ALEXANDRE ALONSO ALVES

LINKAGE ANALYSIS AND QTL MAPPING IN SIMULATED POPULATIONS

Thesis submitted to the Federal University of Viçosa, in partial fulfillment of the requirements of the Genetics and Breeding Graduate Program, for the Degree of Doctor Scientiae.

VIÇOSA
MINAS GERAIS – BRAZIL
2010

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
A474L
2010

Alves, Alexandre Alonso, 1983-
Linkage analysis and QTL mapping in simulated populations / Alexandre Alonso Alves. – Viçosa, MG, 2010.
xvi, 97f. : il. ; 29cm.

Orientador: Acelino Couto Alfenas.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Genética - Métodos estatísticos. 2. Melhoramento genético - Simulação por computador. 3. Genética quantitativa. 4. Marcadores genéticos. I. Universidade Federal de Viçosa. II. Título.

CDD 22. ed. 576.50285

ALEXANDRE ALONSO ALVES

LINKAGE ANALYSIS AND QTL MAPPING IN SIMULATED POPULATIONS

Thesis submitted to the Federal University of Viçosa, in partial fulfillment of the requirements of the Genetics and Breeding Graduate Program, for the Degree of Doctor Scientiae.

Approved: October 25, 2010.

Cosme Damião Cruz
(Co-Adivisor)

Marcos Deon Vilela de Resende
(Co-Adivisor)

Leonardo Lopes Bhering

Lúcio Mauro da Silva Guimarães

Acelino Couto Alfenas
(Advisor)

"The only limit to our realization of tomorrow will be our doubts of today. Let us move forward with strong and active faith."

“There are many ways of going forward, but only one way of standing still.”

Franklin D. Roosevelt

To my fiancee, Gisele P. Domiciano

I DEDICATE

ACKNOWLEDGEMENTS

I am grateful to the Federal University of Viçosa, particularly to the Genetics and Breeding Graduate Program for the opportunity that was provided to me in these last years.

I wish to thank the Brazilian National Research Council, CNPq, for the concession of the PhD scholarship.

I am very grateful to my PhD advisor, Prof Acelino Couto Alfenas, for his friendship, enthusiasm, patience, and for teaching me the exciting science of Forest Pathology.

I am also very grateful to Prof Cosme Damião Cruz (PhD co-advisor), for his unconditional support, for introducing me in the fields of statistical genomics and biometry, and of course, for helping me in the development of this work.

I wish to especially thank Dr. Lúcio Mauro da Silva Guimarães, Prof. Leonardo Lopes Bhering and Dr. Marcos Deon Vilela de Resende (PhD co-advisor) who were always available to discuss ideas and results, and for their friendship.

I am grateful to Dr. Dario Grattapaglia, for his support over the last years and for introducing me to *Eucalyptus* genomic research.

I have to express my gratitude to all my professors, in special Prof Sergio H. Brommonschenkel, for making me see science from a new perspective.

A special thanks to my friends at the Lab of Forest Pathology – BIOAGRO/UFV for the pleasant times, especially Márcia, Talyta, Marcelo, Rodrigo, and Ricardo. I also wish to thank my friends at the Labs of Genomics and Bioinformatics, both from BIOAGRO/UFV.

I am grateful to my friends in the Graduate programs in Genetics and Breeding, and Plant Pathology, both from UFV, for providing me an exciting environment of research, especially Márcio, Ricardo, Leandro and Caio.

Above all, many thanks to my fiancee, Gisele P Domiciano, for being such a special person in my life and for being always by my side, I love you very much; to my cockatiels Kiki and Kim (Seseco); to my parents José Donizeti and Vânia Aparecida; brothers Aléssio and Guilherme; as well as my aunts, Vera, Míriam, Clélia and Sônia for their unconditional support and motivation.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

LIST OF FIGURES

RESUMO

ALVES, Alexandre Alonso. D.Sc., Universidade Federal de Viçosa, outubro de 2010. **Análise de ligação e mapeamento de QTLs em populações simuladas.** Orientador: Acelino Couto Alfenas. Co-orientadores: Cosme Damião Cruz e Marcos Deon Vilela de Resende.

Como os recentes avanços na tecnologia têm levado ao desenvolvimento de novas tecnologias de genotipagem, no futuro, é mais provável que os mapas de ligação de alta densidade serão construídos a partir de marcadores dominantes e co-dominantes. Recentemente, uma abordagem estritamente genética foi proposta para a estimação da freqüência de recombinação (*r*) entre marcadores co-dominantes em famílias de irmãos completos. O conjunto completo de estimadores quase foi obtido, mas infelizmente, uma configuração envolvendo a estimativa da distância entre os marcadores dominantes, que segregam na proporção 3:1 e marcadores co-dominantes, não foi levada em consideração. Aqui novos nove estimadores são acrescentados ao conjunto previamente publicado, tornando possível cobrir todas as combinações de marcadores moleculares com dois a quatro alelos (sem epistasia) em uma família de irmãos completos. Isso inclui a segregação em um ou ambos os genitores, dominância e todas as configurações de fases de ligação. Como populações de retrocruzamentos (RC) são frequentemente utilizadas como populações de mapeamento, tanto em espécies autógamas, quanto em espécies alógamas foi conduzido um estudo de simulação para testar as implicações do tamanho da população, herdabilidade da característica, propriedades do QTL (r^2, a e posição) e densidade de marcadores no poder de detecção e precisão do mapeamento de

QTLs. Para tanto foram simuladas populações com diferentes tamanhos, com diferentes características (h^2, número de QTLs e posição) e os dados analisados com dois métodos de mapeamento de QTLs comumente utilizados (mapeamento por intervalo simples (MIS) e mapeamento por intervalo composto (MIC)). Verificou-se que o tamanho da amostra tem uma grande implicação no poder de detecção e como conseqüência na estimação da magnitude da variação explicada pelo QTL e no efeito genético, em função de populações pequenas não permitirem o mapeamento de QTLs de pequeno efeito, principalmente quando esses estão envolvidos no controle genético de características de baixa herdabilidade. Também foi verificado que o posicionamento de QTLs baseados em MIC é mais acurado que MIS e que em média os QTLs mapeados estavam próximos as suas posições simuladas. Um resultado interessante é que o MIC tende a subestimar os valores de magnitude (r^2) especialmente em populações grandes/características de baixa herdabilidade e superestimá-la em populações pequenas, o que pode ser um reflexo do pequeno coeficiente de variação do erro utilizado, ou devido ao fato de quando os marcadores não se encontram na exata posição do QTL, esse parâmetro é de fato esperado ser subestimado. Destaca-se também, o fato que quando marcadores estão amplamente distribuídos ao longo do genoma (~10cM), e desse modo cobrindo a região do QTL, se um dos marcadores já estiver próximo ao QTL, um maior número de marcadores (~1cM) não melhora a precisão do mapeamento do QTL em populações suficientemente grandes. Baseado nesses resultados recomenda-se o uso de populações de tamanho adequado, ≥500, se a intenção é mapear QTLs em populações de RC, porque nessa situação, mesmo mapas de média densidade podem ser usados para mapear QTLs de grande ou pequeno efeito com grande confiabilidade. Finalmente, como os procedimentos de mapeamento de ligação e mapeamento de QTLs em famílias de irmãos completos (FIC) de espécies alógamas são bastante diversos, foi conduzido um estudo comparando o método de mapeamento por pseudo-testcross modificado (PST) (usando microsatélites), com o método de mapeamento baseado na FIC;

em termos de ordenamento dos marcadores, distância entre os marcadores, comprimento total do mapa, variância das estimativas de distância e estresse. Investigou-se também o poder de detecção e a precisão de métodos de mapeamento de QTLs por intervalos baseados nos mapas PST ou no mapa para a FIC. Verificou-se que em geral as duas estratégias geram mapas altamente correlacionados com comprimentos dos grupos de ligação proporcionais. Verificou-se também que independentemente da abordagem de mapeamento de QTLs utilizadas, o poder de detecção é reduzido em populações pequenas, especialmente em situações onde a herdabilidade da característica ou magnitude do QTL é pequena. Também foi verificado que apesar dos dois métodos serem aparentemente equivalentes em termos de posicionamento do QTL para características de alta herdabilidade/QTLs de grande efeito, o MIC baseado nos mapas pseudo-testcross prove dados mais acurados para características de baixa herdabilidade/QTLs de pequeno efeito. Como relação à magnitude dos QTLs, notou-se que ambos os métodos parecem ser equivalentes, sendo os valores superestimados para características de alta herdabilidade e subestimados para características de baixa herdabilidade, independentemente do tamanho amostral. Assim para espécies alógamas com médio nível de recursos genômicos, propõem-se que tanto a abordagem de PST quanto a abordagem baseada na FIC, e métodos de mapeamento de QTLs relacionados, possam ser utilizados para gerar mapas genéticos e mapear QTLs com alta confiabilidade. É importante ressaltar, entretanto, que outros estudos, usando diferentes cenários, *i.e.* diferentes coeficientes de variação do erro, diferentes números de QTLs, diferentes distribuições de marcadores, que coletivamente podem tornar a simulação um pouco mais realística, são necessários para verificar que os resultados deste trabalho se mantêm em todas as situações.

ABSTRACT

ALVES, Alexandre Alonso. D.Sc., Universidade Federal de Viçosa, October, 2010. **Linkage analysis and QTL mapping in simulated populations.** Advisor: Acelino Couto Alfenas. Co-advisors: Cosme Damião Cruz and Marcos Deon Vilela de Resende.

As high-throughput genomic tools have led to the development of novel genotyping procedures, it is likely that, in the future, high density linkage maps will be constructed from both dominant and co-dominant markers. Recently, a strictly genetic approach was described for estimating the recombination frequency (*r*) between co-dominant markers in full-sib families. The complete set of maximum likelihood estimators for *r* in full-sib families was almost obtained, but unfortunately, one particular configuration involving dominant markers, segregating in a 3:1 ratio and co-dominant markers, was not considered. Here we add nine further estimators to the previously published set, thereby making it possible to cover all combinations of molecular markers with two to four alleles (without epistasis) in a full-sib family. This includes segregation in one or both parents, dominance and all linkage phase configurations. As backcross (BC) populations are often used as mapping populations both in self pollinating species, and in out-breeding species we also undertook a simulation study to test implications of population size, trait heritability, QTL properties (r^2, a and position) and marker density in the power and precision of QTL mapping. For that we have simulated populations with different sizes, with different characteristics (h^2, QTL number and location) and the data analyzed with two

QTL mapping methods (simple interval mapping (SIM) and composite interval mapping (CIM)). We found that sample size has a major implication in the detection power and as consequence in the estimation of the magnitude and additive genetic effect, as small populations do not allow mapping of low effect QTLs, especially if these QTLs are involved in the genetic control of traits with low heritability. We also found that the positioning of the QTLs based on CIM is more accurate than SIM and that on average the mapped QTLs are close to their simulated position. The results showed that CIM tend to underestimate the magnitude (r^2) values especially in large population sizes/low heritabilities traits and overestimate it in smaller populations, which can be a reflection of the low coefficient of variation of the error used, or due to fact that when markers aren´t in the same of the QTL, this parameter is indeed expected to be underestimated. We also highlight the fact, that when markers are evenly distributed across the genome (~10 cM), and therefore covering the QTL region, if one of the markers is already close to the QTL, larger number of markers (~ 1cM) do not improve the precision of QTL mapping in sufficiently large mapping populations. Based on our results we recommend the use of adequate sample size, say ≥500, if the intention is map QTLs in backcross populations, because in this situation even mid-density genetic maps can be used to map QTLs of large or small effect with high confidence. Finally, as the procedures for linkage and QTL mapping in full-sib families of outbreeding species are quite diverse, we also undertook a simulation study comparing the modified pseudo-testcross (using SSR markers) and the full-sib mapping designs in terms of marker ordering, distance between markers, total map size, distance variance and stress. We also investigated the power and precision of interval mapping procedures based on the full-sib and on the modified pseudo-testcross maps. We show that in general the modified pseudo-testcross and the full-sib mapping designs generate highly correlated maps with proportional linkage groups length. That independent of the QTL mapping approach used, detection power is reduced in small populations, especially in situations where trait heritability or QTL magnitude are low. We also

found that although both methods appear to be equivalent in terms of QTL positioning for high heritability traits/major effect QTLs, the CIM based on modified pseudo-testcross maps provide more accurate data for low heritability traits/minor effect QTLs in larger populations. With regard to QTLs magnitude, we show that both methods appear to be equivalent, and that the magnitude values tended to be overestimated for the high heritability trait, and underestimated for the low heritability trait, independent of the sample size. Thus, for outbreeding species with mid-level of genomic resources we propose that either the modified pseudo-testcross or the single full-sib mapping design and the related QTL mapping strategies can be used to generate genetic maps and map QTLs with high confidence. It is important to highlight however, that, other studies, using different scenarios, *i.e.* different coefficients of variation of the error, different number of QTLs, different marker distributions, which collectively may make the simulation a bit more realistic, are needed in order to see if the results of our work hold true in every situation.

GENERAL INTRODUCTION

A key development in the field of complex trait analysis was the establishment of large collections of molecular/genetic markers, which could be used to construct detailed genetic maps of both experimental and domesticated species. These maps provided the foundation for the modern-day Quantitative Trait Loci (QTLs) mapping methodologies (Doerge 2002). A linkage map may be thought of as a 'roadmap' of the chromosomes derived from two different parents. Linkage maps indicate the position and relative genetic distances between markers along chromosomes, which is analogous to signs or landmarks along a highway. The most important use for linkage maps is to identify chromosomal locations containing genes and QTLs associated with traits of interest; such maps may then be referred to as genetic maps. Genetic mapping is based on the principle that genes and, or markers segregate via chromosome recombination (called crossing-over) during meiosis (*i.e.* sexual reproduction), thus allowing their analysis in the progeny (Collard et al. 2005). Genes or markers that are tightly-linked will be transmitted together from parent to progeny more frequently than genes or markers that are located further apart (Schuster and Cruz 2008).

Linkage maps are constructed from the analysis of many segregating markers. The three main steps of linkage map construction are: (i) production of a mapping population; (ii) identification of polymorphism and (iii) linkage analysis of markers. The basis of polymorphism identification, the classes of molecular markers, as well as the new technologies recently developed for high-throughput

genotyping can be found elsewhere (Wenzl et al. 2004; Zhu and Salmeron 2007). The construction of a linkage map requires a segregating plant population (*i.e.* a population derived from sexual reproduction). Several different populations may be utilized for mapping within a given plant species, with each population type possessing advantages and disadvantages (Collard et al. 2005; Schuster and Cruz 2008). F_2 populations, derived from F_1 hybrids, and backcross (BC) populations, derived by crossing the F_1 hybrid to one of the parents, are the simplest types of mapping populations developed for self pollinating species, as well as the most used. The parents selected to generate the mapping population must differ for one or more traits of interest, as to allow further QTL/gene mapping. Population sizes used in preliminary genetic mapping studies generally range from 50 to 250 individuals (Collard et al. 2005), however larger populations are required for high-resolution mapping. If the map will be used for QTL studies (which is often the case), then an important point to note is that the mapping population must be phenotypically evaluated (*i.e.* trait data must be collected) before subsequent QTL mapping. The first maximum likelihood estimators of recombination frequency for a variety of genetic situations in BC and F_2 populations were developed in the early 1950's (Liu 1997). Linkage theory that subsidizes the construction of accurate genetic maps has been extensively dealt with in controlled crosses, and a comprehensive summary of the methods and techniques can be found in Liu, (1997), Schuster and Cruz (2008) and in excellent reviews such as Mackay (2001), Doerge (2002) and Collard et al. (2005).

Mapping populations used for mapping cross pollinating species have been derived from a cross between a heterozygous parent and a haploid or homozygous parent, or a F_1 population, developed by pair crossing heterozygous parental plants that are distinctly different for important traits (Collard et al. 2005). In cross pollinating species, genetic mapping, however, is more complicated since most of these species do not tolerate inbreeding. Linkage analysis in outbreed pedigrees is also complicated by the varying numbers of marker alleles (up to four) that may be present at each marker locus. This

situation generally gives rise to mixed segregation types (one or both parents may be heterozygous at each locus), and the linkage phases of markers are generally unknown. The information content of markers can therefore vary from one marker locus to the next, depending on the type and dominance of the marker system used and the type of mapping population (Kirst et al. 2004). These limitations were partially overcome by new mapping approaches, such as the pseudo-testcross mapping design (Grattapaglia and Sederoff 1994). This strategy take advantage of the fact that single-plant genetic linkage maps can be constructed in outbreed plant species based on single-dose markers that segregate in testcross configurations in heterozygous individuals (Kirst et al. 2004). In this design each parental derived population is treated as a traditional backcross (as the genotypes of the individuals can only be 1 or 0) and thus traditional linkage theory is used either in linkage mapping as in QTL mapping (discussed on the final part of this section). Linkage analysis of other types of crosses, *i.e.* full-sib families and half-sib families derived from highly heterozygous individuals, was first dealt with by Ott (1985); Ritter et al. (1990); Arús et al. (1994); Ritter and Salamini (1996); Maliepaard et al. (1997). Together these papers provided useful formulas for estimating recombination frequency in almost every situation. Recently, in an extensive work with full-sib families, Bhering et al. (2008) obtained estimators that differed from those obtained by Maliepaard et al. (1997), for recombination frequency of different marker configurations, by using a strictly genetic approach, *i.e.* the expected proportion of each phenotypic class in terms of recombination frequency. Based on the latter, an exogamic population mapping module was implemented in GQMOL (Cruz 2010) software, extensively used in Brazil for genetic mapping and QTL analysis.

As previously mentioned, linkage maps provided the foundation for the modern-day QTL mapping methodologies (Doerge 2002). These techniques include single-marker mapping, simple interval mapping (SIM) (Lander and Botstein 1989) and composite interval mapping (CIM) either based on Haley and

Knott (1992) regression or maximum likelihood methods (Zeng 1993, 1994). These are the main methods used to detect statistical associations between phenotype and genotype for the purpose of understanding and dissecting the regions of a genome that affect complex traits in controlled crosses (Doerge 2002; Mackay 2001). In outbreeding species, the associations between phenotype and genotype have been analyzed in either full-sib or half-sib families through techniques developed by Fulker and Cardon (1994), Hayashi and Awata (2004) and, or on random models such as those developed by Goldgar (1990), Schork (1993) and Xu and Atchley (1995). When the pseudo-testcross design is used to construct individual genetic maps (Grattapaglia and Sederoff 1994), the techniques developed for controlled crosses (backcrosses) are often used (Grattapaglia and Kirst 2008).

Chapter 1

Recent advances in microarray technology and the increasing availability of genomic information have provided an opportunity to use microarrays to scan effectively for genetic variations at the whole-genome scale, enabling the production of high-definition gene-based genetic maps. In a context where, marker technologies are undergoing a transition from predominantly serial assays that measure the size of DNA fragments to hybridization-based assays with high multiplexing levels, three hybridization-based technologies have emerged: SNP (Single Nucleotide Polymorphisms) (Fan et al. 2006; Ganal et al. 2009; Rafalski 2002), SFP (Single Feature Polymorphism) (Borevitz et al. 2003; Drost et al. 2009; Zhu and Salmeron 2007) and DArT (Diversity Arrays Technology) (Jaccoud et al. 2001; Wenzl et al. 2004). As these techniques generate dominant markers, in the future then, it is most likely that high density linkage maps will be constructed from both dominant and co-dominant markers (*e.g.* SSRs). Such maps will facilitate well-defining the genetic location of functional markers through flanking high-density co-dominant/dominant markers. Nevertheless, due to dominance, the genotype of an individual at a

dominant marker is often ambiguous, thereby increasing complexity in analysis. Consequently, the accurate estimation of recombination fractions between dominant markers and between dominant and co-dominant markers becomes important.

For F_2 with dominant markers, Tan and Fu (2007) recently improved two-point estimates by taking averages from three-point maximum likelihood estimates, whereas Jansen (2009) developed another method for ordering dominant markers by minimizing the number of recombinations between adjacent markers, as a simple alternative to multi-point maximum likelihood. Recently, in an extensive work with full-sib families, Bhering et al. (2008) obtained estimators for recombination frequency of different marker configurations, by using a strictly genetic approach, *i.e.* the expected proportion of each phenotypic class in terms of recombination frequency. Unfortunately, one particular configuration was not dealt with in the mentioned paper, since distance estimation between dominant markers segregating in a 3:1 ratio and co-dominant markers, was not taken into consideration. Here, an extension of Bhering´s work (Bhering et al. 2008) is provided, enabling the estimation of the recombination frequency between dominant markers segregating in a 3:1 ratio, and co-dominant markers in full-sib families. The estimators and algorithm were developed based on the expected frequencies for each genotypic class. These frequencies were used for building likelihood functions for each possible marker configuration.

Chapter 2

Quantitative trait loci (QTLs) mapping has been in wide use for nearly two decades during which molecular markers have become available in conjunction with interval mapping methods (Borevitz and Chory 2004). Over the past 10 years there has been a tenfold increase in the number of QTL studies published annually. The goal of QTL mapping is to determine the loci that are responsible for variation in complex, quantitative traits. In some situations, the ultimate goal

is the determination of the number, location and the interaction of these loci (Borevitz and Chory 2004). Given plentiful markers and high-throughput genotyping technologies nowadays available (Zhu and Salmeron 2007), QTL studies have been limited by the need of adequate populations and reliable phenotypic measures. Experimental design is therefore paramount. As the accuracy of locating QTL is limited by the number of recombinants that are identified based on the genotypic states of the markers, sample size and accurate genotyping becomes important. With this in mind, a commonly asked question is: *Should I genotype more markers on fewer individuals, or score more individuals (for genotype and phenotype) on fewer markers?* Recent methods of high-throughput genotyping are providing a reliable and cheap mean to genotype hundreds of individuals with an elevated number of markers, coupled with high precision. However, because observed recombinants provide the information, scoring more individuals shall address both previously mentioned concerns (Doerge 2002). Then one of the most important issues when designing experimental populations seems to be the sample size.

As backcross (BC) populations are often used as mapping populations both in self pollinating species (Collard et al. 2005; Doerge 2002), and in out-breeding species, by means of pseudo-testcross mapping design (Grattapaglia and Sederoff 1994), a study was undertaken to test the implications of population size, trait heritability, QTL position and magnitude. For that, populations of different sizes were simulated, along with three traits with different heritabilities. Each trait was set to be partially controlled by six QTLs, each explaining different proportions of the phenotypic variance (from minor to major effect, linked or unliked). The resulting mapping populations were analyzed through simple interval mapping (SIM) and composite interval mapping (CIM) to assess the concern of mapping a *ghost* QTL in place of the two linked QTLs.

Chapter 3

Genetic mapping with outbreeding species, is far more difficult than with inbreeding species, due to the number of segregating alleles per locus/parent and the unknown linkage phase of the loci (Bhering et al. 2008). There are a number of ways to circumvent these complications (Maliepaard et al. 1997). For highly heterozygous species, such as most of the forest trees (*e.g. Eucalyptus* species and hybrids), genetic maps have been developed based mainly on markers segregating in a double pseudo-testcross configuration in F_1 full-sib families (Grattapaglia and Sederoff 1994). It is now possible however, to construct a single genetic map for a full-sib family derived of a cross between two highly heterozygous individuals, based on the information of all markers and individuals, as it is usually done with populations derived from a cross between two fully homozygous diploid parents (Alves et al. 2010; Bhering et al. 2008; Maliepaard et al. 1997). Associations between phenotype and genotype have been analyzed in full-sib families either based on pseudo-testcross maps or on single full-sib maps. When the analysis is based on the pseudo-testcross maps, in most of the cases, interval mapping procedures, such as Composite Interval Mapping (CIM) (Zeng 1994), are used. When the analysis is based on a full-sib map, procedures based on Haseman and Elston (1972) regression, such as the interval mapping technique developed by Fulker and Cardon (1994) are often used.

As the procedures for linkage and QTL mapping in full-sib families of outbreeding species are quite diverse, and not readily comparable, we undertook a simulation study comparing the full-sib and the pseudo-testcross mapping designs in terms of marker ordering, distance between markers, total map size, distance variance and stress. We also investigated the power and precision of interval mapping procedures based on full-sib and on pseudo-testcross maps. We highlight the implications of population size in linkage and QTL mapping, along with the implications of trait heritability and QTL properties in QTL mapping.

References

Alves AA, Bhering LL, Cruz CD, Alfenas AC (2010) Linkage analysis between dominant and co-dominant makers in full-sib families of out-breeding species. Genetics and Molecular Biology 33:499-506

Arús P, Olarte C, Romero M, Vargas F (1994) Linkage analysis of ten isozyme genes in F segregating almond progenies. Journal of America Society of Horticulture Science 119:339-344

Bhering LL, Cruz CD, God PIVG (2008) Estimation of recombination frequency in genetic mapping of full-sib families. Pesquisa Agropecuária Brasileira 43:363-369

Borevitz JO, Chory J (2004) Genomics tools for QTL analysis and gene discovery. Curr Opin Plant Biol 7:132-136

Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Research 13:513-523

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196

Cruz CD (2010) GQMOL: a software for quantitative and genetics analysis. Universidade Federal de Viçosa, Viçosa, MG, Brazil

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43-52

Drost DR, Novaes E, Boaventura-Novaes C, Benedict CI, Brown RS, Yin TM, Tuskan GA, Kirst M (2009) A microarray-based genotyping and genetic mapping approach for highly heterozygous outcrossing species enables localization of a large fraction of the unassembled Populus trichocarpa genome sequence. Plant Journal 58:1054-1067

Fan JB, Chee MS, Gunderson KL (2006) Highly parallel genomic assays. Nat Rev Genet 7:632-644

Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. American Journal of Human Genetics 54:1092-1103

Ganal MW, Altmann T, Roder MS (2009) SNP identification in crop plants. Curr Opin Plant Biol 12:211-217

Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. American Journal of Human Genetics 47:957-967

Grattapaglia D, Kirst M (2008) *Eucalyptus* applied genomics: from gene sequences to breeding tools. New Phytologist 179:911-929

Grattapaglia D, Sederoff R (1994) Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross mapping strategy and RAPD markers. Genetics 137:1121-1137

Haley CS, Knott SA (1992) A simple regressionmethod for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324

Haseman JK, Elston RC (1972) Investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3-19

Hayashi T, Awata T (2004) Eficient method for analysis of QTL using F1 progenies in an outcrossing species. Genetica 122:173-183

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:E25

Jansen J (2009) Ordering dominant markers in F_2 populations. Euphytica 165:401-417

Kirst M, Myburg A, Sederoff R (2004) Genetic mapping in forest trees: markers, linkage analysis and genomics. In: Setlow JK (ed) Genetic Engineering, Principles and Methods Kluwer Academic/Plenum Publishers, pp 105-142

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative trait using RFLP linkage maps. Genetics 121:185-199

Liu B-H (1997) Statistical genomics: linkage, mapping and QTL analysis. CRC Press, Boca Raton, Florida

Mackay TFC (2001) The genetic architecture of quantitative traits. Annual Review of Genetics 35:303-339

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genetical Research 70:237-250

Ott J (1985) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore

Rafalski JA (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Science 162:329-333

Ritter E, Gebhardt C, Salamini F (1990) Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. Genetics 125:645-654

Ritter E, Salamini F (1996) The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. Genetical Research 67:55-65

Schork NJ (1993) Extended multipoint identy-by-descendent analysis of human quantitative traits: efficiency, power, and modeling considerations. American Journal of Human Genetics 53:1306-1393

Schuster I, Cruz CD (2008) Estatística Genômica aplicada a populações derivadas de cruzamentos controlados, 2th edn. Editora UFV, Viçosa

Tan Y-D, Fu Y-X (2007) A new strategy for estimating recombination fractions between dominant markers from an F_2 population. Genetics 175:923-931

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proceedings of the National Academy of Sciences of the United States of America 101:9915-9920

Xu S, Atchely WR (1995) A random model approach to interval mapping of quantitative trait loci. . Genetics 141:1189-1197

Zeng ZB (1993) Theoretical basis of precision mapping of quantitative trait loci. Proceedings of the National Academic of Science 90:10972-10976

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457-1468

Zhu T, Salmeron J (2007) High-definition genome profiling for genetic marker discovery. Trends in Plant Science 12:196-202

CHAPTER 1

LINKAGE ANALYSIS BETWEEN DOMINANT AND CO-DOMINANT MARKERS IN FULL-SIB FAMILIES OF OUT-BREEDING SPECIES

Paper published in Genetics and Molecular Biology 33: 499-506 (2010)

Linkage analysis between dominant and co-dominant makers in full-sib families of out-breeding species

Alexandre Alonso Alves[1], Leonardo Lopes Bhering[2], Cosme Damião Cruz[3] and Acelino Couto Alfenas[1]

[1]Department of Plant Pathology, Federal University of Viçosa, Viçosa, MG, Brazil. [2]Embrapa Agroenergy, Parque Estação Biológica, Brasília, DF, Brazil. [3]Department of General Biology, Federal University of Viçosa, Viçosa, MG, Brazil.

Send Correspondence to Acelino Couto Alfenas. Department of Plant Pathology, Federal University of Viçosa, 36571-000 Viçosa, MG, Brazil. E-mail aalfenas@ufv.br

Abstract

As high-throughput genomic tools, such as the DNA microarray platform, have led to the development of novel genotyping procedures, such as Diversity Arrays Technology (DArT) and Single Nucleotide Polymorphisms (SNPs), it is likely that, in the future, high density linkage maps will be constructed from both dominant and co-dominant markers. Recently, a strictly genetic approach was described for estimating the recombination frequency (*r*) between co-dominant markers in full-sib families. The complete set of maximum likelihood estimators for *r* in full-sib families was almost obtained, but unfortunately, one particular configuration involving dominant markers, segregating in a 3:1 ratio and co-dominant markers, was not considered. Here we add nine further estimators to the previously published set, thereby making it possible to cover all combinations of molecular markers with two to four alleles (without epistasis) in a full-sib family. This includes segregation in one or both parents, dominance and all linkage phase configurations.

Keywords: statistical genomics, exogamic populations, recombination frequency and maximum likelihood.

Introduction

The first maximum likelihood estimators of recombination frequency for a variety of genetic situations in BC_1 and F_2 populations were developed in the early 1950's. For F_2 with dominant markers, Tan and Fu (2007) recently improved two-point estimates by taking averages from three-point maximum likelihood estimates, whereas Jansen (2009) developed another method for ordering dominant markers by minimizing the number of recombinations between adjacent markers, as a simple alternative to multi-point maximum likelihood. Three-point estimates of recombination frequencies were previously used by Ridout *et al.* (1998) for out-breeding species. Nevertheless, linkage analysis of crosses with out-breeders was first dealt with by Ott (1985); Ritter *et al.* (1990); Arús *et al.* (1994); Ritter and Salamini (1996); Maliepaard *et al.* (1997). Together these papers provided useful formulas for estimating recombination frequency in almost every situation. In some cases, the formulas represent the actual estimators, whereas in others they are likelihood equations requiring implementation in numerical maximization methods, such as an EM algorithm, Newton-Raphson, or solved by a graphic method. Recently, in an extensive work with full-sib families, Bhering *et al.* (2008) obtained estimators that differed from those obtained by Maliepaard *et al.* (1997), for recombination frequency of different marker configurations, by using a strictly genetic approach, i.e. the expected proportion of each phenotypic class in terms of recombination frequency. Based on the latter, an exogamic population mapping module was implemented in GQMOL (GQMOL, 2009) software, extensively used in Brazil for genetic mapping and QTL analysis. Unfortunately, one particular configuration was not dealt with in the mentioned paper, since distance estimation between dominant markers segregating in a 3:1 ratio and co-dominant markers, was not taken into consideration. With the advent of high-throughput genomic tools,

such as the DNA microarray platform, new dominant genotyping technology has been developed, such as DArTs (Wenzl et al. 2004) and SNPs. In the future, it is most likely that high density linkage maps will be constructed from both dominant and co-dominant markers. Such maps will facilitate well-defining the genetic location of functional markers through flanking high-density co-dominant/dominant markers. Nevertheless, due to dominance, the genotype of an individual at a dominant marker is often ambiguous, thereby increasing complexity in analysis. Consequently, the accurate estimation of recombination fractions between dominant markers and between dominant and co-dominant markers, becomes important (Tan and Fu 2007).

Here, we provide an extension of Bhering´s work, which enables the estimation of the recombination frequency between dominant markers segregating in a 3:1 ratio, and co-dominant markers in full-sib families. Our estimators and algorithm were developed based on the expected frequencies for each genotypic class. These frequencies were used for building likelihood functions for each possible marker configuration. Based on intrinsic properties and their implementation in free linkage software (GQMOL, 2009), this should be of exceptional use for research groups, whose scope is mapping and the use of molecular markers for selecting monogenic traits, such as disease resistance, plant height, and early flowering, amongst other important dominant traits which are subject to breeding in out-crossing species or constructing high density genetic maps of both dominant and co-dominant markers.

Methods

Estimation of recombination frequency

In full-sib families, markers may vary in the number of segregating alleles (up to four), by one or both parents being heterozygous, markers being dominant or co-dominant, and usually the linkage phases of marker pairs are unknown. Different types of categories and crossings may occur in the general case of multi-allelic systems with four or more alleles (Haseman and Elston

1972). When considering an A locus with i, j, k and l alleles, there are seven possible types of crosses (Bhering et al. 2008), but only four are considered to be informative, since they segregate for at least one parent. Another particularity of genetic mapping in out-breeding species is that the linkage phase is not known *a priori*, as full-sib families are two generation pedigrees. Thus, one has to considerer four combinations, in order to define the correct linkage phase. Alleles might be linked by coupling to one of the parents and undefined for the other, linked by repulsion to one of the parents and undefined for the other, linked by coupling to both parents, or linked by repulsion to both parents (Maliepaard et al. 1997). Therefore, the correct linkage phase is usually determined *a posteriori* by comparing LOD scores obtained for each combination (Bhering et al. 2008).

When considering these particularities, the estimation of recombination frequency (*r*) in full-sib families may be achieved by using the maximum likelihood method. With this method, the expected frequencies for each genotypic class (p_i), which are, in turn, dependent on the recombination frequency between markers (*r*), are used to built likelihood functions [$L(r;n_i)$], which, after being maximized for *r*, give the proper estimator for recombination frequency. For this, let the genotypes of two individuals of an outbreed population for a particularly marker, be A_1A_2 and A_3A_4, respectively. If these two individuals are bred to form a full-sib family the expected segregation pattern is: $1A_1A_3$:$1A_1A_4$:$1A_2A_3$:$1A_2A_4$. Now, let the genotypes of these same two individuals be B_1B_2 and B_3B_4 for another marker. If these two individuals are also bred to form a full-sib family the expected segregation pattern is: $1B_1B_3$:$1B_1B_4$:$1B_2B_3$:$1B_2B_4$.

On considering the haplotypes for the markers in the first parent in the coupling phase, the produced gametes and their frequencies are: $f(A_1B_1) = f(A_2B_2) = (1-r)/2 = P$; $f(A_1B_2) = f(A_2B_1) = r/2 = R$, whereas for the second parent, the expected gametes and frequencies are:$f(A_3B_3) = f(A_4B_4) = (1-r)/2 = P$; $f(A_3B_4) = f(A_4B_3) = r/2 = R$.

On now considering gametes produced by these two individuals, 16 genotypic classes are to be expected in the progeny. The genotypic frequencies for these 16 classes are provided in Table S1. If one now considers that $B_1 = B_3 = B$ and $B_2 = B_4 = b$, and that BB and Bb are indistinguishable, which typically makes the B marker dominant, the estimation of recombination frequency between these two markers can be made by applying the maximum likelihood method. The likelihood function can be written as:

$L(r, ni) = \prod_{i=a}^{h} p_i^{ni}$ which is

$L(r;ni) = [N!/(n_A!....n_H!)] \times (P^2+PR+PR)^{na} \times (R^2)^{nb} \times (P^2+PR+R^2)^{nc} \times (PR)^{nd} \times (P^2+PR+R^2)^{ne} \times (PR)^{nf} \times (PR+PR+R^2)^{ng} \times (P^2)^{nh}$,

and in its simplified form as:

$L(r;ni) = \lambda(1/4-R^2)^{na}(R^2)^{nb}(1/4-PR)^{nc}(PR)^{nd}(1/4-PR)^{ne}(PR)^{nf}(1/4-P^2)^{ng}(P^2)^{nh}$

where: PP is $(1-r)^2/4$, PR is $r(1-r)/4$, RR is $r^2/4$, n_a is the total number of individuals with genotypes $A_1A_3B\_$, n_b is the total number of individuals with genotypes A_1A_3bb, n_c is the total number of individuals with genotypes $A_1A_4B\_$, n_d is the total number of individuals with genotypes A_1A_4bb, n_e is the total number of individuals with genotypes $A_2A_3B\_$, n_f is the total number of individuals with genotypes A_2A_3rr, n_g is the total number of individuals with genotypes $A_2A_4B\_$, n_h is the total number of individuals with genotypes A_2A_4bb and N is the total number of individuals.

The estimate of the recombination fraction is then obtained by the usual method of maximizing the logarithm of the likelihood function (Table 1).

However, as previously mentioned, different types of crossings may occur in a full-sib family (Haseman and Elston 1972). Thus, in order to develop general formulas for estimators of recombination frequency between dominant marker segregating in a 3:1 ratio and co-dominant makers in full-sib families, one has to consider all the different segregation patterns and linkage phases for the co-dominant marker. While the genotypes for the dominant will always be Bb (for both parents), on considering the different types of crosses mentioned above, the genotypes for the co-dominant marker may be: 2 alleles - $A_1A_1xA_1A_2$,

$A_1A_2xA_2A_2$, $A_1A_2xA_1A_2$; 3 alleles - $A_1A_1xA_2A_3$, $A_1A_2xA_3A_3$, $A_1A_2xA_1A_3$, $A_1A_2xA_2A_3$; 4 alleles - $A_1A_2xA_3A_4$.

So in order to provide an extension of Bhering´s work which would enable the estimation of recombination frequency between dominant markers segregating in a 3:1 ratio and co-dominant makers in full-sib families we have built likelihood functions to estimate the recombination frequency for each possible marker configuration based on the expected frequencies for each genotypic class as described above (Tables S2 and S3).

Average Information content and variance of recombination frequency estimators

Bias and variance are important characteristics describing how close one can get to the true value (Maliepaard et al. 1997). Variances of estimated recombination fractions can be estimated from average information content (Liu, 1997). Within that context, the general formula for estimating information content per observation for any single likelihood parameter (θ) is

$$I(\theta) = E_\theta \left[\left[\frac{\partial}{\partial\theta}\log L\,(\theta|x)\right]^2\right]$$

$$= -E_\theta \left[\frac{\partial^2}{\partial\theta^2}\log L\,(\theta|x)\right]$$

which is -1 times the expectation of the second derivative of the log likelihood function or the support function with respect to the parameter (θ).

The variance of a maximum likelihood estimate from a sample size of *N* is then:

$$\sigma^2\left(\hat{\theta}\right) = \frac{1}{N\,I(\theta)}$$

Since the variances of ML-estimators are approximately equal to the inverse of Fisher's information, i.e. the expectation of minus the second derivative of the log-likelihood function (Maliepaard et al. 1997; Schuster and Cruz 2008), we used this approach to obtain the respective functions.

Algorithm integration in GQMOL and mapping procedures

A computer algorithm capable of recognizing the different types of crosses, segregation and linkage phases, and of calculating recombination frequency between dominant markers, as well as the co-dominant markers linked to it based on the likelihood functions here described, was implemented into GQMOL software (GQMOL, 2009). This first requires the construction of an integrated linkage map without the dominant marker, according to traditional methods as described by Ott (1985); Ritter *et al.* (1990); Arús *et al.* (1994); Ritter and Salamini (1996); Maliepaard *et al.* (1997) and Bhering *et al.* (2008). Recombination frequency between the dominant marker and the previously mapped co-dominant marker, according to the likelihood functions here described, is then calculated (see results section). In order to define the correct linkage phase, recombination frequencies are estimated for each of the possible phases predicted in Table S3, and then compared in terms of LOD scores. By comparing scores, the algorithm determines the correct linkage phase, and, in turn, the correct recombination frequency, by identifying the phase and the associated *r* that reached the highest LOD score. After determining the recombination frequency between dominant marker and each of the co-dominant markers, its position on the previously constructed linkage map is defined by traditional alignment methods, such as SARF (Sum of Adjacent Recombination Frequencies) and RCD (Rapid Chain Delineation).

Simulation design and testing

Two hundred (200) individuals segregating for 30 loci were generated according to Mendelian inheritance at a given recombination frequency. The simulated genome consisted of 30 markers distributed at an equal distance throughout three linkage groups. Parents were generated randomly, with four alleles in equal frequency – 25%, and markers segregated in various configurations (Haseman and Elston 1972). To build the simulated map, recombination frequency and LOD scores were calculated using formulas as

described by Bhering *et al.* (2008). So as to test the algorithm, data of one specific marker derived from cross A_1A_2 x A_1A_2 was later re-coded as a dominant marker. Considering that the A1 allele is dominant, data for individuals of genotypes A_1A_1 and A_1A_2 were retyped as 4, and for individuals A_2A_2 were retyped as 2 (4 and 2 are the codes used in GQMOL for the genotypes A_ and aa, respectively). An integrated map without this marker was constructed, as described by Bhering *et al.* (2008). Linkage analysis between the dominant and co-dominant markers was then undertaken, using the functions as presented in Table 1. Comparisons between the *simulated-map* and *algorithm-map* were carried out in terms of marker ordering, distance between markers, total map size, distance variance and stress, in order to evaluate whether the algorithm was efficient as a mapping procedure for dominant markers in full-sib families. The GQMOL simulation module was used for analysis. Simulation was based on 1000 population replicates.

Results

The genotypic frequencies expected for each marker configuration/linkage phase combination, including those predicted by Haseman and Elston (1972), are given in Table S3. Likelihood functions, as well as estimators of recombination frequency between dominant and co-dominant markers, for all types of crosses and segregations in full-sib families of out-breeding species, are given in Table 1. For practical purposes, it is noteworthy that estimators, which are mainly complex polynomials, have a limited value due to their high degree. However, with GQMOL, it is possible to circumvent this limitation by using a graphic method, so that *r* is calculated directly from likelihood functions. Hence, different values are attributed to *r* (in the 0 to 0.5 interval), and LOD score areas calculated for each value. By plotting these scores on a graph having *r* values in its x-coordinate, and LOD scores in the y-coordinate, the highest LOD score is identified on the graph, and the corresponding *r* value on the abscissa (Schuster and Cruz 2008).

The average information content functions relative to all marker configurations involving dominant markers and co-dominant markers in full-sib families of out-breeding species, i.e. different types of crosses, linkage phases, marker configurations and segregations, is presented in Table 2. These functions are useful for evaluating the accuracy of recombination frequency by means of the variance of the estimates. Figure 1 shows that the combinations of dominant and co-dominant markers in configurations 6, 7, 8 and 9 provided a relatively large amount of information. These configurations represent crosses between heterozygous individuals which, according to Haseman and Elston (1972), are the most informative (Bhering et al. 2008). As to co-dominant markers in configurations 1, 2, 3, 4 and 5 (some of which are equivalent and have the same information content function), the functions provided relatively little information. As in configurations 1 and 2, half the progeny is absolutely noninformative, the low information content was indeed expected. Nevertheless, although these latter configurations of dominant and co-dominant markers appear to provide little information, the variance of its estimators was quit low. The variances of estimated recombination frequencies (0.05, 0.10 and 0.20), relative to all marker configurations involving dominant markers and co-dominant markers in full-sib families of out-breeding species and different population size, are given in Table 3. Here, one can observe that the highest efficiency is achieved for completely informative co-dominant markers and crosses (configurations 6, 7, 8 and 9), independent of map saturation, and that with adequate population sizes (≥ 200 individuals), even non-completely informative co-dominant markers, together with dominant markers, may be used for constructing maps. However, if expectation is to obtain a less saturated map, ideally only co-dominant markers in configurations 6, 7, 8 and 9 should be selected, in order to correctly map the dominant markers.

The algorithm was tested through simulation. The *simulated map* is presented in Figure 2A. Data of one specific locus (marker number 5), derived from cross type $A_1A_2xA_1A_2$, and that segregated in a 1:2:1 ratio as evaluated by a

chi-square (χ^2) test, was then re-coded as a dominant marker, as previously described. As expected, linkage analysis without marker 5 data generated a map without the marker itself (data not shown). The linkage map generated with our algorithm and showing marker 5, therein denominated B correctly located in linkage group 1, is shown in Figure 2B. Comparisons between the *simulated-map* and *algorithm-map* indicated that only linkage group 1 was affected, since linkage groups 2 and 3 remained exactly the same on both maps. This shows that the algorithm did not disturb the alignment of the non-involved linkages groups. Linkage group 1 of the *simulated genome* was 100.82 cM long, whereas the algorithm-based map was 100.98 cM. Marker ordering remained unaltered on the *algorithm map*, with a mean marker distance of 12.63 cM, while on the *simulated map*, the mean distance between markers was 12.60 cM. Map variance increased from 15.97 on the *simulated map* to 17.66 on the *algorithm-based*. Spearman correlation, which measures map ordering consistence, was near 1, thereby indicating that the algorithm, and, in turn, the functions and estimators, were efficient in locating dominant markers. On the other hand, Pearson correlation, which measures correlations between marker distances, was 0.93, thereby also indicating the efficiency of both algorithm and formulas. However, as can be seen in Figures 2A and 2B, the distances between the so called B marker and the 4 and 6 markers are slightly different from those estimated between marker 5 and 4 and 6 on the *simulated map*.

Discussion

Since most of the computer packages used for genetic mapping are not capable of analyzing out-breeding populations, with the exception of JoinMap (Stam, 1993), over the past years, we have been developing a free genetic software named GQMOL (GQMOL, 2009), apt at analyzing, through genetic mapping, QTL mapping and simulation, not only controlled crosses, but also full-sib and half-sib families. So as to implement an out-breeding population mapping module in GQMOL, Bhering *et al.* (2008) developed likelihood functions and

estimators for different marker configurations. However, GQMOL was still inept at estimating the distance between dominant and co-dominant markers. Here, we provide an extension of Bhering´s work, apt at estimating recombination frequency between a dominant marker segregating in a 3:1 ratio and co-dominant markers in full-sib families. Likelihood functions, used for estimating recombination frequency between the dominant marker and co-dominant markers for each possible marker configuration predicted by Haseman and Elston (1972), were built based on the expected frequencies for each genotype class in a strictly genetic approach. By maximizing the natural logarithm of the log-likelihood functions, the estimators for the recombination frequency between the two markers were obtained. It is noteworthy that our estimators (including those presented in Bhering *et al.* 2008) are quite different from those obtained by Maliepaard *et al.* (1997). These differences are due to the fact that we have applied a strictly genetic approach, rather than a genetic-statistical approach (iterative procedure - EM algorithm) as used by Maliepaard *et al.* (1997). Both methods appear to be equivalent, since the same data packages analyzed by JoinMap and GQMOL resulted in nearly alike integrated maps (AA Alves – unpublished data). However, in situations where the likelihood function is very flat (i.e., the data provide little information due to dominance and markers being in the repulsion phase), the estimates obtained by the EM algorithm may depend on the starting value for recombination frequency. An overall view of likelihood through graphic procedures, or the explicit likelihood function solution, could possibly give rise to recombination frequency associated with the true maximum in a more reliable way. Our method, apart from being simple, may then be more applicable to a wider range of situations than the methods currently available.

A simple simulation approach was chosen to test our algorithm. A simulated full-sib family was generated for the purpose, and data from one specific marker re-coded for dominance, followed by linkage analyses with our algorithm. The dominant marker was correctly located in the linkage map

generated with the algorithm, and Spearman and Pearson correlations indicated its efficiency in locating the dominant marker without disturbing nearby markers or other linkage groups. Nevertheless, we noticed that the distances between the dominant marker and flanking markers were slightly different from those previously obtained between marker 5 and markers 4 and 6. This was probably due to the loss of information with re-coded data. Whereas three genotypic classes (2 heterozygotes and one homozygote) can be analyzed with co-dominant markers, with dominant markers one can analyze only two (dominant and recessive). This may have affected estimates of recombination frequencies, thereby resulting in different map distances. However, for practical purposes, *e.g.*, MAS – marker assisted selection, bias in distance is not expected to be a problem. Traditional mapping strategies based on co-dominant markers also locate markers near their real position, with an expected bias (Schuster and Cruz 2008). Our algorithm then, proved to be very fast and precise, and its only prior requirement is a linkage map without the dominant marker constructed following traditional methods as described by Bhering *et al.* (2008) or Maliepaard *et al.* (1997).

As to the accuracy of estimates, it has long been recognized that dominant markers in the repulsion linkage phase supply low linkage information content in F_2 populations. Nowadays, this problem is receiving additional attention, as high-throughput genomic tools, such as the DNA microarray platform, have lead to the development of up-to-date genotyping procedures resulting in new dominant markers. Novel methods for mapping such markers circumventing this issue have been described (Jansen 2009; Tan and Fu 2007). Nevertheless, in full-sib families of out-breeding species, dominant markers appear to be unimpeachable, if used together with co-dominant markers. Our variances estimates for three distinct values of recombination frequency (0.05, 0.10 and 0.20), in all marker configurations involving dominant markers and co-dominant markers in full-sib families of out-breeding species and different population size indicates that variances of recombination frequency estimates

are very low, ranging from $0.060878318 \times 10^{-4}$ for completely informative markers in a large population (n=1000) to 8.816327×10^{-4} for partially informative markers in a small population (n=100). These values are very similar to the estimates obtained from co-dominant markers in F_2 populations, and considerable lower when compared to estimates from both co-dominant and dominant markers in F_2. For example, for recombination frequencies of 0.05, 0.10 and 0.20, variance estimates for co-dominant markers in an F_2 of 200 individuals were 1.25×10^{-4}, 2.53×10^{-4} and 5.23×10^{-4}, respectively (Liu 1997; Schuster and Cruz 2008). The variance estimates for co-dominant and dominant markers in the very same F_2 were 2.47×10^{-4}, 4.91×10^{-4} and 9.69×10^{-4}, respectively, (Liu 1997; Schuster and Cruz 2008). As recombination frequency estimator variance is comprised of two main components, viz., the number of recombination events that created the progeny sample and the (in) ability with which these events can be detected for a certain configuration of two loci, it is reasonable to speculate that the first is defined by recombination frequency itself and progeny size, and the second by the segregation types of loci and linkage phases in the parents (Maliepaard et al. 1997). Hence, although the particularities of out-breeding species (number of segregating alleles and different linkage phases) represent an enormous challenge for genetic mapping, these may, on the other hand, contribute to more accurate estimates of recombination frequency.

Finally, it is noteworthy that Bhering *et al.* (2008) nearly obtained the complete set of maximum likelihood estimators for recombination frequency between molecular markers in full-sib families. With the addition of a further nine, all combinations of molecular markers with two to four alleles (without epistasis) in a full-sib family are now accounted for. This includes segregation in one or both parents, dominance and all linkage phase configurations. In summary, by this paper and Bhering *et al.* (2008) an overview of the whole range of situations of molecular markers in crosses with out-breeding species (full-sib families), has been presented from a genetic perspective. Based on its properties and implementation into free linkage software, our approach should be useful

for those interested in using molecular markers for mapping, or as an aid in selecting out-crossing species.

Acknowledgements

We are grateful to Phil Cannon, for his constructive comments on the manuscript. The Bioinformatics Lab of the Federal University of Viçosa, Brazil provided the facilities for the development of this work. This work was also supported by the Brazilian National Research Council, CNPq, with a Ph.D. fellowship to AAA and a research fellowship to ACA and CDC.

References

Arús P, Olarte C, Romero M, Vargas F (1994) Linkage analysis of ten isozyme genes in F segregating almond progenies. Journal of America Society of Horticulture Science 119:339-344

Bhering LL, Cruz CD, God PIVG (2008) Estimation of recombination frequency in genetic mapping of full-sib families. Pesquisa Agropecuária Brasileira 43:363-369

Haseman JK, Elston RC (1972) Investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3-19

Jansen J (2009) Ordering dominant markers in F_2 populations. Euphytica 165:401-417

Liu B-H (1997) Statistical genomics: linkage, mapping and QTL analysis. CRC Press, Boca Raton, Florida

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genetical Research 70:237-250

Ott J (1985) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore

Ridout MS, Tong S, Vowden CJ, Tobutt KR (1998) Three point linkage analysis in crosses of allogamous plant species. Genet Res 72:111-121

Ritter E, Gebhardt C, Salamini F (1990) Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. Genetics 125:645-654

Ritter E, Salamini F (1996) The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. Genetical Research 67:55-65

Schuster I, Cruz CD (2008) Estatística Genômica aplicada a populações derivadas de cruzamentos controlados, 2th edn. Editora UFV, Viçosa

Tan Y-D, Fu Y-X (2007) A new strategy for estimating recombination fractions between dominant markers from an F_2 population. Genetics 175:923-931

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proceedings of the National Academy of Sciences of the United States of America 101:9915-9920

Internet Resources

GQMOL (2009) Quantitative and Molecular Genetics Software. (October 9, 2009) http://www.ufv.br/dbg/gqmol/gqmol.htm.

Supplementary material

The following online material is available for this article:

- Table S1 Genotypic frequencies for a progeny derived from a cross between two fully informative co-dominant markers linked in coupling with four alleles

- Table S2 Probability classes and their respective estimates used in likelihood functions

- Table S3 Genotypic frequencies for progenies derived from crosses between different types of co-dominant markers and a dominant marker for different linkage phases.

This material is made available as part of the online article from http://www.scielo.br.gmb

Table 1. Likelihood functions and expressions for calculating recombination frequency between dominant and co-dominant markers in full-sib families of out-breeding species (different types of crosses, linkage phases – LP and segregations are considered).

Crosses	***LP***	***MC***	***Likelihood functions***	***Estimators***
$A_1A_1xA_1A_2$ $A_1A_1xA_2A_3$	*C*	*1*	$L(r;i) = \lambda(1/4+P/2)^a(R/2)^b(1/4+R/2)^c(P/2)^d$	$r^3 \cdot(N) - r^2 \cdot(2\cdot b + 3\cdot c + d) - r\cdot(a + b - 2\cdot(c - d)) + 2\cdot b = 0$
$A_1A_2xA_2A_2$ $A_1A_2xA_3A_3$	*R*	*2*	$L(r;i) = \lambda(1/4+R/2)^a(P/2)^b(1/4+P/2)^c(R/2)^d$	$r^3 \cdot(N) - r^2 \cdot(3\cdot a + b + 2\cdot d) + r\cdot(2\cdot a - 2\cdot b - c - d) + 2\cdot d = 0$
$A_1A_2xA_1A_2$	*C*	*3*	$L(r;i) = \lambda(1/4-R^2)^A(R^2)^b(1/4+P^{2+}R^2)^c(2PR)^d(1/4-P^2)^e(P^2)^f$	$2\cdot r^7 \cdot(N) - r \cdot(2\cdot a + 2\cdot b + c + d + 4\cdot f) - 2\cdot r^6 \cdot(4\cdot a + 5\cdot b + 6\cdot c + 5\cdot d + 4\cdot e + f) + r^5 \cdot(14\cdot a + 16\cdot b + 10\cdot c + 11\cdot d + 2\cdot(5\cdot e + 3\cdot f)) - r^4 \cdot(14\cdot a + 6\cdot b - 8\cdot c + 3\cdot d + 2\cdot(e + 2\cdot f)) + r^3 \cdot(4\cdot a - 10\cdot b - 9\cdot c - 2\cdot(5\cdot d + 4\cdot e + f)) + r^2\cdot(14\cdot b + 2\cdot c + 9\cdot d + 4\cdot(2\cdot e + f)) - 2\cdot(2\cdot b + d + e) = 0$
	C-R	*4*	$L(r;i) = \lambda(1/4-PR)^a(PR)^b(1/4+2PR)^c(P^{2+}R^2)^d(1/4-PR)^e(PR)^f$	$(2\cdot r - 1)\cdot(2\cdot r\ 4\cdot(N) - 4\cdot r\ 3\cdot(N) + r\ 2\cdot(3\cdot a + 5\cdot b + 4\cdot c + 4\cdot d + 3\cdot e + 5\cdot f) - r\cdot(a + 3\cdot b + 2\cdot c + 2\cdot d + e + 3\cdot f) + b + f) = 0$
	R	*5*	$L(r;i) = \lambda(1/4-P^2)^a(P^2)^b(1/4+P^{2+}R^2)^c(2PR)^d(1/4-R^2)^e(R^2)^f$	$2\cdot r^7 \cdot(N) - r^6 \cdot(4\cdot b + c + d + 2\cdot(e + f)) - 2\cdot r^5 \cdot(4\cdot a + b + 6\cdot c + 5\cdot d + 4\cdot e + 5\cdot f) + r^4 \cdot(10\cdot a + 6\cdot b + 10\cdot c + 11\cdot d + 2\cdot(7\cdot e + 8\cdot f)) - r^3 \cdot(2\cdot a + 4\cdot b - 8\cdot c + 3\cdot d + 2\cdot(7\cdot e + 3\cdot f)) - r^2 \cdot(8\cdot a + 2\cdot b + 9\cdot c + 2\cdot(5\cdot d - 2\cdot e + 5\cdot f)) + r\cdot(8\cdot a + 4\cdot b + 2\cdot c + 9\cdot d + 14\cdot f) - 2\cdot(a + d + 2\cdot f) = 0$

	C	*6*	$L(r;i) = \lambda(1/4-R^2)^a(R^2)^b(1/4-PR)^c(PR)^d (1/4-PR)^e(PR)^f(1/4-P^2)^g(P^2)^h$	$2\cdot r^7\cdot(N) - r^6\cdot(2\cdot a + 2\cdot b + c + d + e + f + 4\cdot h) - 2\cdot r^5\cdot(4\cdot a + 5\cdot b + 6\cdot c + 5\cdot d + 6\cdot e + 5\cdot f + 4\cdot g + h) + r^4\cdot(14\cdot a + 16\cdot b + 10\cdot c + 11\cdot d + 10\cdot e + 11\cdot f + 2\cdot(5\cdot g + 3\cdot h)) - r^3\cdot(14\cdot a + 6\cdot b - 8\cdot c + 3\cdot d - 8\cdot e + 3\cdot f + 2\cdot(g + 2\cdot h)) + r^2\cdot(4\cdot a - 10\cdot b - 9\cdot c - 10\cdot d - 9\cdot e - 2\cdot(5\cdot f + 4\cdot g + h)) + r\cdot(14\cdot b + 2\cdot c + 9\cdot d + 2\cdot e + 9\cdot f + 4\cdot(2\cdot g + h)) - 2\cdot(2\cdot b + d + f + g) = 0$
$A_1A_2xA_1A_3$ $A_1A_2xA_2A_3$ $A_1A_2xA_3A_4$	*C-R*	*7*	$L(r;i) = \lambda(1/4-PR)^a(PR)^b(1/4-R^2)^c(R^2)^d (1/4- P^2)^e(P^2)^f(1/4-PR)^g(PR)^h$	$2\cdot r^7\cdot(N) - r^6\cdot(a + b + 2\cdot c + 2\cdot d + 4\cdot f + g + h) - 2\cdot r^5\cdot(6\cdot a + 5\cdot b + 4\cdot c + 5\cdot d + 4\cdot e + f + 6\cdot g + 5\cdot h) + r^4\cdot(10\cdot a + 11\cdot b + 14\cdot c + 16\cdot d + 10\cdot e + 6\cdot f + 10\cdot g + 11\cdot h) + r3\cdot(8\cdot a - 3\cdot b - 14\cdot c - 6\cdot d - 2\cdot e - 4\cdot f + 8\cdot g - 3\cdot h) - r^2\cdot(9\cdot a + 10\cdot b - 4\cdot c + 10\cdot d + 8\cdot e + 2\cdot f + 9\cdot g + 10\cdot h) + r\cdot(2\cdot a + 9\cdot b + 14\cdot d + 8\cdot e + 4\cdot f + 2\cdot g + 9\cdot h) - 2\cdot(b + 2\cdot d + e + h) = 0$
	R-C	*8*	$L(r;i) = \lambda(1/4-PR)^a(PR)^b(1/4-P^2)^c(P^2)^d (1/4- R^2)^e(R^2)^f(1/4-PR)^g(PR)^h$	$2\cdot r^7\cdot(N) - r^6\cdot(a + b + 4\cdot d + 2\cdot e + 2\cdot f + g + h) - 2\cdot r^5\cdot(6\cdot a + 5\cdot b + 4\cdot c + d + 4\cdot e + 5\cdot f + 6\cdot g + 5\cdot h) + r^4\cdot(10\cdot a + 11\cdot b + 10\cdot c + 6\cdot d + 14\cdot e + 16\cdot f + 10\cdot g + 11\cdot h) + r^3\cdot(8\cdot a - 3\cdot b - 2\cdot c - 4\cdot d - 14\cdot e - 6\cdot f + 8\cdot g - 3\cdot h) - r^2\cdot(9\cdot a + 10\cdot b + 8\cdot c + 2\cdot d - 4\cdot e + 10\cdot f + 9\cdot g + 10\cdot h) + r\cdot(2\cdot a + 9\cdot b + 8\cdot c + 4\cdot d + 14\cdot f + 2\cdot g + 9\cdot h) - 2\cdot(b + c + 2\cdot f + h) = 0$
	R	*9*	$L(r;i) = \lambda(1/4-P^2)^a(P^2)^b(1/4-PR)^c(PR)^d (1/4-PR)^e(PR)^f(1/4-R^2)^g(R^2)^h$	$2\cdot r^7\cdot(N) - r^6\cdot(4\cdot b + c + d + e + f + 2\cdot(g + h)) - 2\cdot r^5\cdot(4\cdot a + b + 6\cdot c + 5\cdot d + 6\cdot e + 5\cdot f + 4\cdot g + 5\cdot h) + r^4\cdot(10\cdot a + 6\cdot b + 10\cdot c + 11\cdot d + 10\cdot e + 11\cdot f + 2\cdot(7\cdot g + 8\cdot h)) - r^3\cdot(2\cdot a + 4\cdot b - 8\cdot c + 3\cdot d - 8\cdot e + 3\cdot f + 2\cdot(7\cdot g + 3\cdot h)) - r^2\cdot(8\cdot a + 2\cdot b + 9\cdot c + 10\cdot d + 9\cdot e + 2\cdot(5\cdot f - 2\cdot g + 5\cdot h)) + r\cdot(8\cdot a + 4\cdot b + 2\cdot c + 9\cdot d + 2\cdot e + 9\cdot f + 14\cdot h) - 2\cdot(a + d + f + 2\cdot h) = 0$

Table 2. Information content functions relative to all marker configurations involving dominant and co-dominant markers in full-sib families of out-breeding species (different types of crosses, linkage phases – LP, marker configurations -MC and segregations are considered).

Crosses	*LP*	*MC*	*Function*
$A_1A_1xA_1A_2$ $A_1A_1xA_2A_3$ $A_1A_2xA_2A_2$	*C*	*1*	$-[12\cdot r^2 - 12\cdot r - 2] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)]$
$A_1A_2xA_3A_3$	*R*	*2*	$-[12\cdot r^2 - 12\cdot r - 2] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)]$
	C	*3*	$-[84\cdot r^6 - 60\cdot r^5 - 250\cdot r^4 + 268\cdot r^3 - 63\cdot r^2 - 70\cdot r + 37] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)\cdot(r^2 - r + 1)\cdot(r^2 + 2\cdot r - 1)]$
$A_1A_2xA_1A_2$	*C-R*	*4*	$-[120\cdot r^4 - 240\cdot r^3 + 216\cdot r^2 - 96\cdot r + 16] / [r\cdot(r-1)\cdot(r^2 - r + 1)\cdot(2\cdot r^2 - 2\cdot r + 1)]$
	R	*5*	$-[84\cdot r^6 - 60\cdot r^5 - 250\cdot r^4 + 268\cdot r^3 - 63\cdot r^2 - 70\cdot r + 37] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)\cdot(r^2 - r + 1)\cdot(r^2 + 2\cdot r - 1)]$
	C	*6*	$-[4\cdot(28\cdot r^6 - 18\cdot\text{^}^5 - 90\cdot r^4 + 88\cdot r^3 - 12\cdot r^2 - 27\cdot r + 12)] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)\cdot(r^2 - r + 1)\cdot(r^2 + 2\cdot r - 1)]$
$A_1A_2xA_1A_3$ $A_1A_2xA_2A_3$ $A_1A_2xA_3A_4$	*C-R*	*7*	$-[112\cdot r^6 - 72\cdot r^5 - 360\cdot r^4 + 352\cdot r^3 - 48\cdot r^2 - 108\cdot r + 48] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)\cdot(r^2 - r + 1)\cdot(r^2 + 2\cdot r - 1)]$
	R-C	*8*	$-[112\cdot r^6 - 72\cdot r^5 - 360\cdot r^4 + 352\cdot r^3 - 48\cdot r^2 - 108\cdot r + 48] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)\cdot(r^2 - r + 1)\cdot(r^2 + 2\cdot r - 1)]$
	R	*9*	$-[4\cdot(28\cdot r^6 - 18\cdot r^5 - 90\cdot r^4 + 88\cdot r^3 - 12\cdot r^2 - 27\cdot r + 12)] / [r\cdot(r+1)\cdot(r-1)\cdot(r-2)\cdot(r^2 - r + 1)\cdot(r^2 + 2\cdot r - 1)]$

Table 3. Variance of estimated recombination frequencies relative to all marker configurations involving dominant and co-dominant markers in full-sib families of out-breeding species and population size.

Marker configuration	***Population size (n)***				
r = 0.05	***100***	***200***	***400***	***800***	***1000***
1 and *2***	3.78429*	1.892145	0.946072	0.473036	0.378428988
3 and *5*	0.249117	0.124558	0.062279	0.03114	0.024911692
4	0.349641	0.174821	0.08741	0.043705	0.034964109
6, 7, 8 and *9*	0.195527	0.097763	0.048882	0.024441	0.019552669
r = 0.1	**100**	**200**	**400**	**800**	**1000**
1 and *2*	6.107143	3.053571	1.526786	0.763393	0.610714286
3 and *5*	0.456649	0.228324	0.114162	0.057081	0.045664893
4	0.806025	0.403012	0.201506	0.100753	0.080602496
6, 7, 8 and *9*	0.365124	0.182562	0.091281	0.04564	0.036512396
r = 0.2	**100**	**200**	**400**	**800**	**1000**
1 and *2*	8.816327	4.408163	2.204082	1.102041	0.881632653
3 and *5*	0.731963	0.365981	0.182991	0.091495	0.073196286
4	2.462069	1.231034	0.615517	0.307759	0.246206897
6, 7, 8 and *9*	0.608783	0.304392	0.152196	0.076098	0.060878318

*Values were multiplied by 10^4.

**Configuration 1 refers to crosses $A_1A_1xA_1A_2$; $A_1A_1xA_2A_3$; $A_1A_2xA_2A_2$; $A_1A_2xA_3A_3$ in coupling; configuration 2, to crosses $A_1A_1xA_1A_2$; $A_1A_1xA_2A_3$; $A_1A_2xA_2A_2$; $A_1A_2xA_3A_3$ in repulsion; configuration 3 to cross in A1A2xA1A2 coupling, configuration 4 to cross in $A_1A_2xA_1A_2$ coupling-repulsion; configuration 5 to cross in $A_1A_2xA_1A_2$; configuration 6 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in coupling; configuration 7 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in coupling-repulsion; configuration 8 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in repulsion-coupling and configuration 9 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in repulsion.

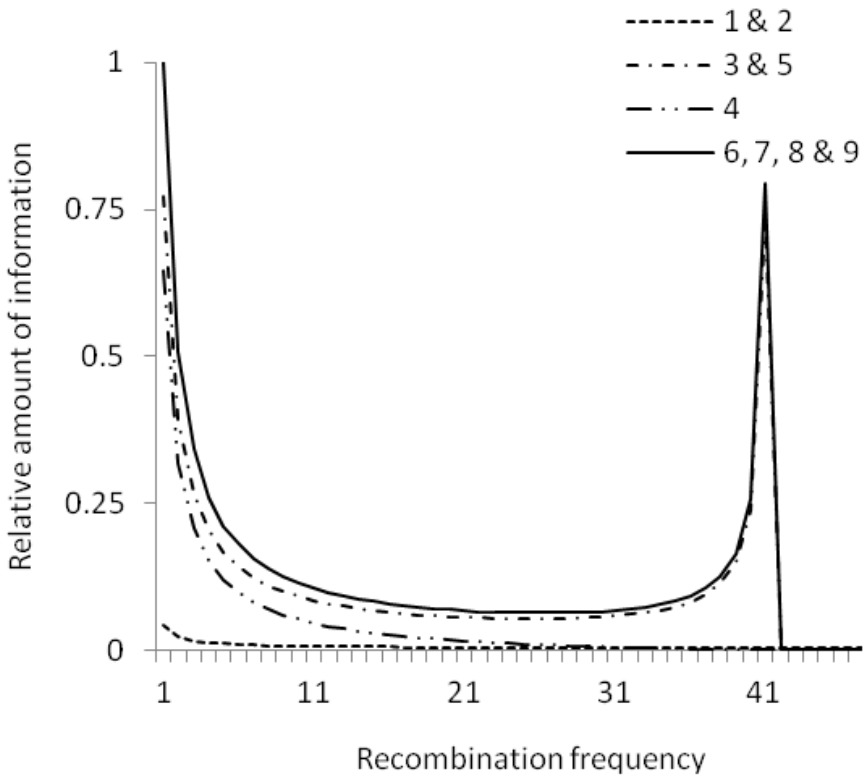

Figure 1. Information content functions relative to all marker configurations involving dominant markers and co-dominant markers in full-sib families of out-breeding species. Configuration 1 refers to crosses $A_1A_1xA_1A_2$; $A_1A_1xA_2A_3$; $A_1A_2xA_2A_2$; $A_1A_2xA_3A_3$ in coupling; configuration 2, to crosses $A_1A_1xA_1A_2$; $A_1A_1xA_2A_3$; $A_1A_2xA_2A_2$; $A_1A_2xA_3A_3$ in repulsion; configuration 3 to cross in A1A2xA1A2 coupling, configuration 4 to cross in $A_1A_2xA_1A_2$ coupling-repulsion; configuration 5 to cross in $A_1A_2xA_1A_2$; configuration 6 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in coupling; configuration 7 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in coupling-repulsion; configuration 8 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in repulsion-coupling and configuration 9 to crosses $A_1A_2xA_1A_3$; $A_1A_2xA_2A_3$ and $A_1A_2xA_3A_4$ in repulsion.

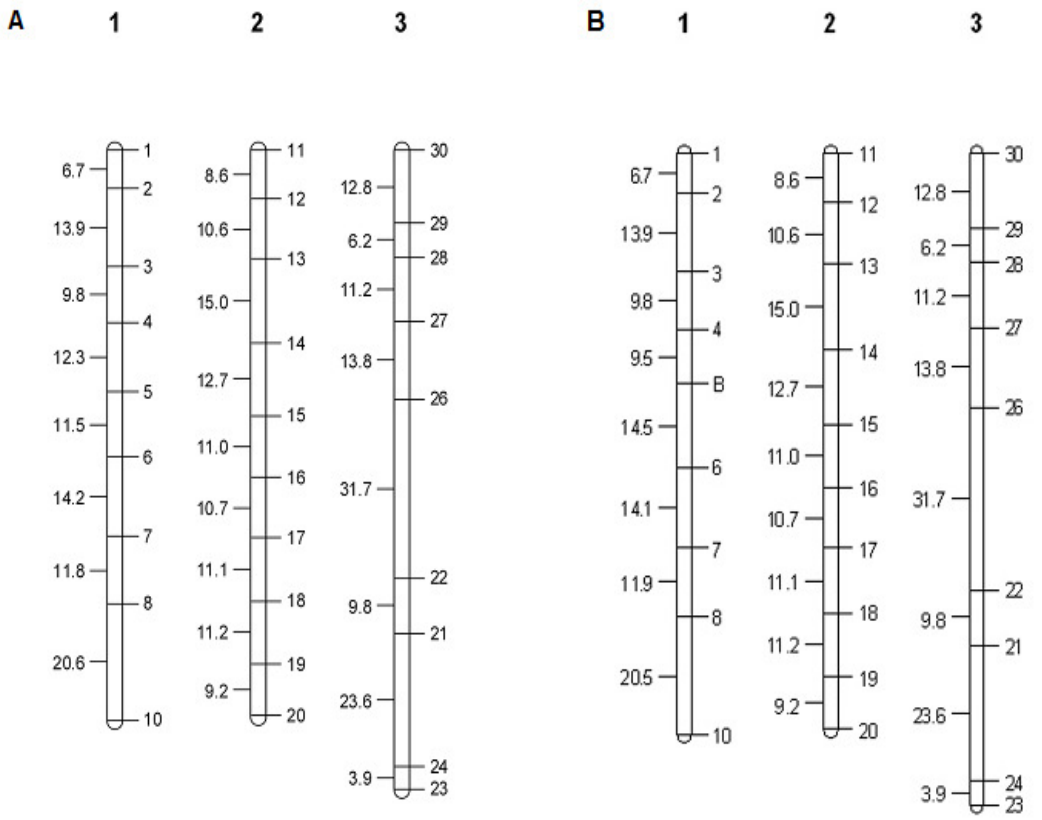

Figure 2. A - simulated genetic map of a full-sib family consisting of three linkage groups and 30 co-dominant markers. **B** - algorithm-based map of a simulated full-sib family showing the correctly located dominant marker (Marker B – which corresponds to marker 5 in the simulated map).

Supplementary Material

Table S 1. Genotypic frequencies for a progeny derived from a cross between two fully informative co-dominant markers linked in coupling with four alleles *.

Individuals	***Class***	***Genotypic frequency***
A_1A_3 B_1B_3	PP	$(1-r)^2/4$
A_1A_3 B_1B_4	PR	$r(1-r)/4$
A_1A_4 B_1B_3	PR	$r(1-r)/4$
A_1A_4 B_1B_4	PP	$(1-r)^2/4$
A_1A_3 B_2B_3	PR	$r(1-r)/4$
A_1A_3 B_2B_4	RR	$r^2/4$
A_1A_4 B_2B_3	RR	$r^2/4$
A_1A_4 B_2B_4	PR	$r(1-r)/4$
A_2A_3 B_1B_3	PR	$r(1-r)/4$
A_2A_3 B_1B_4	RR	$r^2/4$
A_2A_4 B_1B_3	RR	$r^2/4$
A_2A_4 B_1B_4	PR	$r(1-r)/4$
A_2A_3 B_2B_3	PP	$(1-r)^2/4$
A_2A_3 B_2B_4	PR	$r(1-r)/4$
A_2A_4 B_2B_3	PR	$r(1-r)/4$
A_2A_4 B_2B_4	PP	$(1-r)^2/4$

*P= (1-r)/2 ; R= r/2; P+R = 0,5

Table S 2. Probability classes and their respective estimates used in likelihood functions*.

Probabilities	***Estimates***
P^2	$(1-r)^2/4$
R^2	$r^2/4$
P/2	(1-r)/4
R/2	r/4
PR	$(r- r^2)/4$
2PR	$(r- r^2)/2$
¼ - P^2	$(2r - r^2)/4$
¼ - R^2	$(1- r^2)/4$
¼ - PR	$(r^2 – r +1)/4$
¼ + P/2	(2-r)/4
¼ + R/2	(r+1)/4
¼ + 2PR	$(2r^2 -2r -1)/4$
PR + PR + P^2	$(1- r^2)/4$
PR + PR + R^2	$(2r - r^2)/4$
¼ + $P^{2\,+}R^2$	$(r^2 -r +1)/4$
$P^2 + R^2$	$(2r^2 -2r +1)/4$

*P= (1-r)/2 ; R= r/2; P+R = 0,5

Table S 3. Genotypic frequencies for progenies derived from crosses between different types of co-dominant markers (A locus) and a dominant marker (B locus) for different linkage phases. (In each cross both parents are heterozygous for B locus).

Cross	*Segregation*	*Coupling*	*Cou-Rep*	*Rep-Cou*	*Repulsion*
$A_1A_1xA_1A_2$	$A_1A_1B\_$	¼ +P/2	-	-	¼ +R/2
	A_1A_1bb	R/2	-	-	P/2
	$A_1A_2B\_$	¼ + R/2	-	-	¼ + P/2
	A_1A_2bb	P/2	-	-	R/2
$A_1A_1xA_2A_3$	$A_1A_2B\_$	¼ +P/2	-	-	¼ +R/2
	A_1A_2bb	R/2	-	-	P/2
	$A_1A_3B\_$	¼ + R/2	-	-	¼ + P/2
	A_1A_3bb	P/2	-	-	R/2
$A_1A_2xA_2A_2$	$A_1A_2B\_$	¼ +P/2	-	-	¼ +R/2
	A_1A_2bb	R/2	-	-	P/2
	$A_2A_2B\_$	¼ + R/2	-	-	¼ + P/2
	A_2A_2bb	P/2	-	-	R/2
$A_1A_2xA_3A_3$	$A_1A_3B\_$	¼ +P/2	-	-	¼ +R/2
	A_1A_3bb	R/2	-	-	P/2
	$A_2A_3B\_$	¼ + R/2	-	-	¼ + P/2
	A_2A_3bb	P/2	-	-	R/2
$A_1A_2xA_1A_2$	$A_1A_1B\_$	¼ - R^2	¼ - PR	-	¼ - P^2
	A_1A_1bb	R^2	PR	-	P^2
	$A_1A_2B\_$	¼ + P^2 + R^2	¼ + 2PR	-	¼ + P^2 + R^2
	A_1A_2bb	2PR	P^2 + R^2	-	2PR
	$A_2A_2B\_$	¼ - P^2	¼ - PR	-	¼ - R^2
	A_2A_2bb	P^2	PR	-	R^2
$A_1A_2xA_1A_3$	$A_1A_1B\_$	¼ - R^2	¼ - PR	¼ - PR	¼ - P^2
	A_1A_1bb	R^2	PR	PR	P^2

	$A_1A_3B\_$	¼ - PR	¼ - R^2	¼ - P^2	¼ - PR
	A_1A_3bb	PR	R^2	P^2	PR
	$A_1A_2B\_$	¼ - PR	¼ - P^2	¼ - R^2	¼ - PR
	A_1A_2bb	PR	P^2	R^2	PR
	$A_2A_3B\_$	¼- P^2	¼ - PR	¼ - PR	¼ - R^2
	A_2A_3bb	P^2	PR	PR	R^2
$A_1A_2xA_2A_3$	$A_1A_2B\_$	¼ - R^2	¼ - PR	¼ - PR	¼ - P^2
	A_1A_2bb	R^2	PR	PR	P^2
	$A_1A_3B\_$	¼ - PR	¼ - R^2	¼ - P^2	¼ - PR
	A_1A_3bb	PR	R^2	P^2	PR
	$A_2A_2B\_$	¼ - PR	¼ - P^2	¼ - R^2	¼ - PR
	A_2A_2bb	PR	P^2	R^2	PR
	$A_2A_3B\_$	¼- P^2	¼ - PR	¼ - PR	¼ - R^2
	A_2A_3bb	P^2	PR	PR	R^2
$A_1A_2xA_3A_4$	$A_1A_3B\_$	¼ - R^2	¼ - PR	¼ - PR	¼ - P^2
	A_1A_3bb	R^2	PR	PR	P^2
	$A_1A_4B\_$	¼ - PR	¼ - R^2	¼ - P^2	¼ - PR
	A_1A_4bb	PR	R^2	P^2	PR
	$A_2A_3B\_$	¼ - PR	¼ - P^2	¼ - R^2	¼ - PR
	A_2A_3bb	PR	P^2	R^2	PR
	$A_2A_4B\_$	¼- P^2	¼ - PR	¼ - PR	¼ - R^2
	A_2A_4bb	P^2	PR	PR	R^2

CHAPTER 2

QTL MAPPING IN SIMULATED BACKCROSS POPULATIONS: IMPLICATIONS OF POPULATION SIZE, TRAIT HERITABILITY, QTL PROPERTIES AND MARKER DENSITY

Paper to be submitted to Genetics and Molecular Biology

QTL mapping in simulated backcross populations: implications of population size, trait heritability, QTL properties and marker density

Alexandre Alonso Alves[1,2], Leonardo Lopes Bhering[3], Lúcio Mauro da Silva Guimarães[1], Cosme Damião Cruz[2,3] and Acelino Couto Alfenas[1,2]

[1]Department of Plant Pathology, Federal University of Viçosa, Viçosa, MG, Brazil. [2]Graduate Program in Genetics and Breeding, Federal University of Viçosa, Viçosa, MG, Brazil. [3]Department of General Biology, Federal University of Viçosa, Viçosa, MG, Brazil.

Send Correspondence to Acelino Couto Alfenas. Department of Plant Pathology, Federal University of Viçosa, 36571-000 Viçosa, MG, Brazil. E-mail aalfenas@ufv.br

Abstract

As backcross (BC) populations are often used as mapping populations both in self pollinating species, and in out-breeding species, by means of pseudo-testcross mapping design, we undertook a simulation study to test implications of population size, trait heritability, QTL properties and marker density in the power and precision of QTL mapping. We found that sample size has a major implication in the detection power and as consequence in the estimation of the magnitude and additive genetic effect, as small populations do not allow mapping of low effect QTLs, especially if these QTLs are involved in the genetic control of traits with low heritability. We also found that the positioning of the QTLs based on CIM is more accurate than SIM and that on average the mapped QTLs are close to their simulated position. An interesting result is that CIM tend to underestimate the magnitude (r^2) values especially in large population sizes/low heritabilities traits and overestimate it in smaller populations, which can be a reflection of the low coefficient of variation of the error used, or due to

fact that when markers aren´t in the same position of the QTL, this parameter is indeed expected to be underestimated. We also highlight the fact, that when markers are evenly distributed across the genome, and therefore covering the QTL region, if one of the markers is already close to the QTL, larger number of markers do not improve the precision of QTL mapping in sufficiently large mapping populations.

Keywords: Detection power, QTL mapping precision, backcross, simple interval mapping and composite interval mapping.

Introduction

Quantitative trait locus (QTL) mapping has been in wide use for nearly two decades during which molecular markers have become available (Borevitz and Chory 2004). Over the past years there has been a tenfold increase in the number of QTL studies published annually (Price 2006). The goal of QTL mapping is to determine the loci that are responsible for variation in complex, quantitative traits. However, the task of identifying all such regions that are associated with a specific complex phenotype is difficult because of the sheer number of QTLs, the possible epistasis or interactions between QTLs, and because of the many additional sources of variation (Doerge 2002), including population design and data collection.

Many QTL mapping methods have been developed over the last two decades addressing different concerns. Of the most popular are simple interval mapping (SIM) (Lander and Botstein 1989) and Composite Interval Mapping (CIM) (Zeng 1993, 1994). Both methods use an estimated genetic map as a framework for locating QTLs. The intervals, defined by a pair of flanking markers, are searched in linear increments (one-dimensional scan), and statistical methods are used to test whether a QTL is likely to be present at the location within the interval or not. Interval mapping methods searches through the ordered genetic markers, testing the same null hypothesis and using the same

form of likelihood at each increment (Schuster and Cruz 2008). CIM, however, combines simple interval mapping with linear regression by including additional genetic markers in the statistical model in addition to an adjacent pair of linked markers for interval mapping (Zeng 1994). The idea is that the inclusion of additional markers as cofactors - outside a defined window of analysis - helps removing the variation that is associated with other (linked or unlinked) QTLs in the genome. This fact makes CIM more effective at mapping QTLs, both by effectively locating and estimating its effect, compared to simple interval mapping, especially when linked QTLs are involved, by distinguishing between true QTLs and the so called *ghost QTL* (Schuster and Cruz 2008).

Given plentiful markers and high-throughput genotyping technologies nowadays available (Zhu and Salmeron 2007), and efficient QTL mapping procedures and softwares (Cruz 2010b; Lander and Botstein 1989; Wang et al. 2010; Zeng 1994) QTL studies have been limited by the need of adequate populations and reliable phenotypic measures. Experimental design is therefore paramount. Crosses between completely inbred lines, which differ in the trait of interest, offer an ideal setting for detecting and mapping QTLs by marker-trait associations. The reason is that by doing that all F_1s are genetically identical and show complete linkage disequilibrium for genes differing between the inbred lines. A number of designs, *e.g.* backcrosses or F_{2s}, have been proposed to exploit these features. As the accuracy of QTL mapping is limited by the number of recombinants that are identified based on the genotypic states of the markers, sample size and good quality genotyping becomes important. With this in mind, an important concern is to define if it is better to genotype more markers on fewer individuals, or score more individuals (for genotype and phenotype) on fewer markers. Because observed recombinants provide the information, scoring more individuals shall address both previously mentioned concerns (Doerge 2002). Recent methods of high-throughput genotyping (Wenzl et al. 2004; Zhu and Salmeron 2007) are also providing a reliable and cheap mean to genotype hundreds of individuals with an elevated number of markers, coupled with high

precision. Then the most important issue when designing experimental populations is thought to be sample size. Other issues that have implications in the reliability and precision of QTL mapping are trait heritability and QTL properties, *i.e.* number, magnitude of variation (r^2), effect and location. Minor effect QTLs responsible for the genetic determinism of traits that display low heritabilities are generally harder to map than major effect QTLs for high heritability traits (Schuster and Cruz 2008). Linked QTLs (of minor or major effect) also represent a serious challenge, as often a *ghost QTL* is located instead of the two true QTLs (Doerge 2002).

As the QTLs information such as number, locations, and effects is generally unknown, computer simulation is necessary to check the performance of QTL mapping methods in different scenarios. As backcross (BC) populations are often used as mapping populations, both in self pollinating species (Collard et al. 2005; Doerge 2002), and in out-breeding species, by means of pseudo-testcross mapping design (Grattapaglia and Sederoff 1994), we undertook a simulation study testing the implications of population size, trait heritability, QTL properties and marker density in the power and precision of QTL mapping. For that, populations of different sizes were simulated, along with traits with different heritabilities. Each trait was set to be partially controlled by QTLs explaining different proportions of the phenotypic variance (from minor to major effect, linked or unlinked QTLs). The resulting mapping populations were analyzed through simple interval mapping and composite interval mapping.

Methods

Simulations

The GQMOL software (Cruz 2010b) was used in the simulation of the data used in this study. This software allow the generation of information regarding genomes, parents, individuals genotypes (of different mapping populations) as well as quantitative traits, based on random simulation. Using GQMOL simulation module, different scenarios were simulated based on different

population sizes, different traits (with different heritabilities), different QTL magnitudes and different QTLs configurations, *i.e.* two QTLs, of major or minor effect, in a single linkage group or a single QTL, of major or minor effect, on a linkage group, along with different marker densities.

Designing of the genomes and parents

To generate the genome (genome *A*) information, a fictitious species with 2n=2x=10 was taken into account. Each linkage group was set to have a length of 100cM, with 11 co-dominant markers (*e.g.* microsatellites) evenly distributed, totalizing 55 markers. Two minor effect QTLs were placed on linkage group 1 at pre-determined positions (33.3cM and 66.6cM) explaining 5 and 8% of the genotypic variation, respectively. Two major effect QTLs were placed on linkage group 2 at the same pre-determined positions explaining 15 and 18% of the genotypic variation, respectively. These QTLs configurations, allowed us to test the concern of mapping a *ghost QTL* instead of the true QTLs via interval mapping procedures. One major effect QTL, explaining 25% of the genotypic variation was placed on linkage group 3 at a pre-determined position (50cM), and one minor effect QTL, explaining 7% of the genotypic variation, was placed on linkage group 4, also at a pre-determined position (50cM) to test the accuracy of interval mapping procedures in detecting, locating and estimating QTLs effect and magnitude when only one QTL is located in the linkage group. No QTLs were placed in linkage group 5 in order to test the rate of false positives. Parents were generated by attributing a genotype for each marker. The mean degree of dominance was set to zero in all QTLs, to allow the estimation of the additive effect, rather than the confounded effect (additive plus dominance effects). The parents were design to be completely contrasting and homozygotes.

An additional genome (*genome B*) was generated to test the effect of increased marker density in QTL mapping in backcross populations. For that a fictitious species with 2n=2x=6 was taken into account. Two different scenarios were simulated. One in which each linkage group was set to have a length of

100cM, with 100 co-dominant markers evenly distributed, totalizing 300. And one in which each linkage group was set to have a length of 100cM, with 11 co-dominant markers evenly distributed, totalizing 33. On both scenarios one major effect QTL, explaining 25% of the genotypic variation was placed on linkage group 1 at a pre-determined position (50cM), and one minor effect QTL, explaining 5% of the genotypic variation, was placed on linkage group 2, also at a pre-determined position (50cM). No QTLs were placed in linkage group 3. The mean degree of dominance was set to zero in all QTLs. The parents were design to be completely contrasting and homozygotes.

Populations design

Based on the simulated *genome A* and parents, we generated via random simulation, backcross mapping populations with different sizes (100, 200, 500 and 1000). The population individuals were derived from a pool of 1000 gametes/individual derived from the parents previously simulated. The different population sizes were designed so as to test the accuracy of interval mapping procedures in detecting, locating and estimating QTLs effect in different sample sizes. Ten replicates (different populations based on the same genome and parents) were generated for each population size.

Additional mapping populations of n=1000 were generated to test the effect of increased marker density in QTL mapping in backcross population based on simulated *genome B*. Ten replicates were generated for each scenario.

Quantitative traits design

Based on the 40 populations derived from *genome A* (10 replicates for each sample size), data for three quantitative traits were generated. Trait 1 was set to have a heritability of 80%, mean of 100 and coefficient of variation of the error of 2%. Trait 2 was set to have a heritability of 50%, mean of 100 and coefficient of variation of the error of 2%. Trait 3 was set to have a heritability of 20%, mean of 100 and coefficient of variation of the error of 2%. All traits were

set to be partially controlled by 6 QTLs, two located on linkage group 1, two on linkage group 2, one on linkage group 3 and one on linkage group 4, as previously described. Altogether the six QTLs should account for 78% of the genotypic variation. Based on the additional 20 populations derived from *genome B,* data for two quantitative traits were generated. Trait 1 was set to have a heritability of 80%, mean of 100 and coefficient of variation of the error of 2%. Trait 2 was set to have a heritability of 20%, mean of 100 and coefficient of variation of the error of 2%. Both traits were set to be partially controlled by 2 QTLs one on linkage group 1 and one on linkage group 2, as previously described. Altogether the QTLs accounted for 30% of the genotypic variation. A summary of quantitative traits and QTLs properties is presented in Tables 1 and 2, respectively.

Linkage and QTL mapping

Linkage analyses for each population were performed using GQMOL (Cruz 2010b). Chi-square tests were performed in order to check the existence of marker distortion using the Bonferroni protection ($p \leq 0.1$). Linked markers were placed onto linkage groups with a threshold LOD score of 3.0 and a maximum recombination fraction (ϑ) of 0.30. Recombination fractions were transformed to estimated map distances by the Kosambi map function, which assumes that recombination events influence the occurrence of adjacent recombination events (interference between crossing-over events) (Schuster and Cruz 2008).

Simple interval mapping (SIM) (Lander and Botstein 1989) and Composite Interval Mapping (CIM) (Zeng 1994) were used for QTL mapping on the simulated backcross populations. The parameters used for both methods were a threshold likelihood ratio (LR) of 12 (~LOD = 3) to declare significant QTLs and a 2cM genome scanning step. In CIM, we adopted a *p*-value of ≤ 0.1 for entering variables and a *p*-value ≤ 0.1 for removing variables in the forward and backward stepwise regression of the residual phenotype on marker variables. The stepwise

regression was used to select the co-factors included in CIM. All QTL analyses were performed using GQMOL (Cruz 2010b).

Statistical analysis

The data of QTL mapping, *i.e.* likelihood ratio (LR), magnitude of variation explained (r^2) and effect (a – additive effect), were analyzed through ANOVA using an unbalanced completely randomized design. Means were further analyzed with a Tukey test ($p \leq 0.05$) to allow the statistical evaluation of the interval mapping procedures (SIM and CIM) in locating and estimating QTLs effect and magnitude considering the different population sizes, trait heritabilities and QTL properties. The additional populations generated based on *genome B* were evaluated through *t*-tests ($p \leq 0.05$) so as to allow the assessment of the impact of increased marker density in QTL mapping in backcross populations. ANOVAs, Tukey tests and *t*-tests were performed with the aid of the software GENES (Cruz 2010a).

Results

Although it was expected that family size would have a dramatic effect on the precision of QTL mapping, we found that one of its major impact is on the QTL detection power. When we tested smaller sample sizes (*i.e.* 100 and 200) we found that power was greatly reduced when compared to larger population sizes (*i.e.* 500 and 1000). The same occur for trait heritability and QTL magnitude, being the major effect QTLs for high heritability traits detected with higher likelihood ratio (LR) estimates in the testing positions than the major or even minor effect QTLs for low heritability traits (Tables 3 and 4, and Figures 1A and 1B). For instance, considering the trait 3, which was set to have a low heritability (20%) and the CIM approach, while the detection power of the minor effect QTL located in linkage group 4 (QTL41) was of 70% and 80%, when using samples of 500 and 1000, respectively, it dropped to 0% and 10% when using populations of 100 or 200 individuals, respectively (Table 3). This same trend, *i.e.* elevated

detection power in large families and lower detection power in smaller populations, was observed for the large effect QTL located in linkage group 3 (QTL31) and as well as to linked QTLs of large and low effect we when used either CIM or SIM (Table 3). No QTLs were detected on linkage group 5, *i.e.* no false positive QTLs were detected as expected.

When we consider the precision of QTL mapping it interesting to note that, the lower detection power is not always accompanied by a lower precision, at least in terms of QTL positioning. If we consider the data of the QTLs detected in linkage groups 3 and 4 by CIM, the QTL position is on average within 5cM of its parametric position, independent of the population size. Moreover, if we consider only the populations with 200, or more individuals, the QTLs where located within 2.5cM of its original position (Table 4), indicating that the CIM is pretty accurate with regards to QTL positioning. We found that in terms of QTL positioning the major obstacle is the existence of two linked QTLs, *i.e.* two QTLs in the same linkage group. In such situations, one is likely to map a *ghost QTL* instead of the two true QTLs. We noted, as expected, that SIM is more prone to this type of problem than CIM (Table 5 and Figure 2), independent if major or minor effect QTLs pairs are involved. In general our results indicate that linked QTLs are easily mapped when sample size, trait heritability and the genotypic variation explained by the QTL are elevated. It is interesting to note that even using CIM, if the population is considerable small, or trait heritability is low, the possibility of mapping a *ghost QTL* is relatively high (Table 5), thus demonstrating that population design is paramount.

One important issue that it is noteworthy is that when we scored the QTLs in the linkage groups one and two, we considered that every QTL detected before the 50cM mark should correspond to first QTL of that linkage group and that very QTL detected after the 50cM mark to the second QTL. We choose to score markers that way because, in a real situation, where one does not know that there are indeed two linked QTLs, one would consider the QTL mapped as true. In that way if only one QTL was detected, *e.g.* at position 48cM, we

considered that QTL 1 was mapped and that QTL 2 wasn´t. However, as the parametric positions of both QTLs were 33 and 66, when only one QTL was detected, near position 50cM, it is more likely that this QTL correspond to a *ghost QTL.* Because of that in some scenarios analyzed with SIM, the average QTL position is far away from its parametric value (Table 6), *e.g.* QTL 11 and QTL 21 in traits 1 and 2. Such problem is virtually unseen in CIM as the number of ghost QTLs mapped is too small to have a profound impact on the average QTL position (Table 5).

With regards to the magnitude of the genotypic variation explained by the QTLs (r^2), we found that CIM tend to underestimate the estimates especially in large population sizes/low heritabilities traits and overestimate it in smaller populations (Table 5). It is interesting to highlight that, significant differences between sample sizes with regards to r^2 values occurred more often for low heritability traits (Tables 4 and 6). However, as the r^2 values tended to be underestimated when using large population sizes and overestimated when using smaller sample sizes, estimates derived from larger populations are relatively less accurate than those obtained in smaller populations, when comparing to the parametric value. Another interesting point is that, as higher the genotypic variation explained by the QTL is, the more accurate becomes the estimation of r^2 values, by both mapping approaches (Tables 4 and 6). We noted that SIM also tended to overestimate r^2 values when linked QTLs occurred, markedly in smaller populations (Table 6).

In terms estimation of the genetic additive effect (*a*), larger sample sizes seem to provide more accurate data, as the smaller population sizes tend to overestimate this parameter. In contrast, the larger population sizes tended to provide more reliable values, especially if trait heritability is low. For example, considering the minor effect QTL41 and trait 3 (h^2=20%) the estimated additive effect was 1.13 with 100 individuals and 0.57 with 1000 individuals, *i.e.* in the smaller sample size the *a* estimate is nearly 100% overestimated (Table 4). If one considers that the parametric value in this situation is 0.53, it is clear that, as

larger the sample size is, more accurate becomes the estimation of the additive genetic effect of QTL involved in the genetic control of low heritability traits. Considering the high heritability trait the differences are not significant (Tables 4 and 6), and even small populations sizes provide accurate data. A important issue to consider here is that when a *ghost QTL* is located instead of the true QTLs, which occurred most often when data was analyzed with SIM, not only the positioning, magnitude estimates are compromised (as detailed above) but obviously also the additive effect values. Indeed, as we have considered those QTLs mapped near the center of the linkage group as *ghost QTLs*, the higher number of *ghost QTLs* mapped by SIM had a pronounced effect on the estimation *a* values (Table 6). For example, considering QTL11, its additive effect in the population of 1000 individuals was estimated as 2.24, 1.04 and 0.73 with CIM and as 3.24, 1.47 and 1.47 with SIM for traits 1, 2 and 3, respectively. As the parametric values are respectively 1.79, 0.90 and 0.45, it is clear that CIM, by relieving the detection of *ghost QTLs*, provide more accurate data of additive effect when linked QTLs occur.

An important issue that we have addressed is whether or not high density maps provide a framework to more accurate QTL analysis. For that we compared the performance of CIM in the same population using marker densities of 1cM and of 10cM. Our data suggest that the larger number of markers did not improved the precision of QTL mapping (Table 7), as the detected QTLs are, on average, within 2.5cM and 1.5cM of its simulated position, for 10cM and 1cM marker density, respectively. No significant difference was found between these values (Table 8 and Figures 3A and 3B). Thus there is evidence that when one is dealing with adequate sample size (≥ 1000) and adequate genome coverage even mid-density genetic maps can be used to map QTLs, of large or small effect, with high confidence. As before, we noted that the r^2 values tended to be underestimated by CIM, especially for the QTLs of the low heritability trait, and that the additive genetic effect is estimated accurately in both marker densities (Table 7). Another interesting point is that, even using a huge amount of

markers, spaced evenly, we could not detect the minor effect QTL for the low heritability trait (Table 7 and Figure 3A), thus demonstrating that detection power is restricted by sample size, and not due to low marker density, considering a adequate genome coverage.

Taken together our results indicate that: (i) sample size has a major implication in the detection power and as consequence in the estimation of the magnitude and additive genetic effect; (ii) small populations do not allow mapping of low effect QTLs, especially if these QTLs are involved in the genetic control of traits with low heritability; (iii) the positioning of the QTLs based on CIM is more accurate than SIM and that on average the mapped QTLs are close to their simulated position; (iv) CIM tend to underestimate the r^2 values especially in large population sizes/low heritabilities traits and overestimate it in smaller populations; (v) SIM tended to overestimate r^2 values when linked QTLs occurred, markedly in smaller populations; (vi) when a *ghost QTL* is located instead of the true QTLs, the positioning and as consequence the estimation of r^2 and *a* values is compromised; (vii) larger number of markers do not improved significantly the precision of QTL mapping in sufficiently large mapping populations with adequate genome coverage; (viii) when one is dealing with adequate sample size even mid-density genetic maps can be used to map QTLs of large or small effect with high confidence and finally that (ix) that detection power is restricted by sample size, and not due to low marker density, if the genome coverage is adequate.

Discussion

We have attempted to dissect the influence of sample size, trait heritability, QTL properties and marker density in the power and precision of QTL mapping in backcross populations using a simulation approach. Although there is plenty of data on QTL mapping that could be used to evaluate the precision and power, computer simulation data offer the advantage of knowing the true parameters that can be used to compare with the estimates obtained. We

choose to work with backcross (BC) populations because they offered a strait framework model to work with. As BC populations provide the simplest genetic model in QTL mapping, linkage analysis methods and QTL mapping strategies are well developed and implemented in various well known softwares. Another reason is because they have been extensively used as mapping populations not only in selfing species (Collard et al. 2005) but also in out-breeding species by means of pseudo-testcross mapping design (Grattapaglia and Sederoff 1994).

Overall, the QTL mapping literature has shown that when a mapping population of 100-150 individuals derived from an backcross population between two inbreeds, is used along with reasonably good phenotypic data for the traits of interest, and genotype, an analysis of the phenotypic and marker data with an appropriate statistical method will almost always lead to the identification of at least a few markers associated with each trait of interest (Bernardo 2008). However, we found that the detection power is greatly reduced when mapping populations are small (≤ 200 individuals), as these mapping populations do not allow mapping of minor effect QTLs for low heritability traits at all when SIM is used. These same QTLs are poorly detected when CIM is used for its turn. Even major effect QTLs for low heritability traits are often undetected by both SIM and CIM. In this way, since most of the QTLs mapping studies published to date uses populations of ≤200 individuals (for sake of simplicity or due to limited experimental resources), with a few exceptions, the majority of minor and major effect QTLs for low heritability traits may have been undetected due to inappropriate population design. This is probably one of the reasons why marker assisted selection (MAS) has faced enormous difficulties to be implement in plant breeding programs. As pointed out by Lande and Thompson (1990) MAS should apply better to low heritability traits only if the additive genetic variance explained by the molecular markers (g) exceeds the trait heritability (h^2). In a situation where many, if not the majority, of the minor and major effect QTLs are overlooked, it is unlikely that g could exceed h^2. Thus, larger populations (≥ 500 individuals) are required in both QTL mapping and MAS experiments, when one

is dealing with low heritability traits. In practice, however, large mapping populations may not be feasible because breeders tend to improve several populations simultaneously. The use of larger mapping populations would lead to fewer populations being improved, and many breeders prefer to select in a large number of populations with relatively few individuals, instead of in a few populations with many progenies (Bernardo 2004). Then one option is to increase the heritability of the small QTLs, by reducing environmental variation, by having more replicates or by combining analysis of several traits that the gene affects pleiotropically (Price 2006).

Since power of detection suffers badly with smaller population sizes, perhaps Type I error control should not be a major concern. Rather, the major concern might actually be Type II error and the bias on estimating QTL effects. Despite this, declaring the presence of a QTL always carries some risk that such declaration is false. In that context, there has been a great deal of interest in recent years in evaluating what proportion of the declared QTLs in plants are false (Bernardo 2004). As some degree of missing data is likely to occur, as well as errors in the phenotyping and genotyping, coupled with insufficient large populations, to prevent false QTL declaring, Bernardo (2004) suggest that a detected QTL should, in general, be reported as a QTL only if it was identified at a stringent significance level, alpha values of 0.0001 or FDR (False Discovery Rate) levels of 0.01 or lower. However, as we haven´t detected any false positive QTLs in our study, it is reasonable to speculate that if good quality data is employed, most of the QTLs mapped should indeed correspond to loci involved in the genetic control of quantitative traits of interest, and not to false positives.

It is generally believed that QTL mapping does not accurately pinpoint the position of genes underlying quantitative traits on the genome, as QTL studies have typically identified broad genomic regions which are likely to comprise several hundred genes or cis regulatory elements. Theory suggests that the positioning of a QTL in a primary mapping population (those that were not designed for fine mapping experiments) is not accurate, covering a region up to

and over 20 cM (Kearsey and Farquhar 1998; Luo et al. 2002). The 1 likelihood of odds (LOD) support interval with which QTLs are commonly reported is often a large region covering 10 to 30 cM. In most plant species, this could include hundreds or even thousands of genes. van Ooijen (1992) showed that there is a reasonable probability of detecting QTLs that explain at least 5% of the total variance by SIM, if the population (BC or a F_2) is composed of a minimum of 200 individuals, and that, on average, a QTL that explain between 5 and 10% of the genotypic variance is mapped on a interval of 40 or 20cM, respectively. However, as many QTLs have been cloned recently (Price 2006; Salvi and Tuberosa 2005), based on QTL mapping data, there is a least evidence that QTL mapping is more accurate than it was previously supposed. In that respect, by comparing the QTL mapping position to the actual position where the underlying gene was cloned Price (2006) showed that on average the position of genes underlying major QTLs ranges from 0.0 to 1.9 cM and the mean is less than 0.7 cM, whereas for small QTLs the range is 0 to ≤3 cM and the mean is ≤1.2 cM. Here we show that major and minor effect QTLs position in backcross populations, estimated by CIM, is on average within 2.5cM of its parametric position, when populations of more than 200 individuals are used. This fact might indicate that the position of a QTL, obtained from a primary mapping population, is indeed more accurate than is often stated. It is important to note however that, in a 2.5cM window hundreds of genes can occur, depending on the genome size and organization, and thus although QTL mapping proved to be pretty accurate, still there are a lot of genes to be tested if the objective is to identify the gene that underlie a QTL for a quantitative trait of interest or clone it. Recently linkage mapping of several thousand genes and of expression QTLs (eQTLs) have provided opportunities to verify co-localization of genes with QTLs, thus supplying more likely positional candidates to be tested in association genetics experiments. Larger population sizes and marker densities (discussed below) could also effectively improve the resolution of QTL mapping, thus facilitating the identification of candidate genes.

The primary goal of genetic mapping experiments is to identify the locations of genes that affect variable expression of a trait among individuals. But most researchers also use the data to estimate the genetic effects of QTLs (Xu 2003). Beavis (1994) designed a large-scale simulation experiment to evaluate the efficiency of interval analysis for detecting and estimating QTLs effects. These experiments showed that the average estimates of genotypic variances associated with correctly identified QTL were greatly overestimated if only 100 individuals were evaluated, slightly overestimated if 500 individuals were evaluated, and fairly close to the actual magnitude when 1000 individuals were evaluated (Xu 2003). The bias occurs mainly due to sampling of small populations; the true QTLs that are not detected (most of them in small sample sizes) have effects that are detected at the regions that are detected as having QTLs. So the declared QTL regions have effects that appear much larger than they really are. This phenomenon is known as "The Beavis Effect". This study was based on a simulation of F_2 populations. Here we show that in backcross populations, CIM tend to underestimate the r^2 values especially in large population sizes/low heritabilities traits and overestimate it in smaller populations. We also show that SIM tended to overestimate r^2 values when linked QTLs occurred, markedly in smaller populations. We didn´t find a situation, where r^2 values were overestimated when a linkage group held a single QTL, of major or minor effect, independent of the sample size. This could have occurred because of the low value used for the coefficient of variation of the error. In other situations, were this parameter is higher, and possibly more realistic, these result may not hold true. Future studies are then necessary to evaluate this. Overestimation only occurred in the case of linked QTLs, and this is obviously resulting of mapping a *ghost QTL* instead of the true QTLs. In this case, the effect of the true QTLs not detected are detected confounded at other regions, and thus all estimation is completely compromised. As Beavis used a SIM approach and F_2 populations, and we have used SIM/CIM and BC populations, the results are no readily comparable and no conclusion can be

drawn. Another interesting point to highlight is that parametric values for r^2 refers to the QTL, while the r^2 values estimated by mapping procedures uses information of the markers that flank the putative QTL. In this way, it is indeed expected, if the markers are not located exactly in QTL position, that the r^2 estimates would be underestimated. However, we noted that the behavior of the additive effect in our data closely resembles the Beavis effect, as larger populations are more accurate than smaller populations in the estimation of this parameter. On average we found that the additive effect is 0.5-1 fold overestimated when a single (major or minor effect) QTL for a low heritability traits is mapped in small populations (≤ 200 individuals). When the SIM approach is adopted the additive effect of major and minor QTLs for low heritability traits are overestimated even in large populations (1000 individuals). This may reflect the fact, that the variation of other linked or unlinked QTLs, is not removed by this method, a thus can inflate the estimates obtained.

It is interesting to note that if QTL data is to be used in MAS, adequate populations are required to correctly estimate the efficiency of MAS, based on accurate values of r^2 and *a.* Ideally one would choose to include in MAS breeding program the major effect QTLs (high r^2 values then) that have the largest and more accurate predicted additive effect. However, as r^2 values are underestimated in large populations, the efficiency of MAS for its turn should be also, because as pointed out by Lande and Thompson (1990), the MAS efficiency depends upon the genetic variation explained by the molecular markers (*g*). Since *g* equals to m^2/h^2, m^2 and h^2 are, respectively, the proportion of the phenotypic variation explained by the markers and the heritability of the trait, the largest the proportion of the phenotypic variation explained by the QTLs the more effective should be MAS. Xu (2003) developed a theory that helps predict the potential bias in the estimated effect of QTLs. The theory may also be used to correct the bias but, as it is very sensitive to the sample size (n), it works only when n is sufficiently large, say n ≥ 500. Therefore, the theoretical prediction of the bias may not be used retrospectively to correct for the bias when the sample

size is small. In an ideal scenario it would be desirable to utilize all QTL affecting the trait in marker-assisted selection no matter if they are classified as major or minor effect QTLs. But as previously discussed, detection power is low for minor effect QTLs affecting low heritability traits. In that context, we have shown that the detection power is restricted by sample size, and not due to low marker density. Even with a high density map (1 marker per cM) we could not improve the power and positioning of the QTL in a large population (1000 individuals). This is important fact, because as high-throughput genomic tools, such as the DNA microarray platform, are enabling the development of novel genotyping procedures, such as Diversity Arrays Technology (DArT) (Wenzl et al. 2004), Single Nucleotide Polymorphisms (SNPs) (Zhu and Salmeron 2007) and Single Feature Polymorphism (SFPs) (Rostoks et al. 2005), it is likely that, in the future, high density linkage maps can be constructed for any species with a mid-level of genomic information, providing a new framework for new MAS approaches such as genomic selection (Meuwissen et al. 2001). However, one has to keep in mind, no matter if the objective is to fine map QTLs or genomic selection, that designing larger and appropriate populations to fully exploit the advantages of using an enormous amount of molecular markers is paramount.

As mentioned before other studies, using different scenarios, *i.e.* different coefficients of variation of the error, different number of QTLs, different marker distributions, which collectively may make the simulation a bit more realistic, are needed in order to see if the results of our work hold true in every situation.

Acknowledgments

We are grateful to Caio Césio Salgado, for his constructive comments on the manuscript and for the assistance in the numerous simulations. The Bioinformatics Lab of the Federal University of Viçosa, Brazil provided the facilities for the development of this work. This work was supported by the Brazilian National Research Council, CNPq, with a Ph.D. fellowship to AAA, post-doctoral fellowship to LMSG and a research fellowship to ACA and CDC.

References

Beavis WD (1994) The power and deceit of QTL experiments: Lessons from comparative QTL studies. Corn Sorghum Ind Res Conference. Am. Seed Trade Association, Chicago, IL, pp 250-266

Bernardo R (2004) What proportion of declared QTL in plants are false? Theor Appl Genet 109:419-424

Bernardo R (2008) Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Sci 48:1649-1664

Borevitz JO, Chory J (2004) Genomics tools for QTL analysis and gene discovery. Curr Opin Plant Biol 7:132-136

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196

Cruz CD (2010a) Genes: a software for genetics analysis. Universidade Federal de Viçosa, Viçosa, MG, Brazil

Cruz CD (2010b) GQMOL: a software for quantitative and genetics analysis. Universidade Federal de Viçosa, Viçosa, MG, Brazil

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43-52

Grattapaglia D, Sederoff R (1994) Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross mapping strategy and RAPD markers. Genetics 137:1121-1137

Kearsey MJ, Farquhar AGL (1998) QTL analysis in plants: where are we now? . Heredity 80:137-142

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative trait using RFLP linkage maps. Genetics 121:185-199

Luo ZW, Wu C-I, Kearsey MJ (2002) Precision and high-resolution mapping of quantitative trait loci by use of recurrent selection, backcross or intercross schemes. Genetics 161:915-929

Meuwissen T, Hayes B, Goddard M (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819-1829

Price AH (2006) Believe it or not, QTLs are accurate! Trends in Plant Science 11:213-216

Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, Morris J, Cardle L, Marshall DF, Waugh R (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biology 6

Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. Trends in Plant Science 10:297-304

Schuster I, Cruz CD (2008) Estatística Genômica aplicada a populações derivadas de cruzamentos controlados, 2th edn. Editora UFV, Viçosa

van Ooijen JW (1992) Accuracy of mapping quantitative traits loci in autogamous species. Theor Appl Genet 84:803-811

Wang S, Basten CJ, Zeng ZB (2010) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC.

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proceedings of the National Academy of Sciences of the United States of America 101:9915-9920

Xu S (2003) Theoretical Basis of the Beavis Effect. Genetics 165:2259-2268

Zeng ZB (1993) Theoretical basis of precision mapping of quantitative trait loci. Proceedings of the National Academic of Science 90:10972-10976

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457-1468

Zhu T, Salmeron J (2007) High-definition genome profiling for genetic marker discovery. Trends in Plant Science 12:196-202

Table 1. Summary of quantitative traits properties in simulated backcross populations.

Trait*	H^2 (%)$^{+}$	Nº of QTLs	CVe^{++}	Mean	Evr. Var.§	Gen. Var.§§
1	80	6	2	100	4	16
2	50	6	2	100	4	4
3	20	6	2	100	4	1
Trait**	H^2 (%)$^{+}$	Nº of QTLs	CVe^{++}	Mean	Evr. Var.§	Gen. Var.§§
1	80	2	2	100	4	16
2	20	2	2	100	4	1

*Traits designed based on the backcross populations. **Traits design based on the additional backcross populations. $^{+}$Heritability. $^{++}$Coefficient of variation of the error. §Environmental variance. §§Genetic variance.

Table 2. Summary of QTLs responsible for genetic control of traits designed based on the backcross populations properties.

QTL*	Position[=]	PVE[+]	*mdd*[++]	*a1*[§]	*a2*[§§]	*a3*[§§§]
11[¯]	33	5	0	1.79	0.89	0.45
12	66	8	0	2.26	1.13	0.57
21	33	15	0	3.10	1.55	0.78
22	66	18	0	3.39	1.70	0.85
31	50	25	0	4.00	2.00	1.00
41	50	7	0	2.12	1.06	0.53
QTL**	Position	PVE[+]	*mdd*[++]	*a1*[§]	*a2*[§§]	*a3*[§§§]
11	50	25	0	4.00	1.00	-
21	50	5	0	4.39	0.45	-

*QTLs responsible for genetic control of traits designed based on the backcross populations. **QTLs responsible for genetic control of traits based on the additional backcross populations. [=]QTL position in cM from leftmost marker. [¯]QTLxy – x=linkage group; y=QTL number. [+]Percentage of the genotypic variation explained by the QTL. [++]Mean degree of dominance. [§]Additive effect of the QTL with regards to trait 1. [§§]Additive effect of the QTL with regards to trait 2. [§§§]Additive effect of the QTL with regards to trait 3.

Table 3. Power of QTL detection (%) with regards to family size and trait heritability by composite interval mapping (CIM) and simple interval mapping (SIM).

Trait/Pop size	QTL					
	11	12	21	22	31	41
Trait 1 (H^2=0.8)						
100	20/30*	80/60	70/80	90/100	100/90	70/0
200	80/60	90/80	80/60	100/100	100/100	90/60
500	90/50	100/70	100/20	100/80	100/100	100/100
1000	80/30	100/80	100/40	100/70	100/100	100/100
Trait 2 (H^2=0.5)						
100	10/10	40/10	50/50	80/90	80/70	20/0
200	30/50	70/100	60/70	90/100	90/90	60/40
500	90/60	80/60	90/40	100/60	100/100	100/80
1000	70/40	90/100	90/40	100/70	100/100	100/90
Trait 3 (H^2=0.2)						
100	10/0	10/20	10/10	40/70	30/20	0/0
200	10/0	60/20	30/30	40/70	60/50	10/0
500	30/40	50/60	80/40	70/90	100/100	70/70
1000	90/50	50/90	70/40	90/90	100/100	80/70

*Power of QTL detection by CIM/power of QTL detection by SIM.

Table 4. Influence of family size and trait heritability in the precision of QTL mapping by composite interval mapping (CIM).

Trait/Pop size	QTL11				QTL12				QTL21				QTL22				QTL31				QTL41			
	Pos[+]	LR	$R^{2§}$	$a^{§§}$	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*
Trait 1 (H^2=0.8)																								
100	40.1a*	31.8a	11.4b	3.37a	61.1a	34.1a	13.6a	3.62a	37.9a	26.5a	12.4a	3.70a	67.7a	27.7a	12.8a	3.77a	54.5a	40.0a	17.4a	4.08a	54.7a	18.3a	7.0ab	2.64a
200	32.1a**	35.0a	7.4ab	2.59a	68.1a	43.0a	8.6a	2.81a	32.9a	48.3ab	12.6a	3.49a	62.2a	70.1ab	17.2a	4.14a	52.0a	82.1b	19.0a	4.13a	51.9a	24.5a	4.9b	2.09b
500	35.9a	46.2a	4.0a	2.02a	67.7a	95.5b	9.4a	3.01a	35.9a	113.9b	12.0a	3.58a	67.0a	117.4b	11.7a	3.78a	51.55a	175.8c	17.1a	4.14a	52.3a	56.4b	4.8b	2.19b
1000	36.0a	107.1a	5.2ab	2.24a	67.7a	183.1c	8.7a	2.84a	34.9a	202.7c	10.0a	3.34a	68,6a	255.1c	12.6a	3.77a	52.5a	356.0d	17.5a	4.06a	51.4a	106.6c	4.6b	2.10b
Parametric	33	-	5	1.79	66	-	8	2.26	33	-	15	3.10	66	-	18	3.39	50	-	25	4.00	50	-	7	2.12
Trait 2 (H^2=0.5)																								
100	33.2a	30.0a	17.4a	2.63a	77.4a	15.7a	8.6bc	1.79bc	38.3a	26.9a	17.5a	2.57a	64.1a	34.4a	21.9a	2.93a	57.4a	19.9a	11.0ab	1.88a	28.1a	13.3a	7.1a	1.67a
200	36.6a	23.8a	7.4b	1.66b	62.2b	35.4a	10.7c	1.94c	31.9a	20.5a	6.5b	1.72b	69.0a	32.5a	11.2b	2.04b	52.7b	44.1b	13.8b	2.21a	47.4b	16.7a	5.1b	1.38a
500	39.2a	42.3a	5.8b	1.44b	65.7b	30.7ª	4.1a	1.24a	34.6a	61.2a	8.7b	1.83b	65.0a	79.2b	11.4b	2.10b	51.3b	83.3c	11.1ab	1.97a	56.3b	21.18a	2.7c	0.97b
1000	33.6a	42.2a	3.0b	1.04b	64.3b	91.5b	6.6ab	1.54ab	34.2a	116.3b	8.8a	1.89b	68.5a	95.6b	7.2b	1.71b	52.8b	145.5d	10.0b	1.90a	510b	43.97b	2.9c	0.94b
Parametric	33	-	5	0.90	66	-	8	1.13	33	-	15	1.55	66	-	18	1.70	50	-	25	2.00	50	-	7	1.06
Trait 3 (H^2=0.2)																								
100	43.8a	21,4a	15.0a	3.29a	65.3a	14.1a	13.6a	1.84a	51.3a	12.3a	9.6ab	1.45a	68.5a	14.33a	11.3a	1.56a	49.1a	13.8a	10.7a	1.53a	ND	ND	ND	ND
200	22.2a	19.2a	7.5b	1.29b	66.9a	15.6a	6.3b	1.24b	42.8a	28.5a	12.0b	1.66a	66.6a	27.9a	11.7a	1.59a	49.3a	22.8ab	8.9a	1.40a	28.2a	14.7a	6.4a	1.13a
500	40.5a	19.8a	3.5c	0.89c	63.5a	22.6a	3.9c	0.87bc	32.9a	31.3a	5.3ab	0.94a	66.8a	21.6a	3.9b	1.0a	50.7a	29.5b	5.3a	1.03b	53.2b	19.9a	3.3b	0.82b
1000	37.4a	24.0a	2.2c	0.73c	66.8a	32.7a	2.9c	0.56d	38.8a	31.6a	2.9a	0.75a	67.9a	51.9a	4.8b	1.0a	52.4a	59.3c	5.2a	1.07b	54.4a	17.6a	1.2c	0.57c
Parametric	33	-	5	0.45	66	-	8	0.57	33	-	15	0.77	66	-	18	0.85	50	-	25	1.00	50	-	7	0.53

*Means followed by the same letter in the columns do not differ statistically between them, by the Tukey test ($p \leq 0.05$). ** Average values for the n replicates. [+] QTL position in cM from leftmost marker. [§]Genotypic variation explained by the QTL (%). [§§]Additive effect of the QTL. CV – coefficient of variation. LR – likelihood ratio. ND – no data.

Table 5. Number of times that simple interval mapping (SIM) or composite interval mapping (CIM) detected a ghost QTL instead of the two true QTLs in linkage groups (LG) one and two.

Trait/Pop size	SIM/LG		CIM/LG	
	1	2	1	2
Trait 1 (H^2=0.8)				
100	0	0	2	1
200	0	2	0	1
500	4	5	0	0
1000	4	1	0	0
Trait 2 (H^2=0.5)				
100	0	2	0	2
200	3	1	4	1
500	2	6	1	1
1000	3	5	2	1
Trait 3 (H^2=0.2)				
100	1	0	0	0
200	0	3	1	1
500	3	3	3	1
1000	2	4	1	0

Table 6. Influence of family size and trait heritability in the precision of QTL mapping by simple interval mapping (SIM).

Trait/Pop size	QTL11				QTL12				QTL21				QTL22				QTL31				QTL41			
	Pos[+]	LR	$R^{2§}$	$a^{§§}$	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*	Pos	LR	R^2	*a*
Trait 1 (H^2=0.8)																								
100	34.4a*	22.2a	21.8a	4.59a	65.1a	20.0a	19.0a	4.29a	40.1a	31.5a	29.1a	5.29a	66.8a	31.3a	28.5a	5.11a	52.8a	23.6a	22.0a	4.52a	ND	ND	ND	ND
200	34.7a**	24.9a	12.4b	3.40b	65.2a	29.2a	14.3b	3.63b	33.5a	50.8a	24.1a	4.69a	61.9ab	63.1b	29.2a	5.14a	52.7a	42.1b	19.7ab	4.20a	52.4a	16.0a	8.3a	2.74a
500	46.6a	49.2b	10.5b	3.18b	65.0a	53.0b	11.0b	3.30b	38.4a	124.3b	24.1a	4.91a	59.9b	157.2c	28.9a	5.36a	51.7a	94.2c	17.6b	4.16a	54.9a	28.7b	5.8a	2.37ab
1000	44.2a	103.3 c	10.5b	3.24b	63.6a	114.6c	11.9b	3.34b	42.9a	287.0c	27.3b	5.06a	67.5a	292.1d	27.4a	5.11a	52.8a	181.9d	16.8b	3.97a	51.4a	45.8c	4.7b	2.09b
Parametric	33	-	5	1.79	66	-	8	2.26	33	-	15	3.10	66	-	18	3.39	50	-	25	4.00	50	-	7	2.12
Trait 2 (H^2=0.5)																								
100	37.2a	34.4a	29.5a	3.31a	84.6a	13.1a	13.2a	2.09a	34.3ab	24.3a	22.3a	2.72a	63.5a	33.4a	29.9a	3.22a	58.9a	15.0a	15.3a	2.37a	ND	ND	ND	ND
200	34.9a	18.2a	9.3b	1.84b	64.7b	22.6a	11.1a	1.97a	32.8a	27.5a	13.5b	2.22b	67.5a	33.7a	16.7c	2.43c	52.2b	31.3b	14.9a	2.29a	56.6a	13.6a	6.9a	1.54a
500	37.2a	31.3a	6.3b	1.47b	62.5b	37.0a	7.7a	1.64a	44.1b	87.5b	17.5ab	2.39ab	59.4a	119.8b	22.9b	2.89b	51.8b	58.8c	11.3b	1.99b	57.4b	20.3a	4.1b	1.20b
1000	35.7a	57.1b	6.1b	1.47b	64.2b	71.1b	7.5a	1.64a	40.7ab	184.1c	18.5ab	2.62ab	63.2a	169.0c	17.0c	2.47c	53.0b	102.4d	9.9b	1.89b	52.6a	30.7b	3.2b	1.07b
Parametric	33	-	5	0.90	66	-	8	1.13	33	-	15	1.55	66	-	18	1.70	50	-	25	2.00	50	-	7	1.06
Trait 3 (H^2=0.2)																								
100	37.2a	34.4a	29.5a	3.31a	84.6a	13.1a	13.2a	2.09a	34.8ab	24.3a	22.3a	2.72a	63.5a	33.4a	29.9a	3.22a	58.9a	15.0a	15.3a	2.37a	ND	ND	ND	ND
200	34.9a	18.2a	9.3b	1.84b	64.7b	22.5a	11.1a	1.97a	32.8a	27.5a	13.5b	2.22b	67.5a	33.7a	16.7c	2.43c	52.2b	31.3b	14.9a	2.29a	56.4a	13.6a	6.9a	1.54a
500	39.4a	31.3a	6.3a	1.47b	62.5b	37.0a	7.8a	1.64a	44.1b	87.5b	17.5ab	2.39ab	59.4a	119.8b	22.9b	2,89b	51.8b	58.8c	11.3b	1.99b	56.6a	19.4a	3.9b	1.17b
1000	35.7a	57.1b	6.1a	1.47b	64.2b	71.0b	7.5a	1.64a	40.7ab	184.1c	18.1ab	2.62ab	63.2a	169.0c	17.0c	2.47c	53.0b	102.4d	9.9b	1.89b	52.8a	32.7b	3.3b	1.0
Parametric	33	-	5	0.45	66	-	8	0.57	33	-	15	0.77	66	-	18	0.85	50	-	25	1.00	50	-	7	0.53

*Means followed by the same letter in the columns do not differ statistically between them, by the Tukey test ($p \leq 0.05$). ** Average values for the n replicates. [+] QTL position in cM from leftmost marker. [§]Genotypic variation explained by the QTL (%). [§§]Additive effect of the QTL. CV – coefficient of variation. LR – likelihood ratio. ND – no data.

Table 7. Power and precision of QTL mapping by composite interval mapping (CIM) in high density genetic maps compared to mid density maps.

Trait/Marker density	QTL11					QTL22				
	Pow[+]	Pos[++]	LR	$R^{2§}$	$a^{§§}$	Pow	Pos	LR	R^2	*a*
Trait 1 (H^2=0.8)										
11/LG	100	52.6a	199.1a	16.89a	4.06a	100	51.9a	43.7a	3.56a	1.75a
100/LG	100	51.6a	193.4a	16.16a	4.14a	100	49.8a	50.6a	3.91a	1.97a
Parametric	-	50	-	25	4	-	50	-	5	1.7889
Trait 2 (H^2=0.2)										
11/LG	100	52.2a	44.5a	4.48a	0.96a	40	56.5a	21.51a	2.18a	0.72a
100/LG	100	48.7a	42.9a	4.10a	0.94a	50	51.7a	18.5a	1.72a	0.68a
Parametric	-	50	-	25	1	-	50	-	5	0.4472

*Means followed by the same letter in the columns do not differ statistically between them, by a t-test ($p \leq 0.05$). ** Average values for the n replicates. [+] Power of detection (%). [++] QTL position in cM from leftmost marker. [§]Genotypic variation explained by the QTL. [§§]Additive effect of the QTL. LR – likelihood ratio. LG – linkage group.

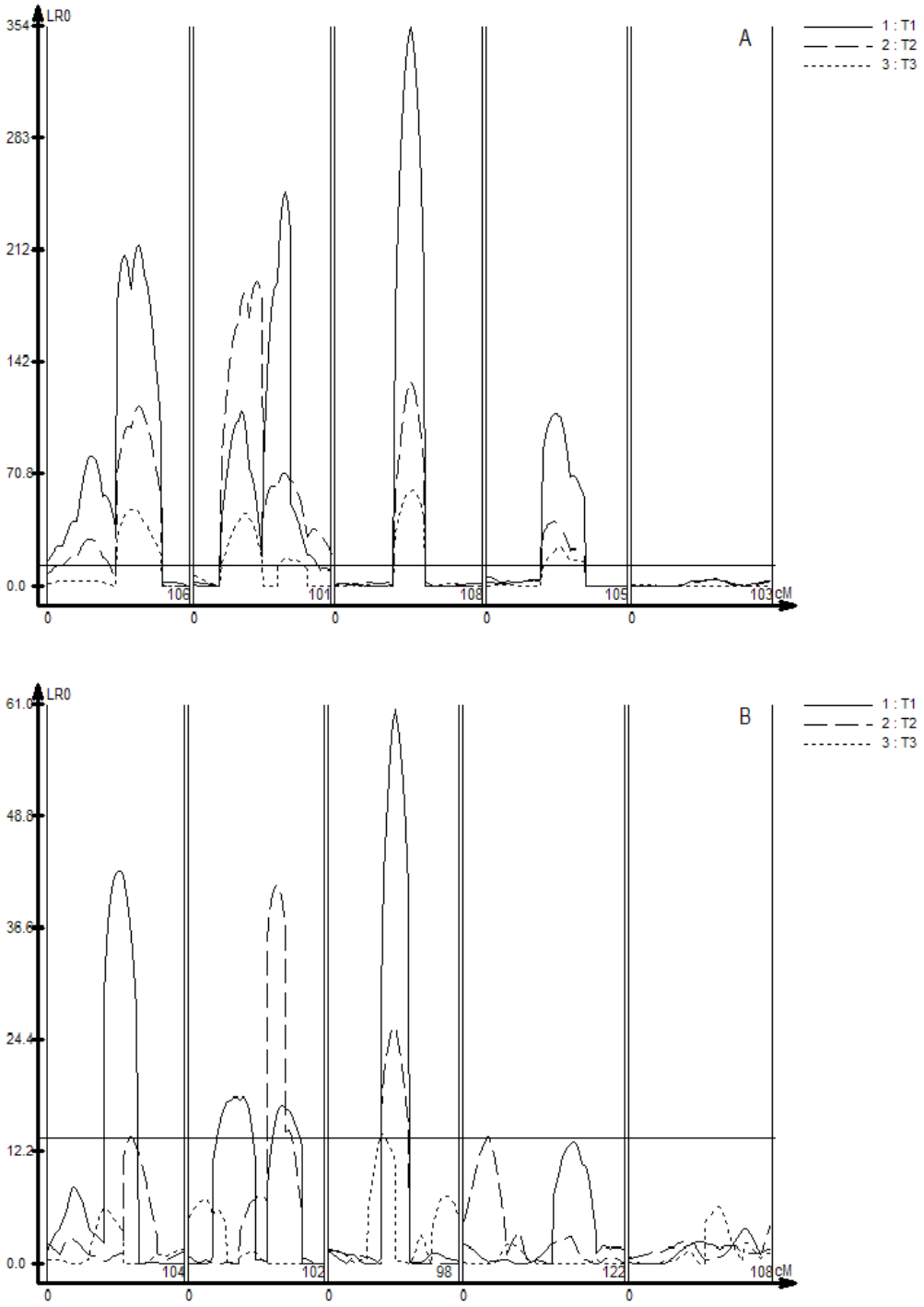

Figure 1. Influence of trait heritability and family size in the accuracy of QTL mapping by composite interval mapping (CIM). **A** Results here presented refer to the family with n=1000, replicate number 1. Trait 1 (–) (H^2=0.8), Trait 2 (- - -) (H^2=0.5) and Trait 3 (.....) (H^2=0.2). **B** Results here presented refer to the family with n=100, replicate number 1. Trait 1 (–) (H^2=0.8), Trait 2 (- - -) (H^2=0.5) and Trait 3 (.....) (H^2=0.2). In the X coordinate is shown the linkage groups separated by a double line. Distances are shown in cM. In the Y coordinate is shown the LR scores for each genomic position. A threshold LR score of 12.0 was set (solid horizontal line) to declare significant QTLs.

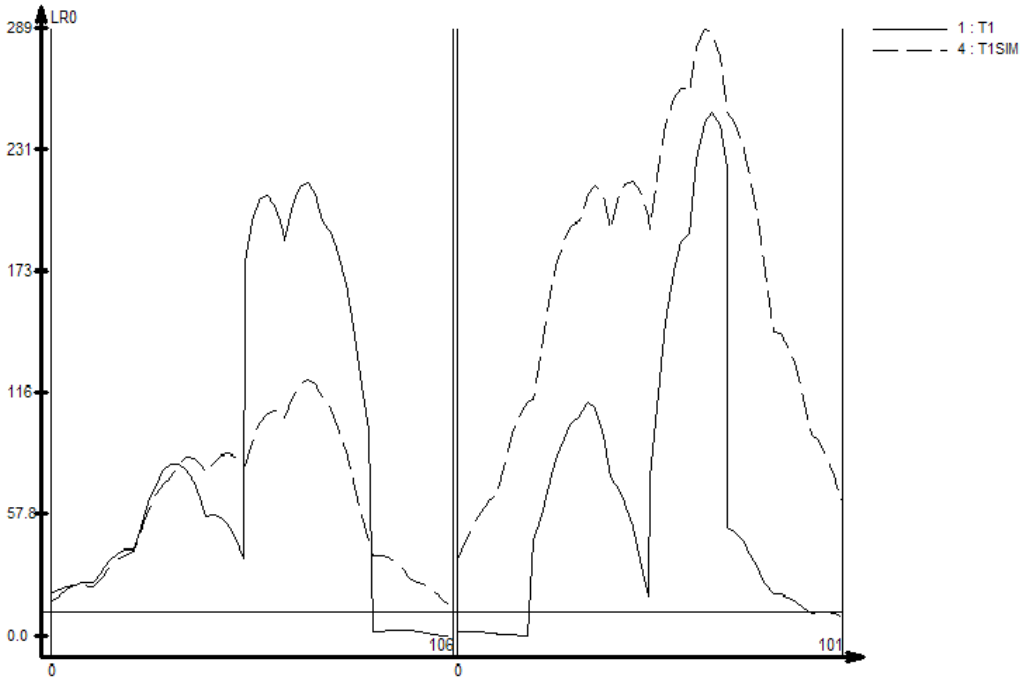

Figure 2. CIM correctly locate the two true QTLs instead of one ghost QTL indentified by SIM. Trait 1(H^2=0.8) analyzed with CIM (–) and Trait 1 analyzed with SIM (- - -). In the X coordinate is shown the linkage groups separated by a double line. Distances are shown in cM. In the Y coordinate is shown the LR scores for each genomic position. A threshold LR score of 12.0 was set (solid horizontal line) to declare significant QTLs. Results here presented refer to the family with n=1000, replicate number 1.

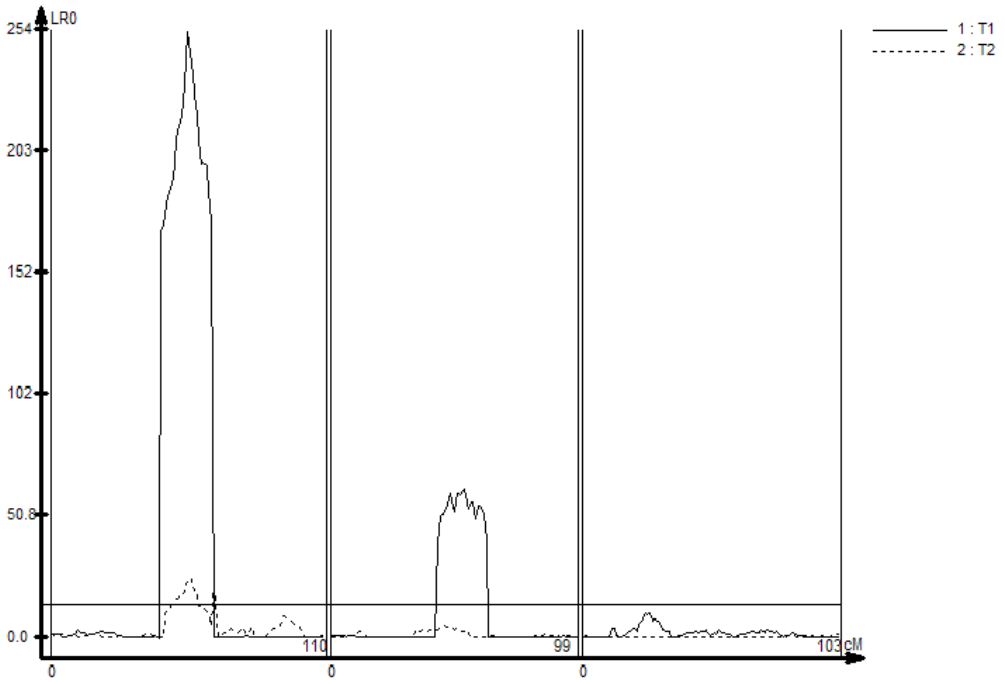

Figure 3. High density maps do not provide a framework to more accurate QTL mapping when compared to mid density maps. Trait 1 (–) (H^2=0.8) and Trait 2 (.....) (H^2=0.2). In the X coordinate is shown the linkage groups separated by a double line. Distances are shown in cM. In the Y coordinate is shown the LR scores for each genomic position. A threshold LR score of 12.0 was set (solid horizontal line) to declare significant QTLs. Results here presented refer to the family with n=1000, replicate number 1.

CHAPTER 3

LINKAGE AND QTL MAPPING IN FULL-SIB FAMILIES OF OUTBREEDING SPECIES

Paper to be submitted to Genetics and Molecular Biology

Linkage and QTL mapping in full-sib families of outbreeding species

Alexandre Alonso Alves[1,2], Leonardo Lopes Bhering[3], Lúcio Mauro da Silva Guimarães[1], Cosme Damião Cruz[2,3] and Acelino Couto Alfenas[1,2]

[1]Department of Plant Pathology, Federal University of Viçosa, Viçosa, MG, Brazil. [2]Graduate Program in Genetics and Breeding, Federal University of Viçosa, Viçosa, MG, Brazil. [3]Department of General Biology, Federal University of Viçosa, Viçosa, MG, Brazil.

Send Correspondence to Acelino Couto Alfenas. Department of Plant Pathology, Federal University of Viçosa, 36571-000 Viçosa, MG, Brazil. E-mail aalfenas@ufv.br

Abstract

As the procedures for linkage and QTL mapping in full-sib families of outbreeding species are quite diverse, we undertook a simulation study comparing the modified pseudo-testcross and the full-sib mapping designs in terms of marker ordering, distance between markers, total map size, distance variance and stress. We also investigated the power and precision of interval mapping procedures based on the full-sib and on the pseudo-testcross maps. We show that in general the pseudo-testcross and the full-sib mapping designs generate highly correlated maps with proportional linkage groups length. That independent of the QTL mapping approach used, *i.e.* CIM based on pseudo-testcross maps or Fulker and Cardon regression based on the full-sib map, detection power is reduced in small populations, especially in situations where trait heritability or QTL magnitude are low. We also found that although both methods appear to be equivalent in terms of QTL positioning for high heritability traits/major effect QTLs, the CIM based on pseudo-testcross maps provide more accurate data for low heritability traits/minor effect QTLs in larger populations. With regard to QTLs magnitude,

we show that both methods appear to be equivalent, and that the magnitude values tended to be overestimated for the high heritability trait, and underestimated for the low heritability trait, independent of the sample size. In terms of the QTLs additive effect, we found, however, that the estimates tended to be overestimated by CIM in the pseudo-testcross maps and underestimated by Fulker and Cardon regression in the full-sib map. Thus, for outbreeding species with mid-level of genomic resources, *i.e.* availability of larger set of multiallelic co-dominant markers, we propose that either the pseudo-testcross or the single full-sib mapping design and the related QTL mapping strategies can be used to generate genetic maps and map QTLs with high confidence.

Keywords: QTL mapping, full-sib families, pseudo-testcross, composite interval mapping and Fulker and Cardon regression.

Introduction

Genetic mapping has become a fundamental part of science and technology development in plant genetics. The availability of linkage maps has made possible a wide range of studies, including the identification and positional cloning of important genes and QTLs (Price 2006; Salvi and Tuberosa 2005). The wide array of scientific opportunities has motivated the development of increasingly dense and genetically informative linkage maps for a wide range of plant species. During the last decades, a number of genetic maps based on DNA markers (Collard et al. 2005), and recently RNA markers (Rostoks et al. 2005) have been constructed for nearly all of the most widely planted plant species.

Genetic mapping with outbreeding species, however, is far more difficult than with inbreeding species, due to the number of segregating alleles per locus/parent and the unknown linkage phase of the loci (Bhering et al. 2008). There are a number of ways to circumvent these complications (Maliepaard et al. 1997). For highly heterozygous species, such as most of the forest trees (*e.g. Eucalyptus* species and hybrids), genetic maps have been developed based

mainly on markers segregating in a double pseudo-testcross configuration in F_1 full-sib families (Grattapaglia and Sederoff 1994). In a cross between heterozygous parents, many dominant markers (*e.g.* RAPD markers) may be heterozygous in one parent and null in the other, and vice versa, therefore segregating in a testcross configuration (1:1 in their progeny). This enables the construction of individual genetic maps. Moreover, in a cross between heterozygous parents, data of many co-dominant markers (*e.g.* microsattelites) can also be re-coded so as the markers segregation reflect the parental origin of the alleles, enabling the positioning of such markers on the pseudo-testcross maps previously generated with dominant markers, or even the construction of pseudo-testcross maps exclusively based on highly polymorphic co-dominant markers (modified pseudo-testcross design) (Brondani et al. 2006). It is now possible however, to construct a single genetic map for a full-sib family derived from a cross between two highly heterozygous individuals, based on the information of all markers and individuals, as it is usually done with populations derived of a cross between two fully homozygous diploid parents. New mapping procedures now account for different markers type, different allele number and different linkage phases (Alves et al. 2010a; Bhering et al. 2008; Maliepaard et al. 1997).

Associations between phenotype and genotype have been analyzed in full-sib families either based on pseudo-testcross maps and on modified pseudo-testcross maps, hereafter referred collectively as pseudo-testcross maps or on single full-sib maps, hereafter referred as full-sib maps. Both types of maps apparently provide sufficient genome coverage for the identification of quantitative trait loci (QTLs) with significant effects on the expression of important traits. When the analysis is based on the pseudo-testcross maps, in most of the cases, interval mapping procedures, such as Composite Interval Mapping (CIM) (Zeng 1994), are used as if one were dealing with a conventional backcross population. Therefore, two separate QTL analyses are performed, one for each parental individual based on the two separate pseudo-testcross maps.

When the analysis is based on a full-sib map, procedures based on Haseman and Elston (1972) regression, such as the interval mapping technique developed by Fulker and Cardon (1994) are often used, along with random model approaches (Xu and Atchely 1995).

As the procedures for linkage and QTL mapping in full-sib families of outbreeding species are quite diverse, and not readily comparable, we undertook a study comparing the full-sib map and the modified pseudo-testcross mapping designs in terms of marker ordering, distance between markers, total map size, distance variance and stress. We also investigated the power and precision of interval mapping procedures based on full-sib and on pseudo-testcross maps. We highlight the implications of population size in linkage and QTL mapping, along with the implications of trait heritability and QTL properties in QTL mapping.

Methods

Simulations

The GQMOL software (Cruz 2010b) was used in the simulation of the data used in this study. This software allow the generation of information regarding genomes, parents, individual genotypes (of different mapping populations) as well as quantitative traits, based on random simulation. Using GQMOL simulation module, different scenarios were simulated based on different population sizes, different traits (with different heritabilities) and different QTL magnitudes. A total of eight scenarios (2 population sizes, 2 traits and 2 QTLs) were simulated and latter analyzed through 2 linkage and QTL mapping strategies.

Genome and parents design

To generate the genome information, a fictitious species with 2n=2x=6 was taken into account. Each linkage group was set to have a length of 100cM,

with 11 co-dominant markers (*e.g.* microsatellites) evenly distributed, totalizing 33 markers. One major effect QTL, explaining 25% of the genotypic variation was placed on linkage group 1 at a pre-determined position (50cM), and one minor effect QTL, explaining 5% of the genotypic variation, was placed on linkage group 2, also at a pre-determined position (50cM) to test the accuracy of interval mapping procedures based on a full-sib map (Fulker and Cardon regression) and on pseudo-testcross maps (composite interval mapping) in detecting, locating and estimating QTLs effect. No QTLs were placed in linkage group 3 in order to test the rate of false positives. Parents were generated to be completely informative (*i.e.* all loci with four alleles) and contrasting, by attributing a genotype for each marker. The mean degree of dominance was set to zero in all QTLs, to allow the estimation of the additive effect.

Full-sib populations design

Based on the simulated genome and parents, we generated via random simulation, full-sib populations with n=200 and n=1000. The population individuals were derived from a pool of 1000 gametes/individual derived from the parents previously simulated. The different population sizes were designed so as to test the influence of sample size in linkage procedures based on all marker information (full-sib map approach) and on pseudo-testcross mapping design, as well as in the accuracy of interval mapping procedures based on a full-sib map and on pseudo-testcross maps. Ten replicates (different populations based on the same genome and parents) were generated for each population size.

Quantitative traits design

Based on the 20 populations derived from the simulated genome (10 replicates for each size) data for two quantitative traits were generated. Trait 1 was set to have a heritability of 80%, mean of 100 and coefficient of variation of the error of 2%. Trait 2 was set to have a heritability of 20%, mean of 100 and

coefficient of variation of the error of 2%. All traits were set to be partially controlled by 2 QTLs, one located on linkage group 1 and one on linkage group 2 as previously described. Altogether the two QTLs accounted for 30% of the genotypic variation. A summary of quantitative traits and QTLs properties is presented in Tables 1 and 2, respectively.

Linkage and QTL mapping

Linkage analyses for each population were performed based on two distinct strategies using GQMOL (Cruz 2010b). Chi-square tests were performed in order to check the existence of marker distortion using the Bonferroni protection ($p \leq 0.1$). First we used the modified pseudo-testcross mapping approach (Grattapaglia and Sederoff 1994). For that, separate linkage maps for each parent were constructed by re-coding the markers data following the procedure describe hereafter. Considering a cross between two completely informative parents, *i.e.* 12x34, the progenies genotypes maybe either 13, 14, 23 or 24. To construct the pseudo-testcross map of the first parent the genotypes 13 and 14 must be re-coded as 1 and the genotypes 23 and 24 as 0. In this way only segregation of alleles from this parental is accounted for. To construct the pseudo-testcross map of the second parent the genotypes 13 and 23 must be re-coded as 1 and the genotypes 14 and 24 as 0. In this way only segregation of alleles from this parental is accounted for. After this procedure the maps are constructed in the same way that it is done to a backcross population. We then used a full-sib mapping approach, that is based on all marker information, and that results in one genetic map to each cross. For that, we have used the maximum likelihood functions described by Bhering et al. (2008). These likelihood functions are based on the expected recombination frequency between markers in different configurations, *i.e.* different linkage phases, for co-dominant markers in full-sib families. In both strategies, linked markers were placed onto linkage groups with a threshold LOD score of 3.0 and a maximum recombination fraction (ϑ) of 0.30. Recombination fractions were transformed to

estimated map distances by the Kosambi map function, which assumes that recombination events influence the occurrence of adjacent recombination events (interference between crossover events) (Schuster and Cruz 2008). Based on the pseudo-testcross maps composite interval mapping (Zeng 1994) was used for QTL mapping. Based on the full-sib map, the Fulker and Cardon regression (Fulker and Cardon 1994) was used for QTL mapping. Both methods work based on marker intervals. CIM, however, combines simple interval mapping with linear regression by including additional genetic markers in the statistical model in addition to an adjacent pair of linked markers for interval mapping (Zeng 1994). The parameters used for both methods were a threshold LR = 12 (~LOD = 3) to declare significant QTLs and a 2cM genome scanning step. In CIM, we adopted a *p*-value of ≤ 0.1 for entering variables and a *p*-value ≤ 0.1 for removing variables in the forward and backward stepwise regression of the residual phenotype on marker variables. The stepwise regression was used to select the co-factors included in CIM. All linkage and QTL analyses were performed using GQMOL (Cruz 2010b).

Comparisons between the pseudo-testcross maps and the full-sib map

The comparisons between the pseudo-testcross maps and the full-sib map were performed in terms of linkage group size (in cM); map variance; marker ordering, measured by the Spearman correlation coefficient; distance between markers, measured by the Pearson correlation and stress, which measures the effect of the changes in the marker distances. All comparisons tests were performed with GQMOL (Cruz 2010b).

Statistical analysis

All data of genome comparison, *i.e.* linkage group size, map variance, Spearman and Pearson correlation coefficients and stress along with data of QTL mapping, *i.e.* position, magnitude of variation (r^2) and additive effect (a), were analyzed through ANOVA using an unbalanced completely randomized design.

Means were further analyzed with a Tukey test ($p \leq 0.05$) to statistically test map consistency and the interval mapping procedures in locating and estimating QTLs effect and magnitude considering the different population sizes, trait heritabilities, QTL magnitude, position and effect. ANOVAs and Tukey tests were performed with the aid of the software GENES (Cruz 2010a).

Results

Linkage mapping analysis

Two distinct linkage mapping approaches often used in full-sib families, *i.e.* the modified pseudo-testcross and the single full-sib mapping designs, were confronted in order to test how one relate to another. Genetic maps, constructed using the pseudo-testcross and the single full-sib mapping design are presented in Figure 1. In this figure, the dotted lines connect the same markers in the different maps generated for the same family. It is possible to verify that in this particular case (replicate 1) no single marker ordering change was observed. Indeed, the Spearman correlation, which measures marker ordering consistency, was 1.0 in all situations (Table 3), indicating that not even in a single case a marker ordering change was observed. In terms of markers distance consistency, we found that on average the correlation between the markers distance in the full-sib map and the markers distance in the pseudo-testcross maps for the first and second parent were, respectively, of 66 and 67%, when size n=200 is considered and of 66 and 62%, when size n=1000 is considered (Table 3). These results indicate that sample size did not have any detectable effect in the markers distance consistency, and that in general the pseudo-testcross and the full-sib mapping designs generate highly correlated maps. However, we noted that pseudo-testcross maps of larger populations, provided lower values of stress when compared to the smaller populations when compared to the full-sib map (of a large and small population, respectively) (Table 3). On average stress values were 116 and 114% higher in smaller populations, for parents 1 and 2,

respectively. As the stress coefficient measures the effect of the changes in the marker distances, smaller populations seem to be more prone to suffer changes.

The length of the linkage groups in the pseudo-testcross maps was proportional to the length of the linkage groups in the full-sib map (Table 4 and Figure 1). Although in general the linkage groups of the pseudo-testcross map of the second parent were longer (105.8cM on average) than, those of the full-sib map (103.8cM on average), and for its turn the linkage groups of the full-sib map longer than those of the pseudo-testcross map of the first parent (102.7cM), no statistical significant differences were found (Table 4). Our data, however, suggest that the pseudo-testcross mapping design provide genetic maps with elevated map variance when compared to the single full-sib map (Table 4). On average the variance in the pseudo-testcross maps were estimated in 4.97 and 0.99 for populations of size n=200 and n=1000, respectively. The variance estimates for the full-sib maps were 2.93 and 0.47 for n=200 and n=100, respectively, *i.e.* on average 50% lower in both cases. We noted that sample size did not have any detectable effect on linkage groups length and on map variance (Table 4).

QTL mapping analysis

Independent of the QTL mapping approach used, *i.e.* CIM based on pseudo-testcross maps or Fulker and Cardon regression based on the full-sib map, when we tested a relatively small population (*i.e.* 200 individuals) we found that power is reduced when compared to larger populations (*i.e.* 1000 individuals), especially in situations where trait heritability or QTL magnitude are low (Table 5). It is interesting to note that, even using a large population, the minor effect QTL was detected on average only in 50% of times. For the high heritability trait/major effect QTL situation, population size did not interfered with QTL detection power. Comparing the detection power of the CIM with the detection power of Fulker and Cardon regression, it is verified that the detection power of the full-sib map approach is slightly increased for low heritability

traits/minor effect QTL while on the other hand the pseudo-testcross map approach power is a bit higher for high heritability traits/minor effect QTLs. The CIM approach detected QTLs only in the first parental pseudo-testcross map, and no QTLs were detected on linkage group 3 by both methods, *i.e.* no false positive QTLs were detected, as expected.

Considering the precision of QTL mapping it interesting to note that, only in one case (low heritability trait/minor effect QTL), we found statistical significant differences in the QTL positioning by both approaches (Table 5). In this case, population size appears to be the major obstacle, as the positioning of QTL in the smaller population tended to be biased by both methods. However, considering only the large population, the QTL mapping approach based on pseudo-testcross maps seem to provide more accurate data, as the major and minor QTLs detected, respectively via CIM and Fulker and Cardon regression, are positioned within 2.5 and 1.15cM, 2.35 and 3.55cM of the original position. In general, however, QTL mapping approaches either based on pseudo-testcross or on full-sib maps proved to be pretty accurate in terms of QTL positioning as, on average the mapped QTLs are within 2cM of its simulated position. Taken together our results suggest that both methods appear to equivalent in terms of QTL positioning for high heritability traits/major effect QTLs, and that the QTL mapping approach based on pseudo-testcross maps provide more accurate data for low heritability traits/minor effect QTLs in larger populations.

Also considering the precision of QTL mapping, it is noteworthy that, in terms of estimated QTLs magnitude, *i.e.* the percentage of the genotypic variation explained by the QTL, r^2 values tended to be overestimated for the high heritability trait, and underestimated for the low heritability trait, independent of the sample size and of the mapping approach (Table 5). However, r^2 values are considerable smaller for the larger population, especially considering the major effect QTL. For example, the estimated magnitude of the major effect QTL, for traits with high and low heritability, respectively, were estimated as 29.1 and 8.8 for sample size n=1000 and 33.4 and 12.15 for sample size n=200, *i.e.* 12,8% and

27.5% lower. This same trend also occurred for the minor effect QTL of the low heritability trait (64.5% of reduction on average) (Table 5). Comparing the approaches used for QTL mapping, only in one situation (small population and major effect QTL for the low heritability trait) we found significant statistical differences between the estimated r^2 values (Table 5). In general then, our results indicate that with regards to the estimation of the QTLs magnitude, the QTL mapping approach based on pseudo-testcross maps appear to be equivalent to QTL mapping approach based on a full-sib map.

With regards to the precision of the QTL mapping approaches, in terms of the QTLs additive effect, we found that *a* estimates tended to be overestimated by CIM in the pseudo-testcross maps and underestimated by Fulker and Cardon regression in the full-sib map (Table 5), with one only exception. In the case of the major effect QTL for the high heritability trait, the *a* estimates obtained by the Fulker and Cardon regression were on average 39% higher than the *a* estimates obtained by CIM. We also noted that, only in one case (minor effect QTL for the low heritability trait), the sample size had a significant effect on the *a* estimates. In this particular case, the estimate for the additive effect based on the larger population seems to be more accurate when considering CIM while considering the Fulker and Cardon regression, the smaller population provided more accurate data.

Discussion

We have attempted to make a comparison between linkage and QTL mapping procedures commonly used in full-sib families. First we compared the modified pseudo-testcross mapping design (Grattapaglia and Sederoff 1994) with a full-sib mapping design that uses all marker information (Bhering et al. 2008; Maliepaard et al. 1997). It is well known that in full-sib families, markers may vary in the number of segregating alleles (up to four), by one or both parents being heterozygous, markers being dominant or co-dominant, and usually the linkage phases of marker pairs are unknown. Because of these particularities

genetic mapping in outbreeding is not as straightforward as it is in inbreeding species. One of the strategies used for genetic mapping a full-sib family is then the pseudo-testcross mapping design that is based on the fact that in allogamous plant species for which only heterozygous individuals are available, one can make use of single dose polymorphic markers, segregating in a 1:1 ratio, to generate linkage maps for individual plants. The pseudo-testcross mapping strategy was initially implemented with dominant RAPD markers (Grattapaglia and Sederoff 1994), but as highly polymorphic, multiallelic, co-dominant markers that detected all four allelic variants of the mating configuration (*e.g.* microsatellites) contains more genetic information, the strategy was later modified to make use of such markers. For that, marker data must be re-coded in two separate groups, so as that data reflect the parental origin of the alleles in each group. Pseudo-testcross maps can be integrated to construct a consensus map, and as nowadays multiallelic co-dominant markers with alleles segregating from both parents are often used, the set of common loci is used as bridges. Complete maps resulting from merging of pseudo-testcross maps have been reported recently (Brondani et al. 2006). Another strategy is to build a single full-sib family map. Maliepaard et al. (1997), Bhering et al. (2008) and Alves et al. (2010a) provided useful formulas for estimating recombination frequency in every situation in full-sib families. This includes segregation in one or both parents, dominance and all linkage phase configurations. This later strategy offers the advantage of constructing a single full-sib map, based on data of both parents and markers, as it is usually done in inbreeding populations.

Using a simulation approach we compared the pseudo-testcross and the full-sib mapping approach using two different sample sizes. It is noteworthy that we have used only fully informative markers, *i.e.* all markers detected four alleles, and both parental individuals were completely informative and contrasting (thus of genotype 12 or 34). In this situation information content is elevated and the variance of the estimated recombination frequencies are minimal, thus allowing markers to be placed onto linkage groups with high

confidence (Alves et al. 2010a; Bhering et al. 2008). Another advantage is that when the data is re-coded so as to allow the use of the pseudo-testcross mapping design, all markers are represented in both parental maps. In this ideal situation, our data suggest that in general the pseudo-testcross and the full-sib mapping designs generate highly correlated maps, as indicated by the Spearman and Pearson correlations coefficients, with proportional linkage groups length independent of the sample size. We did note, however, that map variance was lower when we used a larger population, indicating that although closely resembled, maps generated with elevated number of progenies are more accurate than maps generated based on fewer individuals in terms of markers recombination fractions. However, in a situation, where markers detect less than four alleles and, or if the parental individuals are not completely contrasting for marker loci, markers are mapped with reduced confidence. Moreover, in the case of the pseudo-testcross maps, a set of markers maps onto parent 1 map, while another set of markers (which can share some, but not all markers) maps onto parent 2 map. In this situation it is unlikely that the marker ordering will hold, and that the distances between markers will be comparable between the pseudo-testcross maps with the full-sib map. Even knowing this, we have chosen to use only completely informative co-dominant markers, and completely informative parental individuals, because most of the outbreeding species is highly heterozygous and thus there are increased chances that co-dominant markers, such as microsatellites, segregate in both parents by detecting four alleles, if this is the case. Thus, for outbreeding species with mid-level of genomic resources, *i.e.* availability of large sets of multiallelic co-dominant markers, we propose that either the pseudo-testcross or the single full-sib mapping design can be used to generate genetic maps with high confidence, especially if sample size is elevate, to be used as a framework for QTL mapping.

With regards to the QTL methods commonly used in full-sib families we have chosen to work with composite interval mapping (Zeng 1993, 1994) that uses the pseudo-testcross maps and the Fulker and Cardon regression (Fulker

and Cardon 1994) that use the single full-sib map, because both methods use an estimated genetic map as the framework for locating QTLs, apart from being simple and straightforward. The intervals, defined by a pair of flanking markers, are searched in linear increments (one-dimensional scan), and statistical methods are used to test whether a QTL is likely to be present at the location within the interval or not in both methods. The Fulker and Cardon regression is an extension of the sib-pair regression developed by Haseman and Elston (1972). In this method the QTL IBD (identity by descent) is estimated between two sibs, based on the proportion of the adjacent markers IBD. Then the squared difference of the sib's phenotype is regressed onto the QTL IBD. As the method estimate the QTL IBD along the marker interval, the QTL position and effect are estimated separately. Composite interval mapping (Zeng 1994) combines simple interval mapping with linear regression by including additional genetic markers in the statistical model in addition to an adjacent pair of linked markers for interval mapping. The idea is that the inclusion of additional markers as cofactors - outside a defined window of analysis - helps removing the variation that is associated with other (linked or unlinked) QTLs in the genome. This fact makes CIM more effective at mapping QTLs, both by effectively locating and estimating its effect (Schuster and Cruz 2008). In that context, our data suggest that independent of the QTL mapping approach used, detection power is reduced when small populations are used, especially in situations where trait heritability or QTL magnitude are low. In this way, since most of the QTLs mapping studies published to date, in out-breeding species, uses populations of ≤200 individuals (most of the times due to limited experimental resources), the majority of minor effect QTLs for low heritability traits may have been undetected due to inappropriate population design. Therefore, if the objective is to map and use, if not all, the majority of QTLs involved in the genetic control of a trait of interest, on marker assisted selection (MAS) breeding programs the use of large populations is paramount.

As to the QTL positioning we found that both methods also appear to equivalent, at least for major effect QTLs of high heritability traits. For low heritability traits/minor effect QTLs, mapping approach based on pseudo-testcross maps provide more accurate data in larger populations. Although it might seem intriguing, since parents were design to be contrasting for the QTLs, by analyzing only the alleles that segregate for one of the parents, and by using co-factors so as to remove the variation that is associated with other unlinked QTLs in the genome, CIM generated more accurate data. If the objective is then to fine map or clone the gene(s) that underlie minor effect QTLs, the CIM approach based on pseudo-testcross maps should be preferentially used if one suspect that the QTL alleles are fixed in one of parental individuals. In general, however, both QTL mapping approaches proved to be pretty accurate in terms of QTL positioning, as mapped QTLs are on average 2cM far from its original simulated position. It is important to note, however, that in a 2cM window, hundreds of genes can occur, depending on the genome size and organization, and thus although the QTL mapping approaches proved to be pretty accurate, still there are a lot of genes to be tested if the objective is to identify the gene that underlie a QTL for a quantitative trait of interest or clone it.

The primary goal of genetic mapping experiments is to identify the locations and but also estimate the effects of the QTLs that affect the expression of a trait of interest (Xu 2003). The “Beavis effect” (Beavis 1994) state that average estimates of genotypic variances associated with correctly identified QTLs are greatly overestimated if only 100 individuals are evaluated in a F_2 population and fairly close to the actual magnitude when 1000 individuals are evaluated (Xu 2003). Here, we show that QTLs magnitude estimates tended to be overestimated for the high heritability trait, especially in the smaller population, and underestimated for the low heritability trait, especially in the large population, independent of the sample size and the mapping approach. As Beavis used a simple interval mapping (SIM) approach and F_2 populations, and we have used CIM and pseudo-testcross populations, or Fulker and Cardon

regression and a full-sib map, the results are no readily comparable and no conclusion can be drawn. However, our results are in agreement with those obtained by Alves et al. (2010b) that found that the magnitude estimates tended to be underestimated by CIM in backcrossing populations of large size, noted when analyzing traits with low heritability. As the pseudo-testcross maps closely resemble a BC map, maybe there is evidence that CIM indeed underestimate r^2 values, in such populations. In terms of the QTLs additive effect, we found that the estimates tended to be overestimated by CIM in the pseudo-testcross maps and underestimated by Fulker and Cardon regression in the full-sib map with one only exception (in the case of the major effect QTL for the high heritability trait). It is interesting to note here that, if QTL data is to be used in MAS, as the r^2 values are underestimated in large populations, the efficiency of MAS for its turn should be also, because as pointed out by Lande and Thompson (1990), the MAS efficiency depend upon the genetic variation explained by the molecular markers (*g*). Since *g* equals to m^2/h^2, where m^2 and h^2 are, respectively, the proportion of the phenotypic variation explained by the markers and the heritability of the trait, the largest the proportion of the phenotypic variation explained by the QTLs the more effective should be MAS. Ideally, however, one would choose to include in a MAS breeding program only the major effect QTLs (high r^2 values then) that have the largest and more accurate predicted additive effect. In this situation the QTLs mapped by CIM in pseudo-testcross maps could have an upward bias, while the QTLs mapped by Fulker and Cardon regression a downwards bias.

Finally, for outbreeding species with mid-level of genomic resources, *i.e.* availability of larger set of multiallelic co-dominant markers, we propose then that either the pseudo-testcross or the single full-sib mapping design and the related QTL mapping strategies can be used to generate genetic maps and map QTLs with high confidence. It is important to highlight however, that, other studies, using different scenarios, *i.e.* different coefficients of variation of the error, different number of QTLs, different marker distributions, which collectively

may make the simulation a bit more realistic, are needed in order to see if the results of our work hold true in every situation.

Acknowledgments

We are grateful to Caio Césio Salgado, for his constructive comments on the manuscript and for the assistance in the numerous simulations. The Bioinformatics Lab of the Federal University of Viçosa, Brazil provided the facilities for the development of this work. This work was supported by the Brazilian National Research Council, CNPq, with a Ph.D. fellowship to AAA, post-doctoral fellowship to LMSG and a research fellowship to ACA and CDC.

References

Alves AA, Bhering LL, Cruz CD, Alfenas AC (2010a) Linkage analysis between dominant and co-dominant makers in full-sib families of out-breeding species. Genetics and Molecular Biology 33:499-506

Alves AA, Bhering LL, Guimarães LMS, Cruz CD, Alfenas AC (2010b) QTL mapping in simulated backcrossing populations: implications of population size, trait heritability, QTL properties and marker density. Genetics and Molecular Biology in press

Beavis WD (1994) The power and deceit of QTL experiments: Lessons from comparative QTL studies. Corn Sorghum Ind Res Conference. Am. Seed Trade Association, Chicago, IL, pp 250-266

Bhering LL, Cruz CD, God PIVG (2008) Estimation of recombination frequency in genetic mapping of full-sib families. Pesquisa Agropecuária Brasileira 43:363-369

Brondani RPV, Williams ER, Brondani C, Grattapaglia D (2006) A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus. BMC Plant Biology 6:16

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196

Cruz CD (2010a) Genes: a software for genetics analysis. Universidade Federal de Viçosa, Viçosa, MG, Brazil

Cruz CD (2010b) GQMOL: a software for quantitative and genetics analysis. Universidade Federal de Viçosa, Viçosa, MG, Brazil

Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. American Journal of Human Genetics 54:1092-1103

Grattapaglia D, Sederoff R (1994) Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross mapping strategy and RAPD markers. Genetics 137:1121-1137

Haseman JK, Elston RC (1972) Investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3-19

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genetical Research 70:237-250

Price AH (2006) Believe it or not, QTLs are accurate! Trends in Plant Science 11:213-216

Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, Morris J, Cardle L, Marshall DF, Waugh R (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biology 6

Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. Trends in Plant Science 10:297-304

Schuster I, Cruz CD (2008) Estatística Genômica aplicada a populações derivadas de cruzamentos controlados, 2th edn. Editora UFV, Viçosa

Xu S (2003) Theoretical Basis of the Beavis Effect. Genetics 165:2259-2268

Xu S, Atchely WR (1995) A random model approach to interval mapping of quantitative trait loci. . Genetics 141:1189-1197

Zeng ZB (1993) Theoretical basis of precision mapping of quantitative trait loci. Proceedings of the National Academic of Science 90:10972-10976

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457-1468

Table 1. Summary of quantitative traits properties in simulated full-sib populations.

Trait*	H^2 (%)$^{+}$	Nº of QTLs	CVe^{++}	Mean	Evr. Var.§	Gen. Var.§§
1	80	2	2	100	4	16
2	20	2	2	100	4	1

*Traits designed based on the full-sib populations. $^{+}$Heritability. $^{++}$Coefficient of variation of the error. §Environmental variance. §§Genetic variance.

Table 2. Summary of QTLs responsible for genetic control of traits designed based on the full-sib populations properties.

QTL*	Position[=]	PVE[+]	*mdd*[++]	*a1*[§]	*a2*[§§]
11[¯]	50	25	0	2.83	0.71
12	50	5	0	1.26	0.32

*QTLs responsible for genetic control of traits designed based on the full-sib populations. [=]QTL position in cM from leftmost marker. [¯]QTLxy – x=linkage group; y=QTL number. [+]Percentage of the genotypic variation explained by the QTL. [++]Mean degree of dominance. [§]Additive effect of the QTL with regards to trait 1. [§§]Additive effect of the QTL with regards to trait 2.

Table 3. Spearman and Pearson correlations, and stress between the pseudo-testcross maps and the full-sib map.

Maps	Spearman Correlation			Pearson Correlation			Stress		
	1	2	3	1	2	3	1	2	3
200 – PTM1^{+}/200 - FSM	1a	1a	1a	0.63a	0.77a	0.62a	14.93a	14.56a	14.01a
200 – PTM2^{++}/200 - FSM$^{¬}$	1a	1a	1a	0.67a	0.63a	0.70a	14.21a	14.42a	14.82a
1000 – PTM1/1000 - FSM	1a	1a	1a	0.70a	0.67a	0.63a	6.55b	6.66b	7.06b
1000 – PTM1/1000 - FSM	1a	1a	1a	0.64a	0.72a	0.51a	6.40b	6.62b	7.02b

*Means followed by the same letter in the columns do not differ statistically between them, by a *t*-test ($p \leq 0.05$). ** Average values for the 10 replicates. $^{+}$ PTM1 – pseudo-testcross map parent 1. $^{++}$PTM2 – pseudo-testcross map parent 2. $^{¬}$FSM – full-sib map. LG – linkage group.

Table 4. Mean size of and mean variance of linkage groups, of both pseudo-testcross maps and of the full-sib map.

Maps	Mean size/LG			Mean variance/LG		
	1	2	3	1	2	3
200 – PTM1^{+}	101.57a	102.34a	98.12a	4.16a	6.10a	4.84a
200 – PTM2^{++}	110.16a	104.42a	104.56a	4.88a	4.56b	5.28a
200 – FSM $^{¯}$	103.85a	102.99a	100.97a	2.09b	2.76c	2.33b
1000 – PTM1	104.92a	104.87a	104.63a	1.05c	1.30d	0.90bc
1000 – PTM1	105.01a	106.18a	105.02a	0.87c	0.94d	0.92bc
1000 – FSM	104.78a	105.49a	104.81a	0.44c	0.61d	0.36c

*Means followed by the same letter in the columns do not differ statistically between them, by a Tukey test ($p \leq 0.05$). ** Average values for the 10 replicates. $^{+}$ PTM1 – pseudo-testcross map parent 1. $^{++}$PTM2 – pseudo-testcross map parent 2. $^{¯}$FSM – full-sib map. LG – linkage group.

Table 5. Power and precision of QTL mapping by composite interval mapping (CIM) in the pseudo-testcross maps and by Fulker and Cardon regression in the full-sib map.

Trait/Marker density	QTL1					QTL2				
	Pow[+]	Pos[++]	LR	$R^{2§}$	$a^{§§}$	Pow	Pos	LR	R^2	*a*
Trait 1 (H^2=0.8)										
200 – PTM	100	50.9a	89.5a	33.0a	5.64a	90	47.1a	23.3a	7.35a	2.84a
200 – FSM	100	50.9a	581.0a	33.8a	7.96b	60	49.6a	30.1a	7.31a	1.78b
1000 – PTM	100	52.2a	356.7a	28.1b	5.64a	100	50.8a	92.5a	6.24a	2.63a
1000 – FSM	100	51.8a	11768.5b	30.1ab	7.76b	100	52.5a	589.4b	6.48a	1.64b
Parametric	-	50	-	25	4	-	50	-	5	1.7889
Trait 2 (H^2=0.2)										
200 – PTM	70	51.8a	25.8a	11.2ab	1.50a	10	61.4a	13.1a	6.04a	1.06a
200 – FSM	80	54.8a	82.3a	13.1b	0.65b	10	46.4c	25.6a	7.05a	0.46c
1000 – PTM	100	53.2a	89.8a	8.5a	1.43a	60	51.5bc	22.9a	2.13b	0.67b
1000 – FSM	100	52.9a	1126.2b	9.1a	0.50b	70	54.6b	62.2a	2.15b	0.11d
Parametric	-	50	-	25	1	-	50	-	5	0.4472

*Means followed by the same letter in the columns do not differ statistically between them, by a Tukey test ($p \leq 0.05$). ** Average values for the n replicates. [+] Power of detection (%). [++] QTL position in cM from leftmost marker. [§]Genotypic variation explained by the QTL. [§§]Additive effect of the QTL. LR – likelihood ratio. LR – likelihood ratio. PTM – pseudo-testcross map. FSM – full-sib map.

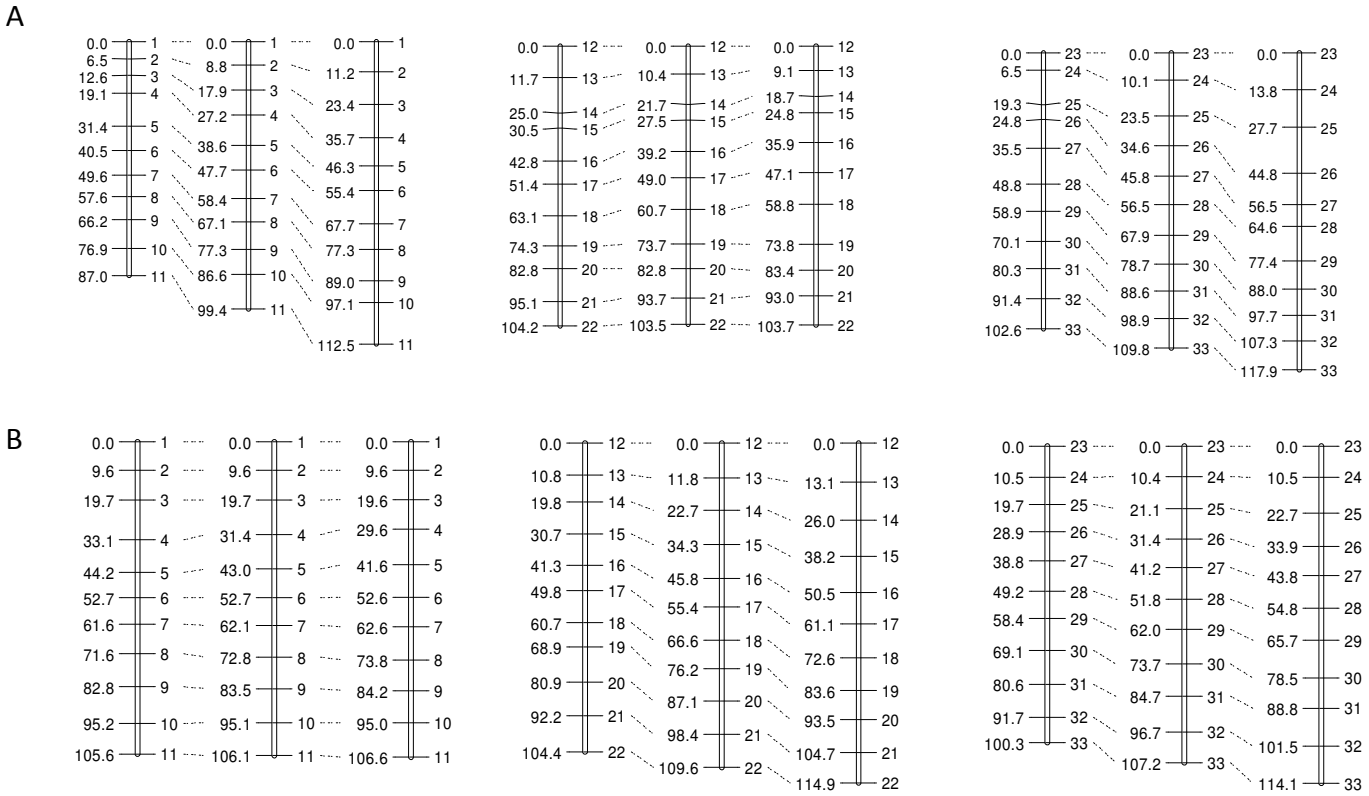

Figure 1. A Genetic maps based on a full sib of size n=200 and **B** Genetic maps based on a full sib of size n=1000. Linkage groups one, two and three are shown from left to right (in groups). On the left of each group is shown the pseudo-testcross map for the first parent, in the center the single map constructed based on all marker data, and on the right the pseudo-testcross map for the second parent. Dotted lines connect the same markers in different maps. Distances in centiMorgan (cM) Kosambi are indicated on the left of each linkage group. Marker names are shown in the right. Maps here shown refer to replicate number 1 of the populations.

GENERAL CONCLUSIONS

Chapter 1

- The complete set of maximum likelihood estimators for recombination frequency between molecular markers in full-sib families is now complete with the addition of the nine estimators.
- All combinations of molecular markers with two to four alleles (without epistasis) in a full-sib family are now accounted for. This includes segregation in one or both parents, dominance and all linkage phase configurations.
- Based on its properties and implementation into free linkage software, the approach presented in *Chapter 1* should be useful for those interested in using dominant and co-dominant molecular markers for mapping, or as an aid in selecting out-crossing species.

Chapter 2

- Sample size has a major implication in the detection power and as consequence in the estimation of the magnitude and additive genetic effect.
- Small populations do not allow mapping of low effect QTLs, especially if these QTLs are involved in the genetic control of traits with low heritability.
- The positioning of the QTLs based on CIM is more accurate than SIM and that on average the mapped QTLs are close to their simulated position.

- CIM tend to underestimate the magnitude r^2 values especially in large population sizes/ low heritabilities traits and overestimate it in smaller populations.
- SIM tended to overestimate r^2 values when linked QTLs occurred, markedly in smaller population.
- When a *ghost QTL* is located instead of the true QTLs, the positioning and as consequence the estimation of the r^2 and of the genetic additive effect (*a*) values is compromised
- Giving adequate genome coverage the addition of more molecular markers does not seem to improve the precision of QTL mapping.
- When one is dealing with adequate sample size and genome coverage even mid-density genetic maps can be used to map QTLs of large or small effect with high confidence.
- Detection power is restricted by sample size, and not due to low marker density, giving adequate genome coverage.

Chapter 3

- In general the modified pseudo-testcross and the full-sib mapping designs generate highly correlated maps with proportional linkage groups length.
- Independent of the QTL mapping approach used, *i.e.* CIM based on modified pseudo-testcross maps or Fulker and Cardon regression based on the full-sib map, detection power is reduced in small populations, especially in situations where trait heritability or QTL magnitude are low.
- Although both methods appear to equivalent in terms of QTL positioning for high heritability traits/major effect QTLs, the CIM based on pseudo-testcross maps provide more accurate data for low heritability traits/minor effect QTLs in larger populations.
- With regard to QTLs magnitude, both methods appear to be equivalent, and the magnitude values tended to be overestimated for the high

heritability trait, and underestimated for the low heritability trait, independent of the sample size.

- The estimates of the QTLs additive effect tended to be slightly overestimated by CIM in the pseudo-testcross maps and slightly underestimated by Fulker and Cardon regression in the full-sib map.
- Thus, for outbreeding species with mid-level of genomic resources, *i.e.* availability of larger set of multiallelic co-dominant markers, either the modified pseudo-testcross or the single full-sib mapping design and the related QTL mapping strategies can be used to generate genetic maps and map QTLs with high confidence.