GABRIEL BORGES MUNDIM

EFFICIENCY OF GENOME-WIDE ASSOCIATION STUDY IN OPEN-POLLINATED POPULATIONS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS - BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
M965e
2016

Mundim, Gabriel Borges, 1987-
Efficiency of genome-wide association study in open-pollinated populations / Gabriel Borges Mundim. – Viçosa, MG, 2016.
ix, 48f. : il. ; 29 cm.

Orientador: José Marcelo Soriano Viana.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.26-35.

1. Genética vegetal. 2. Genética quantitativa. 3. Milho - Melhoramento genético. I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 581.35

GABRIEL BORGES MUNDIM

EFFICIENCY OF GENOME-WIDE ASSOCIATION STUDY IN OPEN-POLLINATED POPULATIONS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 12 de fevereiro de 2016.

Vinícius Ribeiro Faria

Marcos Deon Vilela de Resende

Fabyano Fonseca e Silva
(Coorientador)

Antônio Carlos Baião de Oliveira

José Marcelo Soriano Viana
(Orientador)

Aos meus pais, Edilson e Maria Perpétua,

À minha amada esposa Vanessa,

Aos meus avós, familiares e amigos.

DEDICO

AGRADECIMENTOS

Agradeço a Deus, pela sabedoria e por guiar os meus passos.

Aos meus pais, Edilson e Maria Perpétua, pelo amor, carinho, apoio e confiança em mim depositados.

À minha irmã Luísa, pela amizade e pelo carinho.

Aos meus avós, Oldemar (in memorian) e Purcínia, Jovino e Felinda, pelo carinho e orações.

À minha amada esposa Vanessa, pelo amor, companheirismo, paciência e por me apoiar em todos os momentos. Aos seus pais, José Francisco e Olga, pela amizade e confiança.

Aos meus demais familiares, pelo apoio, incentivo e amizade.

À Universidade Federal de Viçosa (UFV) e ao Programa de Pós-Graduação em Genética e Melhoramento, pela oportunidade.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG), pelo apoio financeiro.

Ao meu orientador, professor José Marcelo Soriano Viana, por sua orientação segura, amizade, incentivo, dedicação e pelos seus ensinamentos.

Aos professores Fabyano Fonseca e Silva, Vinícius Ribeiro Faria, ao Dr. Marcos Deon Vilela de Resende e ao Dr. Antônio Carlos Baião de Oliveira pela disponibilidade de participação na banca e pelas valiosas contribuições.

Aos colegas do Programa Milho-Pipoca, pelo trabalho em equipe, pela amizade e colaboração.

Aos amigos da Dow Agrosciences Sementes & Biotecnologia, pela oportunidade de trabalho e crescimento profissional.

A todos os meus amigos com os quais eu convivi desde a graduação, e a todos os companheiros de república, pela amizade, confiança e apoio.

A todos que contribuíram para o desenvolvimento desse trabalho e para a minha formação profissional.

MUITO OBRIGADO!

BIOGRAFIA

GABRIEL BORGES MUNDIM, filho de Edilson Borges da Silva e Maria Perpétua Mundim Borges, nasceu em 15 de novembro de 1987, em Patos de Minas, Minas Gerais.

Estudou na Escola Estadual "Coronel Cristiano", em Lagoa Formosa, Minas Gerais e concluiu o Ensino Médio em 2005, no Colégio Marista de Patos de Minas, Minas Gerais.

Em maio de 2006, ingressou no curso de Agronomia da Universidade Federal de Viçosa, obtendo o título de Engenheiro Agrônomo em janeiro de 2011.

Em março de 2011, iniciou o curso de Mestrado em Genética e Melhoramento pela Universidade Federal de Viçosa, obtendo o título de Magister Science em fevereiro de 2013.

Em março de 2013, iniciou o curso de Doutorado em Genética e Melhoramento pela Universidade Federal de Viçosa, submetendo-se à defesa de tese em fevereiro de 2016.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

MUNDIM, Gabriel Borges, D. Sc., Universidade Federal de Viçosa, fevereiro de 2016. **Eficiência do estudo de associação genômica ampla em populações de polinização aberta.** Orientador: José Marcelo Soriano Viana. Coorientadores: Fabyano Fonseca e Silva e Rodrigo Oliveira de Lima.

A maioria dos estudos de associação genômica ampla (GWAS) com espécies vegetais publicados até agora têm empregado painel de linhagens e quase nenhuma informação sobre os GWAS em populações de polinização aberta foi encontrada na literatura. Portanto, os objetivos deste trabalho foram: (i) apresentar aspectos teóricos, potencial e limitações dos GWAS em populações de polinização aberta; (ii) analisar a influência da herdabilidade do QTL e do tamanho populacional sobre os GWAS com populações de polinização aberta; e (iii) comparar a eficácia dos GWAS na detecção de QTL em populações de polinização aberta, painel de linhagens e linhagens endogâmicas recombinantes (RILs). Cinquenta amostras de populações com desequilíbrio de ligação (LD) foram simuladas, considerando os tamanhos populacionais de 400 e 200 indivíduos, e 10.000 SNPs, 10 QTL e 90 genes menores foram aleatoriamente distribuídos em 10 cromossomos, com uma densidade média de 1 SNP a cada 0,1 cM. Os valores fenotípicos simulados referem-se à três características de milho-pipoca com diferentes graus de dominância, considerando herdabilidades em sentido amplo de 0,4 e 0,8. Os cenários foram comparados com base no poder de detecção de QTL, no número de associações falso-positivas, no viés na posição estimada do QTL e na amplitude do intervalo dos SNPs significativos para o mesmo QTL. Os resultados evidenciaram que, quando o LD entre um QTL e um ou alguns marcadores é restrito a SNPs muito próximos do QTL, os GWAS em populações de polinização aberta podem ser altamente eficientes (até 80% de poder de detecção, com reduzido número de associações espúrias), dependendo principalmente do tamanho populacional e da herdabilidade da característica. Para o painel de linhagens, corrigindo para a estrutura populacional, os GWAS alcançaram o maior poder de detecção de QTL (cerca de 96%), associado com o menor número de associações espúrias e viés. Na condição de baixa herdabilidade e tamanho populacional reduzido, os GWAS são ineficazes para as populações de polinização aberta, painel de linhagens e RILs.

ABSTRACT

MUNDIM, Gabriel Borges, D. Sc., Universidade Federal de Viçosa, February, 2016. **Efficiency of genome-wide association study in open-pollinated populations**. Adviser: José Marcelo Soriano Viana. Co-advisers: Fabyano Fonseca e Silva and Rodrigo Oliveira de Lima.

Most papers about genome-wide association studies (GWAS) with plant species published until now have employed inbred lines panel and almost no information on GWAS in open-pollinated populations was found in literature. Therefore, the objectives of this study were (i) to present theoretical aspects, potential and limitations of GWAS in open-pollinated populations; (ii) to analyze the influence of QTL heritability and population sample size on GWAS with open-pollinated populations; and (iii) to compare the efficacy of GWAS on QTL detection in open-pollinated populations, inbred lines panel and recombinant inbred lines (RILs). Fifty samples of populations with linkage disequilibrium (LD) were simulated, considering sample sizes of 400 and 200 individuals, and 10,000 SNPs, 10 QTL and 90 minor genes were randomly distributed in 10 chromosomes, with an average SNP density of 0.1 cM. The phenotypic values simulated refer to three popcorn traits with different degrees of dominance, considering broad sense heritabilities of 0.4 and 0.8. The scenarios were compared based on power of QTL detection, number of false-positive associations, bias in the estimated QTL position and range of the significant SNPs for the same QTL. Results evidenced that when the LD between a QTL and one or few markers is restricted to SNPs very close or within the QTL, the GWAS in open-pollinated populations can be highly efficient (up to 80% power of QTL detection with reduced number of spurious associations), depending mainly on the population sample size and trait heritability. For inbred lines panel, correcting for population structure, the GWAS achieved the highest power of QTL detection (around 96%), associated with the smallest number of spurious associations and bias. Under low heritability and reduced sample size, GWAS are ineffective for open-pollinated populations, inbred lines panel and RILs.

1. Introduction

Association analysis, also known as linkage disequilibrium (LD) mapping or association mapping, is a relatively new population-based approach used to identify marker-trait associations based on LD. Linkage disequilibrium, also known as gametic phase disequilibrium, gametic disequilibrium or allelic association, can be simply stated as the "non-random association of alleles at different loci" or the correlation between polymorphisms that is caused by their shared history of mutation and recombination (Flint-Garcia et al., 2003). Association analysis have been successful in detecting genes associated with diseases in humans (Kerem et al., 1989; Hastbacka et al., 1992; Sladek et al., 2007; Weiss et al., 2009), animals (Barendse et al., 2007; Kijas et al., 2009; Bolormaa et al., 2011; Fan et al., 2011) and different quantitative traits in plants (Thornsberry et al., 2001; Tian et al., 2011; Schaefer & Bernardo, 2013; Suwarno et al., 2015). There are two main association mapping strategies: the candidate gene approach, which focuses on polymorphisms in specific genes controlling traits of interest, and the genome-wide association studies (GWAS), which survey the entire genome for polymorphisms associated with complex traits (Risch & Merkangas, 1996).

With the advent of dense genetic linkage maps, geneticists and breeders has exploited the GWAS to identify genes underlying quantitative trait variation. The conventional quantitative trait loci (QTL) mapping approach based on linkage analysis in a biparental population highly structured with known pedigrees (such as F2 and backcross' populations) have exhibited some limitations, such as the limited number of recombination events resulting in poor resolution (in the range of 10 to 30 cM) for quantitative traits and only two alleles at any given locus can be studied simultaneously. As the GWAS are based on LD among individuals not closely related, all meiotic and recombination events between those individuals contribute to improve mapping

resolution (Stuber et al., 1999; Flint-Garcia et al., 2005). Other advantages of GWAS are the reduction in cost and time to develop a mapping population (Yu & Buckler, 2006) and the possibility of evaluating a large number of alleles in diverse populations (Krill et al., 2010). Although these advantages of LD mapping in comparison with the traditional linkage analysis mapping, a joint linkage and LD mapping strategy was proposed by Wu & Zeng (2001) in order to take the advantages of both methods. This strategy can simultaneously capture the information about linkage between marker and QTL and the LD degree created at a historic time, which implies in a greater reliability of fine QTL mapping and facilitates the development of functional markers to be used in marker-assisted selection and map-based cloning genes (Gupta et al., 2005; Lu et al., 2010; Li et al., 2015).

The first studies of association analysis were performed to dissect human diseases, most notably Alzheimer's disease (Corder et al., 1994) and cystic fibrosis (Kerem et al., 1989). Recently, several studies involving association analysis have also been published with plant species, initially with Arabidopsis (Nordborg et al., 2002; Horton et al., 2014), and further barley (Pasam et al., 2012), sorghum (Morris et al., 2013), rapeseed (Li et al., 2014), wheat (Maccaferri et al., 2015), rice (Yang et al., 2015), sugarcane (Gouy et al., 2015), soybean (Zhang et al., 2015) and maize. Especially with maize, Schaefer & Bernardo (2013) identified major QTL for flowering time (19 QTL), kernel composition (13 QTL), resistance to northern corn leaf blight (13 QTL) and Goss's wilt and blight (9 QTL) through GWAS involving a collection of 284 historical maize inbred lines and 39,166 single nucleotide polymorphism (SNP) markers. Thirunavukkarasu et al. (2014) evaluated 240 elite inbred lines of subtropical maize under water stress and used a set of 29,619 high-quality SNP markers to perform a GWAS which identified 50 SNP markers consistently associated with agronomic traits related to functional traits which could lead

to drought tolerance. Pace et al. (2015) carried out a GWAS with 384 inbred lines evaluated regarding 22 seedling root architecture traits and genotyped with 681,257 SNP markers, which resulted in 268 marker-trait associations identified. Some of these SNP markers were located within or near (<1 kb) to gene models which identify possible candidate genes involved in root development at the seedling stage.

Most papers about GWAS with plant species published until now have employed inbred lines panel and almost no information on GWAS in open-pollinated populations was found in literature. Thus, the objectives of this study were (i) to present theoretical aspects, potential and limitations of GWAS in open-pollinated populations; (ii) to analyze the influence of QTL heritability and population sample size on GWAS with open-pollinated populations; and (iii) to compare the efficacy of GWAS on QTL detection in open-pollinated populations, inbred lines panel and recombinant inbred lines (RILs).

2. Materials and Methods

2.1. Quantitative genetics theory for GWAS in open-pollinated populations

Consider a biallelic QTL (alleles **B**/**b**) and a SNP (alleles **C**/**c**) located in the same chromosome, and a population (generation 0) of an open-pollinated species. Assuming linkage disequilibrium (LD), the joint genotype probabilities in the population are (for simplicity, we omitted the superscript (0) - for generation 0 - in all parameters that depend on the LD measure of generation –1)

$$f_{22} = p_b^2 p_c^2 + 2p_b p_c \Delta_{bc}^{(-1)} + \left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{21} = 2p_b^2 p_c q_c + 2p_b \left(q_c - p_c\right)\Delta_{bc}^{(-1)} - 2\left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{20} = p_b^2 q_c^2 - 2p_b q_c \Delta_{bc}^{(-1)} + \left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{12} = 2p_b q_b p_c^2 + 2\left(q_b - p_b\right)p_c \Delta_{bc}^{(-1)} - 2\left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{11} = f_{11g} + f_{11n} = 4p_b q_b p_c q_c + 2\left(q_b - p_b\right)\left(q_c - p_c\right)\Delta_{bc}^{(-1)} + 4\left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{10} = 2p_b q_b q_c^2 - 2\left(q_b - p_b\right)q_c \Delta_{bc}^{(-1)} - 2\left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{02} = q_b^2 p_c^2 - 2q_b p_c \Delta_{bc}^{(-1)} + \left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{01} = 2q_b^2 p_c q_c - 2q_b \left(q_c - p_c\right)\Delta_{bc}^{(-1)} - 2\left[\Delta_{bc}^{(-1)}\right]^2$$

$$f_{00} = q_b^2 q_c^2 + 2q_b q_c \Delta_{bc}^{(-1)} + \left[\Delta_{bc}^{(-1)}\right]^2$$

where f_{ij} is the probability of the individual with i and j copies of the allele **B** of the QTL and allele **C** of the SNP (i, j = 2, 1, or 0), p is the frequency of the major allele (**B** or **C**), $q = 1 - p$ is the frequency of the minor allele (**b** or **c**), and $\Delta_{bc}^{(-1)} = P_{\mathbf{BC}}^{(-1)} P_{\mathbf{bc}}^{(-1)} - P_{\mathbf{Bc}}^{(-1)} P_{\mathbf{bC}}^{(-1)}$ is the measure of LD in the gametic pool of generation −1 (Kempthorne, 1957). The indices g and n identify the double heterozygotes in coupling and repulsion phases. Notice that $\Delta_{bc}^{(-1)} = r_{bc}^{(-1)} \sqrt{p_b q_b p_c q_c}$, where $r_{bc}^{(-1)}$ is the correlation between the values of the alleles at the two loci (one for **B** and **C**, and zero for **b** and **c**) in the gametic pool of generation −1 (Hill and Robertson, 1968).

The QTL genotypic values are $G_{\mathbf{BB}} = m_b + a_b$, $G_{\mathbf{Bb}} = m_b + d_b$, and $G_{\mathbf{bb}} = m_b - a_b$, where m_b is the mean of the genotypic values of the homozygotes, a_b is the deviation between the genotypic value of the homozygote of higher expression and m_b, and d_b is the dominance deviation (the deviation between the genotypic value of the heterozygote and m_b). The average genotypic values of individuals with the genotypes **CC**, **Cc**, and **cc** are

$$G_{\mathbf{CC}} = \frac{1}{p_c^2}\left(f_{22} G_{\mathbf{BBCC}} + f_{12} G_{\mathbf{BbCC}} + f_{02} G_{\mathbf{bbCC}} \right)$$
$$= M + 2q_c \kappa_{bc} \alpha_b + \left(-2q_c^2 \kappa_{bc}^2 d_b \right) = M + 2\alpha_{\mathbf{C}} + D_{\mathbf{CC}} = M + A_{\mathbf{CC}} + D_{\mathbf{CC}} = m_c + a_c$$

$$G_{\mathbf{Cc}} = \frac{1}{2p_c q_c}\left(f_{21} G_{\mathbf{BBCc}} + f_{11} G_{\mathbf{BbCc}} + f_{01} G_{\mathbf{bbCc}} \right)$$
$$= M + \left(q_c - p_c \right) \kappa_{bc} \alpha_b + 2p_c q_c \kappa_{bc}^2 d_b = M + \left(\alpha_{\mathbf{C}} + \alpha_{\mathbf{c}} \right) + D_{\mathbf{Cc}} = M + A_{\mathbf{Cc}} + D_{\mathbf{Cc}} = m_c + d_c$$

$$G_{\mathbf{cc}} = \frac{1}{q_c^2}\left(f_{20} G_{\mathbf{BBcc}} + f_{10} G_{\mathbf{Bbcc}} + f_{00} G_{\mathbf{bbcc}} \right)$$
$$= M + \left(-2p_c \kappa_{bc} \alpha_b \right) + \left(-2p_c^2 \kappa_{bc}^2 d_b \right) = M + 2\alpha_{\mathbf{c}} + D_{\mathbf{cc}} = M + A_{\mathbf{cc}} + D_{\mathbf{cc}} = m_c - a_c$$

where $M = m_b + (p_b - q_b)a_b + 2p_b q_b d_b$ is the population mean, $\kappa_{bc} = \left[\frac{\Delta_{bc}^{(-1)}}{p_c q_c}\right]$, $\alpha_b = a_b + (q_b - p_b)d_b$ is the average effect of a gene substitution, $\alpha_{\mathbf{C}} = q_c \kappa_{bc} \alpha_b$ and $\alpha_{\mathbf{c}} = -p_c \kappa_{bc} \alpha_b$ are the average effects of the SNP alleles, and A and D are the SNP additive and dominance values. The average effect of substituting the allele **C** for **c** is $\alpha_{SNP} = \alpha_{\mathbf{C}} - \alpha_{\mathbf{c}} = \kappa_{bc} \alpha_b$. The dominance deviation for the SNP is $d_{SNP} = \kappa_{bc}^2 d_b$. The other SNP parameters are $m_c = M + (q_c - p_c)\alpha_{SNP} - (1 - 2p_c q_c)d_{SNP}$, $a_c = \alpha_{SNP} - (q_c - p_c)d_{SNP}$, and $d_c = d_{SNP}$.

Notice that assuming no QTL in LD with the SNP, $G_{\mathbf{CC}} = G_{\mathbf{Cc}} = G_{\mathbf{cc}} = M$. Thus, the identification of the QTL can be based on the test of the hypothesis that there is no difference between these genotypic means. Assuming thousands of SNPs, it is necessary to employ a Bonferroni-type procedure to control the type I error when there are multiple-comparisons, as that proposed by Benjamini and Hochberg (1995).

Alternatively, the QTL identification can be made by testing that there is no relationship between the genotypic values for the individuals **CC**, **Cc**, and **cc** with the number of copies of one SNP allele. The parameters of the additive-dominance model can be derived by fitting the model $G = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where x_1 = 1, 0, or -1 if the individual is **CC**, **Cc,** or **cc**, and x_2 = 0 or 1 if the individual is homozygous or heterozygous, respectively, and G is the QTL genotypic value. The model can be expressed as $y_{(9 \times 1)} = X_{(9 \times 3)}.\beta_{(3 \times 1)} + \varepsilon_{(9 \times 1)}$, where y is the vector of QTL genotypic values, conditional on the SNP genotype, X is the incidence matrix, β is the parameter vector and ε is the error vector. The matrix of genotype probabilities is $P_{(9 \times 9)} = \text{diagonal}\{f_{ij}\}$. Thus, for the complete model or a reduced model, $\beta = (X'PX)^{-1}(X'Py)$. Fitting the

complete model, $\beta_0 = m_c$, $\beta_1 = a_c$, and $\beta_2 = d_{SNP}$. Assuming no QTL in LD with the SNP, $\beta_1 = \beta_2 = 0$ and $\beta_0 = M$. Fitting the additive model, $G = \beta_0 + \beta_1 x + \varepsilon$ or $G = \beta_0 + \beta_1 x_1 + \varepsilon$ (no dominance), $\beta_1 = \alpha_{SNP}$.

The alternative additive-dominance model can be fitted based on the orthogonal contrasts derived by Cockerham (1954) and expressed as $G = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ (x = 2, 1, or 0). In this case, the parameters for the complete model are

$$\beta_0 = M - 2p_c\alpha_{SNP} - 2p_c^2 d_{SNP}$$

$$\beta_1 = \alpha_{SNP} + (1 + 2p_c)d_{SNP}$$

$$\beta_2 = -d_{SNP}$$

If there are two QTL (alleles **B**/**b** and **E**/**e**) in LD with the SNP (alleles **C**/c), it can be demonstrated that

$$\alpha_{SNP} = \kappa_{bc}\alpha_b + \kappa_{ce}\alpha_e$$

$$d_{SNP} = \kappa_{bc}^2 d_b + \kappa_{ce}^2 d_e$$

where $\kappa_{ce} = \left[\frac{\Delta_{ce}^{(-1)}}{p_c q_c}\right]$.

If there is population structure, this must be corrected in the GWAS to avoid spurious associations due to admixture LD. Consider, for simplicity, two subpopulations in Hardy-Weinberg equilibrium and one SNP (alleles **C**/**c**) and a QTL (alleles **B**/**b**) in linkage equilibrium in both subpopulations. Assuming that p and q are the allelic frequencies in one subpopulation and r and s are the allelic frequencies in the other subpopulation, the average genotypic value of individuals **CC**, **Cc**, and **cc** are

$$G_{\mathbf{CC}} = m_b + \left(\frac{1}{u_1 p_c^2 + u_2 r_c^2} \right) \left\{ \left[u_1 (p_b - q_b) p_c^2 + u_2 (r_b - s_b) r_c^2 \right] a_b + \left(u_1 2 p_b q_b p_c^2 + u_2 2 r_b s_b r_c^2 \right) d_b \right\}$$

$$G_{\mathbf{Cc}} = m_b + \left(\frac{1}{u_1 2 p_c q_c + u_2 2 r_c s_c} \right) \left\{ \left[u_1 (p_b - q_b) 2 p_c q_c + u_2 (r_b - s_b) 2 r_c s_c \right] a_b + \left(u_1 2 p_b q_b 2 p_c q_c + u_2 2 r_b s_b 2 r_c s_c \right) d_b \right\}$$

$$G_{\mathbf{cc}} = m_b + \left(\frac{1}{u_1 q_c^2 + u_2 s_c^2} \right) \left\{ \left[u_1 (p_b - q_b) q_c^2 + u_2 (r_b - s_b) s_c^2 \right] a_b + \left(u_1 2 p_b q_b q_c^2 + u_2 2 r_b s_b s_c^2 \right) d_b \right\}$$

where u_1 and u_2 are the proportions of individuals from subpopulations 1 and 2 (probabilities of an individual belongs to subpopulations 1 and 2). Only if there is not population structure (u_1= 1 or 0), $G_{\mathbf{CC}} = G_{\mathbf{Cc}} = G_{\mathbf{cc}} = M$ (and $\beta_1 = \beta_2 = 0$ and $\beta_0 = M$).

2.2. Quantitative genetics theory for GWAS with inbred lines panel

In general, the inbred lines in a panel represent the genetic variability for the traits under assessment. Therefore, an inbred lines panel includes inbreds from distinct populations or heterotic groups. Assume again a QTL (alleles **B**/**b**) and a SNP (alleles **C**/**c**) located in the same chromosome, and that they are in LD in a population (generation 0). Assuming n ($n \to \infty$) generations of selfing, the probabilities of the inbreds are (for simplicity, we omitted again the superscript (0) - for generation 0 - in all parameters that depend on the LD measure of generation −1)

$$\lim f_{22}^{(n)} = f_{22} + \frac{1}{2}\left(f_{21} + f_{12} \right) + \frac{1}{4} f_{11} + \frac{1}{2}\left(\frac{1-2r_{bc}}{1+2r_{bc}} \right)\Delta_{bc}^{(-1)}$$

$$\lim f_{20}^{(n)} = f_{20} + \frac{1}{2}\left(f_{21} + f_{10} \right) + \frac{1}{4} f_{11} - \frac{1}{2}\left(\frac{1-2r_{bc}}{1+2r_{bc}} \right)\Delta_{bc}^{(-1)}$$

$$\lim f_{02}^{(n)} = f_{02} + \frac{1}{2}\left(f_{01} + f_{12} \right) + \frac{1}{4} f_{11} - \frac{1}{2}\left(\frac{1-2r_{bc}}{1+2r_{bc}} \right)\Delta_{bc}^{(-1)}$$

$$\lim f_{00}^{(n)} = f_{00} + \frac{1}{2}\left(f_{01} + f_{10} \right) + \frac{1}{4} f_{11} + \frac{1}{2}\left(\frac{1-2r_{bc}}{1+2r_{bc}} \right)\Delta_{bc}^{(-1)}$$

where r_{bc} is the frequency of recombinant gametes. The haplotypes are $P_{\mathbf{BC}}^{(n)} = p_b p_c + \Delta_{bc}^{(n)}$, $P_{\mathbf{Bc}}^{(n)} = p_b q_c - \Delta_{bc}^{(n)}$, $P_{\mathbf{bC}}^{(n)} = q_b p_c - \Delta_{bc}^{(n)}$, and $P_{\mathbf{bc}}^{(n)} = q_b q_c + \Delta_{bc}^{(n)}$, where $\Delta_{bc}^{(n)} = \left(\frac{1}{1+2r_{bc}} \right)\Delta_{bc}^{(-1)}$. Thus, if there is crossing-over ($0 < r_{bc} \leq 0.5$), the LD in this inbred population is lower than the LD in the generation –1. If the SNP and QTL are completely linked ($r_{bc} = 0$), the LD in the inbred population is the same LD in the generation –1.

For the inbred lines derived from a population, we have

$$G_{\mathbf{CC}}^{(n)} = \frac{1}{f_{.2}^{(n)}}\left[f_{22}^{(n)}\left(m_b + a_b \right) + f_{02}^{(n)}\left(m_b - a_b \right) \right] = M_{IL} + 2q_c\alpha_{SNP}^{(n)} = M_{IL} + A_{\mathbf{CC}}^{(n)}$$

$$G_{\mathbf{cc}}^{(n)} = \frac{1}{f_{.0}^{(n)}}\left[f_{20}^{(n)}\left(m_b + a_b \right) + f_{00}^{(n)}\left(m_b - a_b \right) \right] = M_{IL} - 2p_c\alpha_{SNP}^{(n)} = M_{IL} + A_{\mathbf{cc}}^{(n)}$$

where $M_{IL} = m_b + \left(p_b - q_b \right)a_b$ is the inbred population mean, $\alpha_{SNP}^{(n)} = \left(\frac{1}{2+4r_{bc}} \right)\kappa_{bc}a_b$ is the SNP average effect of allele substitution in the inbred

population, and A is the SNP additive value for an inbred line. Assuming no QTL in LD with the SNP, $G^{(n)}_{\mathbf{CC}} = G^{(n)}_{\mathbf{cc}} = M_{IL}$.

The haplotypes of an inbred lines panel including inbreds from N populations are

$P^{(n)'}_{\mathbf{BC}} = \bar{p}_b\bar{p}_c + \Delta^{(n)'}_{bc}$, $P^{(n)'}_{\mathbf{Bc}} = \bar{p}_b\bar{q}_c - \Delta^{(n)'}_{bc}$, $P^{(n)'}_{\mathbf{bC}} = \bar{q}_b\bar{p}_c - \Delta^{(n)'}_{bc}$, and $P^{(n)'}_{\mathbf{bc}} = \bar{q}_b\bar{q}_c + \Delta^{(n)'}_{bc}$, where

$$\Delta^{(n)'}_{bc} = \sum_{i=1}^{N} u_i\left[\Delta^{(n)}_{bc_i} + p_{b_i}p_{c_i}\right] - \left(\sum_{i=1}^{N} u_i p_{b_i}\right)\left(\sum_{i=1}^{N} u_i p_{c_i}\right) = \bar{\Delta}^{(n)}_{bc} + \overline{p_b p_c} - \bar{p}_b\bar{p}_c$$

and u_i is the probability of an inbred line belongs to population i. Because this function is too complex to interpret, the analysis of the LD value in an inbred lines panel, relative to the LD in the inbreds from each population, will be presented further, using the simulated data.

Due to population structure, spurious associations involving SNP and QTL in linkage equilibrium in the non-inbred populations can be declared. Assume, for simplicity, an inbred lines panel with inbreds from two populations where a SNP (alleles **C**/**c**) and a QTL (alleles **B**/**b**) are in linkage equilibrium. Let u_1 and u_2 be the proportions of inbreds from these populations. Assuming that p and q are the allelic frequencies in one population, that r and s are the allelic frequencies in the other population, and that p $\neq$ q or r $\neq$ s,

$$G^{(n)}_{\mathbf{CC}} = m_b + \left(\frac{1}{u_1p_c + u_2r_c}\right)\left[u_1\left(p_b - q_b\right)p_c + u_2\left(r_b - s_b\right)r_c\right]a_b$$

$$G^{(n)}_{\mathbf{cc}} = m_b + \left(\frac{1}{u_1q_c + u_2s_c}\right)\left[u_1\left(p_b - q_b\right)q_c + u_2\left(r_b - s_b\right)s_c\right]a_b$$

If there is no population structure (u_1 = 1 or 0), $G^{(n)}_{\mathbf{CC}} = G^{(n)}_{\mathbf{cc}} = M_{IL}$.

2.3. Simulation

We simulated 50 samples of populations with linkage disequilibrium using the software REALbreeding (Azevedo et al., 2015; Viana et al., 2013). This software has been developed using the program REALbasic 2009. Population 1, generation 10r, is the advanced generation of a composite of two populations in linkage equilibrium (population 1, generation 0), obtained after 10 generations of random crosses, assuming sample size of 400 individuals. Population 1, generations 10s and 10r10s, were obtained from Population 1, generation 0, assuming 10 generations of selfing and 10 generations of random crosses followed by 10 generations of selfing, respectively, also assuming sample size 400. Populations 2, 3, and 4, generations 10s, are also inbred populations (10 generations of selfing) derived from composites of two populations, also assuming sample size 400. The parents of populations 2 and 3 were assumed be non-improved and improved populations, respectively. An improved population was defined as having frequencies of favorable genes greater than 0.5, while a non-improved population was defined as having frequencies lower than 0.5. A composite is a Hardy-Weinberg equilibrium population with LD only for linked markers and genes. In the case of a composite of two populations in linkage equilibrium, $\Delta_{bc}^{(-1)} = \left(\frac{1-2r_{bc}}{4}\right)\left(p_b^1 - p_b^2\right)\left(p_c^1 - p_c^2\right)$, where the indices 1 and 2 refer to the parental populations.

Based on our input, the software REALbreeding randomly distributed 10,000 SNPs, 10 QTL (of higher effect) and 90 minor genes (QTL of lower effect) in 10 chromosomes (1,000 SNPs and 10 genes by chromosome). The average SNP density was 0.1 cM. The genes were distributed in the regions covered by the SNPs. Four, three, two, and one QTL

were inserted in chromosomes 1, 5, 9, and 10, respectively. We also specified one SNP within each QTL and a minimum distance between linked QTL of 10 cM. To allow REALbreeding computing the phenotypic value for each genotyped individual, we informed minimum and maximum genotypic values for homozygotes, proportion between the parameter a for a QTL and the parameter a for a minor gene (a_{QTL}/a_{mg}), degree of dominance ($(d/a)_i$, i = 1, ..., 100), direction of dominance, and the broad sense heritability. The REALbreeding program saves two main files, one with the marker genotypes and the other with the additive, dominance, and phenotypic values (non-inbred populations) or the genotypic and phenotypic values (inbred populations). The true additive and dominance genetic values or genotypic values are computed from the population gene frequencies (random values), LD values, average effects of gene substitution or a deviations, and dominance deviations. The phenotypic values are computed from the true population mean, additive and dominance values or genotypic values, and from error effects sampled from a normal distribution. The error variance is computed from the broad sense heritability.

We simulated three popcorn traits. The minimum and maximum genotypic values of homozygotes for grain yield, expansion volume, and days to maturity were 30 and 180 g/plant, 15 and 65 mL/g, and 100 and 170 days, respectively. We defined positive dominance for grain yield ($0 < (d/a)_i \leq 1.2$), bidirectional dominance for expansion volume ($-1.2 \leq (d/a)_i \leq 1.2$), and no dominance for days to maturity ($(d/a)_i = 0$). The broad sense heritabilities were 0.4 and 0.8. These values can be associated with individual and progeny assessment, respectively. Assuming $a_{QTL}/a_{mg} = 10$, each QTL explained approximately 4 and 8% of the phenotypic variance for heritabilities of 0.4 and 0.8. The GWAS was performed in population 1, generations 10r and 10r10s, and in the inbred lines panel obtained from inbreds of the populations 1 to 4, generation 0. To allow the

assessment of the influence of the sample size on the GWAS efficacy, we considered sample sizes of 400 and 200. Thus, we used 100 or 50 inbreds from populations 1 to 4 to generate the inbred lines panel. To assess the influence of the QTL heritability on the GWAS efficacy, we converted four QTL (QTL 3, 7, 8, and 10 on chromosomes 1, 5, 9, and 10, respectively) to minor genes and assumed QTL heritability of 12% (for trait heritability of 0.7). Then, the GWAS was performed in population 1, generation 10r.

2.4. Statistical analyses

For the population structure analysis, we used the Structure software (Falush et al., 2003) and compared the admixture model with correlated allelic frequencies to the no admixture model with independent allelic frequencies (Pritchard et al., 2000). The number of SNPs, sample size, burn-in period, and number of MCMC (Monte Carlo Markov chain) iterations were 100 (10 random SNPs by chromosome), 400 (simulation 1), 10,000, and 40,000, respectively. The number of populations assumed (K) ranged from 1 to 7 and the most probable K value was determined based on the inferred plateau method (Viana et al., 2013). The following quadratic-plateau (QP) model was used:

$$y_i = \begin{cases} a + bK_i + cK_i^2 + e_i, & \text{if } K_i < K_0 \\ p + e_{i,} & \text{if } K_i \geq K_0 \end{cases} \quad \text{(Fuller and Gallant, 1974),}$$

where y_i is the $\log \Pr(X|K_i)$ provided by $K_i = 1, 2, \ldots, 7$; a, b and c are the parameters of the quadratic function; K_0 is the true K value; p is the plateau (stabilized value of $\log \Pr(X|K_i)$ after K_0 and e_i is the residual term, assumed as $e_i \sim N(0, \sigma_e^2)$. The QP model is nonlinear when K_0 is treated as unknown. Thus, the NLIN procedure of SAS software (SAS Institute, 2007) was used to fit this model using an iterative least squares procedure based on Gauss-Newton algorithm.

The analyses of LD and association were performed with the software PowerMarker (Liu and Muse, 2005) and Tassel (Bradbury et al., 2007) for the inbred lines panel and RILs. For open-pollinated populations, we used a single-locus F-test with the software PowerMarker, where each marker is regarded as a factor in a one-way ANOVA, according to the model $y = X\beta + \varepsilon$, where y is the vector of phenotypic values, X is the incidence matrix of the SNP genotypes, β is the vector of SNP additive and dominance effects and ε is the residual vector. The F-test is then performed and reports a raw p-value for each marker (Liu and Muse, 2005). It is important to know that these p-values are not adjusted for multiplicity. As there is no relationship between the inbred lines, the GWAS with the inbred lines panel was based on the general linear model (GLM) using the software Tassel and including the Q-matrix of population membership estimates as a covariate to correct for population structure (Bradbury et al., 2007). In both cases, we used a Benjamini-Hochberg false discovery rate (FDR) of 5 and 1% (Benjamini and Hochberg, 1995) to control the type I error.

To classify each significant association as true or false, we used a program developed in REALbasic 2009. The classification criterion was based on the difference between the position of the SNP and the position of a true QTL (candidate gene). If the difference was less than or equal to 2.5 cM (Yu et al., 2008), the association was classified as true. The scenarios were compared based on power of QTL detection (probability of reject H_0 when H_0 is false; control of type II error), calculated as the ratio between the number of true SNP correctly identified and the total number of QTL simulated and then averaged over the 50 simulations; number of false-positives (control of type I error); bias in the estimated QTL position (precision of mapping), calculated as the difference

between the position of the SNP identified and the position of the true QTL simulated (candidate gene); and range of the significant SNPs for the same QTL.

3. Results

The results for assessing the efficacy of GWAS in open-pollinated populations refer to population 1, generation 10r. In generation 0, the degree of linkage disequilibrium is so high that several significant associations are observed along the length of a chromosome with one or more QTL or in one or more large chromosome regions (Figure 1). As will be discussed, these several significant associations are not false-positives (at least most of them). For sure, this is due to the degree of LD and presence of QTL. Even assuming a FDR of 1%, it is worthless for the identification of candidate genes to infer that there are one or more QTL in a chromosome region spanning 20 cM. When the linkage disequilibrium between a QTL and one or few markers is restricted to SNPs very close or within the QTL, the analysis can be highly efficient, depending mainly on the QTL effect, sample size, and trait heritability. Assuming heritability of 0.8 and sample size 400 (simulation 1), the significant associations for expansion volume observed in chromosome 1 evidenced five QTL with a FDR of 5% or four QTL with a FDR of 1% (Figure 1). This implies in a power of QTL detection of 100%. Three of the four true QTL (candidate genes) were identified by SNPs located within the QTL and one by five or four SNPs in a region spanning approximately 2.0 or 1.7 cM, depending on the FDR. The significant associations at a FDR of 5 or 1% for SNPs 223 (at position 21.7 cM), 243 (at position 23.3), 245 (at position 23.4 cM), and 252 (at position 23.7 cM) are mainly attributable to their linkage disequilibrium with QTL 2. The absolute LD values are 0.1488, 0.1494, 0.1747, and 0.1416, respectively (P values highly significant by the chi-square test). The significant association at a FDR of 5% for SNP 627 (at position 61.8 cM) is not a false-positive association, since it is in LD with QTL 2 ($|\Delta| = 0.0366$, P value of the chi-square test = 3.22E-6) and QTL 3 ($|\Delta| = 0.0302$, P value of the chi-square test = 7.55E-5). Then, the result is interpreted as a fifth QTL.

Only for high heritability and greater sample size the results from GWAS were clearly different between days to maturity and the other two traits, except for the power of QTL detection (Table 1). The number of significant associations, the number of false-positives, the bias in QTL position, and the average range of chromosome regions with one or more QTL were greater in the absence of dominance. With a FDR of 5%, the power of detection ranged from 88 to 93%, but associated with a high number of false-positive associations. Further, on average, each true QTL was identified based on two to three (for days to maturity) SNPs, in chromosomes regions spanning 0.8 to 1.2 cM. The bias in QTL position ranged from 0.5 to 0.6 cM. Increasing the control of the type I error provided better results, greatly reducing the number of false-positive associations. The power of QTL detection ranged from 75 to 80% and each QTL was identified based on one to two SNPs, in chromosome regions spanning 0.4 to 0.6 cM. The bias in QTL position ranged from 0.3 to 0.4 cM.

Assuming QTL of lower effect, heritability of 0.8 and sample size 200 or heritability of 0.4 and sample size 400, it is better to assume a FDR of 5% ensuring greater power of QTL detection and lower number of false-positive associations. However, especially due to lower QTL effect, the power of detection ranged from 33 to 39% (Table 1). Under lower QTL effect, low heritability and reduced sample size, GWAS are ineffective, showing an average power of QTL detection less than or equal to 5%. This scenario does not improve increasing the FDR to 10% (data not shown).

Increasing the QTL heritability from 4 and 8% to 12% determined an increase in the power of QTL detection, specially assuming sample size of 200 individuals (Table 2). The bias in the QTL position, the range of the regions with an identified QTL, the number of false-positives, and the number of significant associations in chromosomes with one

to two QTL also increased, mainly with greater sample size. Notice that assuming 200 individuals, the power of QTL detection reached 70-75%, regardless of the trait.

We also provided results for comparing GWAS in open-pollinated population and in inbred lines panel. An impressive result from GWAS with inbred lines panel is the efficacy of discarding spurious associations due to population structure (Figure 2). From the analysis of expansion volume, assuming a FDR of 1%, heritability of 0.8, and sample size 400 (simulation 1), the number of spurious associations in chromosome 3 (no QTL) were reduced from 477 to zero. Further, correcting for population structure decreased the number of significant associations in chromosome 1 (four QTL) from 464 to 9. This implies in a power of QTL detection of 100%, but three to five false-positive associations. The population structure analysis evidenced four subpopulations (Figure 3). In general, the efficacy of GWAS was greater with inbred lines panel (Table 3). The power of QTL detection was higher and the number of false-positive associations was lower. Further, in general only SNPs within the QTL showed significant associations. Notice also that no differences were observed between the traits and, similarly for open-pollinated populations, the analysis is ineffective assuming lower QTL effect and reduced heritability and sample size.

The analysis of the parametric linkage disequilibrium in the populations and in the inbred lines panel, based on a random 10 cM segment of chromosome 1 (100 SNPs), evidenced: higher LD in population 1 (average of the absolute values equal to 0.0403; 627 values greater than 0.1; generation 0) and lower LD in population 3 (average of the absolute values equal to 0.0203; 48 values greater than 0.1; generation 0), slight decrease in the LD with selfing (5-6%), and lowest LD in the inbred lines panel (average of the absolute values equal to 0.0249; 8 values greater than 0.1) (Figures 4 and 5). The LD decay due to recombination was approximately 25%, regardless of the population. In

population 1, for example, the number of absolute LD values greater than 0.1 decreased 60%.

The GWAS with RILs (recombinant inbred lines) from population 1, generation 10r (lower parametric LD), evidenced a high number of significant associations at a FDR of 1% along the length of one or more chromosomes with one to four QTL, under high heritability and greater sample size (Table 4). As explained, this makes the GWAS ineffective, reducing the power of QTL detection. Assuming heritability of 0.4 and 400 individuals and heritability of 0.8 and 200 individuals, the results are similar to those observed for population 1, generation 10r, but with a FDR of 1% and a greater number of significant associations in chromosomes with one to four QTL. Compared to GWAS in generation 10r, the lower efficacy of GWAS with RILs can be attributable to higher heritability, due to increase in the genotypic variance for the same error variance, and higher LD. Based on simulation 1, the estimated heritability with RILs was approximately 0.9 for the three traits, assuming heritability of 0.8 and 400 individuals assessed in generation 10r10s (12.5% greater than the heritability at generation 10r). Due to sampling, the estimated LD was greater with RILs than with non-inbred plants in generation 10r (Figure 6). Based on simulation 1, the average estimated Δ and r^2 values were, respectively, 0.0252 and 0.0241 for the RILs and 0.0235 and 0.0225 for generation 10r. Although these average values are equivalent, the Δ values with RILs were on average four times greater than the Δ values in generation 10r. Once again, the GWAS were ineffective assuming low heritability and reduced sample size.

4. Discussion

4.1. GWAS in open-pollinated populations: theoretical aspects, potential and limitations

One of the main contributions of this research is to show the quantitative genetics theory for GWAS in open-pollinated populations. The theoretical aspects of quantitative genetics showed that considering there is no QTL in LD with one given SNP allele, the QTL identification can be based on a F-test where the null hypothesis is there is no difference between the SNP genotypic means of the individuals with different SNP genotypes from the population, or by a regression analysis to check if there is no relationship between the genotypic values of the individuals from the population and the number of copies of one given SNP allele. In both cases, it is clear that the QTL identification depends on the presence and the degree of LD between the QTL and the SNP locus.

When analyzing 10,000 SNPs across the simulated genome, the biggest challenge for QTL identification is to achieve a great detection power of true significant associations with a reduced number of false-positive associations. The degree of LD observed in population 1 – generation 0 was so high that several significant associations were identified in this scenario. Most of these associations are true, so that they are in a range of 5.0 cM of the QTL present in this chromosome region and were identified due to its linkage disequilibrium with these QTL. However, there still exist false-positive associations, even increasing the control of the type I error, which difficult the search for candidate genes based on significant SNP loci declared within an extent of around 20 cM (cases of QTL 3 and 4, for example). According to Larsson et al. (2013), false-positive associations can arise from other sources, which despite being rare, are typically unaccounted for in association studies, such as markers that are in long-range LD with

causative polymorphisms. Additionally, causative polymorphisms for one trait may not necessarily be causal for another highly correlated trait (and, hence a spurious association), but will be statistically associated with both traits. Both of these types of false-positive associations do not occur randomly across the genome and thus, they are very challenging to eliminate.

The LD decay due to recombination after ten generations of random crosses was beneficial to QTL identification in population 1 – generation 10r. With around 25% decrease in LD, the GWAS was highly efficient, so the LD becomes restricted to the true QTL and one or few SNP loci very close or within the QTL. This situation implied in a good average power of QTL detection with lower false-positive associations and average bias in the QTL position than observed in population 1 – generation 0, disregarding the influence of population sample size, trait heritability and degree of dominance, and level of control of the type I error. These results are significantly comparable to those obtained by Yu et al. (2008) in a simulation study which investigated the genetic and statistical properties (average power of QTL detection, FDR and R^2) of the nested association mapping (NAM) design currently being implemented in maize to dissect the genetic basis of complex quantitative traits. With 5,000 genotypes, these authors achieved 57% of average power of QTL detection (ranging from 30 to 85%), considering two trait heritabilities (0.4 and 0.7), two different numbers of QTL controlling the trait (20 and 50) and two different genotyping schemes (complete marker information and common-parent-specific markers only).

4.2. GWAS in open-pollinated populations: influence of QTL heritability and sample size

The power of QTL detection in GWAS is determined by several factors including the population sample size as well as the genetic architecture and heritability of the trait

under evaluation (Yu et al. 2008). Importantly for genetic mapping applications, the heritability corresponds to the amount of phenotypic variation that can be attributed to genetic effects and thus to the cumulative effects of QTL (Buckler et al., 2009; Kump et al., 2011). High estimates of heritability depends on the quality of phenotypic data assessment, which results in genotypic values estimated with high accuracy and, consequently, should facilitate QTL detection with substantial power (Liu et al., 2011). The influence of QTL heritability on GWAS in open-pollinated populations was evidenced by the increase in power of QTL detection, associated with a little increase in the number of spurious associations and in the bias in QTL position. As in previous studies, a higher heritability always gave higher QTL detection power, particularly for QTL with moderate to small effect (Yu et al. 2008). Hung et al. (2012) assessed 19 quantitative traits in maize and achieved heritabilities greater than 0.8 for traits related to flowering time and plant architecture, resulting in a good power to detect QTL for these traits. In contrast, traits which had lower heritabilities (up to 0.6) and were more strongly affected by environmental variation allow only a reasonable power of QTL detection. Similar results were obtained by Kump et al. (2011), which evaluated resistance to southern leaf blight (SLB) disease in maize and obtained a heritability of 87%, indicating the potential for accurate mapping of SLB-resistance genes. These authors identified 32 QTL with predominantly small and additive effects on SLB resistance and many of the SNP within and outside of QTL intervals are also within or near to genes previously shown to be involved in plant disease resistance in other studies (Poland et al., 2009; Broglie et al., 2009).

The success of GWAS depends a lot on population sample size also (Flint-Garcia et al., 2005) and results have demonstrated that GWAS are best carried out with a large sample size (Yu & Buckler, 2006). According to Flint-Garcia et al. (2005), increasing the

population size necessarily increases the number of individuals with rare alleles, thus improving the power to test the association between these rare alleles and the trait of interest. The influence of population sample size on GWAS in our study was demonstrated with the increase in power of QTL detection, increase in the number of spurious associations (mainly in chromosomes with one to four QTL), and in the bias in QTL position, disregarding the trait, heritability and FDR. The increase in number of false-positive associations due to increase in population sample size was much more pronounced with high heritabilities (0.08 and 0.12) of each QTL. Yu et al. (2008) showed that the gain in accuracy by increasing sample size was evidenced by increased power of QTL detection and smaller FDR, mainly with heritability of 0.7 in comparison with a heritability of 0.4. Long & Langley (1999) performed a simulation study which demonstrated that sufficient power to detect marker-trait associations for QTL that account for as little as 5% of the phenotypic variation occurs when approximately 500 individuals are genotyped for approximately 20 SNP loci within the candidate gene region. These authors declared that more power is achieved by increasing the population size than by increasing the SNP density within the candidate gene.

It is clearly remarkable that the efficiency of GWAS in open-pollinated populations is determined by the equilibrium of QTL heritability and sample size together. If the heritability per locus is proportional to the number of loci affecting the trait, consequently, the sample size needed to detect association of a marker to each locus will be proportional to the number of loci affecting the trait also (Weir, 2010). Increasing both the QTL heritability and population sample size implies in a greater power of QTL detection, however, increased the number of false-positive associations and bias in the QTL position also, given a same FDR. The increase in power of QTL detection associated with less increase in spurious association and bias in the QTL position can be achieved by

increasing the heritability of each QTL (from 0.04 to 0.12) for low population sample size (200 individuals) or by increasing the control of type I error (from 5 to 1%) with high QTL heritabilities (0.08 and 0.12) and population sample size (400 individuals).

4.3. GWAS in open-pollinated populations, inbred lines panel and RILs

Most papers about GWAS with maize published until now have employed inbred lines panel (Samayoa et al., 2015; Van Inghelandt et al., 2012; Yang et al., 2010) or nested association mapping (NAM) populations (Bian et al., 2014; Tian et al., 2011; Kump et al., 2011) and almost no information on GWAS in open-pollinated populations was found in literature. According to Flint-Garcia et al. (2005), the inbred lines panel exploits the rapid breakdown of LD in diverse maize lines, enabling very high resolution for QTL mapping via association analysis. One of the most important maize inbred panel was assembled by Flint-Garcia et al. (2005) for association mapping and consists of 302 inbred lines from temperate and tropical regions, both current breeding lines and historically important lines, which represents a large fraction of the global genetic diversity in maize breeding (Yang et al., 2010). This population has been successfully used by the maize community to perform GWAS in economically important quantitative traits such as kernel composition (Cook et al., 2012) and *Fusarium* ear rot resistance (Zila et al., 2013).

As a result of constructing an inbred lines panel using lines from various breeding programs, distinct origins, heterotic groups and genetic arrangement is the presence of confounding structure in these panels, which can cause false-positive marker-trait associations if the data was not corrected for population structure (Yan et al. 2009). Analyzing our results from GWAS with inbred lines panel, it is impressive the number of spurious association identified when ignoring population structure and the efficacy of discarding these false-positive associations when correcting for population structure. The

efficiency of GWAS with inbred lines panel was significant, since the power of QTL detection was much higher than with open-pollinated population and RILs, associated with a lower number of false-positive associations (close to zero) and bias in the QTL position, disregarding the trait, heritability, population sample size and FDR. The lowest parametric LD values for the inbred lines panel are comparable to other studies already published (Yan et al., 2009; Remington et al., 2001). Moreover, with the inbred lines panel, in general only SNP loci within the QTL showed significant association, which is a highlighted result from GWAS that can serve as basis for a fine mapping strategy to be used in marker-assisted selection and map-based cloning genes (Gupta et al., 2005).

The most important example of a NAM population synthesized for GWAS is the set of 5,000 RILs derived from crosses between the reference maize inbred line B73 and 25 other founder inbreds. This maize NAM panel captures a substantial proportion of the global genetic diversity of maize inbred lines and the high allele diversity and large sample size provide a great power of QTL detection (McMullen et al., 2009; Kump et al., 2011). According to Yu et al. (2008), NAM populations have the advantages of lower sensitivity to genetic heterogeneity and higher power of QTL detection, as well as higher efficiency in using the genome sequence or dense markers while still maintaining high allele richness due to diverse founders. By choosing diverse founders, LD within chromosome segments resulting from historical/evolutionary recombination was mostly preserved in RILs due to the small probability of recombination within the short genetic distances between flanking common-parent-specific markers, which leads to a great power of QTL detection.

When comparing these statements with our results from GWAS in open-pollinated populations versus RILs, it is possible to confirm the increase in QTL detection power when using RILs in the scenario of high QTL heritability, high population sample size

and high control of type I error. However, the number of spurious association increased proportionally much more, disregarding the trait, making the GWAS with RILs (generation 10r10s) less efficient than with non-inbred population (generation 10r). This excess of false-positive associations can be attributed to the higher LD estimates with RILs in comparison with non-inbred plants of an open-pollinated population.

5. Conclusion

In short, the genome-wide association studies (GWAS) with open-pollinated populations demonstrated its higher potential in terms of power of QTL detection, associated with lower number of false-positive associations and bias in the estimated QTL position than the GWAS with recombinant inbred lines (RILs). The GWAS with inbred lines panel were more efficient than with open-pollinated populations and RILs, achieving the highest power of QTL detection, associated with the smallest number of spurious associations and bias. The GWAS with non-inbred populations were mainly affected by the population sample size and trait heritability. The degree of dominance almost no affected the GWAS. It is clearly remarkable that the efficiency of GWAS in open-pollinated populations is determined by the equilibrium of QTL heritability and sample size together. Under low heritability and reduced sample size, GWAS are ineffective for open-pollinated populations, inbred lines panel and RILs.

6. Acknowledgments

We thank the National Council for Scientific and Technological Development (CNPq), the Brazilian Federal Agency for Support and Evaluation of Graduate Education (Capes), and the Foundation for Research Support of Minas Gerais State (Fapemig) for financial support.

7. References

Azevedo, C. F.; Resende, M. D. V.; Silva, F. F.; Viana, J. M. S.; Valente, M. S. F.; Resende Jr., M. F. B.; Muñoz, P. (2015) Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genetics, 16:105-118. doi 10.1186/s12863-015-0264-2

Barendse, W.; Reverter, A.; Bunch, R. J.; Harrison, B. E.; Barris, W.; Thomas, M. B. (2007) A validated whole-genome association study of efficient food conversion in cattle. Genetics, 176:1893-905.

Benjamini, Y.; Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57:289-300.

Bian, Y.; Yang, Q.; Balint-Kurti, P. J.; Wisser, R. J.; Holland, J. B. (2014) Limits on the reproducibility of marker associations with southern leaf blight resistance in the maize nested association mapping population. BMC Genomics, 15:1068-1082.

Bolormaa, S.; Hayes, B.; Savin, K.; Hawken, R.; Barendse, W.; Arthur, P.; et al. (2011) Genome-wide association studies for feedlot and growth traits in cattle. Journal of Animal Science, 89:1684-1697.

Bradbury, P. J.; Zhang, Z.; Kroon, D. E.; Casstevens, T. M.; Ramdoss, Y.; Buckler, E. S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics, 23(19):2633-2635. doi 10.1093/bioinformatics/btm308

Broglie, K. E.; Butler, K. H.; Butruille, M. G.; Conceição, A. S.; Frey, T. J.; Hawk, J. A.; Jaqueth, J. S.; Jones, E. S.; Multani, D. S.; Wolters, P. J. (2009) Polynucleotides and methods for making plants resistant to fungal pathogens. United States Patent number 7,619,133.

Buckler, E. S.; Holland, J. B.; Bradbury, P. J.; Acharya, C. B.; Brown, P. J.; Browne, C.; Ersoz, E.; Flint-Garcia, S. A.; Garcia, A.; Glaubitz, J. C.; Goodman, M. M.; Harjes, C.; Guill, K.; Kroon, D. E.; Larsson, S.; Lepak, N. K.; Li, H. H.; Mitchell, S. E.; Pressoir, G.; Peiffer, J. A.; Rosas, M. O.; Rocheford, T. R.; Romay, M. C.; Romero, S.; Salvo, S.; Villeda, H. S.; da Silva, H. S.; Sun, Q.; Tian, F.; Upadyayula, N.; Ware, D.; Yates, H.; Yu, J. M.; Zhang, Z. W.; Kresovich, S.; McMullen, M. D. (2009) The genetic architecture of maize flowering time. Science, 325:714-718.

Cockerham, C. C. (1954) An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. Genetics, 39:859-882.

Cook, J. P.; McMullen, M. D.; Holland, J. B.; Tian, F.; Bradbury, P.; Ross-Ibarra, J.; Buckler, E. S.; Flint-Garcia, S. A. (2012) Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels. Plant Physiology, 158:824-834.

Corder, E. H.; Saunders, A. M.; Risch, N. J.; Strittmatter, W. J.; Schmechel, D. E.; et al. (1994) Protective effect of apolipoprotein-E type-2 allele for late-onset Alzheimer-disease. Nature Genetics, 7:180-184.

Falush, D.; Stephens, M.; Pritchard, J. K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164:1567-1587.

Fan, B.; Onteru, S. K.; Du, Z. Q.; Garrick, D. J.; Stalder, K. J.; Rothschild, M. F. (2011) Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. PLoS One, 6:e14726.

Flint-Garcia, S. A.; Thornsberry, J. M.; Buckler, E. S. (2003) Structure of linkage disequilibrium in plants. Annual Reviews of Plant Biology, 54:357-374. doi 10.1146/annurev.arplant.54.031902.134907

Flint-Garcia, S. A.; Thuillet, A. C.; Yu, J.; Pressoir, G.; Romero, S. M.; Mitchell, S. E.; Doebley, J.; Kresovich, S.; Goodman, M. M.; Buckler, E. S. (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. The Plant Journal, 44:1054-1064. doi 10.1111/j.1365-313X.2005.02591.x

Fuller, W. A.; Gallant, A. R. (1974) Fitting segmented polynomial regression models whose join points have to be estimated. Journal of American Statistician Association, 68:144-147.

Gouy, M.; Rousselle, Y.; Thong Chane, A.; Anglade, A.; Royaert, S.; Nibouche, S.; Costet, L. (2015) Genome wide association mapping of agro-morphological and disease resistance traits in sugarcane. Euphytica, 202(2):269-284. doi 10.1007/s10681-014-1294-y

Gupta, P. K.; Rustgi, S.; Kulwal, P. L. (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Molecular Biology, 57:461-485. doi 10.1007/s11103-005-0257-z

Hastbacka, J.; de la Chapelle, A.; Kaitila, I.; Sistonen, P.; Weaver, A.; Lander, E. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nature Genetics, 2:204-211.

Hill, W. G.; Robertson, A. (1968) Linkage disequilibrium in finite populations. Theoretical and Applied Genetics, 38(6):226-231. doi 10.1007/BF01245622

Horton, M. W.; Bodenhausen, N.; Beilsmith, K.; Meng, D.; Muegge, B. D.; Subramanian, S.; Vetter, M. M.; Vilhjálmsson, B. J.; Nordbor, M.; Gordon, J. I.; Bergelson, J. (2014)

Genome-wide association study of Arabidopsis thaliana leaf microbial community. Nature Communications, 5:5320-5326. doi 10.1038/ncomms6320

Hung, H. Y.; Browne, C.; Guill, K.; Coles, N.; Eller, M.; Garcia, A.; Lepak, N.; Melia-Hancock, S.; Oropeza-Rosas, M.; Salvo, S.; Upadyayula, N.; Buckler, E. S.; Flint-Garcia, S. A.; McMullen, M. D.; Rocheford, T. R.; Holland, J. B. (2012) The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. Heredity, 108:490-499.

Kempthorne, D. (1957) An introduction to genetic statistics. John Wiley & Sons Inc, New York.

Kerem, B. S.; Rommens, J. M.; Buchanan, J. A.; Markiewicz, D.; Cox, T. K.; et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. Science, 245:1073-1080.

Kijas, J. W.; Townley, D.; Dalrymple, B. P.; Heaton, M. P.; Maddox, J. F.; McGrath, A.; et al. (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One, 4:e4668.

Krill, A. M.; Kirst, M.; Kochian, L. V.; Buckler, E. S.; Hoekenga, O. A. (2010) Association and linkage analysis of aluminum tolerance genes in maize. PLoS One, 5(4):e9958.

Kump, K. L.; Bradbury, P. J.; Wisser, R. J.; Buckler, E. S.; Belcher, A. R.; Oropeza-Rosas, M. A.; Zwonitzer, J. C.; Kresovich, S.; McMullen, M. D.; Ware, D.; Balint-Kurti, P. J.; Holland, J. B. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nature Genetics, 43(2):163-169. doi:10.1038/ng.747

Larsson, S. J.; Lipka, A. E.; Buckler, E. S. (2013) Lessons from Dwarf8 on the Strengths and Weaknesses of Structured Association Mapping. PLOS Genetics, 9(2): e1003246.

Li, C.; Li, Y.; Bradbury, P. J.; Wu, X.; Shi, Y.; Song, Y.; Zhang, D.; Rodgers-Melnick, E.; Buckler, E. S.; Zhang, Z.; Li, Y.; Wang, T. (2015) Construction of high-quality recombination maps with low-coverage genomic sequencing for joint linkage analysis in maize. BMC Biology, 13:78-89. doi 10.1186/s12915-015-0187-4

Li, F.; Chen, B.; Xu, K.; Wu, J.; Song, W.; Bancroft, I.; Harper, A. L.; Trick, M.; Liu, S.; Gao, G.; Wang, N.; Yan, G.; Qiao, J.; Li, J.; Li, H.; Xiao, X.; Zhang, T.; Wu, X. (2014) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (Brassica napus L.). DNA Research, 1:1-13. doi 10.1093/dnares/dsu002

Liu, K.; Muse, S. V. (2005) PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics, 21:2128-2129.

Liu, W.; Gowda, M.; Steinhoff, J.; Maurer, H. P.; Wurschum, T.; Longin, C. F. H.; Cossic, F.; Reif, J. C. (2011) Association mapping in an elite maize breeding population. Theoretical and Applied Genetics, 123:847-858. doi 10.1007/s00122-011-1631-7

Long, A. D.; Langley, C. H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research. 9: 720-731.

Lu, Y.; Zhang, S.; Shah, T.; Xie, C.; Hao, Z.; Li, X.; Farkhari, M.; Ribaut, J. M.; Cao, M.; Rong, T.; Xu, Y. (2010) Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. PNAS, 107(45):19585-19590. doi 10.1073/pnas.1006105107

Maccaferri, M.; Zhang, J.; Bulli, P.; Abate, Z.; Chao, S.; Cantu, D.; Bossolini, E.; Chen, X.; Pumphrey, M.; Dubcovsky, J. (2015) A genome-wide association study of resistance to stripe rust (Puccinia striiformis f. sp. tritici) in a worldwide collection of

hexaploid spring wheat (Triticum aestivum L.). G3: Genes, Genomes & Genetics, 5:449-465. doi 10.1534/g3.114.014563

McMullen, M. D.; Kresovich, S.; Villeda, H. S.; Bradbury, P.; Li, H.; Sun, Q.; Flint-Garcia, S. A.; Thornsberry, J.; Acharya, C.; Bottoms, C.; Brown, P.; Browne, C.; Eller, M.; Guill, K.; Harjes, C.; Kroon, D.; Lepak, N.; Mitchell, S. E.; Peterson, B.; Pressoir, G.; Romero, S.; Oropeza-Rosas, M.; Salvo, S.; Yates, H.; Hanson, M.; Jones, E.; Smith, S.; Glaubitz, J. C.; Goodman, M.; Ware, D.; Holland, J. B.; Buckler, E. S. (2009) Genetic properties of the maize nested association mapping population. Science, 325(5941):737-740. doi: 10.1126/science.1174320

Morris, G. P.; Ramub, P.; Deshpandeb, S. P.; Hashc, C. T.; Shahb, T.; Upadhyayab, H. D.; Riera-Lizarazub, O.; Brownd, P. J.; Acharyae, C. B.; Mitchelle, S. E.; Harrimane, J.; Glaubitze, J. C.; Buckler, E. S.; Kresovicha, S. (2013) Population genomic and genome-wide association studies of agro climatic traits in sorghum. PNAS, 110(2):453-458. doi 10.1073/pnas.1215985110

Nordborg, M.; Borevitz, J. O.; Bergelson, J.; Berry, C. C.; Chory, J.; et al. (2002) The extent of linkage disequilibrium in Arabidopsis thaliana. Nature Genetics, 30:190-193.

Pace, J.; Gardner, C.; Romay, C.; Ganapathysubramanian, B.; Lübberstedt, T. (2015) Genome-wide association analysis of seedling root development in maize (Zea mays L.). BMC Genomics, 16:47-58. doi 10.1186/s12864-015-1226-9

Pasam, R. K.; Sharma, R.; Malosetti, M.; van Eeuwijk, F. A.; Haseneyer, G.; Kilian, B.; Graner, A. (2012) Genome-wide association studies for agronomical traits in a worldwide spring barley collection. BMC Plant Biology, 12:16-37.

Poland, J. A.; Balint-Kurti, P. J.; Wisser, R. J.; Pratt, R. C.; Nelson, R. J. (2009) Shades of gray: The world of quantitative disease resistance. Trends in Plant Science, 14:21-29.

Pritchard, J. K.; Stephens, M.; Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics, 155:945-959.

Remington, D. L.; Thornsberry, J. M.; Matsuoka, Y.; Wilson, L. M.; Whitt, S. R.; Doebley, J.; Kresovich, S.; Goodman, M. M.; Buckler, E. S. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. PNAS, 98(20):11479-11484. doi 10.1073ypnas.201394398

Risch, N.; Merkangas, K. (1996) The future of genetic studies of complex human diseases. Science, 273(5281):1516-1517.

Samayoa, L. F.; Malvar, R. A.; Olukolu, B. A.; Holland, J. B.; Butrón, A. (2015) Genome-wide association study reveals a set of genes associated with resistance to the Mediterranean corn borer (Sesamia nonagrioides L.) in a maize diversity panel. BMC Plant Biology, 15:35-49. doi 10.1186/s12870-014-0403-3

SAS Institute (2007) The SAS System for Windows, version 9.2. SAS Institute Inc., Cary NC.

Schaefer, C. M.; Bernardo, R. (2013) Genome-wide association mapping of flowering time, kernel composition, and disease resistance in historical Minnesota maize inbreds. Crop Science, 53:2518-2529. doi 10.2135/cropsci2013.02.0121

Sladek, R.; Rocheleau, G.; Rung, J.; Dina, C.; Shen, L.; Serre, D.; et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature, 445:881-885.

Stuber, C. W.; Polacco, M.; Senior, M. L. (1999) Synergy of empirical breeding, marker assisted selection and genomics to increase crop yield potential. Crop Science, 39:1571-1583.

Suwarno, W. B.; Pixley, K. V.; Palacios‐Rojas, N.; Kaeppler, S. M.; Babu, R. (2015) Genome-wide association analysis reveals new targets for carotenoid biofortification

in maize. Theoretical and Applied Genetics, 128:851-864. doi 10.1007/s00122-015-2475-3

Thirunavukkarasu1, N.; Hossain, F.; Arora, K.; Sharma, R.; Shiriga, K.; Mittal, S.; Mohan, S.; Namratha, P. M.; Dogga, S.; Rani, T. S.; Katragadda, S.; Rathore, A.; Shah, T.; Mohapatra, T.; Gupta, H. S. (2014) Functional mechanisms of drought tolerance in subtropical maize (Zea mays L.) identified using genome-wide association mapping. BMC Genomics, 15:1182-1193.

Thornsberry, J. M.; Goodman, M. M.; Doebley, J.; Kresovich, S.; Nielsen, D.; et al. (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics, 28:286-289.

Tian, F.; Bradbury, P. J.; Brown, P. J.; Hung, H.; Sun, Q.; Flint-Garcia, S. A.; et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics, 43:159-162.

Van Inghelandt, D.; Melchinger, A. E.; Martinant, J. P.; Stich, B. (2012) Genome-wide association mapping of flowering time and northern corn leaf blight (Setosphaeria turcica) resistance in a vast commercial maize germplasm set. BMC Plant Biology, 12:56-70.

Viana, J. M. S.; Valente, M. S. F.; Silva, F. F.; Mundim, G. B.; Paes, G. P. (2013) Efficacy of population structure analysis with breeding populations and inbred lines. Genetica, 141:389-399. doi 10.1007/s10709-013-9738-1

Weir, B. (2010) Statistical genetic issues for genome-wide association studies. Genome, 53:869-875. doi 10.1139/G10-062

Weiss, L. A.; Arking, D. E.; Daly, M. J.; Chakravarti, A.; Brune, C. W.; West, K.; et al. (2009) A genome-wide linkage and association scan reveals novel loci for autism. Nature, 461:802-808.

Wu, R.; Zeng, Z. B. (2001) Joint linkage and linkage disequilibrium mapping in natural populations. Genetics, 157:899-909.

Yan, J. B.; Shah, T.; Warburton, M.; Buckler, E. S.; McMullen, M. D.; Crouch, J. (2009) Genetic characterization of a global maize collection using SNP markers. PLoS ONE 4:e8451.

Yang, W.; Guo, Z.; Huang, C.; Wang, K.; Jiang, N.; Feng, H.; Chen, G.; Liu, Q.; Xiong, L. (2015) Genome-wide association study of rice (Oryza sativa L.) leaf traits with a high-throughput leaf scorer. Journal of Experimental Botany, 66(18):5605-5615. doi 10.1093/jxb/erv100

Yang, X.; Yan, J.; Shah, T.; Warbuton, M. L.; Li, Q.; Li, L.; Gao, Y.; Chai, Y.; Fu, Z.; Zhou, Y.; Xu, S.; Bai, G.; Meng, Y.; Zheng, Y.; Li, J. (2010) Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. Theoretical and Applied Genetics, 121:417-431. doi 10.1007/s00122-010-1320-y

Yu, J. M.; Buckler, E. S. (2006) Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology, 17:1-6.

Yu, J.; Holland, J. B.; McMullen, M. D.; Buckler, E. S. (2008) Genetic design and statistical power of nested association mapping in maize. Genetics, 178:539-551. doi 10.1534/genetics.107.074245

Zhang, J.; Song, Q.; Cregan, P. B.; Nelson, R. L.; Wang, X.; Wu, J.; Jiang, G. L. (2015) Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. BMC Genomics, 16:217-227. doi 10.1186/s12864-015-1441-4

Zila, C. T.; Samayoa, L. F.; Santiago, R.; Butrón, A.; Holland, J. B. (2013) A genome-wide association study reveals genes associated with Fusarium ear rot resistance in a maize core diversity panel. G3: Genes, Genomes & Genetics, 3:2095-3104.

Table 1 Average number of significant associations with a FDR of 1 and 5%, power of QTL detection (%), number of false-positive associations in chromosomes with no QTL and one to four QTL, bias in the QTL position (cM), and average range for the regions with identified QTL, regarding population 1, generation 10r (random cross), three traits (expansion volume (EV; mL/g), grain yield (GY; g/plant), and days to maturity (DM)), two sample sizes, and two heritabilities[1]

Population	FDR	Trait	Sample	h^2	Sig. Assoc.	Power	False+0	False+1-4	Bias	Av. range
Open-pollinated population	5%	EV	400	0.8	32.1 (13; 73)	88.6 (60; 100)	3.3 (0; 17)	7.9 (1; 28)	0.52 (0.09; 0.83)	0.84 (0.11; 2.18)
				0.4	6.7 (0; 22)	37.2 (0; 80)	0.6 (0; 5)	0.9 (0; 8)	0.21 (0.00; 0.98)	0.28 (0.00; 2.45)
			200	0.8	6.2 (0; 33)	37.0 (0; 70)	0.5 (0; 3)	1.1 (0; 12)	0.16 (0.00; 0.88)	0.17 (0.00; 1.42)
				0.4	0.8 (0; 5)	5.2 (0; 40)	0.2 (0; 2)	0.3 (0; 1)	0.22 (0.00; 1.76)	0.14 (0.00; 1.31)
		GY	400	0.8	31.3 (10; 82)	87.8 (70; 100)	3.3 (0; 12)	7.5 (0; 28)	0.51 (0.00; 0.97)	0.77 (0.00; 1.54)
				0.4	5.7 (0; 25)	34.4 (0; 90)	0.4 (0; 2)	0.8 (0; 7)	0.20 (0.00; 0.85)	0.18 (0.00; 1.03)
			200	0.8	5.9 (0; 18)	32.6 (10; 80)	0.8 (0; 8)	1.1 (0; 7)	0.16 (0.00; 0.97)	0.16 (0.00; 1.75)
				0.4	0.8 (0; 16)	3.8 (0; 30)	0.6 (0; 4)	0.4 (0; 6)	0.15 (0.00; 1.39)	0.06 (0.00; 0.94)
		DM	400	0.8	50.7 (12; 119)	92.8 (70; 100)	6.5 (0; 21)	14.9 (3; 48)	0.65 (0.06; 1.02)	1.22 (0.07; 2.82)
				0.4	8.6 (0; 33)	39.0 (0; 100)	1.1 (0; 5)	1.4 (0; 9)	0.29 (0.00; 0.70)	0.38 (0.00; 1.64)
			200	0.8	6.6 (0; 41)	34.0 (0; 70)	0.7 (0; 3)	1.2 (0; 12)	0.23 (0.00; 0.91)	0.27 (0.00; 2.00)
				0.4	1.0 (0; 7)	5.0 (0; 30)	0.5 (0; 2)	0.5 (0; 3)	0.17 (0.00; 0.82)	0.15 (0.00; 2.45)
	1%	EV	400	0.8	15.0 (7; 39)	76.6 (50; 100)	0.4 (0; 3)	2.0 (0; 13)	0.32 (0.00; 0.75)	0.41 (0.00; 1.70)
		GY	400	0.8	14.8 (6; 39)	75.4 (60; 100)	0.3 (0; 3)	2.1 (0; 10)	0.33 (0.00; 0.80)	0.43 (0.00; 1.32)
		DM	400	0.8	20.6 (6; 51)	80.0 (40; 100)	0.6 (0; 3)	4.1 (0; 19)	0.44 (0.00; 0.93)	0.61 (0.00; 1.99)

[1]the values between parentheses are the minimum and maximum.

Table 2 Average number of significant associations with a FDR of 1 and 5%, power of QTL detection (%), number of false-positive associations in chromosomes with no QTL and one to two QTL, bias in the QTL position (cM), and average range for the regions with identified QTL, regarding population 1, generation 10r (random cross), three traits (expansion volume (EV; mL/g), grain yield (GY; g/plant), and days to maturity (DM)), two sample sizes, and QTL heritability of 12%[1]

Population	FDR	Trait	Sample	Sig. Assoc.	Power	False+0	False+1-2	Bias	Av. range
Open-pollinated	5%	EV	400	26.8 (7; 63)	99.2 (60; 100)	2.9 (0; 11)	8.1 (0; 30)	0.59 (0.00; 0.92)	1.27 (0.00; 2.74)
Population			200	6.0 (2; 20)	70.8 (20; 100)	0.5 (0; 4)	0.8 (0; 5)	0.18 (0.00; 0.87)	0.25 (0.00; 1.58)
		GY	400	31.6 (8; 74)	99.6 (80; 100)	3.6 (0; 16)	9.8 (0; 43)	0.68 (0.14; 1.01)	1.39 (0.12; 2.60)
			200	7.2 (3; 19)	74.8 (20; 100)	0.7 (0; 7)	1.2 (0; 6)	0.21 (0.00; 0.75)	0.30 (0.00; 1.49)
		DM	400	50.2 (20; 110)	100.0 (100; 100)	6.5 (0; 23)	18.4 (3; 50)	0.77 (0.49; 1.05)	2.02 (0.76; 3.49)
			200	8.8 (2; 31)	75.2 (40; 100)	1.0 (0; 9)	1.9 (0; 12)	0.26 (0.00; 0.96)	0.35 (0.00; 1.38)
	1%	EV	400	13.3 (4; 38)	98.4 (60; 100)	0.5 (0; 2)	2.7 (0; 16)	0.40 (0.00; 0.79)	0.65 (0.00; 1.92)
		GY	400	15.6 (5; 43)	98.4 (80; 100)	0.6 (0; 7)	3.2 (0; 21)	0.53 (0.03; 0.81)	0.82 (0.03; 1.63)
		DM	400	22.3 (8; 49)	100.0 (100; 100)	0.9 (0; 6)	6.5 (0; 21)	0.58 (0.02; 0.99)	1.14 (0.03; 2.98)

[1]the values between parentheses are the minimum and maximum.

Table 3 Average number of significant associations with a FDR of 1 and 5%, power of QTL detection (%), number of false-positive associations in chromosomes with no QTL and one to four QTL, bias in the QTL position (cM), and average range for the regions with identified QTL, regarding an inbred lines panel, three traits (expansion volume (EV; mL/g), grain yield (GY; g/plant), and days to maturity (DM)), two sample sizes, and two heritabilities[1]

Population	FDR	Trait	Sample	h2	Sig. Assoc.	Power	False+0	False+1-4	Bias	Av. range
Inbred Lines	5%	EV	400	0.8	14.6 (9; 27)	96.0 (90; 100)	0.5 (0; 2)	3.1 (0; 13)	0.12 (0.00; 0.70)	0.14 (0.00; 0.58)
panel				0.4	6.9 (2; 14)	58.0 (20; 100)	0.3 (0; 2)	0.6 (0; 4)	0.04 (0.00; 0.35)	0.04 (0.00; 0.42)
			200	0.8	5.4 (0; 10)	45.2 (0; 80)	0.1 (0; 1)	0.6 (0; 4)	0.04 (0.00; 0.69)	0.05 (0.00; 1.04)
				0.4	0.7 (0; 5)	5.4 (0; 40)	0.1 (0; 1)	0.4 (0; 1)	0.09 (0.00; 0.92)	0.06 (0.00; 1.12)
		GY	400	0.8	13.9 (7; 23)	96.2 (70; 100)	0.3 (0; 2)	2.8 (0; 11)	0.11 (0.00; 0.41)	0.12 (0.00; 0.50)
				0.4	7.9 (2; 17)	61.4 (20; 100)	0.4 (0; 3)	1.0 (0; 5)	0.05 (0.00; 0.28)	0.05 (0.00; 0.33)
			200	0.8	5.1 (0; 13)	42.6 (0; 70)	0.2 (0; 2)	0.5 (0; 5)	0.05 (0.00; 0.72)	0.04 (0.00; 1.04)
				0.4	1.0 (0; 5)	8.2 (0; 30)	0.2 (0; 2)	0.2 (0; 2)	0.00 (0.00; 0.00)	0.00 (0.00; 0.00)
		DM	400	0.8	15.4 (8; 29)	96.0 (80; 100)	0.5 (0; 3)	3.7 (0; 17)	0.14 (0.00; 0.43)	0.17 (0.00; 0.57)
				0.4	9.1 (4; 15)	70.8 (40; 90)	0.4 (0; 2)	1.2 (0; 5)	0.06 (0.00; 0.62)	0.05 (0.00; 0.52)
			200	0.8	5.6 (0; 12)	46.8 (0; 80)	0.1 (0; 2)	0.6 (0; 4)	0.05 (0.00; 0.53)	0.04 (0.00; 0.54)
				0.4	1.1 (0; 6)	9.6 (0; 40)	0.1 (0; 1)	0.1 (0; 1)	0.01 (0.00; 0.23)	0.00 (0.00; 0.05)
	1%	EV	400	0.8	10.8 (6; 21)	91.6 (60; 100)	0.1 (0; 1)	1.0 (0; 10)	0.05 (0.00; 0.63)	0.06 (0.00; 0.61)
		GY	400	0.8	10.2 (7; 15)	91.2 (70; 100)	0.0 (0; 1)	0.7 (0; 6)	0.03 (0.00; 0.23)	0.03 (0.00; 0.26)
		DM	400	0.8	10.7 (7; 16)	91.6 (70; 100)	0.0 (0; 1)	1.0 (0; 5)	0.07 (0.00; 0.39)	0.07 (0.00; 0.47)

[1]the values between parentheses are the minimum and maximum.

Table 4 Average number of significant associations with a FDR of 1 and 5%, power of QTL detection (%), number of false-positive associations in chromosomes with no QTL and one to four QTL, bias in the QTL position (cM), and average range for the regions with identified QTL, regarding population 1, generation 10r10s (random cross and selfing), three traits (expansion volume (EV; mL/g), grain yield (GY; g/plant), and days to maturity (DM)), two sample sizes, and two heritabilities[1]

Population	FDR	Trait	Sample	h^2	Sig. Assoc.	Power	False+0	False+1-4	Bias	Av. range
RILs	1%	EV	400	0.8	34.5 (4; 122)	87.0 (40; 100)	0.3 (0; 2)	12.7 (0; 68)	0.61 (0.00; 1.05)	0.90 (0.00; 2.43)
				0.4	7.3 (1; 31)	39.0 (10; 70)	0.1 (0; 2)	1.7 (0; 12)	0.17 (0.00; 0.73)	0.25 (0.00; 1.25)
			200	0.8	4.9 (0; 24)	27.0 (0; 70)	0.1 (0; 2)	1.1 (0; 9)	0.15 (0.00; 1.15)	0.20 (0.00; 1.42)
		GY	400	0.8	34.1 (5; 123)	86.0 (50; 100)	0.2 (0; 2)	12.9 (0; 73)	0.60 (0.00; 1.05)	0.88 (0.00; 2.29)
				0.4	9.5 (2; 42)	43.0 (10; 70)	0.1 (0; 1)	2.8 (0; 23)	0.30 (0.00; 1.09)	0.41 (0.00; 2.00)
			200	0.8	6.3 (0; 27)	30.0 (0; 60)	0.1 (0; 4)	1.8 (0; 16)	0.22 (0.00; 1.29)	0.30 (0.00; 2.71)
		DM	400	0.8	40.4 (11; 142)	89.0 (70; 100)	0.4 (0; 3)	15.8 (0; 81)	0.66 (0.12; 1.07)	1.01 (0.13; 2.62)
				0.4	16.6 (4; 65)	61.0 (30; 100)	0.2 (0; 2)	5.2 (0; 44)	0.47 (0.00; 1.09)	0.62 (0.00; 1.94)
			200	0.8	7.4 (0; 42)	32.0 (0; 80)	0.2 (0; 3)	2.1 (0; 15)	0.26 (0.00; 1.21)	0.36 (0.00; 2.11)
	5%	EV	200	0.4	1.8 (0; 34)	7.0 (0; 30)	0.2 (0; 1)	1.2 (0; 23)	0.19 (0.00; 1.12)	0.33 (0.00; 2.47)
		GY	200	0.4	2.9 (0; 19)	11.0 (0; 30)	0.3 (0; 1)	1.5 (0; 9)	0.41 (0.00; 1.77)	0.67 (0.00; 3.59)
		DM	200	0.4	4.6 (0; 33)	17.0 (0; 50)	0.3 (0; 2)	2.0 (0; 19)	0.32 (0.00; 1.26)	0.37 (0.00; 1.81)

[1]the values between parentheses are the minimum and maximum.

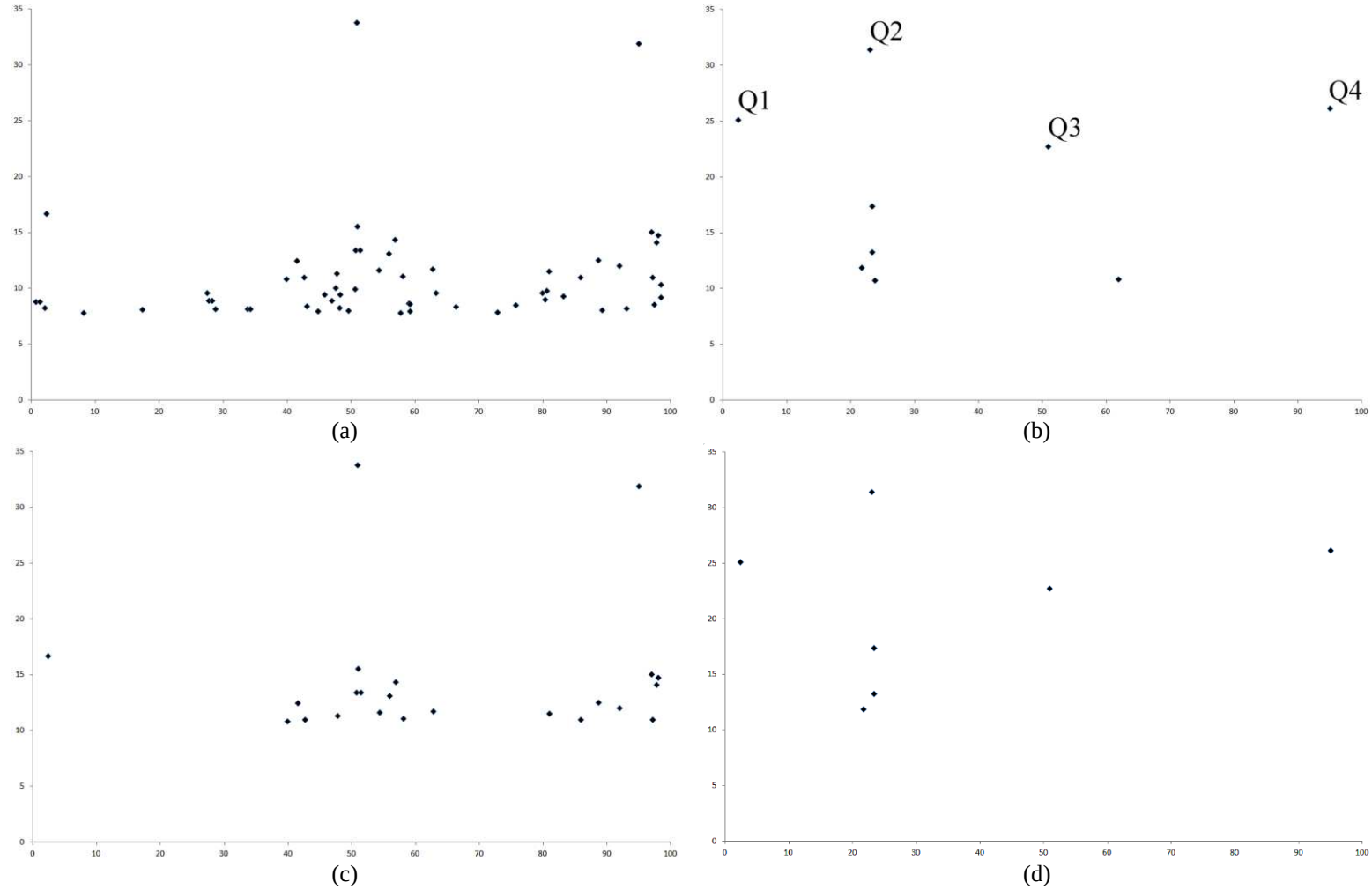


Figure 1 Significant associations at a FDR of 5 (a and b) and 1% (c and d) (F test; Y axe) in chromosome 1 (SNP position (cM); X axe), from the GWAS in population 1, generations 0 (a and c) and 10 (b and d), regarding expansion volume, heritability of 0.8, and sample size 400 (simulation 1) (Q = QTL).

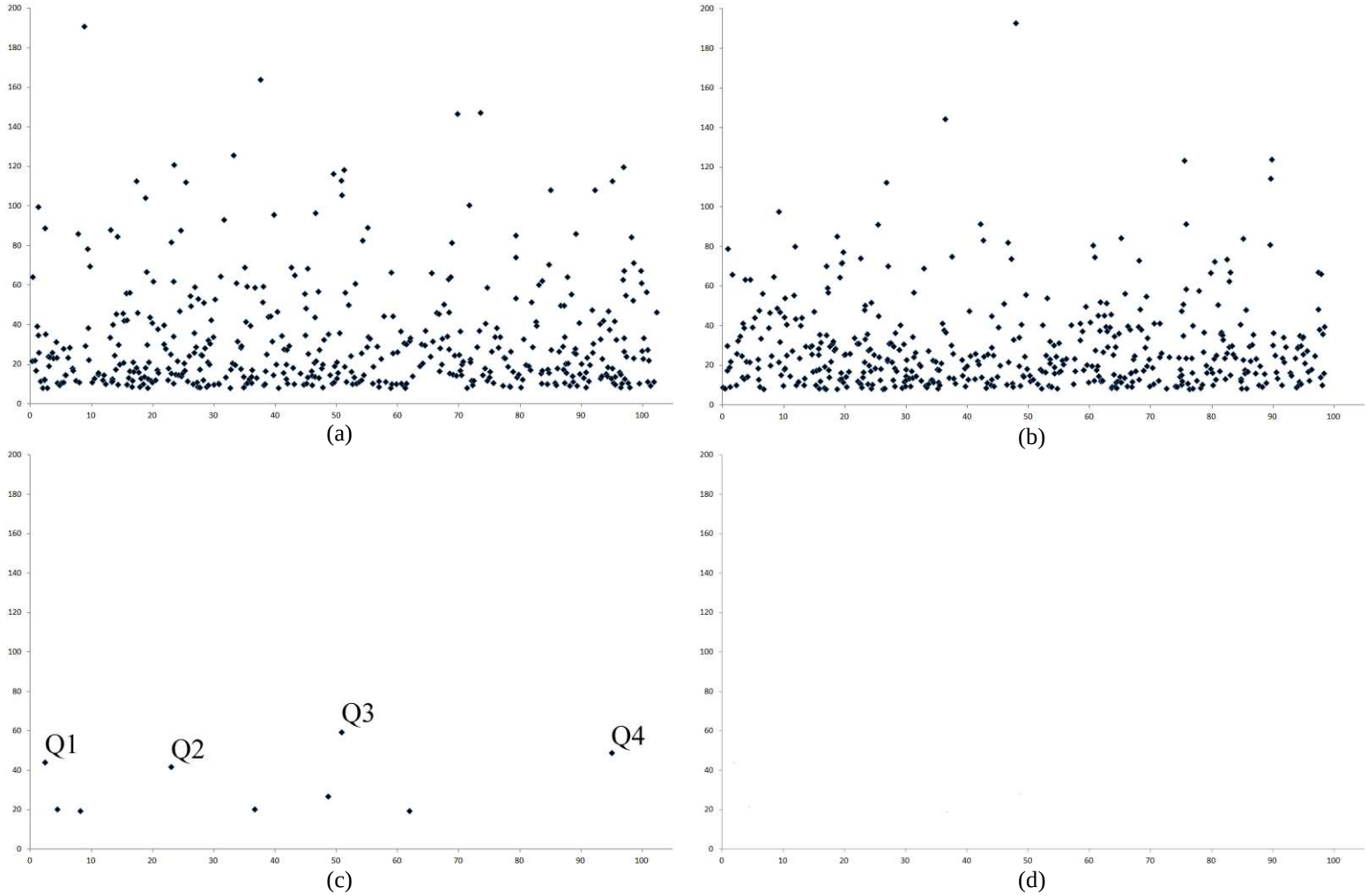
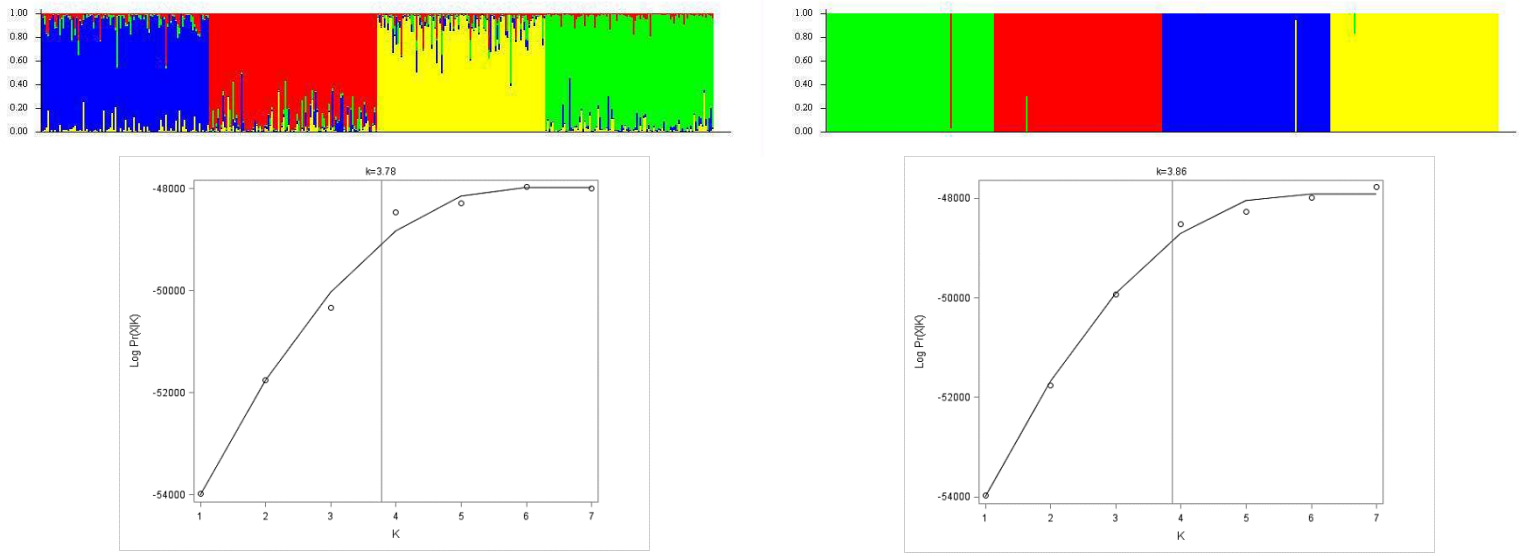


Figure 2 Significant associations at a FDR of 1% (F test; Y axe) in chromosomes 1 and 3 (SNP position (cM); X axe) ignoring (a and b, respectively) and correcting for the population structure (c and d, respectively), from the GWAS in an inbred lines panel regarding expansion volume, heritability of 0.8, and sample size 400 (simulation 1) (Q = QTL).



(a) (b)

Figure 3 Results from the population structure analysis and the inferred plateau method, based on the admixture model with correlated allelic frequencies (a) and the no admixture model with independent allelic frequencies (b).

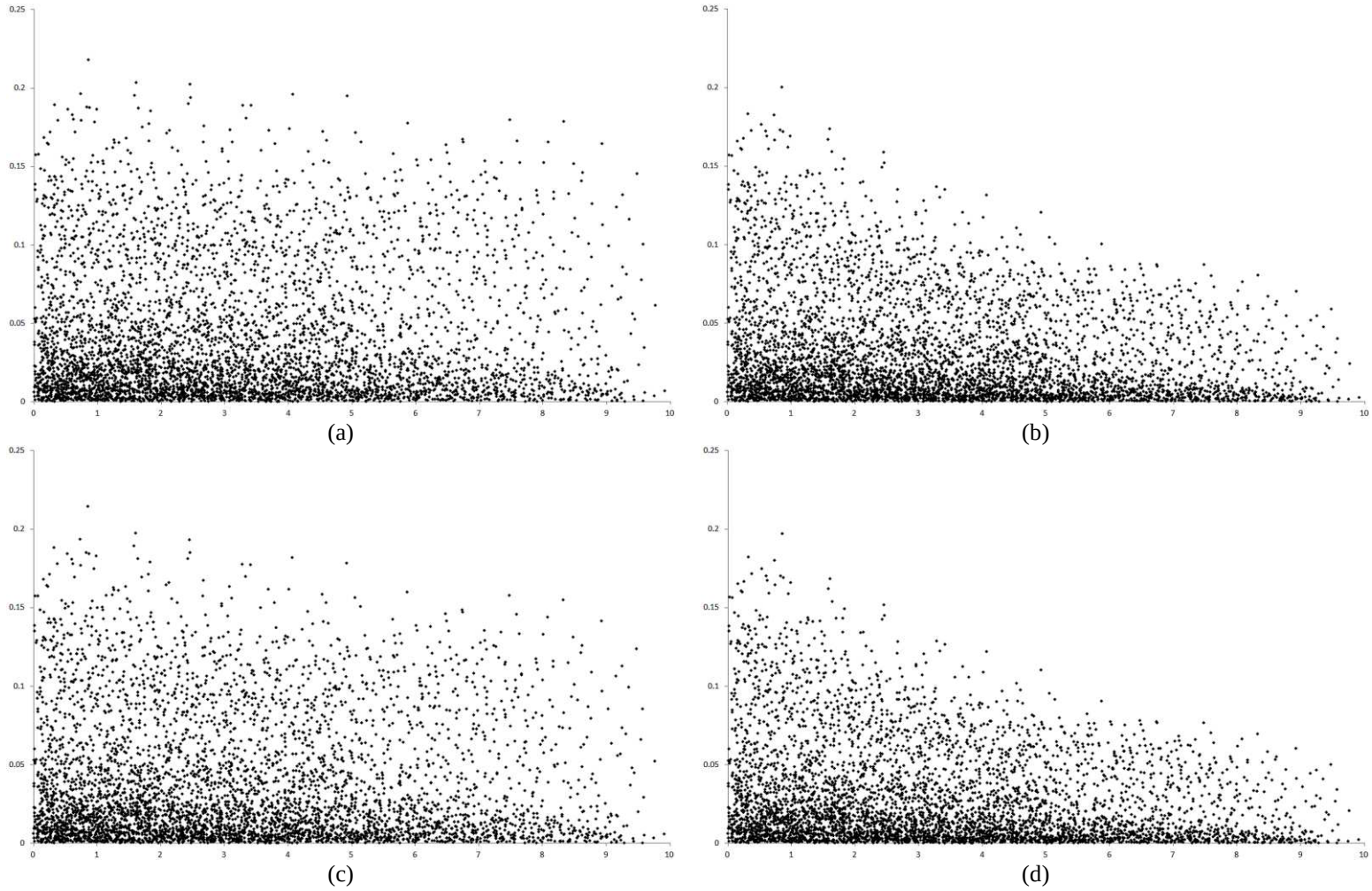


Figure 4 Relationship between the parametric LD value (absolute value; Y axe) and distance (cM; X axe) in population 1, generations 0 (a), 10r (random cross) (b), 10s (selfing) (c), and 10r10s (d), assuming a segment of 10 cM of chromosome 1 (centered on QTL 3).

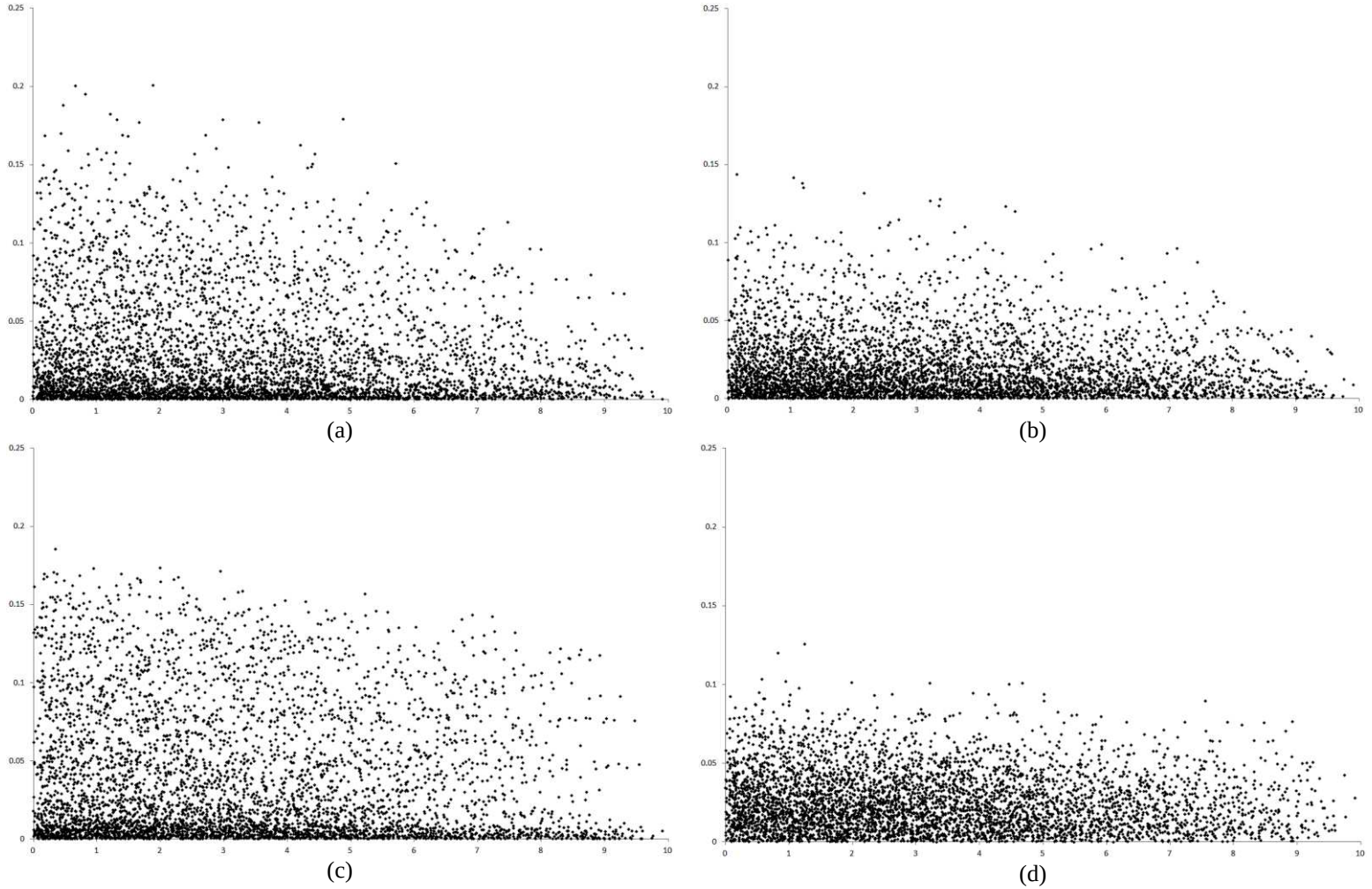


Figure 5 Relationship between the parametric LD value (absolute value; Y axe) and distance (cM; X axe) in populations 2 (a), 3 (b), and 4 (c), generation 10s (selfing), and in the inbred lines panel (d), assuming a segment of 10 cM of chromosome 1 (centered on QTL 3).

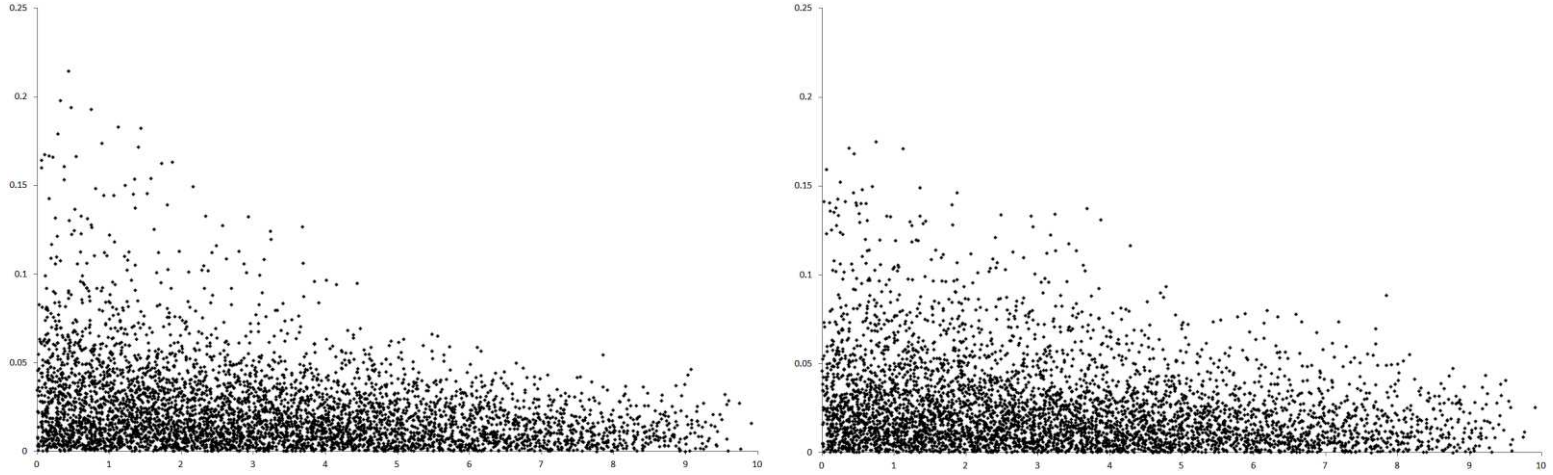


Figure 6 Relationship between the estimated LD value (absolute value; Y axe) and distance (cM; X axe) in population 1, generations 10r (random cross) (a) and 10r10s (random cross and selfing) (b), simulation 1, assuming a segment of 10 cM of chromosome 1 (centered on QTL 3).