Estimativa do desequilíbrio de ligação e determinação da estrutura populacional utilizando SNPs em progênies de ciclo C-9 de seleção recorrente em milho-pipoca da UENF

Ismael Albino Schwantes[(1)]; Janeo Eustáquio de Almeida Filho[(2)]; Fernando Higino de Lima e Silva[(2)]; Samuel Henrique Kamphorst[(1)]; Valter Jário de Lima[(1)]; Antonio Teixeira do Amaral Júnior[(3)]

[(1)] estudante de pós-graduação; Universidade Estadual do Norte Fluminense; Campos dos Goytacazes, RJ; ismael.schwantes31@gmail.com; [(2)] Pós-doutorado; Universidade Estadual do Norte Fluminense; [(3)] Professor Associado; Universidade Estadual do Norte Fluminense.

RESUMO: O objetivo do presente estudo foi estimar o desequilíbrio de ligação (LD) e determinar a estrutura populacional utilizando SNPs no ciclo C-9 de seleção recorrente em milho-pipoca da UENF, a fim de obter informações relevantes para estudos posteriores envolvendo seleção recorrente genômica e associação genômica ampla. O estudo contou com 200 indivíduos, os quais foram genotipados, e são potenciais genitores das famílias endogâmicas S1 que serão utilizadas em ensaios de progênies. Foram utilizados 21442 SNPs, sendo que pouco mais de 10 % foram monomórficos e aproximadamente 35 % com MAF > 0 e <0.05. Foram observados poucos dados perdidos, com magnitude inferior a 5%, para a quase totalidade do marcadores. Ao analisar os resultados (PCA e *heatmap*) pode-se constatar que os indivíduos apresentam um grau de parentesco muito baixo e não há indicação de formação de diferentes estruturas populacionais, que é condizente com a estratégia de seleção recorrente intrapopulacional, a partir da UENF-14. O algoritmo de Evanno permitiu identificar 2 grupos genéticos dentro da população de seleção recorrente da UENF, porém, este resultado não é consentâneo com as estimativas de GRM (*Genomic Relationship Matrix*) e pela origem dos genótipos, em que uma única população foi considerada para obtenção dos ganhos genéticos. A análise de LD constatou que a metade do decaimento ocorreu em 109,940 Kb, o que fornece informações estratégicas importantes para estudos posteriores de seleção genômica.

Termos de indexação: desequilíbrio de ligação, seleção recorrente genômica, estrutura populacional

INTRODUÇÃO

Seleção Recorrente (SR) é uma eficiente estratégia de melhoramento populacional, tendo em vista que a população melhorada pode ser recomendada como cultivar, e/ou ser base para obtenção de linhagens elite. Desta forma a SR é um dos métodos de melhoramento mais utilizados para várias espécies como é o caso do milho-pipoca. Porém, este procedimento é trabalhoso e demanda muito tempo, pois a cada ciclo de seleção geralmente são realizadas três etapas: obtenção das progênies, avaliação e recombinação das famílias superiores. Desta forma, o tempo necessário para completar um ciclo de SR pode requerer de 2,0 a 2,5 anos, o que reduz muito os ganhos genéticos anuais (Hallauer et al., 2010). Com a utilização da Seleção Recorrente Genômica (SRG) – procedimento que utiliza os princípios da Seleção Genômica Ampla (GWS) – pode-se diminuir o tempo demandado para cada ciclo de seleção (Heffner et al., 2009), pois a SRG possibilita realizar as fases de avaliação e recombinação simultaneamente. Esse método proporciona uma forma de seleção precoce direta, ou seja, atua precocemente sobre os genes expressos na idade adulta (Fritsche Neto, 2011). Além do mais, a predição do valor genético utilizando as informações de marcas genéticas tem se mostrado mais acurada, do que as avaliações tradicionais que são baseadas exclusivamente no pedigree (Crossa et al., 2014).

Contudo o sucesso da GWS e dos estudos de associação genômica ampla (GWAS) dependem da extensão do desequilíbrio de ligação (LD) ao longo do genoma. Uma vez que idealmente para a seleção genômica todos os genes que contribuem para a variação do caráter estarão em desequilíbrio de ligação com pelo menos um marcador, na população em estudo (Goddard & Hayes, 2007). Dessa forma o decaimento do desequilíbrio de ligação (LD) sobre a distância física em um população determina a densidade da cobertura por marcadores necessária para realizar análises genômicas. Por exemplo, se o LD decai rapidamente, uma densidade mais elevada de marcadores é necessária para obter marcadores localizados perto o suficiente de *loci* funcionais (Yu & Buckler, 2006).

Outro fator crucial principalmente para a GWAS é a estrutura populacional, uma vez que várias associações espúrias podem ser devido a padrões estruturais (Yang et al., 2014). A estrutura em uma população que ocorre naturalmente devido a processos dispersivos (deriva genética), sistemáticos (mutação, migração e seleção), bem como acasalamentos controlados, comumente observados em populações de melhoramento. No caso da GWS, o efeito da estrutura teria maior impacto na extrapolação do modelo preditivo para outra população não relacionada (Wray et al., 2013).

Portanto, a proposta do presente estudo foi de estimar o desequilíbrio de ligação e determinar a estrutura populacional utilizando SNPs no ciclo C-9 de seleção recorrente em milho-pipoca da UENF, a fim de obter informações relevantes para estudos posteriores envolvendo seleção recorrente genômica e associação genômica ampla.

MATERIAL E MÉTODOS

A população utilizada nesse estudo é referente ao nono ciclo (C-9) de seleção recorrente intrapopulacional de milho-pipoca da UENF, essa população foi concebida pela recombinação das 30 famílias de irmãos completos selecionadas no ciclo anterior (C-8). O presente estudo contou com 200 indivíduos, potenciais genitores das famílias endogâmicas S_1 que serão utilizadas em ensaios de progênies.

A extração do DNA genômico foi realizada utilizando-se o kit de extração QIAGEN (DNeasy® Mini Kit), conforme exigências da empresa de genotipagem, uma vez que a técnica exige concentrações adequadas e excelente integridade molecular. O material vegetal a ser utilizado no processo de extração foi proveniente de tecido foliar jovem, coletado individualmente dos genótipos genitores. A quantificação do DNA foi realizada por meio de kits específicos para o fluorômetro QubitTM (invitrogenTM), a exemplo o Qubit® dsDNA BR (molecular probes® by life technologies). As imagens da integridade das amostras extraídas foram foto documentadas em gel de agarose a 1 %, utilizando o quantificador de peso molecular DNA Lambda (invitrogenTM).

As análises moleculares e de bioinformática foram realizadas na RAPID GENOMICS LLC, sendo esta uma empresa incubada na Florida University, em Gainesville-Flórida-USA. Para a genotipagem, foram utilizados cerca de 22000 SNPs, disponibilizados pela empresa responsável, que realizou o processo de hibridização entre as sondas e as sequências de DNA de milho-pipoca disponibilizadas.

Para determinar a estrutura populacional foi utilizado o software R (R Development Core Team, 2013), com o qual foi estimada a matriz de parentesco genômico, usando todos os marcadores com *Minor Allele Frequency* (MAF) > 0.05 (GRM-*Genomic Relationship Matrix*) usando a função A.mat do pacote rrBLUP (Endelman, 2012), e pela GRM foi realizado a análise de componentes principais (PCA). Para complementar a avaliação da estrutura populacional, foi realizada a inferência Bayesiana implementada pelo programa STRUCTURE (Pritchard et al., 2000), sendo que o número de clusters foi avaliado pelo método proposto por Evanno et al. (2005). Para o STRUCTURE foram utilizados 684 SNPs filtrados com o software PLINK 1.9 (Purcell & Chang, 2015), todos com MAF >0.05, em equilíbrio de Hardy-Weinberg (p-valor<0.05) e também filtrados pelo LD, usando o procedimento *LD pruning* mantendo apenas marcadores com r^2<0.2 em janelas de 1000Kb.

A estimativa do desequilíbrio de ligação foi realizada por meio do quadrado da correlação de Pearson (r^2), o r^2 foi calculado entre todos os pares de marcadores com MAF>0.05 pertencentes ao mesmo cromossomo, utilizando-se o software PLINK. Para o estudo do decaimento, foi ajustado o modelo não linear proposto por Hill & Weir (1988), usando a função *nlm* do software R (R Development Core Team, 2013). A medida r^2 capta, além dos efeitos da fração de recombinação entre os locos sobre o decaimento do LD, as histórias mutacionais ocorridas (Mangin et al., 2011).

RESULTADOS E DISCUSSÃO

Cada um dos 200 genitores foi genotipado para 21442 SNPs, sendo pouco mais de 10 % de monomórficos e aproximadamente 35 % com MAF > 0 e <0.05 (Figura 1). Foi observada reduzida magnitude de dados perdidos, com percentual inferior a 5 % para a maioria dos marcadores.

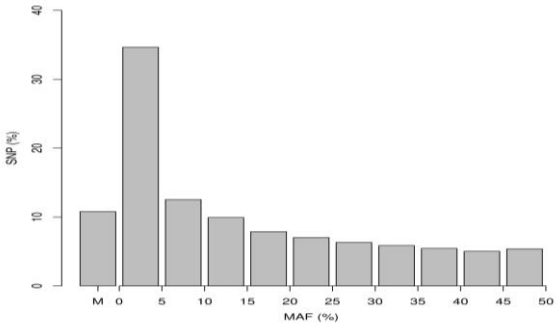

Figura 1. Distribuição de frequência da MAF dos marcadores polimórficos e dos monomórficos (M).

Por meio do uso de marcadores moleculares é possível estimar o coeficiente de parentesco entre os indivíduos e, com isto melhorar

as estimativas dos componentes de variância, aumentar a acurácia na predição dos valores genotípicos dos indivíduos em seleção e também inferior sobre o padrão estrutural da população considerando a identidade por estado (IBS).

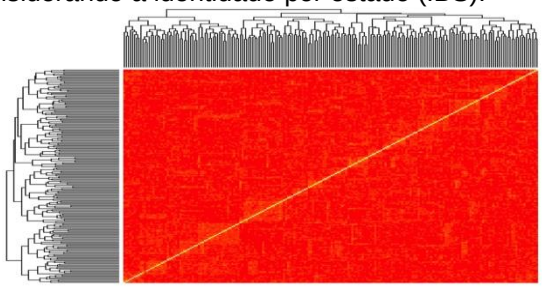

Figura 2: Matriz dos coeficientes de parentesco (GRM)

Analisando-se a matriz de parentesco na Figura 2, onde quanto maior a intensidade do vermelho menor é o parentesco e quanto maior a intensidade do amarelo maior é o coeficiente de parentesco, pode-se inferir que os 200 genitores apresentam um grau de parentesco muito baixo e não há indicação de formação de diferentes estruturas populacionais.

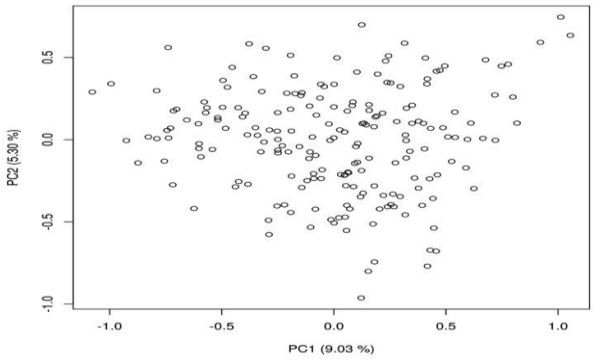

Figura 3: Análise de componentes principais (PCA)

Ao realizar a análise de componentes principais na matriz de parentesco, pode ser verificado que os dois primeiros componentes explicaram menos de 15 % da variação total da matriz GRM, além de verificar que não há formação de grupos populacionais, o que é condizente com a população de seleção recorrente em uso (Figura 3).

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	20	-124996.845000	1.422923	—	—	—
2	20	-123898.635000	5.663761	1098.210000	407.840000	72.008693
3	20	-123208.265000	11.017368	690.370000	221.345000	20.090552
4	20	-122739.240000	620.110547	469.025000	197.920000	0.319169
5	20	-122072.295000	141.074838	666.945000	122.190000	0.866136
6	20	-121527.540000	42.454129	544.755000	62.245000	1.466171
7	20	-121045.030000	44.220513	482.510000	15.805000	0.357413
8	20	-120578.325000	64.930172	466.705000	271.430000	4.180337
9	20	-120383.050000	689.262958	195.275000	113.165000	0.164183
10	20	-120074.610000	585.882651	308.440000	—	—

Figura 3: Saída de dados do método de Evanno gerado via Structure.

O software STRUCTURE é utilizado para inferir sobre a estrutura da população, que abrange a heterogeneidade na distribuição dos genótipos e do grau de endogamia dentro de populações e entre estas. O algoritmo de Evanno et al. (2015) identificou 2 grupos genéticos dentro da população UENF-14, sob seleção recorrente. Entretanto, este resultado não corrobora o indicado pela GRM (PCA e *heatmap*) e pela origem dos genótipos, uma vez que são indivíduos oriundos de um programa de seleção recorrente intrapopulacional, sendo realizadas recombinações entre progênies selecionadas de uma mesma população. Isto pode ser explicado pelo fato do critério delta K não realizar o cálculo de K, para delta K=1. No entanto ao se analisar o *Mean LnP(K),* gerado pelo STRUCTURE (Tabela 1), pode-se verificar que o maior valor negativo foi para K=1. Com isso, pode-se especular que a ausência de uma suposta estrutura, permite realizar estudos de associação genômica ampla para essa população.

Desequilíbrio de ligação (LD) é a associação não aleatória entre alelos de diferentes locos em uma população e utilizando-se marcadores SNPs, a medida r^2 é uma das mais utilizadas (Mangin et al., 2011). Procedendo-se a análise de LD deste estudo (Figura 4) pode-se constatar que a metade do decaimento ocorreu em 109,940 Kb, sendo este um resultado semelhante ao encontrado por Ching et al. (2002) em milho comum, onde relatam um alcance do LD de 100-500 Kb em linhagens endogâmicas. Porém estes valores não são comuns e resultam dos processos específicos de melhoramento a que esta população foi submetida.

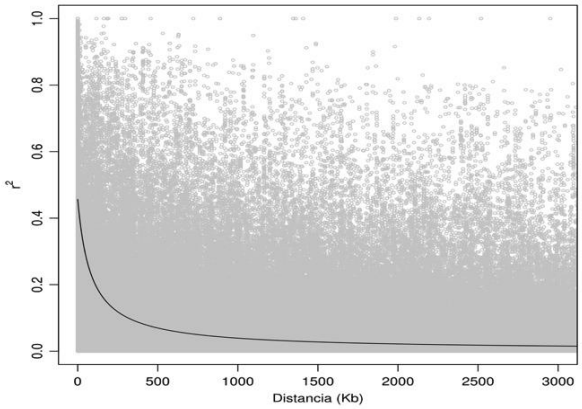

Figura 4: Análise do desequilíbrio de ligação (LD)

Outros estudos realizados com milho comum relatam a extensão do LD como baixa, com rápido decaimento ao longo das distâncias físicas entre locos, com alcance variando, na maioria das

populações, entre 0,5 e 7,0 Kb (Remington et al., 2001; Truntzler et al., 2012). Em milho, o LD encontrado depende muito do tipo de população estudada, mas basicamente o LD possui baixa extensão, devido ao grande genoma e a alta taxa de recombinação, além da grande mobilidade existente no genoma desta espécie (pela ação de transposons e retrotransposons) (Gupta et al., 2005).

Pode-se inferir que a população de seleção recorrente de milho-pipoca em uso possui uma LD com decaimento um tanto mais elevado que o normal e que isto ocorre devido ao efeito da seleção artificial realizada durante o melhoramento. Dessa forma, para a presente população, o número de marcadores necessários tanto para GWS quanto para GWAS é substancialmente menor do que em populações de tamanho efetivo elevado, como por exemplo coleções nucleares. Em contrapartida as inferências realizadas para uma população convencional de melhoramento tanto em termos de GWS como GWAS não podem ser extrapoladas para outras populações, uma vez que determinados blocos haplótipos de tamanho elevado observados em uma população de melhoramento provavelmente não são comuns a outras populações. Isto posto, demonstra a importância do conhecimento da magnitude do LD, já que esta informação estabelece a quantidade de marcas necessárias para a realização de diversos estudos, incluindo os estudos de associação e de seleção assistida (Goldstein, 2001).

CONCLUSÕES

Os genótipos em estudo caracterizam-se por não formar diferentes estruturas populacionais.

A origem dos genótipos em estudo explica os resultados encontrados.

O software STRUCTURE via método de Evanno não fornece o delta K, para casos onde o número ideal de grupos é 1.

O LD foi de cerca de 100 kb, o que fornece informações estratégicas importantes para estudos posteriores de seleção genômica.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro das agências de fomento à pesquisa: CAPES, FAPERJ e CNPq.

REFERÊNCIAS

CHING, A. D. A.; CALDWELL, K. S.; JUNG, M.; DOLAN, M.; SMITH, O. S. H.; TINGEY, S.; MORGANTE, M.; RAFALSKI, A. J. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC genetics**, Londres, v. 3, n. 1, p. 19, 2002.

DESCHAMPS, S.; LLACA, V.; MAY, G. D. Genotyping-by-sequencing in plants. **Biology**, v. 1, p. 460–483, 2012.

FRITSCHE-NETO, R. **Seleção genômica ampla e novos métodos de melhoramento do milho** (Tese de Doutorado). Viçosa: UFV, 2011, 28p.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p. 323-330, 2007.

GOLDSTEIN, D. B. Islands of linkage disequilibrium. **Nature Genetics**, New York, n. 29, p. 109-111, 2001.

GUPTA, P. K.; RUSTGI, S.; KULWAL, P. L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects**. Plant molecular Biology**, Amsterdam, n. 57, p. 461- 485, 2005.

HALLAUER, A.R; MIRANDA FILHO, J.B, CARENA, M.J. **Quantitative genetics in maize breeding**. 3. ed, New York : Springer, 2010, 663p.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J. L. Genomic Selection for Crop Improvement. **Crop Science**, n. 49, p.1, 2009.

HILL, W.G.; WEIR, B.S. Variances and covariances of squared linkage disequilibria in finite populations. **Theor Popul Biol**, v.33, p. 54–78, 1988.

MANGIN, B.; SIBERCHICOT, A.; NICOLAS, S.; DOLIGEZ, A.; THIS, P.; CIERCO-AYROLLES, C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. **Heredity**, New York, v. 108, n. 3, p. 285-291, 2011.

PURCELL, S; CHANG, C. PLINK, Version 1.9. 2015

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. <http://www.R-project.org/>. Acesso em 27 de maio de 2016.

REMINGTON, D. L.; THORNSBERRY, J. M.; MATSUOKA, Y.; WILSON, L. M.; WHITT, S. R.; DOEBLEY, J.; KRESOVICH, S.; GOODMAN, M. M.; BUCKLER, E. S. Structure of linkage disequilibrium and phenotypic associations in the maize genome. **Proceedings of the National Academy of Sciences**, Washington, v. 98, n. 20, p. 11479-11484, 2001.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of Population Structure Using Multilocus Genotype Data. **Genetics**, Heidelberg, v. 155, p. 945-959, 2000.

TRUNTZLER, M.; RANC, N.; SAWKINS, M. C.; NICOLAS, S.; MANICACCI, D.; LESPINASSE, D.; RIBIÈRE, V.; GALAUP, P.; SERVANT, F.; MULLER, C.; MADUR, D.; BETRAN, J.; CHARCOSSET, A.; MOREAU, L. Diversity and linkage disequilibrium features in a composite

public/private dent maize panel: consequences for association genetics as evaluated from a case study using flowering time. **Theoretical and Applied Genetics**, Heidelberg, v. 125, n. 4, p. 731-747, 2012.

YANG, J.; ZAITLEN, N.; GODDARD, M.; VISSCHER, P.; PRICE, A. "Advantages and Pitfalls in the Application of Mixed-Model Association Methods." **Nature Genetics**, v. 46, n. 2, p.100–106, 2014.

YU, J.; BUCKLER, E.S. Genetic association mapping and genome organization of maize. **Curr. Opin. Biotechnol.** v.17, p. 155–160, 2006.

WRAY, N. R.; YANG, J.; HAYES, B.; PRICE, A.; GODDARD, M.; VISSCHER, P. "Pitfalls of Predicting Complex Traits from SNPs." **Nature Reviews. Genetics**, v.14, n.7, p. 507–515, 2013.