 no ssGWAS2 (Tabela 7).

Tabela 7 – Identificação (GO.ID) e descrição (Termos) de termos GO (*Gene Ontology*) enriquecidos significativamente após realização de GWAS na linhagem de corrida da raça Quarto de Milha.

GO.ID	Genes	Termos	Ann. ¹	Sign. ²	Exp. ³	p-value
ssGWAS1						
GO:0006878	SLC31A2, SLC31A1	cellular copper ion homeostasis	2	2	0,17	0,0069
GO:0000724	GEN1, SMC6, PPP4R2	double-strand break repair via homologou...	6	3	0,5	0,0363
GO:0007528	PDZRN3, TNC	neuromuscular junction development	4	2	0,33	0,0369
GO:0072659	SPTBN1, PIGW, EPB41L3	protein localization to plasma membrane	18	3	1,5	0,0371
ssGWAS2						
GO:0045796	ABCG8, ABCG5	negative regulation of intestinal choles...	2	2	0,07	0,0013
GO:0006878	SLC31A1, SLC31A2	cellular copper ion homeostasis	2	2	0,07	0,0013
GO:0007588	ABCG8, ABCG5	excretion	4	2	0,14	0,0073
GO:0033344	ABCG8, ABCG5	cholesterol efflux	5	2	0,18	0,0119
GO:0035434	SLC31A1, SLC31A2	copper ion transmembrane transport	2	2	0,07	0,0354
GO:0030042	PLEKHH2, MICAL1	actin filament depolymerization	6	2	0,22	0,0355
GO:0031145	CDC26	anaphase-promoting complex-dependent cat...	1	1	0,04	0,0361
GO:0070979	CDC26	protein K11-linked ubiquitination	1	1	0,04	0,0361
GO:0035989	COL11A1	tendon development	1	1	0,04	0,0361
GO:0003431	COL27A1	growth plate cartilage chondrocyte devel...	1	1	0,04	0,0361
GO:0033617	COX20	mitochondrial respiratory chain complex ...	1	1	0,04	0,0361
GO:0042953	CUBN	lipoprotein transport	1	1	0,04	0,0361
GO:0009157	DCK	deoxyribonucleoside monophosphate biosyn...	1	1	0,04	0,0361
GO:0036305	ENAM	ameloblast differentiation	1	1	0,04	0,0361
GO:0070175	ENAM	positive regulation of enamel mineraliza...	1	1	0,04	0,0361
GO:0001561	HAO1	fatty acid alpha-oxidation	1	1	0,04	0,0361

¹ Genes anotados para o termo descrito; ² Genes significativos dentre os genes anotados; ³ Esperado.

Como já mencionado, a característica IV max foi caracterizada como quantitativa neste trabalho e espera-se, portanto, grande número de alterações genéticas com pequeno efeito que contribuam para diferenciações no metabolismo, na anatomia e no comportamento de animais da linhagem de corrida da raça Quarto de Milha. Nesse sentido, o excesso de variedade de termos de ontologia genética e genes associados dificulta o enriquecimento, o que é descrito para características poligênicas, sobretudo se tomado o ssGWAS1, em que foram encontradas maior número de regiões de importância (gVar > 1%) com menores efeitos. De modo geral, ambos os métodos (ssGWAS1 e ssGWAS2) não apresentaram enriquecimentos altamente significativos como os observados para características governadas por genes de efeito maior. Entretanto, a

ordenação e a ligação de genes a processos biológicos conduziram à investigação sistemática de funções e vias.

Entre os termos mais enriquecidos estão os ligados aos genes *ABCG5*, *ABCG8*, *SLC31A1* e *SLC31A2*. Os genes *SLC31A1* e *SLC31A2* foram enriquecidos de forma significativa em ssGWAS1 e ssGWAS2. Estes genes codificam as proteínas CTR1 e CTR2 que controlam o fluxo intra e extracelular de cobre (Cu^{2+}) (WEE et al., 2013). O papel biológico do cobre inclui a incorporação em cuproenzimas para: fosforilação oxidativa (citocromo-c oxidase); atividade antioxidante celular (superóxido dismutase); formação de tecido conjuntivo (lisil oxidase); função neurológica (dopamina p-hidroxilase); e metabolismo do ferro (ceruloplasmina) (FESTA e THIELE, 2011). A incapacidade de controlar o equilíbrio de cobre está relacionada à doenças genéticas de sobrecarga ou deficiência deste elemento, o que, recentemente, foi associado a distúrbios neurodegenerativos (KIM (Nevitt) (Thiele) et al., 2008).

Considerando o ssGWAS1, observou-se menor numero de genes associados a termos significativos. Os genes *GEN1* e *SMC6* foram relacionados basicamente ao processo de recombinação cromossômica homologa, atuando na Junção de Holliday (HJ) e quebra de cadeias duplas de DNA. O mecanismo HJ é altamente conservado desde bactérias até humanos (CHAN e WEST, 2015). Os genes *SPTBN1*, *PIGW* e *EPB41L3* foram associados ao direcionamento e localização de proteínas na membrana plasmática. O *SPTBN1* é da família de genes beta espectrina (não eritrocitário). As espectrinas são membros da superfamília de proteínas de ligação cruzada F-actina que compõem uma rede de citoesqueletos de actina subjacente à membrana plasmática e funcionam na determinação da forma celular, disposição de proteínas transmembranares e organização de organelas (JIANG et al., 2010).

Embora não significativos na análise de enriquecimento, alguns genes encontrados pelo método ssGWAS1 podem trazer contribuições importantes para entendimento de fatores genéticos que atuam sobre IV max. Os genes *WHRN* e *SYT4* estão associados à formação da camada de células de *Purkinje* do cerebelo e à regulação de catecolaminas (dopamina), respectivamente. As células de *Purkinje* são encontradas no interior da camada de *Purkinje* no cerebelo (PURVES et al., 2008), e, entre as suas funções, as relacionadas com o movimento são as mais conhecidas. O cerebelo não inicia o movimento, mas contribui para a coordenação, precisão e precisão no tempo, de forma que danos cerebelares produzem transtornos no movimento fino, equilíbrio, postura e aprendizagem motora em humanos (FINE et al., 2002). O gene *SYT4* codifica em mamíferos uma das mais de 17 isoformas conhecidas de sinaptotagmina (syt). Essas

proteínas, localizadas nas vesículas sinápticas, estão ligadas à sinapse e relacionam a sua função à das catecolaminas, que incluem epinefrina, norepinefrina e dopamina, as quais atuam como neurotransmissores excitatórios em diversos locais do sistema nervoso central e periférico (PURVES et al., 2008; FITZGERALD et al., 2011).

O genes *SOCS3* e *ACACA* também descritos no ssGWAS1, têm importante papel no metabolismo energético. O *SOCS3* está envolvido na regulação da leptina de forma específica em diversas células. Estudos conduzidos em camundongos sugerem que a presença da proteína *SOCS3* em alguns tipos de neurônios regula negativamente a sinalização da leptina e desempenha um papel importante na mediação da sensibilidade à leptina, da homeostase da glucose e do gasto energético (ZHANG, et al., 2012). Em outra pesquisa, Yang et al. (2012) demonstraram que *SOCS3* no interior do músculo esquelético é regulador crítico da leptina e da ação da insulina. Além disso, os autores relacionam *SOCS3* à regulação de acetil-CoA carboxilase (*ACC*), proteína transcrita pelo gene *ACACA* e também envolvida no metabolismo de ácidos graxos

Tomando os genes enriquecidos de forma significativa no ssGWAS2, *ABCG5* e *ABCG8*, localizados no ECA 15, apresentaram os menores valores de *p*. Esses genes codificam os semi-transportadores de *ATP binding cassette* (*ABC*) G5 e G8 (*ABCG5* e *ABCG8*) que limitam a absorção intestinal e promovem a excreção biliar de esteróis neutros (GRAFF et al., 2003). A superfamília transportadora de *ABC* contém a maior família de genes transportadores. As proteínas *ABCG5* e *ABCG8* translocam grande variedade de substratos, incluindo açúcares, aminoácidos, íons metálicos, peptídeos e proteínas, e grande número de compostos hidrofóbicos e metabolitos através de membranas extras e intracelulares (DEAN et al., 2001). Estes genes estão ligados a sitosterolemia em humanos, doença hereditária rara em que as gorduras de vegetais se acumulam no sangue e nos tecidos, conduzindo à arteriosclerose, glóbulos vermelhos anormais e depósitos de gordura nos tendões (LEE et al., 2001).

Os genes *PLEKHH2* e *MICAL1* foram associados à despolimerização do filamento de actina em humanos. Contudo, a maior parte dos estudos com *PLEKHH2* foi em relação à nefropatia (GREENE et al., 2008). Por sua vez, *MICAL1* possui evidências mais claras de, a partir de seu produto proteico, participar do processo despolimerização da actina-F (SCHIMIDT et al., 2008), o que está relacionado a muitos processos celulares importantes, incluindo contração muscular, motilidade celular, divisão celular e citocinese, movimento de vesículas e organelas, sinalização celular e estabelecimento e manutenção de junções e forma celular (DOMINGUEZ e HOLMES, 2011).

Ressalta-se ainda os genes *COL11A1* e *COL27A1* que estão ligados a termos de desenvolvimento de tendões e crescimento e desenvolvimento de condrócitos em placas de cartilagem. O gene *COL11A1* codifica uma das duas cadeias alfas do colágeno tipo XI (ZHIDKOVA et al., 1995). Disfunções desse gene estão ligadas à síndrome de Stickler e a síndrome de Marshall em humanos, a quais são caracterizadas por anomalias faciais distintas, problemas oculares, perda auditiva e problemas articulares (STICKLER et al., 1965; ANNUNEM et al., 1999). O gene *COL27A1* codifica membro da família de colágeno fibrilar e atua na calcificação da cartilagem e na transição da cartilagem para osso. Em humanos, disfunções desse gene levam à síndrome de *Steel* que pode ocasionar deformidades ósseas no crânio e escoliose (GONZAGA-JAUREGUI et al., 2015).

O gene *COX20* está localizado na região SS2_REG12, a qual explicou individualmente a maior % da variância de IV max encontrada neste estudo. Este gene foi descrito em leveduras e seu ortólogo em humanos é denominado *FAM36A*. O produto deste gene atua na incorporação de subunidades ao complexo citocromo C oxidase (COX) da cadeia respiratória. Uma mutação *missense* no gene *FAM36A* foi associada a pacientes humanos com ataxia (perda de coordenação) e hipotonia muscular. A interação da proteína FAM36A com COX2 já foi destacada, visto que em pacientes afetados havia o acúmulo de COX1 e ausência de COX2 na montagem do complexo COX, suportando que o FAM36A modificado dificulta o passo inicial na incorporação da subunidade COX2 (SZKLARCZYK et al., 2013). Previamente, a interação das isoformas (1 e 2) de COX4, foi relacionada à melhor ou pior eficiência na respiração celular em situações de hipóxia (FUKUDA et al., 2010). Em estudo realizado por Gu et al. (2010) foi identificada associação fraca, porém significativa, entre o SNP intrônico g.22684390C>T do gene *COX4I2* equino (ECA 22) e a retrospectiva de desempenho em corridas na raça PSI. Entretanto, esse polimorfismo não teve o mesmo efeito em equinos da linhagem de corrida da raça Quarto de Milha, possivelmente pelo fato de o metabolismo muscular ser predominantemente anaeróbio nas corridas de curta distância (PEREIRA et al., 2016b).

A característica velocidade em diversas raças equinas que disputam provas de diferentes distâncias e intensidades é caracterizada por importantes adaptações fisiológicas, sobretudo ligadas ao metabolismo muscular aeróbio e anaeróbio (NIELSEN et al., 2006). Contudo, neste estudo conduzido com a raça Quarto de Milha foi observado grande número de genes relacionados ao sistema neuromotor, desenvolvimento ósseo e de tendões, respostas hormonais, contração muscular, metabolismo energético e de esteróis. Em regiões genômicas possivelmente selecionadas para desempenho dentro da linhagem de corrida, Meira et al. (2015) descreveram muitos genes, mesmo utilizando

níveis de classificação de processos biológicos mais superficiais, como relacionados ao sistema nervoso. Já na raça PSI, Gu et al. (2009) indicaram regiões genômicas responsáveis por funções biológicas como a sinalização de insulina, metabolismo energético, e processos metabólicos de esteroides, entre outros, como sendo importantes para a diferenciação do desempenho na raça. Ressalta-se que vários genes enriquecidos como: *CUBN*; *DCK*; *ENAM*; *HAO1*; *PIGW*; e *EPB41L3* não têm funções bem definidas, mesmo considerando organismos modelo, e por isso não puderam ser discutidos de forma aprofundada. Destaca-se também o fato de que os termos para processos biológicos e funções dos genes descritos aqui tenham sido associados mais claramente ao metabolismo e fisiologia, principalmente, de humanos, e que há pouca ou nenhuma informação sobre as funções destes genes em processos biológicos na espécie equina. Entretanto, humanos e equinos possuem alta similaridade no que diz respeito à fisiologia do exercício (VON ENGELHARDT, 1992; LINDINGER, 1999; DAVIS et al., 2005; CAPOMACCIO et al., 2011), o que pode endossar a importância de genes prospectados neste estudo como candidatos para desempenho em corridas na raça Quarto de Milha.

4. Conclusões

Até o momento poucos trabalhos com enfoque genômico tiveram como objetivo central entender o tipo de controle genético e/ou desvendar regiões genômicas e genes associados à característica de desempenho na linhagem de corrida da raça Quarto de Milha. O método ssGBLUP foi empregado pela primeira vez em estudos deste tipo na raça, o que permitiu prospectar regiões que atuam sobre o IV max. Se consideradas simultaneamente, as duas abordagens empregadas neste estudo (ssGWAS1 e ssGWAS2), permitiram a identificação de 33 janelas de SNPs, com 1 Mb de extensão, em 14 cromossomos, explicando mais que 1% da variância genética total da característica. Ao todo foram prospectados 140 genes nestas regiões. Como esperado, a característica IV max é controlada por grande número de genes de pequeno efeito. De forma geral, os genes *PDZRN3*, *PPP4R2*, *SLC31A1*, *SLC31A2*, *GEN1*, *SMC6*, *WHRN*, *SYT4*, *ABCG5*, *ABCG8*, *PLEKHH2*, *MICAL1*, *COL11A1*, *COL27A1* e *COX20*, os quais foram destacados por meio de análise de enriquecimento, estão relacionados ao sistema neuromotor, ao desenvolvimento ósseo e de tendões, à respostas hormonais, à contração muscular, ao metabolismo energético e de esteróis. Estas funções podem ser consideradas de relevância para o desempenho atlético, corroborando estes genes como candidatos para o desempenho em corridas na raça Quarto de Milha.

5. Referências

- ALEXA, A.; RAHNENFUHRER, J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.26.0. 2016. Disponível em: <<https://bioconductor.org/packages/release/bioc/html/topGO.html>>. Acesso em: 8 jan. 2017.
- ABQM. Associação brasileira dos criadores de cavalos Quarto de Milha. Disponível em: <<http://www.abqm.com.br/item1.htm>>. Acesso em: 15 de março. 2016.
- AGUILAR, I.; MISZTAL, I.; JOHNSON, D.L.; LEGARRA, A.; TSURUTA, S.; LAWLOR, T.J. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, Champaign, v. 93, p.743–752, 2010.
- AGUILAR, I.; MISZTAL, I.; TSURUTA, S.; WIGGANS, G.R.; LAWLOR, T.J. Multiple trait genomic evaluation of conception rate in Holsteins. **Journal of Dairy Science**, Champaign, v. 94, p. 2621-2624, 2011.
- AMERICA'S HORSE DAILY. All About the Racing American Quarter Horse. 2008. Disponível em: <http://americashorsedaily.com/all-about-the-racing-american-quarter-horse/#.Urx8N_RDvik> Acesso em: 12 nov. 2013.
- AMERICAN QUARTER HORSE ASSOCIATION, AQHA. Horse-Breeding History. Disponível em: < <https://www.aqha.com/daily/breeding/2016/breeding-archive/horse-breeding-history/>>. Acesso em: 20 out. 2016.
- ANNUNEN, S.; KORKKO, J.; CZARNY, M.; WARMAN, M. L.; BRUNNER, H. G.; KAARIAINEN, H.; MULLIKEN, J. B.; TRANEBJAERG, L.; BROOKS, D. G.; COX, G. F.; CRUYSBERG, J. R.; CURTIS, M. A.; DAVENPORT, S. L.; FRIEDRICH, C. A.; KAITILA, I.; KRAWCZYNSKI, M. R.; LATOS-BIELENSKA, A.; MUKAI, S.; OLSEN, B. R.; SHINNO, N.; SOMER, M.; VIKKULA, M.; ZLOTOGORA, J.; PROCKOP, D. J.; ALA-KOKKO, L. Splicing mutations of 54-bp exons in the COL11A1 gene cause Marshall syndrome but other mutations cause overlapping Marshall/Stickler phenotypes. **The American Journal of Human Genetics**, v. 65, n. 4, p. 974-983, 2016
- BINNS, M. M.; BOEHLER, D. A.; LAMBERT, D. H. Identification of the myostatin locus (MSTN) as having a major effect on optimum racing distance in the Thoroughbred horse in the USA. **Animal Genetics**, v. 41, supl. 2, p. 154-158, 2010.
- BOSIO, Y.; BERTO, G.; CAMERA, P.; BIANCHI, F.; AMBROGIO, C.; CLAUS, P.; DI CUNTO, F. PPP4R2 regulates neuronal cell differentiation and survival, functionally cooperating with SMN. **European Journal of Cell Biology**, v. 91, p. 662–674, 2012.
- CAPOMACCIO, S.; CAPPELLI, K.; SPINSANTI, G.; MENCARELLI, M.; MUSCETTOLA, M.; FELICETTI, M.; BONIFAZI, M. Athletic humans and horses: Comparative analysis of interleukin-6 (IL-6) and IL-6 receptor (IL-6R) expression in peripheral blood mononuclear cells in trained and untrained subjects at rest. **BMC Physiology**, v. 11, n. 3, 2011. Disponível em: < <http://doi.org/10.1186/1472-6793-11-3>>.
- CHAN, Y. W.; WEST, S. GEN1 promotes Holliday junction resolution by a coordinated nick and counter-nick mechanism. **Nucleic Acids Research**, v. 43, n. 22, p. 10882-10892, 2015.

CHEN, C.Y.; MISZTAL, I.; AGUILAR, I.; TSURUTA, S.; MEUWISSEN, T.H.E.; AGGREY, S.E.; WING, T.; MUIR, W.M. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. **Journal of Animal Science**, Champaign, v. 89, p. 23-28, 2011.

CLAYTON, D. snpStats: SnpMatrix and XSnpmatrix classes and methods. R package version 1.24.0. 2015. Disponível em: <<http://bioconductor.org/packages/release/bioc/html/snpStats.html>>.

CORRÊA, M. J. M.; MOTA, M. D. S. Genetic evaluation of performance traits in Brazilian Quarter Horse. **Journal of Applied Genetics**, v. 48, p. 145-151, 2007.

DAVIS, M. S.; MALAYER, J. R.; VANDEVENTER, L.; ROYER, C. M.; MCKENZIE, E. C.; WILLIAMSON, K. K. Cold weather exercise and airway cytokine expression. **Journal of Applied Physiology**, v. 98, p. 2132-2136, 2005. Disponível em: <10.1152/jappphysiol.01218.2004>.

DEAN, M.; RZHETSKY, A.; ALLIKMETS, R. The Human ATP-Binding Cassette (ABC) Transporter Superfamily. **Genome Research**, v. 11, p. 1156-1166, 2001.

DOMINGUEZ, R.; HOLMES, K. C. Actin structure and function. *Annual Review of Biophysics*, v. 40, p. 169-186, 2011.

DURINCK, S.; SPELLMAN, P.; BIRNEY, E.; HUBER, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. **Nature Protocols**, v.4, p. 1184–1191, 2009.

FESTA, R. A.; THIELE, D. J. Copper: an essential metal in biology. **Current Biology**, v. 21, n. 21, p. 877-883, 2011.

FINE, E. J.; IONITA, C. C.; LOHR, L. The history of the development of the cerebellar examination. *Seminars in Neurology*, v. 22, ed. 4, p. 375–384, 2002. Disponível em: <doi:10.1055/s-2002-36759>.

FITZGERALD, P. A. Adrenal medulla and paraganglia. In: Gardner DG et al, (Ed.9). *Basic & Clinical Endocrinology*. Greenspan's: ed. McGraw-Hill, 2011.

FUKUDA, R.; ZHANG, H.; KIM, J.W.; SHIMODA, L.; DANG, C.V.; SEMENZA, G.L. HIF-1 regulates cytochrome oxidase subunits to optimize efficiency of respiration in hypoxic cells. **Cell**, v. 129, p. 111-122, 2007.

GO, Y. Y.; BAILEY, E.; COOK, D. G.; COLEMAN, S. J.; MACLEOD, J. N.; CHEN, K.-C.; TIMONEY, P. J.; BALASURIYA, U. B. R. Genome-Wide Association Study among Four Horse Breeds Identifies a Common Haplotype Associated with In Vitro CD3+ T Cell Susceptibility/Resistance to Equine Arteritis Virus Infection. **Journal of Virology**, v. 85, n. 24, p. 13174-13184, 2011.

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

GONZAGA-JAUREGUI, C.; GAMBLE, C. N.; YUAN, B. Mutations in COL27A1 cause Steel syndrome and suggest a founder mutation effect in the Puerto Rican population. **European Journal of Human Genetics**, v. 23, n. 3, p. 342-346, 2015.

GRAF, G. A.; YU, L.; LI, W. P.; GERARD, R.; TUMA, P. L.; COHEN, J. C.; HOBBS, H. H. ABCG5 and ABCG8 are obligate heterodimers for protein trafficking and biliary

cholesterol excretion. **The Journal of Biological Chemistry**, v. 278, p. 48275-48282, 2003.

GREENE, C.N.; KEONG, L.M.; CORDOVADO, S.K. MUELLER, P. W. Sequence variants in the PLEKHH2 region are associated with diabetic nephropathy in the GoKinD study population. **Human Genetics**, v. 124, n. 3, p. 255-262, 2008.

GU, J.; MACHUGH, D. E.; MCGIVENY, B. A.; PARK, S. D. E.; KATZ L. M.; HILL E. M. Association of sequence variants in CKM (creatine kinase, muscle) and COX4I2 (cytochrome c oxidase, subunit 4, isoform 2) genes with racing performance in Thoroughbred horses. **Equine Veterinary Journal**, v. 42, p. 569-575, 2010.

HILL, E.M.; GU, J.; EIVERS, S.S.; FONSECA, R.G.; MCGIVENY, B.A.; GOVINDARAJAN, P. et al A Sequence Polymorphism in MSTN Predicts Sprinting Ability and Racing Stamina in Thoroughbred Horses. **PLoS One**, v. 5, p. 1-6, 2010a.

HILL, E. W.; MCGIVNEY, B. A.; GU J.; WHISTON, R.; MACHUGH D. E. A genomewide SNP-association study confirms a sequence variant (g.66493737C>T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. **BMC Genomics**, v. 11, n.552, 2010b. Disponível em: <<http://doi.org/10.1186/1471-2164-11-55>>.

HILL, E.M.; GU, J.; MCGIVENY, B.A.; MACHUGH, D.E. Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance. **Animal Genetics**, v. 41, supl. 2, p. 56-63, 2010c.

JIANG, X.; GILLEN, S.; ESPOSITO, I.; GIESE, N. A.; MICHALSKI, C. W.; FRIESS, H.; KLEEFF, J. Reduced expression of the membrane skeleton protein beta1-spectrin (SPTBN1) is associated with worsened prognosis in pancreatic cancer. **Histology and Histopathology**, v. 25, p. 1497–1506, 2010.

KANG, H. M.; SUL, J. H.; SERVICE, S. K.; ZAITLEN, N. A.; KONG, S. Y.; FREIMER, N. B.; SABATTI, C.; ESKIN, E. Variance component model to account for sample structure in genome-wide association studies. **Nature Genetics**, v. 42, n. 4, p. 348-354, 2010.

KIM, B-E.; NEVITT, T.; THIELE, D. J. Mechanisms for copper acquisition, distribution and regulation. **Nature Chemical Biology**, v. 4, p. 176 -185, 2008.

LEE, M-H.; LU, K.; HAZARD, S.; YU, H.; SHULENIN, S.; HIDAKA, H.; KOJIMA, H.; ALLIKMETS, R.; SAKUMA, N.; PEGORARO, R.; SRIVASTAVA, A. K.; SALEN, G.; DEAN, M.; PATEL, S. B. Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. **Nature Genetics**, v. 27, p. 79-83, 2001.

LEMOES, M. V. A.; CHIAIA, H. L. J.; BERTON, M. P.; FEITOSA, F. L. B.; ABOUJAOUD, C.; CAMARGO, G. M. F.; PEREIRA, A. S. C.; ALBUQUERQUE, L. G.; FERRINHO, A. M.; MUELLER, L. F.; MAZALLI, M. R.; FURLAN, J. J. M.; CARVALHEIRO, R.; GORDO, D. M.; TONUSSI, R.; ESPIGOLAN, R.; SILVA, R. M. DE O.; DE OLIVEIRA, H. N.; DUCKETT, S.; AGUILAR, I.; BALDI, F. Genome-wide association between single nucleotide polymorphisms with beef fatty acid profile in Nelore cattle using the single step procedure. **BMC Genomics**, v. 17, n. 1, p. 1471-2164, 2016.

LIMA, R.A.S.; SHIROTA, R.; BARROS, G.S.C. Estudo do Complexo do Agronegócio Cavalo no Brasil. CEPEA – Centro de Estudos Avançados em Economia Aplicada. Piracicaba: ESALQ/USP, 2006.

LINDINGER, M. I. Exercise in the heat: thermoregulatory limitations to performance in humans and horses. **Canadian Journal of Applied Physiology**, v. 24, p. 152-163, 1999.

MAGALHÃES, A. F. B.; DE CAMARGO, G. M. F.; FERNANDES, G.A.; JUNIOR, GORDO, D. G. M.; TONUSSI, R. L. et al. (2016). Genome-Wide Association Study of Meat Quality Traits in Nellore Cattle. **PLoS ONE**, v. 11, n. 6, 2016. Disponível em: <e0157845. doi: 10.1371/journal.pone.0157845>.

McCUE, M. E.; BANNASCH, D. L.; PETERSEN, J. L.; GURR, J.; BAILEY, E.; BINNS, M. M.; DISTL, O.; GUERIN, G.; HASEGAWA, T.; HILL, E. W.; LEEB, T.; LINDGREN, G.; PENEDO, M. C.; ROED, K. H.; RYDER, O. A.; SWINBURNE, J. E.; TOZAKI, T.; VALBERG, S. J.; VAUDIN, M.; LINDBLAD-TOH, K.; WADE, C. M.; MICKELSON, J. R. A high density SNP array for the domestic horse and extant *Perissodactyla*: utility for association mapping, genetic diversity, and phylogeny studies. **PLoS Genetics**, v. 8, n. 1, 2012. Disponível em: <e1002451. doi: 10.1371/journal.pgen.1002451>.

McQUEEN, C. M.; DOAN, R.; DINDOT, S. V.; BOURQUIN, J. R.; ZLATEV, Z. Z. Identification of Genomic Loci Associated with *Rhodococcus equi* Susceptibility in Foals. **PLoS ONE**, v. 9, n. 6, 2014. Disponível em: <e98710. doi: 10.1371/journal.pone.0098710>.

MEIRA, C. T.; FORTES, M. R. S.; FARAH, M. M.; PORTO NETO, L.; KELLY, M.; MOORE, S. S.; PEREIRA, G. L.; CHARDULO, L. A. L.; CURI, R. A. Speed index in the racing Quarter Horse: a genome-wide association study. **Journal of Equine Veterinary Science**, v. 34, p. 1263-1268, 2014a.

MEIRA, C.T.; CURI, R.A.; FARAH, M.M.; OLIVEIRA, H.N.; BELTRAN, N.A.R.; SILVA, J.A.IV.; MOTA, M.D.S. Prospection of genomic regions divergently selected in racing line of Quarter Horses in relation to cutting line. **Animal**, v. 8, n. 11, p. 1754-1764, 2014b.

MISZTAL, I.; TSURUTA, S.; LOURENÇO, D.; AGUILAR, I.; LEGARA, A.; VITEZICA, Z. Manual for BLUPF90 family of programs. University of Georgia, Athens, 2014.

MISZTAL, I.; TSURUTA, S.; STRABEL, T.; AUVRAY, B.; DRUET, T.; LEE, D. H. BLUPF90 and related programs (BGF90). In The 7th World Congress Genetics Application Livestock Production, p. 28, Montpellier, France, 2002.

NIELSEN, B.D.; TURNER, K.K.; VENTURA, B.A.; WOODWARD, A.D.; O'CONNOR C.I. Racing speeds of quarter horses, thoroughbreds and Arabians. **Equine Veterinary Journal**, v. 36, suppl. 2, p. 128-32, 2006.

PEREIRA, G.L.; DE MATTEIS, R.; REGITANO, L.C.A.; CHARDULO, L.A.L.; CURI, R.A. MSTN, CKM, and DMRT3 Gene Variants in Different Lines of Quarter Horses. **Journal of Equine Veterinary Science**, v. 39, p. 33-37, 2016a.

PEREIRA, G.L.; DE MATTEIS, R.; MEIRA, C.T.; REGITANO, L.C.A.; SILVA, J.A.IV.; CHARDULO, L.A.L.; CURI, R.A. Comparison of Sequence Variants in the PDK4 and COX4I2 Genes Between Racing and Cutting Lines of Quarter Horses and Associations With the Speed Index. **Journal of Equine Veterinary Science**, v. 39, p. 1-6, 2016b.

PETERSEN, J.L.; MICKELSON, J. R.; COTHRAN, E. G.; ANDERSSON, L. S.; AXELSSON, J. Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data. **PLoS ONE**, v.8, n. 1, 2013. Disponível em: <e54997. doi: 10.1371/journal.pone.0054997>.

PRICE AL.; ZAITLEN NA.; REICH D.; PATTERSON N. New approaches to population stratification in genome-wide association studies. **Nature Reviews Genetics**, v. 11, n. 7, p. 459–463, 2010.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M.A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P.I.; DALY, M.J.; SHAM, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, Baltimore, v. 81, p. 559-575, 2007.

PURVES, D.; AUGUSTINE, G. J.; FITZPATRICK, D.; HALL, W. C.; LAMANTIA, A.; MCNAMARA, J.O.; WHITE L.E. **Neuroscience**. 2. ed. Stamford: Sinauer Associates Inc., 2008. p. 432–434.

R Core Team (2016). R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. URL < <https://www.R-project.org/>>

RESENDE, M.D.V.; SILVA, F.F.; LOPES, P.S.; AZEVEDO, C.F. Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial. Viçosa: Universidade Federal de Viçosa, 2012. 291p.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F.S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**. v.15, p.478, 2014.

SCHMIDT E. F.; SHIM S. O.; STRITTMATTER S. M. Release of MICAL autoinhibition by semaphorin-plexin signaling promotes interaction with collapsin response mediator protein. **Journal of Neuroscience**, v. 28, p. 2287–2297, 2008.

SCHRÖDER, W.; KLOSTERMANN, A.; STOCK, K.F.; DISTL, O. A genome-wide association study for quantitative trait loci of show-jumping in Hanoverian Warmblood horses. **Animal Genetics**, p.1-9, 2011.

SEWDUTH, R. N.; JASPARD-VINASSA, B.; PEGHAIRE, C.; GUILLABERT, A.; FRANZL, N.; LARRIEU-LAHARGUE, F. The ubiquitin ligase PDZRN3 is required for vascular morphogenesis through Wnt/planar cell polarity signalling. **Nature Communications**, v. 5, n. 4832, 2014. Disponível em: <doi: 10.1038/ncomms5832. pmid:25198863>.

SHIN, D-H.; LEE, J. W.; PARK, J-E.; CHOI, I-Y.; OH, H-S.; KIM, H. J.; KIM, H. Multiple Genes Related to Muscle Identified through a Joint Analysis of a Two-stage Genome-wide Association Study for Racing Performance of 1,156 Thoroughbreds. **Asian-Australasian Journal of Animal Sciences**, v. 28, n. 6, p. 771-781, 2015.

SILVA, R. M. O.; STAFUZZA, N. B.; FRAGOMENI, B. O.; CAMARGO, G. M. F.; CEACERO, T. M.; CYRILLO, J. N. S. G.; BALDI, F.; BOLIGON, A. A.; MERCADANTE, M. E. Z.; LOURENCO, D. L.; MISZTAL, I.; LUCIA GALVÃO DE ALBUQUERQUE, L. G. Genome-Wide Association Study for Carcass Traits in an Experimental Nelore Cattle Population. **PLoS ONE**, v. 12, n. 1, 2017. Disponível em: <doi: 10.1371/journal.pone.0169860>

STICKLER, G. B.; BELAU P. G.; FARRELL F. J. Hereditary progressive arthropathy. **Mayo Clinic Proceedings**, v. 40, p. 433–455, 1965.

SZKLARCZYK, R.; WANSCHERS, B. F.; NIJTMANS L. G.; RODENBURG, R. J.; ZSCHOCKE, J.; DIKOW N.; VAN DEN BRAND, M. A.; HENDRIKS-FRANSSSEN, M. G.; GILISSEN, C.; VELTMAN, J. A.; NOOTEBOOM, M.; KOOPMAN, W. J.; WILLEMS, P. H.; SMEITINK, J. A.; HUYNEN, M. A.; VAN DEN HEUVEL, L. P. A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase assembly and is associated with ataxia and muscle hypotonia. *Human Molecular Genetics*, v. 22, p. 656-667, 2013.

TOZAKI, T.; MIYAKE, T.; KAKOI, H.; GAWAHARA, H.; SUGITA, S.; HASEGAWA, T.; ISHIDA, N.; HIROTA, K.; NAKANO, Y. A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene. *Animal Genetics*, v. 41, suppl. 2, p. 28–35, 2010.

VANRADEN, P.M. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, v.91, p.4414-23, 2008.

VITEZICA, Z. G.; AGUILAR, I.; MISZTAL, I.; LEGARRA, A. Bias in genomic predict for populations under selection. *Genetic Research*, p. 1- 10, 2011.

VON ENGELHARDT, W. Physical performance: a comparison between horses and men. *Dtsch Tierarztl Wochenschr*, v. 99, p. 24-26, 1992.

WANG, C.; MA, P.; ZHANG, Z.; DING, X.; LIU, J.; FU, W.; WENG, Z.;ZHANG, Q. Comparison of five methods for genomic breeding value estimation for the common dataset of the 15th QTL-MAS Workshop. *BMC Proceedings*, London, v. 6, p. 1-5, 2012.

WANG, H.; MISZTAL, I.; AGUILAR, I.; LEGARRA, A.; FERNANDO, R.L.; VITEZICA, Z.; OKIMOTO, R.; WING, T.; HAWKEN, R.; MUIR W.M. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chicken. *Genetics Research*, v.94, p.73-83, 2014.

WEE, N. K. Y.; WEINSTEIN, D. C.; FRASER, S. T.; ASSINDER, S. J. The mammalian copper transporters CTR1 and CTR2 and their roles in development and disease. *The International Journal of Biochemistry & Cell Biology*, v. 45, n. 5, p. 960–963, 2013.

WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. **Springer-Verlag**, New York, 2009.

YANG, Z.; HULVER, M.; MCMILLAN, R.P.; CAI, L.; KERSHAW, E.E.; YU, L.; XUE, B.; SHI, H. Regulation of Insulin and Leptin Signaling by Muscle Suppressor of Cytokine Signaling 3 (SOCS3). *PLoS ONE*. v.7, 2012. Disponivel em: < doi: 10.1371/journal.pone.0047493>

ZHANG, W. N.; WANG, L.; WANG, Q.; LUO, X.; FANG, D. F.; CHEN, Y.; PAN, X.; MAN, J. H.; XIA, Q.; JIN, B. F. CUEDC2 (CUE Domain-containing 2) and SOCS3 (Suppressors of Cytokine Signaling 3) cooperate to negatively regulate janus kinase 1/signal transducers and activators of transcription 3 signaling. *The Journal of Biological Chemistry*, v. 287, p. 382–392, 2012.

ZHANG, Z.; LIU, J.; DING, X.; BIJMA, P.; DE KONING, D. J. e ZHANG, Q. Best linear unbiased prediction of genomic breeding values using a trait-specific marker derived relationship matrix. *PLoS One*, v. 5, e12648, 2010.

ZHIDKOVA, N. I.; JUSTICE, S. K.; MAYNE, R. Alternative mRNA processing occurs in the variable region of the pro- α 1(XI) and pro- α 2(XI) collagen chains. *The Journal of Biological Chemistry*, v. 270, p. 9486-9493, 1995.

CAPITULO 4 – Sequenciamento do exoma e regiões UTR de equinos em regiões relacionadas ao desempenho em corridas na raça Quarto de Milha

RESUMO – Dentre os equinos selecionados para velocidade, a linhagem de corrida da raça Quarto de Milha se destaca pelo alto desempenho em provas de curtas distâncias, sendo considerados os mais velozes do mundo. Contudo, poucos trabalhos têm pesquisado, de forma mais aprofundada, polimorfismos e genes que possam atuar sobre essa característica. O objetivo deste trabalho foi analisar exomas e regiões UTR de equinos de corrida da raça Quarto de Milha contrastantes para Índice de Velocidade máximo (IV max) em regiões previamente associadas à característica por meio de GWAS, visando a prospecção de polimorfismos gênicos causais, ligados ou em forte desequilíbrio de ligação com o desempenho em corridas. Para isso, foram utilizadas, inicialmente, informações genotípicas e fenotípicas de 360 equinos da linhagem de corrida da raça Quarto de Milha, os quais foram previamente genotipados com chip de SNPs para obtenção das estimativas individuais de valores genéticos genômicos (*Genomic Estimated Breeding Values* – GEBV) de desempenho. A partir dessas informações foram formados dois grupos de animais contrastantes para desempenho em corridas (20 animais de IV max superior e 20 IV max inferior), para ser sequenciados. Foram realizadas duas corridas *rapid run* para dois *pools* de bibliotecas com 20 e 19 amostras escolhidas aleatoriamente entre as 40 de IVs contrastantes. Para a primeira (n= 20) foram obtidas 2.498.929.809 sequências pareadas (*paired-end*) e para a segunda (n= 19) 1.629.737.973. Foram observadas leituras de boa qualidade para toda extensão das *reads* sequenciadas e cobertura média de 43x. Foram identificadas 1.203 variantes (1.105 SNPs e 93 InDels) em 33 regiões de interesse obtidas, anteriormente, por meio de estudo de GWAS, das quais 61,3% não estavam registradas/depositadas no banco de dados de variantes de equinos. Do total de polimorfismos, 29 (24 SNPs e 5 InDels) foram considerados de importância com base em três abordagens distintas e independentes: escores SIFT classificado como deletério (<0,05), grau de impacto na região consenso de cada polimorfismo, e frequências alélicas diferentes, identificadas pelo teste de Fisher ($p < 0,01$), entre os grupos de cavalos contrastantes para IV max. Com isso, oito genes descritos como candidatos funcionais e posicionais em trabalhos anteriores (*ABCG5*, *COL11A1*, *GEN1*, *SOCS3*, *MICAL1*, *SPTBN1*, *EPB41L3* e *SHQ1*), e oito novos genes candidatos (*AKNA*, *ARMC2*, *FKBP15*, *LHX1*, *NOL10*, *TMEM192*, *ZFP37*, *FIG4* e *HNRNPU*), alguns com funções conhecidas e outros não, foram relacionados ao desempenho em corridas de cavalos da raça Quarto de Milha.

Palavras-chave – cavalo, InDel, Índice de Velocidade, polimorfismos, SNP.

EXOME AND UTR SEQUENCING IN REGIONS RELATED TO RACING PERFORMANCE OF QUARTER HORSES

ABSTRACT – Among horses selected for speed, the racing line of Quarter Horses is characterized by high performance in sprint races, with these animals being considered the fastest horses in the world. However, few studies have investigated in more detail the polymorphisms and genes that act on this trait. The objective of this study was to analyze exomes and UTRs in regions previously associated with this trait by GWAS in Quarter Horse racehorses with contrasting maximum speed index (SI_{max}), prospecting causal gene polymorphisms that are related to or are in strong linkage disequilibrium with racing performance. Genotypic and phenotypic records from 360 animals of the racing line of Quarter Horses, previously genotyped with an SNP chip to obtain individual genomic estimated breeding values for performance, were used for the formation and sequencing of two groups of animals with contrasting racing performance (20 animals with superior SI_{max} and 20 with inferior SI_{max}). Two rapid runs were performed using two pools of libraries containing 20 and 19 samples randomly chosen among the 40 samples with contrasting SIs. A total of 2,498,929,809 paired-end sequences were obtained in the first run (n = 20) and 1,629,737,973 in the second (n = 19). Good quality data were obtained throughout the reads sequenced, with an average coverage of 43x. A total of 1,203 variants (1,105 SNPs and 93 InDels) were identified in 33 regions of interest obtained previously by GWAS; of these, 61.3% were not registered/deposited in the horse genomic variant database. Twenty-nine of the polymorphisms (24 SNPs and 5 InDels) were considered to be important based on three different and independent approaches: SIFT scores classified as deleterious (<0.05), degree of impact on the consensus region of each polymorphism, and different allele frequencies identified by Fisher's exact test (p< 0.01) between the groups of horses with contrasting SI_{max}. Thus, eight genes described as functional and positional candidates in previous studies (*ABCG5*, *COL11A1*, *GEN1*, *SOCS3*, *MICAL1*, *SPTBN1*, *EPB41L3*, and *SHQ1*) and eight new candidate genes (*AKNA*, *ARMC2*, *FKBP15*, *LHX1*, *NOL10*, *TMEM192*, *ZFP37*, *FIG4*, and *HNRNPU*), some of them with known function, were related to racing performance in Quarter Horses.

Keywords: horse, InDel, speed index, polymorphisms, SNP.

1. Introdução

Dentre os equinos selecionados para velocidade, a linhagem de corrida da raça Quarto de Milha se destaca pelo alto desempenho em provas de curtas distâncias, sendo considerados os mais velozes do mundo (AMERICA'S HORSE DAILY, 2008). Isso se deve principalmente à adaptações do metabolismo energético e dos tipos de fibras musculares esqueléticas (NIELSEN et al., 2006; PETERSEN et al 2013). Nesse sentido, esse segmento da raça pode ser visto como um organismo modelo importante no estudo genético de alterações ligadas ao esforço intenso em períodos curtos de tempo.

Com o intuito de encontrar genes e polimorfismos de DNA específicos para a linhagem de corrida da raça Quarto de Milha e que possam estar ligados à velocidade, Meira et al. (2015a;b) realizaram estudos envolvendo assinaturas de seleção e associação ampla do genoma (*genome wide association study* – GWAS). Contudo, o pequeno número de animais utilizados nestes experimentos não foi adequado para a detecção de QTLs. Em GWAS realizado com maior número de animais e com abordagem mais apropriada (*single-step GWAS*) para a identificação de regiões genômicas, genes e polimorfismos de efeito menor que possam estar relacionados ao desempenho (índice de velocidade máximo – IV max) em cavalos Quarto de Milha de corrida, Pereira et al. (2017, capítulo 3) encontraram 33 regiões de aproximadamente 1 Mb distribuídas por 14 cromossomos equinos (ECA) que explicaram de forma separada mais de 1% da variância genética total da característica.

Em GWAS realizado na raça Puro Sangue Inglês (PSI), a qual possui grande influência na formação da linhagem de corrida da raça Quarto de milha, Shin et al. (2015) descreveram sete regiões cromossômicas importantes associadas ao desempenho em corridas (ECA 5, 8, 16, 20, 21, 28 e 30) contendo genes relacionados à fisiologia muscular. Destas regiões, cinco, localizadas nos ECA 5, 8, 16, 20 e 30, coincidiram com as encontradas por Pereira et al. (2017, capítulo 3).

Embora, na maioria dos casos, GWAS sejam eficientes para detecção de QTLs, não são capazes de gerar informações adicionais de alterações específicas em genes e sequências em regiões associadas do genoma. Atualmente os custos relacionados ao sequenciamento de regiões genômicas de tamanho significativo, ou mesmo de todo o genoma (*Whole Genome Sequencing* – WGS) tem se tornado baixos (LIU et al., 2012), possibilitando o uso de tecnologias de nova geração (*Next Generation Sequencing* – NGS) para, dentre inúmeras possibilidades, o refinamento de experimentos que utilizam painéis de *Single Nucleotide Polymorphisms* (SNPs). O sequenciamento de todo o exoma (*Whole Exome Sequencing* – WES) pode ser, do mesmo modo, empregado para esse propósito.

Essa abordagem permite a detecção de variantes genéticas (SNPs, *Insertions/Deletions* – InDels e *Copy Number Variation* – CNVs) já mapeadas, ou *de novo*, em regiões expressas e 5' e 3' UTR de todo o genoma.

O objetivo deste trabalho foi analisar exomas e sequências UTR de equinos de corrida da raça Quarto de Milha contrastantes para IV max em regiões previamente associadas com a característica por meio de GWAS (PEREIRA et al., 2017, capítulo 3), visando a prospecção de polimorfismos gênicos causais, ligados ou em forte desequilíbrio de ligação com o desempenho em corridas.

2. Material e métodos

2.1. Amostras, genotipagem e obtenção de GEBVs

Inicialmente foram utilizados 360 equinos da linhagem de corrida da raça Quarto de Milha, 78 machos e 282 fêmeas, registrados na associação brasileira de criadores (Associação Brasileira de Criadores de Cavalo Quarto de Milha – ABQM), os quais foram previamente genotipados com chip de SNPs para obtenção das estimativas individuais de valores genéticos genômicos (*Genomic Estimated Breeding Values* – GEBV) de desempenho. Desse total, 120 cavalos tiveram o sangue coletado no ano de 2011, os quais foram genotipados com painel de 54k (*Equine SNP50 BeadChip* – Illumina, Inc., EUA). Os demais, n= 240, foram coletados em 2015, e, por sua vez, foram genotipados com painel de 65k (*Equine SNP70 BeadChip* – Illumina, Inc., EUA). Todos os procedimentos envolvendo os animais foram realizados segundo as normas brasileiras de bem-estar animal (Protocolo nº 157/2014 – Comissão de Ética no Uso de Animais/CEUA, FMVZ, Unesp, Botucatu, SP, Brasil).

A leitura de ambos os arranjos foi realizada por meio do sistema *HiScan* (Illumina, Inc., EUA). Os controles de qualidade (*quality control* – QC) foram realizados por meio da exclusão de indivíduos (*call rate* $\leq 0,9$) e de SNPs (*call rate* $\leq 0,9$; valor de $p \leq 1 \times 10^{-5}$ para *Hardy-Weinberg Equilibrium*; cromossomo não autossômicos) de forma independente para os arranjos de 54k (n= 120) e 65k (n= 240), utilizando o pacote *snpStats* (CLAYTON, 2015) do programa R v.3.3.1 (R Core Team, 2016). Após o QC, permaneceram 116 e 233 amostras e 46.712 e 56.604 SNPs para os arranjos de 54k e 65k, respectivamente. A partir disso, com o intuito de homogeneização das informações dos genótipos, foi realizada a imputação em duas vias entre os dois arranjos (i.e. do 54k para o de 65k e do 65k para o de 54k) por meio do programa *FLimpute* v.2.2 (SARGOLZAEI et al., 2014). Após a

conferência da acurácia de imputação, o critério de QC *minor allele frequency* – MAF (MAF < 0,05) foi aplicado ao painel de genótipos imputados. A partir disso, os 349 animais que passaram pelo controle de qualidade inicial passaram a contar com 55.114 SNPs em suas matrizes de genótipos.

Os dados de desempenho dos cavalos Quarto de Milha de corrida (Índice de Velocidade máximo – IV max), atualizados até o mês de dezembro de 2015, foram obtidos junto ao Departamento de Estatística ou ao banco de dados online do Jockey Club de Sorocaba/SP e junto ao banco de dados norte americano online Equibase. Com objetivo de garantir a confiabilidade da medida IV max, foram considerados para as análises apenas indivíduos com três ou mais corridas em seu histórico de competição, totalizando assim, 305 animais genotipados com informação de desempenho. Adicionalmente, e de acordo com o permitido pelo método empregado (ssGBLUP), foram incorporadas às análises informações de pedigree de cada animal genotipado (pai, mãe e avós) por meio da montagem de uma de uma matriz de parentesco composta por 620 indivíduos. Também foram adicionados registros de IV max a 187 animais não genotipados que compunham a matriz de parentesco. Dessa forma, para a obtenção dos GEBVs foram utilizados os registros de IV max de 492 animais, seguindo o modelo $y = X\beta + Za + e$, onde: y é o vetor de observações; β é o vetor de efeitos fixos (sexo, idade e ano da corrida referentes à observação do IV max); a é o vetor de efeitos genéticos aditivos diretos; X é a matriz de incidência; Z é a matriz de incidência dos efeitos aleatórios genéticos aditivos diretos; e é o vetor do efeito residual.

2.2. Seleção de amostras com GEBVs contrastantes para desempenho em corridas

Com base em GEBVs contrastantes para IV max foram selecionadas 60 amostras, sendo 30 de IV max superior e 30 de IV max inferior. Esses extremos foram refinados levando em consideração os seus históricos de desempenho e, a partir disso, 20 indivíduos de cada grupo foram escolhidos. Dessa forma, 40 amostras foram consideradas para o sequenciamento do exoma e regiões UTR.

Com o intuito de averiguar a relação genética entre os indivíduos da linhagem de corrida da raça Quarto de Milha contrastantes para IV max foi conduzida análise de estratificação por meio da plotagem do gráfico MDS (escalonamento multidimensional). Para isso foi considerado, para os 40 indivíduos, o número de SNPs comuns aos dois arranjos que passaram pelo controle de qualidade de genotipagem. Esses marcadores foram submetidos à pruning para seleção de SNPs que não estavam em desequilíbrio de

ligação (*linkage disequilibrium* – LD), considerando $r^2 < 0,2$. As probabilidades de compartilhar marcadores idênticos por estado (*identity-by-state* – IBS) entre todas as amostras par a par foram calculadas utilizando os SNPs autossômicos que permaneceram após o *prune*. Esta análise foi realizada por meio do programa PLINK 1.07 (PURCELL et al., 2007).

2.3. Preparo das bibliotecas, captura e sequenciamento dos exomas equino

O preparo de bibliotecas, a captura e o enriquecimento dos exomas e regiões UTR para sequenciamento foram realizados em dois ciclos conforme o *User's Guide* do *SeqCap EZ HyperCap Workflow v1.0* (Roche, Ale). Em cada ciclo foram feitos *pools* com 20 amostras de DNA multiplexadas.

A preparação de bibliotecas teve início a partir de alíquotas de 100 ng de DNA genômico (gDNA), as quais sofreram fragmentação enzimática utilizando o *kit KAPA HyperPlus Library Preparation* (Kapa Biosystems, EUA). A ligação dos adaptadores específicos para cada amostra foi realizada com os *Kits SeqCap Adapter A e B* (Roche, Ale). A amplificação dos fragmentos da biblioteca ocorreu via *Ligation-Mediated – Polymerase Chain Reaction* (LM – PCR). A qualidade dos procedimentos realizados foi verificada por meio da análise de 1 μ L de biblioteca de cada amostra amplificada utilizando o *Agilent Bioanalyzer DNA 1000 chip* (Agilent, EUA), conforme as instruções do fabricante. Com auxílio dos kits *SeqCap EZ HE-Oligo A e B* (Roche, Ale), foram preparados *pools* de bibliotecas com amostras multiplexadas a partir da ligação dos indexes aos adaptadores A e B. A captura dos fragmentos de interesse da biblioteca genômica inicial foram realizadas utilizando os kits *SeqCap EZ Developer Library* e *SeqCap EZ Accessory Kit v2* (Roche, Ale). As sondas para captura foram sintetizadas pela empresa Roche (Roche, Ale) com base nas coordenadas de regiões codificantes (exons), 5' e 3' UTR considerando a montagem EquCab 2.0 do genoma equino. A amplificação do DNA multiplexado capturado foi realizada por meio de LM – PCR. Por fim, a quantificação dos fragmentos enriquecidos das regiões alvo foi realizada via RT-qPCR utilizando o *kit KAPA Library Quantification* (*Library Quantification Primer Premix and qPCR MasterMix* e *Library Quantification DNA Standards* – Kapa Biosystems, EUA) em aparelho termociclador em tempo real *Rotor gene 6000* (Corbett/Qiagen, EUA). A concentração de DNA das bibliotecas foi determinada por meio de regressão linear com base em seis amostras padrões presentes no *kit KAPA Library Quantification*.

As bibliotecas preparadas foram sequenciadas por meio do equipamento *HiSeq 2500* (Illumina Inc., EUA) em duas corridas *Rapid Run* de leituras 2 x 100 nucleotídeos (*paired end*). Na primeira corrida foi sequenciado um *pool* contendo bibliotecas de 20 amostras multiplexadas, as quais foram aplicadas em ambas as linhas da *flow cell*. Posteriormente, o mesmo procedimento foi repetido para o segundo *pool* de bibliotecas. As informações resultantes do sequenciamento foram geradas em forma de imagens. O programa *Consensus Assessment of Sequence and Variation* (CASAVA) versão 1.8.2 (Illumina Inc., EUA) foi utilizado para localizar os *clusters* de fragmentos na *flow cell*, os quais foram sintetizados por meio do *HiSeq Rapid PE cluster kit v2* (Illumina Inc., EUA), e analisar a intensidade e o ruído das imagens brutas. Após o *demultiplex*, processo que retira os indexes que serviram para diferenciar as amostras que foram sequenciadas nas mesmas linhas da *flow cell*, os arquivos com as sequências de nucleotídeos dos *clusters* que estavam no formato *.bcl foram transformados para o formato *.fastq.gz também por meio do CASAVA. A checagem inicial da qualidade das *reads* e da cobertura média do sequenciamento foram obtidas considerando o total de amostras por meio dos programas *FastQC* (ANDREWS, 2010) e *CLC Workbench Genomics v.9.5.2* (Qiagen, EUA), respectivamente.

2.4. Alinhamento das *reads* ao genoma equino de referência

Com base nos dados de qualidade obtidos por meio do *FastQC*, parâmetros para limpeza e corte (*trim*) das *reads* foram aplicados para algumas amostras individuais por meio do programa *CLC Workbench Genomics*, com intuito de ajustá-los para inclusão no seu *workflow*. Para conferir a eficácia da limpeza (*trim*), as *reads* obtidas com o *CLC Workbench Genomics* foram novamente analisadas utilizando o programa *FastQC*, em que foram comparadas *per base quality* e *per sequence content* antes e após a limpeza. Seguindo os parâmetros definidos com base nos testes descritos acima, a limpeza de todas as *reads* foi realizada.

O mapeamento das *reads* em relação ao genoma equino de referência (EquCab 2.0) foi conduzido para cada amostra (uma por vez), seguindo os padrões (*default*) do programa *CLC Workbench Genomics v.9.5.2* (Qiagen, EUA), em que foram gerados arquivos *.bam (*binary alignment/map*), semelhantes aos arquivos *.sam (*sequence alignment/map*), porém em formato comprimido. Estes arquivos foram ordenados e indexados por meio do programa SAMtools v.1.4 (LI et al., 2009) e por meio do programa *CLC Workbench Genomics v.9.5.2* (Qiagen, EUA) foi obtida a cobertura média para todas

as amostras e a percentagem de regiões cobertas para diferentes faixas de coberturas (1x, 10x, 20x e 40x) para cada amostra. Para isso foram consideradas, conjuntamente, todos os éxons, e regiões 5' e 3' UTR descritos no EquCab 2.0.

2.5. Detecção e filtragem de polimorfismos

A identificação de polimorfismos (SNPs e InDels), para cada amostra em relação à referência, foi realizada com o programa CLC Workbench Genomics seguindo os parâmetros que seguem: probabilidade de variante de 90%; frequência mínima de 30%; qualidade de mapeamento mínima de 20; e qualidade phred score $Q \geq 30$. Adicionalmente, utilizou-se um segundo filtro de qualidade (marginal variant filter – MVF) em que foram excluídas variantes com balanço forward-reverse menor que 0,05 para evitar o strand bias, o que afetaria principalmente polimorfismos do tipo InDel. Após a identificação das variantes, arquivos *.vcf (variant call format) contendo SNPs e InDels foram gerados para cada amostra e combinados em arquivo único (merge) por meio do programa vcftools (DANECEK et al., 2011). Os genótipos perdidos após o procedimento de merge (homozigotos referência) foram recuperados utilizando o programa FixVcfMissingGenotypes (LINDENBAUM, 2015).

Com o objetivo de analisar as variantes localizadas/mapeadas nas regiões de interesse, as 33 regiões associadas em estudo de GWAS descritas por Pereira et al. (2017, capítulo 3), cada uma contendo aproximadamente 1 Mb (Tabela 1), foram filtradas. Após a filtragem das regiões de interesse, as variantes foram submetidas à controle de qualidade por meio do programa vcftools (DANECEK et al., 2011). Considerou-se cobertura mínima de sequenciamento de 20x, frequência mínima de amostras genotipadas para uma variante de 0,5, e desvio de HWE $< 0,001$.

Tabela 1 – Localização cromossômica e posições iniciais e finais de 33 regiões genômicas associadas ao desempenho em corridas na raça Quarto de Milha por meio de dois métodos diferentes de single-step GWAS (métodos 1 e 2) para obtenção do efeito de SNPs. (Pereira, 2017, capítulo 3).

Método 1 - ssGWAS1				Método 2 - ssGWAS2			
Região	Chr ¹	Posição inicial	Posição final	Região	Chr ¹	Posição inicial	Posição final
1	1	20.721.039	21.689.628	1	3	63.351.099	64.348.556
2	1	61.826.124	62.818.065	2	5	62.937.896	63.920.127
3	2	69.370.047	70.369.774	3	9	66.879.753	67.826.611
4	3	65.380.573	66.369.357	4	10	58.735.380	59.710.571
5	8	30.526.194	31.504.074	5	15	53.973.327	54.889.258
6	8	62.566.792	63.559.486	6	15	72.126.768	73.121.731
7	10	57.990.860	58.961.242	7	15	75.795.552	76.782.973
8	10	58.997.563	59.976.633	8	20	59.429.342	60.422.986
9	11	03.345.863	04.341.072	9	22	10.600.456	11.589.137
10	11	35.523.097	36.481.549	10	22	14.927.558	15.876.595
11	15	45.572.062	46.563.225	11	25	18.073.334	19.043.039
12	15	46.601.694	47.575.761	12	30	138.095	01.097.585
13	15	76.296.526	77.289.510				
14	15	82.855.154	83.838.250				
15	15	87.766.362	88.753.736				
16	16	16.542.440	17.474.810				
17	16	17.542.693	18.503.480				
18	16	77.697.272	78.681.644				
19	22	10.868.934	11.861.413				
20	25	17.757.118	18.713.444				
21	25	18.768.518	19.722.839				

¹Chr = cromossomo

2.6. Prospecção e anotação dos polimorfismos e genes de interesse

Para a prospecção de SNPs e InDels de importância aplicou-se ao conjunto total de variantes localizadas nas regiões de interesse três abordagens distintas e independentes. Primeiramente, por meio da ferramenta *Variant Effect Predictor* (VEP) do *Ensembl* (<http://www.ensembl.org/>), as localizações dos polimorfismos no genoma e seus efeitos funcionais – SNP sinônimo ou SNP não sinônimo (*missense*), SNP *stopgain* ou *nonsense* (causa *stop* códon) e SNP *stoploss* (elimina o *stop* códon) e InDel *frameshift* ou não *frameshift* – foram analisadas. Também por meio do VEP foram prospectadas variantes de importância com base na predição de scores SIFT (SNPs não sinônimos) e no impacto na estrutura de proteínas (impacto alto). O score SIFT prevê se uma substituição de aminoácidos pode afetar a função da proteína com base na

homologia de sequências e na semelhança físico-química entre os aminoácidos alternativos. Cada substituição de aminoácidos recebe uma pontuação e uma predição qualitativa. A predição qualitativa é derivada da pontuação de *score*, de modo que as substituições com pontuação $<0,05$ são denominadas "prejudiciais" (*deleterious*) e todas as outras são denominadas "toleradas" (*tolerated*). Em seguida, utilizou-se o teste exato de Fisher (FISHER, 1970) com significância de 1% ($p < 0,01$), por meio do programa PLINK 1.07 (PURCELL et al., 2007), para identificar frequências alélicas diferentes entre as amostras contrastantes para IV max.

Por fim, as funções biológicas dos genes em que se encontram localizados os SNPs de interesse foram investigadas por meio da ferramenta PANTHER v.10 (MI et al., 2015).

3. Resultados e discussão

3.1. Amostras sequenciadas

Animais do grupo com desempenho superior apresentaram 5,4 vezes mais vitórias em números absolutos do que os do grupo de pior desempenho (Tabela 2). Considerando a proporção de primeiros lugares em relação ao total de corridas, nota-se mais vez grande diferença entre os grupos contrastantes, em que animais de grupo de melhor desempenho venceram em média uma corrida a cada três que disputaram. No grupo com desempenho inferior essa proporção foi de uma vitória a cada nove provas disputadas. Ao se considerar o número total de corridas, os indivíduos do grupo IV max superior disputaram em média duas vezes mais provas do que os de IV max inferior. Na prática, cavalos que conseguem boas colocações em suas primeiras provas, tendem a ser mantidos em atividade e em disputas por maior tempo, o que não ocorre com indivíduos de desempenho insatisfatório. Com base nessas informações, pode-se afirmar que as estimativas de GEBVs utilizadas nesse estudo mostraram-se adequadas para a separação de grupos contrastantes, sendo mais eficientes para esse propósito, do que EBVs ou fenótipos corrigidos para efeitos fixos, simulados previamente. É importante ressaltar, que em corridas de cavalos da raça Quarto de Milha, as distâncias curtas e as pequenas diferenças de tempo entre animais vitoriosos e de pior desempenho são pequenas e distribuem-se de forma contínua, fato que não permite a separação de dois grupos de animais incontestavelmente contrastantes.

Tabela 2 – Animais extremos para desempenho em corridas (considerando a amostra total de cavalos da linhagem de corrida da raça Quarto de Milha), baseados em valores genômicos (GEBVs) com base em efeitos estimados de marcadores obtidos no ssGWAS2.

Animais com maiores GEBVs					Animais com menores GEBVs				
Id	N	1º	2º	3º	Id	N	1º	2º	3º
C074	36	9	10	7	C156	23	0	4	5
C096	21	5	4	3	C032	19	2	5	5
C208	20	8	2	4	C093	18	3	4	2
C073	19	5	3	3	C341	10	1	2	3
C081	18	9	6	1	C019	9	2	3	1
C361	18	2	2	4	C349	9	1	1	0
C113	16	9	2	3	C280	8	1	3	0
C318	15	5	6	0	C121	8	0	3	2
C199	12	3	2	4	C386	7	2	1	1
C205	12	3	1	3	C025	6	1	1	2
C144	11	6	1	0	C014	5	1	1	1
C157	10	7	2	0	C071	5	1	3	0
C228	10	4	3	1	C024	5	1	1	3
C398	9	3	2	2	C010	5	1	2	0
C109	9	3	0	2	C197	4	1	1	0
C194	9	3	2	3	C132	4	0	0	0
C007	8	4	2	1	C268	4	0	0	2
C305	7	3	1	0	C138	4	0	0	2
C251	6	4	0	0	C294	3	0	1	1
C027	6	2	1	3	C333	3	0	0	1
Totais	272	97	52	44	Totais	159	18	36	31
Proporções em relação ao total de corridas		0,36	0,19	0,16	Proporções em relação ao total de corridas		0,11	0,23	0,19
Proporções em relação ao total de 1º, 2º e 3º lugares		0,50	0,27	0,23	Proporções em relação ao total de 1º, 2º e 3º lugares		0,21	0,42	0,36

Na análise de estratificação do conjunto amostral contrastante não foram identificados *clusters* de forma significativa (Figura 1). Portanto, não foram encontradas subpopulações originadas, principalmente, por diferenças de ancestralidade, o que pode acarretar em diferenças sistemáticas nas frequências alélicas entre os grupos contrastantes, levando a associações genéticas errôneas.

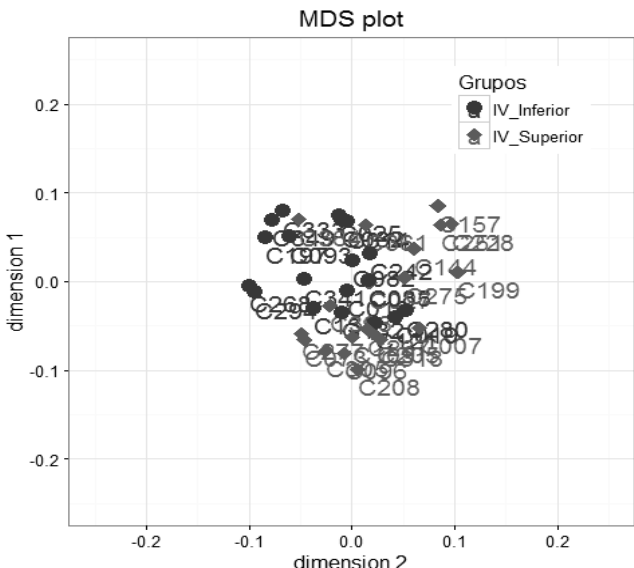


Figura 1 – MDS plot do conjunto total de amostras contrastantes (n= 40) considerando os 39.664 SNPs comuns a ambos os arranjos equinos. Os círculos e os losangos correspondem a animais genotipados com os arranjos de 54k e 65k, respectivamente.

3.2. Sequenciamento e alinhamento

Os dois *pools* aplicados nas corridas de sequenciamento foram formados aleatoriamente. Dessa forma, cada *pool* continha tanto bibliotecas de amostras de indivíduos IV max superior (aproximadamente 50%) quanto de IV max inferior (aproximadamente 50%). Ao longo das preparações das bibliotecas, uma das amostras não atingiu os padrões necessários de amplificação e foi descartada. Assim, foram sequenciados ao todo 39 indivíduos, sendo 20 de IV max superior e 19 de IV max inferior.

Para a primeira corrida de sequenciamento (n= 20) foram obtidas 2.498.929.809 sequências pareadas (*paired-end*) e para a segunda (n= 19) 1.629.737.973. A qualidade inicial das *reads* (antes do controle de qualidade) foi avaliada por meio do programa FastQC considerando de forma conjunta todas as amostras de cada corrida do sequenciamento. Foram checados itens como: qualidade de leitura das *reads*, conteúdo de GC, proporção de contribuição de cada nucleotídeo, entre outros. Foram observadas boas qualidades de leitura para toda extensão das *reads* sequenciadas (até 100pb). Contudo, a proporção da contribuição de cada nucleotídeo e a qualidade *score phred* foram ruins nos primeiros pb na extremidade 5' e nos 3 pb da extremidade 3', porém acima do aceitável

A cobertura do sequenciamento foi obtida com base em todas as regiões codificantes, 5' e 3' UTR para cada amostra. A cobertura média considerando ambas as

corridas de sequenciamento foi de 43x ($\pm 13,8x$), com média de 50x ($\pm 20,1x$) e 35x ($\pm 3,7x$) para a primeira e a segunda corrida, respectivamente. Foram observadas as médias de 74% ($\pm 16,6\%$) e 68% ($\pm 4,07\%$) de regiões alvo cobertas 20x (limiar mínimo de corte para detecção de variantes) na primeira e segunda corrida de sequenciamento, respectivamente (Tabela 3). Embora os dados estatísticos descritivos de cobertura tenham sido obtidos para as regiões codificantes e UTRs, o alinhamento de regiões adjacentes às regiões alvo (intrônicas e intergênicas) ocorreu em menor escala, devido à captura pelas sondas não se limitar estritamente ao alvo. Isso pode trazer ganho considerável de informação, principalmente, pelo fato de aumentar a extensão das regiões sequenciadas.

Tabela 3 – Porcentagens de cobertura em diferentes faixas (1x, 10x, 20x e 40x) de duas corridas de sequenciamento (corridas 1 e 2) considerando o exoma de cada uma das 39 amostras de equinos de corridas da raça Quarto de Milha sequenciadas.

Cobertura - Corrida 1					Cobertura - Corrida 2				
	1x	10x	20x	40x		1x	10x	20x	40x
C07	93,52%	86,43%	73,71%	40,35%	C85	94,04%	85,82%	73,20%	42,97%
C10	94,47%	91,76%	88,46%	79,16%	C93	93,56%	82,14%	64,06%	29,38%
C14	94,17%	90,44%	85,40%	71,44%	C121	94,13%	86,73%	74,56%	45,80%
C19	94,37%	91,34%	87,72%	77,71%	C132	93,94%	85,24%	71,72%	42,27%
C25	94,37%	91,47%	87,80%	77,39%	C138	94,10%	85,65%	72,66%	44,70%
C27	94,51%	91,78%	88,54%	79,66%	C197	93,88%	84,56%	69,16%	36,82%
C32	93,70%	88,26%	80,14%	57,51%	C242	94,04%	85,74%	72,28%	41,61%
C73	94,05%	89,99%	84,37%	68,43%	C268	94,02%	85,32%	71,21%	40,16%
C82	88,98%	48,54%	14,06%	0,42%	C277	93,91%	84,24%	69,74%	40,73%
C96	93,97%	89,77%	83,62%	65,68%	C280	93,97%	85,45%	71,69%	40,75%
C109	94,16%	90,41%	84,95%	67,86%	C294	93,96%	86,00%	71,65%	37,63%
C144	93,49%	86,62%	75,71%	47,76%	C305	93,81%	83,64%	65,86%	29,82%
C157	93,42%	86,22%	74,09%	43,07%	C333	93,64%	83,68%	68,26%	35,86%
C199	93,17%	83,76%	65,89%	28,00%	C341	93,48%	81,92%	63,58%	27,62%
C205	93,63%	90,97%	86,96%	42,19%	C349	93,72%	83,24%	65,49%	29,56%
C208	93,22%	83,96%	68,14%	32,77%	C361	93,43%	81,33%	61,29%	24,37%
C228	93,32%	85,02%	69,77%	32,87%	C364	93,64%	83,10%	65,11%	28,73%
C237	93,05%	82,57%	61,83%	21,70%	C381	93,41%	80,50%	60,77%	25,34%
C251	93,06%	82,07%	62,49%	24,08%	C398	93,69%	83,70%	67,05%	32,14%
C275	93,35%	85,18%	71,43%	37,84%					

3.3. Detecção e filtragem de SNPs e InDels

Por meio programa *CLC Workbench Genomics* foram identificadas 150.106 variantes para todo o exoma, regiões 5' e 3' UTR e regiões adjacente capturadas pelas sondas. Isto, considerando a aplicação do QC final por meio de filtros de qualidade do programa *vcftools*, mas não a filtragem para regiões de interesse (PEREIRA et al., 2017, capítulo 3).

Após a seleção para as 33 regiões de interesse (~33 Mb), foram identificadas 1.206 variantes (1.105 SNPs e 93 InDels). Desse modo, foram localizadas em média um SNP a cada 25 Kb e um InDel a cada 300 Kb, ou seja, média de 40 SNPs e 2,2 InDels para cada região analisada (~1Mb de extensão). O número de SNPs e InDels tomando somente o exoma, 5' e 3' UTR dentro das 33 regiões de interesse foram de 408 SNPs e 43 InDels, respectivamente. Portanto, a densidade média foi de aproximadamente um SNP a cada 1,6 Kb e um InDel a cada 15,1 Kb se considerado o exoma como 2% (~650kb) do tamanho total de todas as regiões de interesse.

Quanto à qualidade do mapeamento dos polimorfismos, de modo geral as proporções de SNPs e InDels de baixa qualidade (0-30) foram pequenas, ficando a maior proporção entre faixas medianas de qualidade (30-90), com um decaimento natural após esses valores (Figura 2). Contudo, como observado por Godoy et al. (2015), a percentagem de InDels mapeadas com baixa qualidade é maior que a de SNPs devido à dificuldade inerente à maior complexidade em mapear *reads* em regiões de InDels (LI et al., 2009).

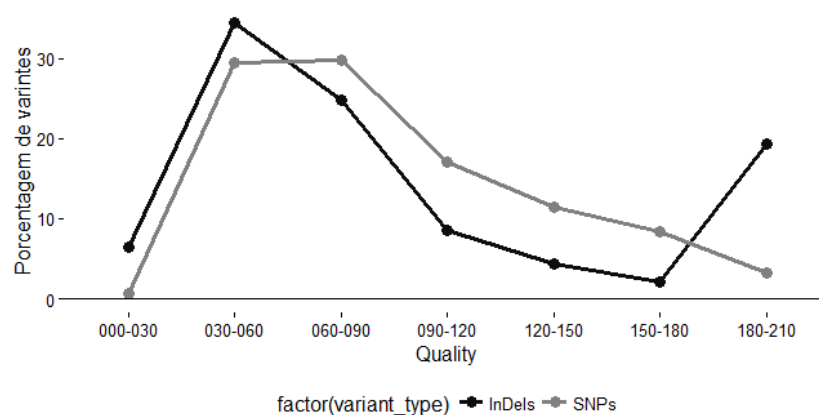


Figura 2 – Percentagens de SNPs e InDels (eixo y) em relação a diferentes intervalos de qualidade (0 a 210) (eixo x) nas regiões de interesse sequenciadas em cavalos Quarto de Milha de corrida contrastantes para desempenho.

Na Figura 3 são observados dados de cobertura nas posições das variantes para cada amostra, que, assim como para o exoma e regiões UTR, foram obtidos considerando somente polimorfismos localizados dentro das 33 regiões de importância. A cobertura média para os polimorfismos foi de 37,8x considerando as duas corridas de sequenciamento. De maneira geral, as amostras da primeira corrida de sequenciamento apresentaram cobertura das variantes menos uniforme se comparadas às amostras da segunda corrida. Dentre as amostras sequenciadas, a C82 teve o menor número de variantes identificadas, devido, provavelmente, à pior qualidade de seu sequenciamento, causado por supostos problemas na preparação de sua biblioteca.

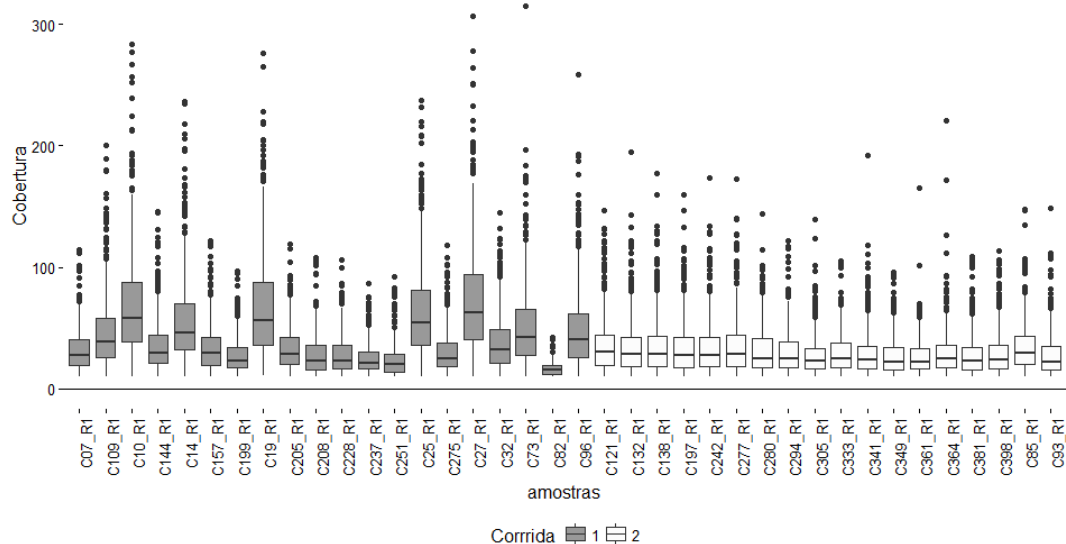


Figura 3 – Box-plot da cobertura das variantes prospectadas em regiões alvo (eixo y) para cada amostra de equinos de corrida da raça Quarto de Milha sequenciada (eixo x), em que caixas da cor cinza e brancas representam as corridas 1 e 2 de sequenciamento, respectivamente. As caixas representam regiões interquartis e a linha central a mediana.

O número de SNPs identificados nas regiões de interesse para cada amostra sequenciada foi sensivelmente maior se comparado ao de InDels (Figura 4). Isso pode ser atribuído, primeiramente, ao fato de o número de InDels ser cerca de oito vezes menor que o de SNPs em genomas eucariotos (LUNTER e GOODSON, 2007; CARTWRIGHT, 2009) e, também, à capacidade de mapeamento desses polimorfismos que pode variar de forma considerável em função da ferramenta utilizada para as suas identificações. Isso se deve à robustez e sensibilidade de

diferentes algoritmos e à variações na cobertura de sequenciamento (KRØIGÅRD et al., 2016).

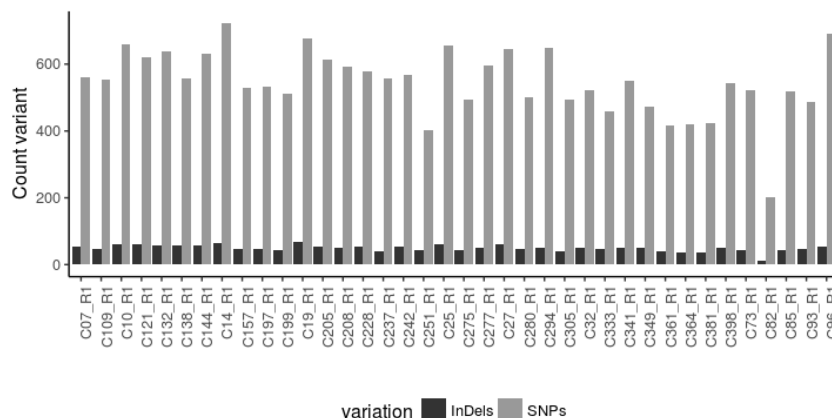


Figura 4 – Número de SNPs e InDels (eixo y) identificados nas regiões sequenciadas em cada amostra de equino da raça Quarto de Milha sequenciada (eixo x).

3.4. Anotação funcional de variantes genômicas

As posições das 1.203 variantes (1.105 SNPs e 93 InDels) prospectadas nas regiões de interesse foram submetidas ao programa VEP. Destas, 61,3% não estavam registradas/depositadas em banco de dados de polimorfismos de DNA equino.

Considerando todas as variantes, 57% foram mapeadas em regiões de íntron (49%) ou inter-gênicas (8%) (Tabela 4). Em um primeiro momento este fato pode ser considerado inesperado, visto que foi realizado sequenciamento de exons e regiões UTR. Contudo, como mencionado anteriormente, as sondas utilizadas capturam fragmentos de tamanhos que, muitas vezes, podem ser maiores que as regiões alvo. Considerando somente as variantes localizadas em regiões codificantes, houve maior proporção de polimorfismos sinônimos (60%). Entretanto, proporção considerável de polimorfismos *missense* foi observada (40%).

Tabela 4. Localização genômica das variantes identificadas nas regiões de interesse sequenciadas em animais Quarto de Milha de IV max superior e de IV max inferior.

Localização no genoma	Número de variantes	Porcentagem de variantes
3 prime UTR variant	30	2.5%
5 prime UTR variant	9	0.7%
Coding sequence variant	2	0.2%
Downstream gene variant	26	2.2%
Frameshift variant	2	0.2%
Frameshift variant , splice region variant	1	0.1%
Frameshift variant , splice region variant , intron variant	2	0.2%
Intergenic variant	28	2.3%
Intron variant	597	49.7%
Missense variant	161	13.4%
Non coding transcript exon variant , non coding transcript variant	8	0.7%
Splice region variant , 3 prime UTR variant	1	0.1%
Splice region variant , 5 prime UTR variant	1	0.1%
Splice region variant , intron variant	33	2.7%
Splice region variant , synonymous variant	9	0.7%
Start lost	1	0.1%
Synonymous variant	250	20.8%
Upstream gene variant	40	3.3%

Para checar o potencial impacto da mudança de aminoácidos em decorrência de polimorfismos não sinônimos (*missense*) dentro das 33 regiões de interesse, o escore SIFT foi obtido por meio da ferramenta VEP. Foram identificados SNPs com escores SIFT menores que 0,05 em seis cromossomos diferentes (ECA 2, 10, 11, 15, 16 e 25) (Tabela 5).

No ECA 15, foram identificados quatro SNPs em três diferentes regiões de importância, nas posições de 54 Mb (ssGWAS2), e 77 Mb e 88 Mb (ssGWAS1). O SNP rs396926159 (ECA15 – 88 Mb) foi mapeado no primeiro exon de gene ainda não caracterizado (ENSECAG00000013638), em posição *start lost*. Assim, foi classificado como de alto impacto, ou seja, a alteração no códon (aTg/aCg) leva à perda do sitio de iniciação da síntese proteica. Embora, assumamos que o início da tradução começa com o no *start* códon, alguns estudos tem mostrado que a síntese proteica poderia começar alguns sítios antes do códon ATG (PARSONS et al., 2013). Contudo, a não caracterização do gene em equinos não permite discussão mais aprofundada. Os demais SNPs do

ECA 15 ainda não foram depositados em banco de dados e estão localizados nos genes *ABCG5* (15:54808938C>G e 15:54813992C>T) e *GEN1* (15:77037894A>T), os quais foram indicados, dentre os genes prospectados por Pereira et al. (2017, capítulo 3), como candidatos funcionais para desempenho em corridas na raça Quarto de Milha. Vale destacar ainda que no trabalho de Pereira e colaboradores, o gene *ABCG5* teve o maior número de termos de ontologia genética enriquecidos significativamente.

No ECA 16 foram identificados três SNPs, dois dos quais (rs782886289 e 16:17363187G>A) localizados no gene *SHQ1*. Este gene foi descrito por SHIN et al. (2015) como candidato para desempenho em corrida em equinos após associação significativa por meio GWAS em cavalos PSI. No ECA 10 foram apontados quatro SNPs, dos quais dois estão localizados no gene *ARMC2*. Esses polimorfismos estão localizados em região concordante entre os métodos ssGWAS1 e ssGWAS2 utilizados por Pereira et al. (2017, capítulo 3). No ECA 15, os polimorfismos 25:17759560T>A e 25:18115651T>A foram mapeados nos genes *ZFP37* e *FKBP15*, respectivamente. O gene *FKBP15*, está localizado em uma região concordante do cromossomo 15 entre os métodos ssGWAS1 e ssGWAS2 (PEREIRA et al., 2017, capítulo 3).

Tabela 5. Descrição dos SNPs de score SIFT menor que 0,05 em relação ao genoma equino de referência encontrados nas regiões de interesse sequenciadas em animais Quarto de Milha de IV max superior e de IV max inferior.

Localização	Alelo ^a	Simbolo HGCN	Gene	Éxon	Variante	Codons	SIFT
2:69626616	A	<i>TMEM192</i>	ENSECAG00000014003	2/6	rs782866029	Cat/Aat	0.02
10:58739720	A	<i>ARMC2</i>	ENSECAG00000024506	18/10	rs395641393	Gtt/Att	0.03
10:58739790	C	<i>ARMC2</i>	ENSECAG00000024506	18/10	rs397493223	gGa/gCa	0.03
10:58985153	A	-	ENSECAG00000026868	4/6	-	tGc/tAc	0.00
10:59027763	G	-	ENSECAG00000011811	3/9	rs394184729	tTt/tGt	0.00
11:3713962	A	-	ENSECAG00000023129	1/3	-	cGa/cAa	0.04
11:3714315	A	-	ENSECAG00000023129	3/3	rs396071787	Gat/Aat	0.0 (LC) ¹
15:54808938	G	<i>ABCG5</i>	ENSECAG00000010864	4/13	-	Cgg/Ggg	0.04
15:54813992	T	<i>ABCG5</i>	ENSECAG00000010864	8/13	-	Cgg/Tgg	0.00
15:77037894	A	<i>GEN1</i>	ENSECAG00000020295	13/13	-	gAa/gTa	0.01
15:88474045*	G	-	ENSECAG00000013638	1/6	rs396926159	aTg/aCg	0.03
16:17363187	A	-	ENSECAG00000000673	4/4	-	Gtg/Atg	0.01
16:17592454	C	<i>SHQ1</i>	ENSECAG00000015673	5/13	-	Gag/Cag	0.01
16:17592482	G	<i>SHQ1</i>	ENSECAG00000015673	5/13	rs782886289	cCc/cGc	0.01
25:17759560	T	<i>ZFP37</i>	ENSECAG00000011666	4/5	-	Ggg/Agg	0.00

25:18115651 T *FKBP15* ENSECAG00000017713 26/29 - Ccg/Acg 0.02 (LC)
^{*} *start_lost* com alto impacto; ^a alelo alternativo; ¹ *low confidence*

Com base na localização consenso de cada variante, foram prospectados polimorfismos de alto impacto no genoma (Tabela 6), ou seja, aqueles que afetam sensivelmente a proteína traduzida, causando truncamento, perda ou decaimento de função. Das seis variantes encontradas, cinco tem potencial para causar *frameshift* e destas, três estão localizadas em regiões de *splice*. Variantes em regiões de *splice* (*splice variant*) estão localizadas na junção exon-íntron, ocorrendo dentro do exon (1-3pb) e/ou dentro do intron (1-8pb) (BUTKIEWICZ e BUSH, 2016). No ECA 15 foram encontradas duas variantes de alto impacto, uma SNP *start_lost* (rs396926159), já descrita anteriormente por apresentar escore SIFT 0,03, e uma inserção de 4pb (15:82864445) em uma região intrônica de *splice* no gene *NOL10*. No ECA 11, dois polimorfismos foram descritos nas posições 11:4215438 e 11:36178208 nos genes *SOCS3* e *LHX1*, respectivamente. O polimorfismo 11:4215438 corresponde a InDel de uma base no exon 6 do gene *SOCS3*, o qual foi descrito anteriormente como candidato para desempenho em Quartos de Milha de corrida por Pereira et al. (2017, capítulo 3). Na posição 63.660.646 do gene *COL11A1* no ECA 5 também foi encontrado um InDel. Este gene teve muitos termos enriquecidos de forma significativa por Pereira et al. (2017, capítulo 3), os quais o relacionaram como candidato para desempenho em corridas na raça Quarto de Milha.

Tabela 6. Descrição de variantes de alto impacto encontradas em regiões de interesse sequenciadas em cavalos Quarto de Milha dos grupos IV max superior e IV max inferior.

Localização	Alelo ^a	Consequence	Simbolo HGCN	Exon	Codons	Variante	SIFT
11:4215438	-	Frameshift	<i>SOCS3</i>	6/7	aCc/ac	-	-
11:36178208	C	Frameshift, splice_region_variant intron_variant	<i>LHX1</i>	-	-/C	-	-
15:82864445	GTAA	Frameshift_variant, splice_region_variant	<i>NOL10</i>	-	gtt/gGTAAtt	-	-
15:88474045	G	Start_lost	-	1/6	aTg/aCg	rs396926159	0.03
25:19192379	G	Frameshift_variant	<i>AKNA</i>	5/23	-/C	-	-
5:63660646	G	Frameshift_variant, splice_region_variant intron_variant	<i>COL11A1</i>	-	-/G	-	-

A partir do teste de Fisher, foram identificados sete SNPs com frequências alélicas significativamente diferentes ($p < 0,01$) em cinco cromossomos (ECA 2, 8, 10, 15, 30) (Tabela 7). Um SNP *missense* (rs782906387) foi identificado no exon quatro do gene *HNRNPU* (ECA 30). Este gene está localizado em região cromossômica que explicou, de forma única, mais de 7% da variância genética total para característica IV max (PEREIRA et al., 2017, capítulo 3). Região à aproximadamente 2 Mb de distância desse gene foi descrita por Shin et al. (2015) como contendo pelo menos um gene significativo para desenvolvimento muscular em cavalos PSI. No ECA 10 foram descritos três SNPs sinônimos, todos inseridos em região concordante (58.997.563–59.710.571) entre as duas abordagens (ssGWAS1 e ssGWAS2) utilizadas por Pereira et al. (2017, capítulo 3) para identificar QTLs associados ao desempenho em Quartos de Milha de corrida. Dentre os três polimorfismos, dois estão localizados em genes conhecidos/anotados (*FIG4* e *MICAL1*), sendo o *MICAL1* descrito por esses autores como candidato posicional e funcional para a característica. Dois SNPs intrônicos, o rs69003211 no ECA 15 e o 8:30591631T>C (não depositado em bancos de dados) no ECA 8, estão localizados nos genes *SPTBN1* e *EPB41L3*, respectivamente. Ambos os genes também foram descritos como candidatos funcionais para desempenho em corrida no Quarto de Milha por Pereira et al. (2017, capítulo 3).

Tabela 7. Descrição de polimorfismos em regiões de interesse sequenciadas com diferença de frequências alélicas, identificados por meio de teste exato de Fisher ($p < 0,01$), entre indivíduos Quarto de Milha dos grupos de IV max superior e de IV max inferior.

Localização	Alelo ¹	Conseq	HGNC*	Exon	Intron	Codons	Identificação de variante	Freq ² IV+	Freq IV-	SIFT
10:59081866	T	Sinônimo	-	5/5	-	gtG/gtT	rs68832518	0,33	0,12	-
10:59190226	G	Sinônimo	<i>MICAL1</i>	11/24	-	acT/acC	rs68832574	0,38	0,12	-
10:59412795	A	Sinônimo	<i>FIG4</i>	16/24	-	gcG/gcA	rs68861307	0,40	0,13	-
15:45580693	C	Intrônico	<i>SPTBN1</i>	-	22/36	-	rs69003211	0,50	0,23	-
2:69784058	T	<i>Upstream</i>	-	-	-	-	rs397340625	0,25	0,06	-
30:187747	A	<i>Missense</i>	<i>HNRNPU</i>	4/13	-	Gca/Aca	rs782906387	0,31	0,10	0.66
8:30591631	T	Intrônico	<i>EPB41L3</i>	-	1/21	-	-	0,45	0,14	-

*=HUGO Gene Nomenclature Comittee; ¹Alelo alternativo; ²frequência do menor alelo (MAF) para as amostras IV superior (caso) e IV inferior (controle).

3.5. Genes prospectados

Considerando os três métodos de prospecção de variantes de importância utilizados no presente trabalho (escore SIFT, impacto e teste de Fisher), 16 genes caracterizados/anotados na espécie equina podem ser destacados. Dentre estes, nove foram encontrados em regiões de interesse para desempenho em corridas no Quarto de Milha exclusivamente por meio do método ssGWAS1 (*AKNA*, *GEN1*, *LHX1*, *NOL10*, *SHQ1*, *SOCS3*, *TMEM192*, *SPTBN1* e *EPB41L3*) e três (*ABCG5*, *COL11A1* e *HNRNPU*) exclusivamente em regiões associadas por meio de ssGWAS2. Com isso, quatro genes (*ARMC2*, *FKBP15*, *MICAL1* e *FIG4*) foram apontados por ambos os métodos (PEREIRA et al., 2017, capítulo 3). Do total dos genes identificados neste trabalho, oito foram indicados como candidatos posicionais e funcionais para desempenho em corridas no Quarto de Milha por Pereira et al. (2017, capítulo), dos quais um foi destacado por Shin et al. (2015) no PSI.

De maneira geral estes genes estão relacionados em humanos ao metabolismo de esteróis (*ABCG5*), ao metabolismo energético de carboidratos e lipídeos (*SOCS3*), ao processo de recombinação cromossômica homologa, atuando na Junção de Holliday (*GEN1*), ao desenvolvimento de tendões e ao crescimento e desenvolvimento de condrócitos em placas de cartilagem (*COL11A1*), e à despolimerização do filamento de actina (*MICAL1*), processo importante na contração muscular. O gene *SHQ1*, também associado ao desempenho em corridas no PSI (SHIN et al., 2015), codifica elemento necessário para a montagem de ribonucleoproteínas da telomerase.

Para os genes não descritos como candidatos funcionais (por meio de análise de enriquecimento) para desempenho em cavalos de corrida Quarto de Milha por Pereira et al. (2017, capítulo 3), termos ontológicos relacionados à processos biológicos foram obtidos com base em anotações do genoma equino a partir da ferramenta PANTHER. Alguns genes (*NOL10*, *TMEM192* e *ARMC2*) não apresentaram termos para processos biológicos e possuem função pouco conhecida, mesmo considerando organismos modelo. O gene *FKBP15* apresentou funções conhecidas em plantas ou organismos menos complexos. O gene *HNRNPU*, descrito no ECA 30, região que explicou individualmente mais de 7% da variância genética total da característica IV max (PEREIRA et al., 2017, capítulo 3), possui termos associados principalmente ao desenvolvimento das células do músculo cardíaco (GO:0055013) e também à regulação circadiana da expressão gênica (GO:0032922). O gene *AKNA* apresentou apenas um termo de ontologia genética – regulação positiva da transcrição a partir do promotor da RNA polimerase II. Este gene

codifica fatores de transcrição *AT-Hook*, o qual é um *DNA-binding motif* presente em muitas proteínas. Estudos destacam o papel fundamental dos fatores de transcrição *AT-hook* como reguladores principais de processos celulares fundamentais envolvidos no desenvolvimento, função imunológica, câncer, diabetes e outras doenças em humanos (MOLITERNO e RESAR, 2011).

O gene *ZFP37* codifica uma proteína *zinc-finger* e é expresso em todos os tecidos do sistema nervoso central, atuando tanto no desenvolvimento de células neuronais, quanto na manutenção da atividade destas em indivíduos adultos. O gene *LHX1* (*LIM homeobox 1*) codifica fator de transcrição importante para o desenvolvimento dos sistemas renal e urogenital. Este gene, também foi descrito como candidato para a síndrome de Mayer-Rokitansky-Kuster-Hauser, desordem caracterizada por anomalias do trato genital feminino (LEDIG et al., 2012). Embora, grande parte das pesquisas envolvendo esse gene em humanos tenha sido relacionada ao aparelho reprodutivo, alguns termos associados a ele estão ligados ao desenvolvimento do cerebelo e diferenciação de células Purkinje cerebelar (GO:0021549 e GO:0021702). A função dessas células está estritamente relacionada à coordenação de movimentos e o cerebelo ao ajuste fino e precisão de movimentos. Por fim, o gene *FIG4* foi relacionado por termos de ontologia à comportamento de locomoção e desenvolvimento neuronal.

Parte dos genes aqui identificados coincidem com alguns dos anteriormente descritos em estudos de GWAS na raça Quarto de Milha (PEREIRA et al., 2017, capítulo 3) e PSI (SHIN, et al., 2015) como candidatos funcionais para desempenho em corridas. Embora, poucos genes candidatos funcionais novos, já caracterizados e anotados, tenham sido identificados, suas funções se assemelham com a de alguns genes já associados anteriormente, como por exemplo: desenvolvimento neuro-motor; comportamento locomotor e regulação hormonal. O fato de boa parte dos genes candidatos novos ainda não terem sido caracterizados, ou mesmo, os já anotados não possuírem funções conhecidas que possam ser relacionados ao desempenho atlético em corridas, abre campo para pesquisas pontuais de genes e proteínas que possam estar relacionados à característica.

4. Conclusões

Neste trabalho foram obtidas sequências de exons, UTR e adjacências de regiões consideradas de interesse por GWAS anteriormente realizado em equinos de corrida da raça Quarto de Milha. Com base em três abordagens distintas e independentes identificou-

se nas sequências de 39 animais de fenótipos contrastantes para desempenho em provas (20 de IV max superior e 20 de IV max inferior) 29 variantes (24 SNPs e 5 InDels) e 16 genes de importância. Com isso, oito genes descritos como candidatos funcionais e posicionais em trabalhos anteriores (*ABCG5*, *COL11A1*, *GEN1*, *SOCS3*, *MICAL1*, *SPTBN1*, *EPB41L3* e *SHQ1*), e oito genes candidatos novos (*AKNA*, *ARMC2*, *FKBP15*, *LHX1*, *NOL10*, *TMEM192*, *ZFP37*, *FIG4* e *HNRNPU*), alguns com funções conhecidas e outros não, foram relacionados ao desempenho em corridas de cavalos da raça Quarto de Milha.

5. Referências

AMERICA'S HORSE DAILY. All About the Racing American Quarter Horse. 26 de fevereiro de 2008. Disponível em: <http://americashorsedaily.com/all-about-the-racing-american-quarter-horse/#.Urx8N_RDvik> , acesso em: 12 nov. 2013.

ANDREWS, S. (2010) FastQC: a quality control tool for high throughput sequence data. Disponível em: <www.bioinformatics.bbsrc.ac.uk/projects/?fastqc/>, acesso em: 06 out. 2016.

BUTKIEWICZ, M.; BUSH, W. S. In Silico Functional Annotation of Genomic Variation. **Current Protocols in Human Genetics**, v. 88, n. 6, 1-17, 2016. Disponível em: <<http://doi.org/10.1002/0471142905.hg0615s88>>

CARTWRIGHT, R. A. Problems and solutions for estimating indel rates and length distributions. **Molecular Biology and Evolution**, Oxford, v. 26, p. 473–480, 2009.

CLAYTON, D. snpStats: SnpMatrix and XSnpmatrix classes and methods. R package version 1.24.0. 2015. Disponível em: <<http://bioconductor.org/packages/release/bioc/html/snpStats.html>>.

DANECEK, P.; AUTON, A.; ABECASIS, G.; ALBERS, C. A.; BANKS, E.; DEPRISTO, M. A.; HANDSAKER, R. E.; LUNTER, G.; MARTH, G. T.; SHERRY, S. T.; MCVEAN, G.; DURBIN, R. 1000 Genomes Project Analysis Group. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156-2158, 2011.

FISHER, R.A. Statistical Methods for Research Workers, Edinburgh: Oliver and Boyd, 14.ed., 1970.

GODOY, T. F.; MOREIRA, G. C. M.; BOSCHIERO, C.; GHEYAS, A. A.; GASPARIN, G.; PADUAN, M.; ANDRADE, S. C. S.; MONTENEGRO, H.; BURT, D. W.; LEDUR, M. C.; COUTINHO, L. L. SNP and INDEL detection in a QTL region on chicken chromosome 2 associated with muscle deposition. **Animal Genetics**, v. 46, p. 158-163, 2015.

KRØIGÅRD, A. B.; THOMASSEN, M.; LÆNKHOLM, A.-V.; KRUSE, T. A.; LARSEN, M. J.. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. **PLoS ONE**, v. 11, n. 3, 2016. Disponível em: <<http://doi.org/10.1371/journal.pone.0151664>>.

LEDIG, S.; BRUCKER, S.; BARRESI, G.; SCHOMBURG, J.; RALL, K.; WIEACKER, P. Frame shift mutation of LHX1 is associated with Mayer–Rokitansky–Küster–Hauser (MRKH) syndrome. **Human Reproduction**, v. 27, n. 9, p. 2872-2875, 2012.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 2009.

LINDENBAUM, P. FixVcfMissingGenotypes. Disponível em: <<https://github.com/lindenb/jvarkit/wiki/FixVcfMissingGenotypes>>, acesso em: 2 jan. 2017.

LIU, L.; LI, Y.; LI, S.; HU, N.; HE, Y.; PONG, R.; LIN, D.; LU, L.; LAW, M.;. Comparison of next-generation sequencing systems. **Journal of Biomedicine and Biotechnology**, v. 2012, 2012. Disponível em: <[doi:10.1155/2012/251364](https://doi.org/10.1155/2012/251364)>.

LUNTER, G.; GOODSON, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. **Genome Research**, New York, v. 21, n. 6, p. 936-939, 2011.

MEIRA, C. T.; FORTES, M. R. S.; FARAH, M. M.; PORTO NETO, L.; KELLY, M.; MOORE, S. S.; PEREIRA, G. L.; CHARDULO, L. A. L.; CURI, R. A. Speed index in the racing Quarter Horse: a genome-wide association study. **Journal of Equine Veterinary Science**, v. 34, p. 1263-1268, 2014a.

MEIRA, C.T.; CURI, R.A.; FARAH, M.M.; OLIVEIRA, H.N.; BELTRAN, N.A.R.; SILVA, J.A.IV.; MOTA, M.D.S. Prospection of genomic regions divergently selected in racing line of Quarter Horses in relation to cutting line. **Animal**, v. 8, n. 11, p. 1754-1764, 2014b.

MI, H.; POUDEL, S.; MURUGANUJAN, A.; CASAGRANDE, J. T.; THOMAS, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. **Nucleic Acids Research**, v. 44, n. 1, p. D336-D342, 2016.

MOLITERNO, A. R.; RESAR, L. M. AKNA: Another AT-hook transcription factor “hooking-up” with inflammation. **Cell Research**, v. 21, n. 11, p. 1528-1530, 2011.

NIELSEN, B.D.; TURNER, K.K.; VENTURA, B.A.; WOODWARD, A.D.; O'CONNOR C.I. Racing speeds of quarter horses, thoroughbreds and Arabians. *Equine Veterinary Journal*, v. 36, p. 128-132, 2006.

O'CONNOR, M. S.; CARLSON, M. E.; CONBOY, I. M. Differentiation rather than aging of muscle stem cells abolishes their telomerase activity. **Biotechnology Progress**, v. 25, p. 1130-1137, 2009.

PARSONS, M. T.; WHILEY, P. J.; BEESLEY, J.; DROST, M.; DE WIND, N.; THOMPSON, BA.; MARQUART, L.; HOPPER, J. L.; JENKINS, M. A.; BROWN M. A. Consequences of germline variation disrupting the constitutional translational initiation codon start sites of MLH1 and BRCA2: Use of potential alternative start sites and implications for predicting variant pathogenicity. **Molecular Carcinogenesis**, v. 54, n. 7, p. 513-522, 2013.

PETERSEN, J.L.; MICKELSON, J. R.; COTHRAN, E. G.; ANDERSSON, L. S.; AXELSSON, J. Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data. **PLoS ONE**, v.8, n. 1, 2013. Disponível em: <e54997. doi: 10.1371/journal.pone.0054997>.

PURCELL, S., NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M.A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P.I.; DALY, M.J.; SHAM, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, Baltimore, v. 81, p. 559-575, 2007.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL < <https://www.R-project.org/> >.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F.S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**. v.15, p.478, 2014.

SHIN, D-H.; LEE, J. W.; PARK, J-E.; CHOI, I-Y.; OH, H-S.; KIM, H. J.; KIM, H. Multiple Genes Related to Muscle Identified through a Joint Analysis of a Two-stage Genome-wide Association Study for Racing Performance of 1,156 Thoroughbreds. **Asian-Australasian Journal of Animal Sciences**, v. 28, n. 6, p. 771-781, 2015

Capítulo 5 – Considerações finais

Nos trabalhos apresentados foram identificadas regiões genômicas e genes que possivelmente atuam sobre o desempenho em corridas de cavalos da raça Quarto de Milha. As abordagens utilizadas para associação revelaram um grande número de regiões e genes ligados ao Índice de Velocidade (IV), indicando que o desempenho em corridas nesta raça é uma característica quantitativa, ou seja, não é controlada por muitos genes.

O IV, além de pedigree, é comumente utilizado por criadores para comércio e seleção de animais de melhor desempenho. Embora, essa característica tenha apresentada baixa herdabilidade, o que pode indicar uma importante influência de ambiente (treinamento, alimentação, criador, etc.), os resultados das associações se mostraram coerentes com os processos biológicos que podem estar envolvidos em diferenças no desempenho em corridas da raça. As regiões genômicas associadas continham genes que apresentaram funções relacionadas ao sistema neuromotor, ao desenvolvimento ósseo e de tendões, à respostas hormonais, à contração muscular, ao metabolismo energético e de esteróis. Deste modo, de maneira prática, alterações nesses genes podem ter efeito sobre o temperamento, coordenação de movimentos, capacidade de arranque, força entre outras características importantes para corridas curtas e de alta intensidade, como são as disputadas pela linhagem de corrida desta raça.

De forma geral, os marcadores associados neste estudo tiveram importante efeito sobre o desempenho, visto que explicaram uma porcentagem relativamente alta da variância genética do Índice de velocidade. Nesse sentido, após a validação dos efeitos dessas alterações genéticas sobre o desempenho, é viável a utilização de painéis de genotipagem de baixa densidade, contendo somente os marcadores associados, como ferramenta de auxílio na seleção de animais geneticamente superiores.