X Simpósio Brasileiro de Melhoramento Animal
Uberaba, MG – 18 a 23 de agosto de 2013
S B M A 2013

Efeito de SNPs em posições equívocas sobre o decaimento do desequilíbrio de ligação

Adam Taiti Harth Utsunomiya[1], Marcos Vinícius Gualberto Barbosa da Silva[2], Daniel Jordan Abreu dos Santos[1], Marco Antônio Machado[2], Rui da Silva Verneque[2], João Cláudio Panetto[2]

[1]Programa de Pós-Graduação em Genética e Melhoramento Animal – FCAV UNESP, Jaboticabal. e-mail: adamtaiti@gmail.com
[2]Embrapa Gado de Leite, Juiz de Fora. e-mail: marcos.vb.silva@embrapa.br

Resumo: Existem marcadores SNPs com posições erradas nos mapas genéticos. Este trabalho avaliou o efeito destes SNPs sobre a curva de decaimento do desequilíbrio de ligação em quatro raças zebuínas e uma população F2 oriunda do cruzamento entre as raças Gir e Holandesa. As médias das correlações entre pares de SNPs estavam deflacionadas a curta distância e inflacionadas à longa distância na raça Sindi e na F2. Para as demais, houve apenas inflação das médias das correlações para as distâncias acima de 200Kb. Estes resultados indicam que é necessário realizar um controle de qualidade dos dados para exclusão de SNPs para estudos de desequilíbrio de ligação.
Palavras–chave: desequilíbrio de ligação, polimorfismo de um único nucleotídeo.

Effect of misplaced SNP on the linkage disequilibrium decay

Abstract: There are SNP markers with erroneous positions in genetic maps. This study evaluated the effect of these SNP on linkage disequilibrium decay in four zebu breeds and in a F2 population originated by Gyr and Holstein crossbred. The average of SNP correlations were deflated at short distances and inflated at long distances in Sindi breed and F2 population. For the others, there was only inflation averages the distances above 200Kb. These results indicate that it is necessary to performe a data quality control to the exclusion of misplaced SNP to linkage disequilibrium studies.
Keywords: linkage disequilibrium, single nucleotide polymorphism

Introdução

A localização física de polimorfismos de um único nucleotídeo (SNP) é determinada pelo alinhamento de sua sonda contra um genoma referência. Em bovinos, existem duas montagens do genoma de referência amplamente utilizadas: o Btau, desenvolvido no *Baylor College of Medicine*, e o UMD, desenvolvido na *University of Maryland*, ambas instituições localizadas nos EUA.

A montagem de um genoma de referência é complexa e pode incorrer em alguns equívocos quanto ao posicionamento de algumas sequências, afetando a determinação da posição de alguns SNPs e até mesmo atribuindo posições erradas a eles (*SNP misplaced*). Alguns trabalhos tem proposto algoritmos de detecção de possíveis *misplaced* (BOHMANOVA *et al*., 2010, PAUSCH *et al*., 2013) e sugerido possíveis implicações em estudos de decaimento do desequilíbrio de ligação (DDL), imputação, entre outros estudos, que necessitem da localização correta dos SNPs.

O objetivo deste estudo foi analisar os possíveis efeitos dos *misplaced* SNPs sobre o decaimento do desequilíbrio de ligação nas raças Gir, Guzerá, Sindi, Nelore e em uma população F2.

Material e Métodos

Foram utilizadas informações de 973 animais da raça Nelore, 1.997 da raça Gir, 1.023 da raça Guzerá, 117 da raça Sindi e 372 de população F2 todos genotipadas utilizando o Illumina® BovineSNP50K BeadChip e coordenadas genômicas baseadas na montagem UMD v3.1. O controle de qualidade (CQ) das amostras e dos SNPsm bem como as correlações entre SNPs (r^2), foram efetuados no software PLINK v1.07 (PURCELL et al., 2007). Foram removidos da análise os animais com call rate < 0.90, valores de heterozigosidade que excederam ± 3 desvios-padrão da média e SNPs com MAF < 0.02, call rate de SNPs < 0.98, p-value do teste exato de Fischer para equilíbrio de Hardy-Weinberg < 1e-6 e SNPs coincidentes e com $r^2 > 0.998$. Os dados resultantes deste CQ foram duplicados e um CQ adicional para exclusão dos *misplaced SNPs* foi aplicado a um dos conjuntos para avaliar seu efeito sobre o DDL. Os pares de SNPs foram ordenados de forma decrescente em 21 janelas, segundo a distância física entre eles, e as médias de r^2 de cada janela para os dois conjuntos de dados foram comparadas pelo teste t.

Resultados e Discussão

Foram detectados 90; 50; 219; 150 e 180 *misplaced SNPs* para as raças Nelore, Gir, Guzerá, Sindi e F2, respectivamente. Todas as médias de r^2 acima de 100Kb foram estatisticamente diferentes para as raças Guzerá, Nelore e F2. Para as raças Sindi e Gir, as médias passaram a ser diferentes a partir de 500Kb (Tabela 1). A raça Nelore apresentou maior sensibilidade à presença de *misplaced SNPs*, com inflação média de 33%. Analisando as raças Sindi (Figura 1) e F2, até 20Mb, as médias das janelas dos dados com *misplaced SNPs* apresentaram-se deflacionadas em relação às médias das janelas dos dados sem *misplaced SNPs*. Para distâncias maiores que 20Mb, as médias com *misplaced SNPs* apresentaram-se inflacionadas em relação as médias sem *misplaced SNPs*. Este padrão é esperado, pois se, durante a montagem do genoma, um conjunto de SNPs é colocado no ínicio do cromossomo, quando deveria estar na posição mais central, por exemplo, ele apresentará suas maiores correlações a longas distâncias e menores correlações a curtas distâncias (Figura 2).

Tabela 1. Valores de p do teste t de Student para diferença entre as médias de cada janela para as diferentes raças estudadas.

Janelas	Gir	Sindi	F2	Guzerá	Nelore
0-40Kb	0.89169^{NS}	0.66851^{NS}	0.75333^{NS}	0.25781^{NS}	0.59079^{NS}
40-60Kb	0.84299^{NS}	0.73823^{NS}	0.96808^{NS}	0.45689^{NS}	0.72907^{NS}
60-100Kb	0.63438^{NS}	0.63351^{NS}	0.16702^{NS}	0.24372^{NS}	0.26992^{NS}
100-200Kb	0.39265^{NS}	0.43471^{NS}	0.03137*	0.02814*	8.56463e-06*
200-500Kb	0.21363^{NS}	0.12759^{NS}	6.66924e-05*	0.00215*	2.96367e-61*
500-1Mb	0.02008*	0.01541*	2.07084e-08*	1.20916e-05*	5.06386e-239*
1-2Mb	0.00024*	0.00295*	9.64300e-13*	1.83266e-10*	0*
2-5Mb	4.71076e-12*	2.50485e-07*	3.91347e-30*	2.27770e-36*	0*
5-10Mb	2.10036e-19*	3.56842e-08*	3.10491e-27*	2.68555e-69*	0*
10-20Mb	2.94529e-36*	5.27957e-05*	0.04742*	5.12230e-195*	0*
20-40Mb	1.62011e-72*	2.74782e-43*	2.59473e-144*	0*	0*
40-60Mb	3.52391e-103*	3.11455e-94*	0*	5.17234e-309*	0*
60-80Mb	1.10241e-93*	1.12299e-131*	0*	3.85534e-147*	0*
80-100Mb	9.11641e-45*	3.05851e-136*	1.45574e-317*	1.64688e-59*	0*
100-120Mb	3.39416e-20*	2.00585e-77*	1.11164e-246*	0.60631^{NS}	0*
>120Mb	6.57541e-06*	8.37623e-11*	2.00773e-100*	1.02233e-27*	0*

* = P<0,01; NS = não significativo

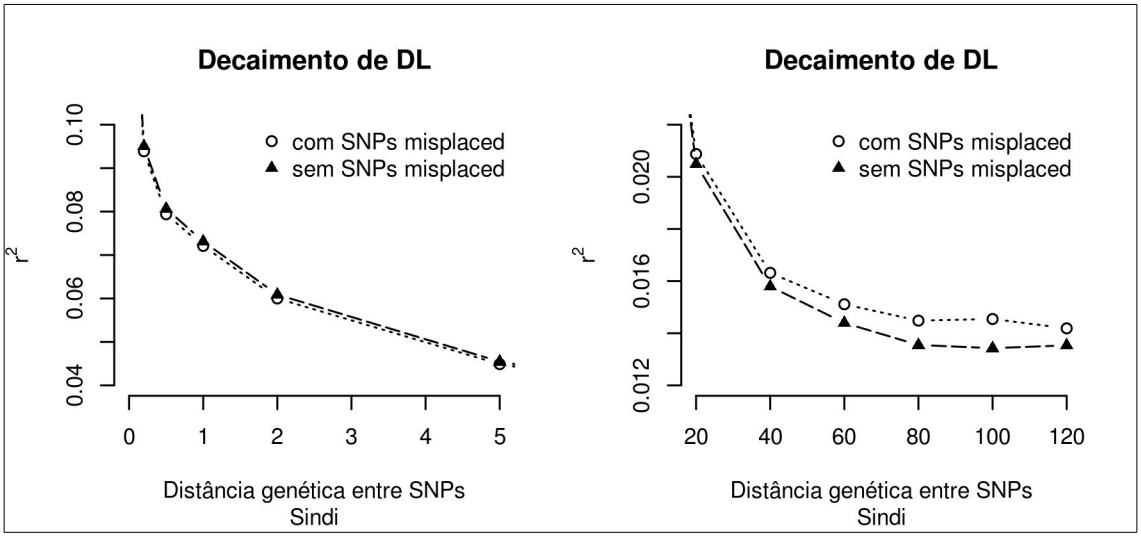

Figura 1. Decaimento do desequilíbrio de ligação da raça Sindi à curta (esquerda) e longa (direita) distância entre SNPs.

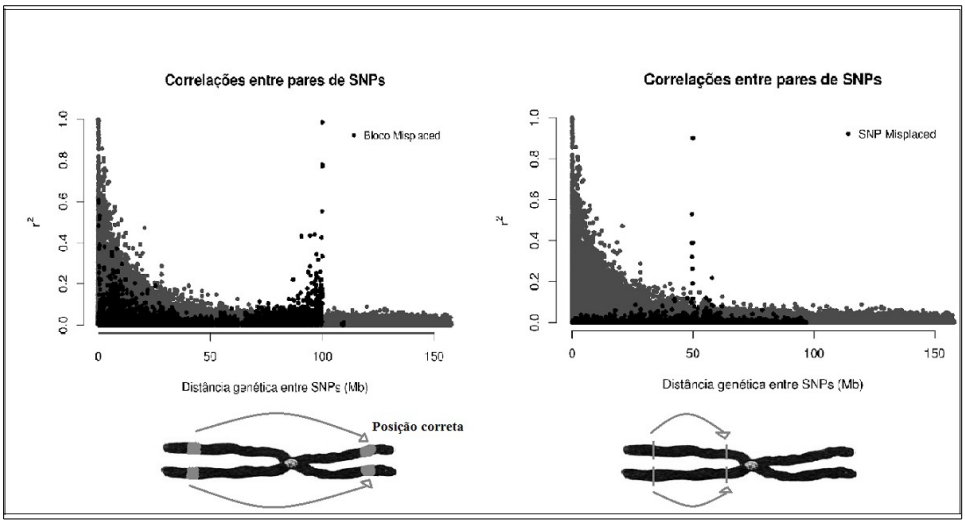

Figura 2. Apresentação do comportamento do DL quando há um bloco de *misplaced SNPs* (esquerda) e quando há um único *misplaced SNP*.

Realizando um estudo das características do desequilíbrio de ligação em bovinos da raça Holandesa, Bohmanova et al. (2010) detectaram correlações mais altas do que as esperadas a longas distâncias em alguns cromossomos, indicando possíveis *misplaced SNPs*. Na tentativa de identificar e corrigir o posicionamento desses marcadores, os autores deste trabalho desenvolveram um algoritmo no qual a maior correlação de um SNP é registrada e, se esta correlação estiver a uma distância maior do que 10 Mb, esse SNP é anotado como um possível *misplaced*. Este algoritmo detectou 223 *misplaced* de 38.590 SNPs analisados, e suas respectivas localizações foram corrigidas assumindo que o marcador deveria estar localizado entre dois outros SNPs com as correlações mais altas com o mesmo. No entanto, a correção do posicionamento de um SNP, ou conjunto deles, é um tanto arbitrária, pois o tipo de correção proposta será dependente do ascertainment bias do SNPchip para cada raça que se deseja estudar, o que pode tornar a estimativa viesada.

Em um estudo de imputação de genótipos de alta densidade em bovinos da raça Fleckvieh, Pausch et al. (2013) detectaram regiões com baixa acurácia de imputação e atribuíram isto aos *misplaced SNPs*. Usando este procedimento, 5.039 de 599.535 SNPs foram identificados como *misplaced*. Pausch et al. (2013) também mencionam que SNPs significativamente associados com fenótipos em estudos de associação ampla do genoma devem ser validados quanto à sua localização para evitar erros de interpretação dos resultados.

Conclusões

A presença de *misplaced SNPs* em um conjunto de dados superestima a correlação entre SNPs a longas distâncias e pode subestimar essa correlação a curtas distâncias. Isto indica a necessidade de se detectar e excluir estes SNPs em estudos de desequilíbrio de ligação e outros estudos que dependem ou utilizam a posição física de um marcador.

Literatura citada

PAUSCH, H.; AIGNER, B.; EMMERLING, R.; CHRISTIAN, E. et al. Imputation of high-density genotypes in the Fleckvieh cattle population. **Genetic Selection Evolution**, v.45, 2013.

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

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L. et al. PLINK: A tool set for whole genome association and population-based linkage analyses. **Am. J. Hum. Genet.**, v.81, 2007.