**University of São Paulo
“Luiz de Queiroz” College of Agriculture**

**Construction of a high-density genetic map of *Acca sellowiana* (Berg.)
Burret based on two connected mapping populations**

Marianella Fernanda Quezada Macchiavello

Thesis presented to obtain the degree of Doctor in Science.
Area: Genetics and Plant Breeding

**Piracicaba
2017**

Marianella Fernanda Quezada Macchiavello
Bsc. in Biological Sciences

**Construction of a high-density genetic map of *Acca sellowiana* (Berg.)
Burret based on two connected mapping populations**

Advisor:

Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2017

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Quezada Macchiavello, Marianella Fernanda

Construction of a high-density genetic map of *Acca sellowiana* (Berg.)
Burret based on two connected mapping populations / Marianella Fernanda
Quezada Macchiavello. – – Piracicaba, 2017 .

61 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura “Luiz de
Queiroz”.

1. *Acca sellowiana* 2. Genotipagem por sequenciamento (GBS) 3. Mapa
genético 4. Myrtaceae . I. Título.

DEDICATORY

*A mi mamá, **Silvia**,
por todo su amor y apoyo incondicional
A **Hugo**,
por acompañarme en este camino
de crecimiento profesional y personal.
Dedico*

ACKNOWLEDGEMENTS

A mi orientador Antonio Augusto Franco Garcia, por haberme recibido en el laboratorio de Genética Estadística, por la orientación en este trabajo, todo su apoyo y enseñanzas fueron esenciales para el desarrollo de este trabajo.

También quería agradecer a Clara Pritsch y todo el Lab. de Biotecnología Vegetal, que a la distancia siempre estuvieron brindando apoyo y aportando ideas para la realización de este trabajo. Agradecer también a Beatriz Vignale y Danilo Cabrera, por el soporte brindado en el desarrollo de las actividades de campo que forman parte de este proyecto. A los estudiantes Gonzalo, Máximo, Valeria, Gustavo y Santiago que trabajaron duramente con los guayabos en el campo. A los profesores del curso de Genética de Facultad de Agronomía.

Quisiera agradecer a la Escola Superior de Agricultura Luiz de Queiroz, de la Universidad de São Paulo, por permitirme realizar las actividades de Doctorado. A los profesores y equipo técnico del programa de postgrado de Genética y Mejoramiento de Plantas, que siempre me apoyaron y auxiliaron en todas las actividades académicas y administrativas.

A mis compañero del Laboratorio de Genética Estadística, con los cuales compartí cuatro excelentes años de crecimiento no sólo académico sino también personal. Primero agradecer a aquellos que me recibieron por primera vez: Maria Marta, Guilherme, Carina, Adriana, João, Marcelo y Graciela. Gracias a ellos y con su ayuda decidí volver por más tiempo para realizar este trabajo. De la turma actual, mis agradecimientos para Leticia, Cris, Cateto, Matheus, Kaio, Gabriel, Jhonathan, Eliane, Felipe y Danilo. Especialmente quería agradecer el cariño y compañerismo compartido con Letícia, con la cual comenzamos y terminamos juntas este recorrido académico. También agradecerle la participación de Cateto que auxilió en la programación de toda la metodología presentada en este trabajo. A Guilherme, que mismo a distancia auxilió los trabajos de SNP calling y construcción de mapas. A Cris, por nuestras entretenidas conversaciones sobre estadística, mapas y la vida. A Carina y João, un gran ejemplo de dedicación y trabajo.

Quisiera agradecer a mi familia, mi mamá, mis hermanas Alejandra y Carina, y a Ezequiel y Nicolás, que siempre me acompañaron y apoyaron en este proceso. Su cariño y apoyo fueron esenciales durante esta etapa.

A Hugo, que se embarcó conmigo en este proyecto académico, que fue también un desafío a nivel personal. Con su apoyo fue posible atravesar por los momentos de mayores dificultades.

Finalmente agradecer a las agencias financiadoras que permitieron la realización de este proyecto. A Capes, por otorgar la beca de Doctorado completo durante estos 4 años. A la ANII por la financiación del proyecto FMV que permitió financiar las actividades de este proyecto. A INIA por el apoyo económico dado a través del premio a mejor escolaridad de posgrado (FA-UdelaR). A la Facultad de Agronomía, Universidad de la República que me permitió continuar con mis actividades de formación profesional.

Obrigada!

SUMMARY

Resumo	6
Abstract	7
List of Figures	8
List of Tables	10
1 Introduction	11
2 Materials and Methods	17
2.1 Mapping populations	17
2.2 DNA extraction and library construction	17
2.3 SNP and genotype calling	17
2.4 Construction of integrated map for individual populations	18
2.5 Construction of a composite integrated genetic map	20
3 Results	23
3.1 GBS libraries	23
3.2 SNP calling	23
3.3 Genotype calling	24
3.4 H5 and H6 genetic map	26
3.5 Composite genetic map	28
4 Discussion	33
References	43
Appendices	53

RESUMO

Construção de um mapa genético de alta densidade em *Acca sellowiana* (Berg.) com base em duas populações de mapeamento geneticamente conectadas

Acca sellowiana, conhecida como feijoa, *pineapple guava* ou goiabeira-serrana, é uma árvore frutífera nativa do Uruguai e do Brasil, pertencente a família Myrtaceae. A espécie destaca-se por suas frutas altamente aromáticas, com reconhecido valor nutracêutico e terapêutico. Apesar do promissor valor agrônomo, os estudos genéticos nesta espécie são limitados. Os mapas genéticos são valiosas ferramentas em tais estudos, sendo empregados no desenvolvimento de estratégias de melhoramento molecular nos programas de melhoramento. No entanto, a falta de um elevado número de marcadores polimórficos nas populações de mapeamento é uma das principais limitações no desenvolvimento de mapas genéticos saturados. Recentemente, novos métodos de genotipagem baseados em tecnologia de sequenciamento de nova geração permitem identificar e genotipar milhares de marcadores em populações de mapeamento. Esta rápida e eficiente estratégia é muito útil para culturas pouco estudadas, com recursos genômicos limitados. Neste estudo, foram construídos mapas genéticos saturados em *A. sellowiana*. Foram usadas duas populações de mapeamento, H5 ('TCO × BR', $n = 160$) e H6 ('TCO × DP', $n = 184$), conectadas geneticamente pelo mesmo genitor feminino. A estratégia de genotipagem por sequenciamento (*genotyping by sequencing*; GBS) foi usada para simultaneamente identificar e genotipar marcadores de polimorfismos de nucleotídeo único (*single nucleotide polymorphism*; SNP) em ambas populações. No processo de detecção de SNPs, duas estratégias foram implementadas: na primeira foi empregado o genoma de referência de espécie relacionada *Eucalyptus grandis*; na segunda, foi empregado uma abordagem *de novo*, que não requer genoma de referência. Após o processo de genotipagem quantitativo, 5350 e 4227 SNPs de alta qualidade foram selecionados para mapeamento em H5 e H6, respectivamente. Os mapas integrados H5 e H6 compreendem 1236 e 1302 marcadores distribuídos no 11 grupos de ligação esperados. Os mapas genético abrangeram um comprimento total de 1593 cM e 1572 cM, com uma distância média entre marcadores de 1.29 cM e 1.21 cM, nas populações H5 e H6, respectivamente. Um alto nível de colinearidade foi observado entre os dois mapas. Além disso, uma grande proporção de marcadores foram mapeados em ambos mapas e posteriormente usados na construção de um mapa genético integrado, considerando a informação de ambas populações simultaneamente. Foi apresentada uma nova abordagem para estimar as frações de recombinação em duas populações conectadas, onde a informação das meioses de todos os indivíduos é capturado num único estimador, usando uma estimativa de máxima verossimilhança multiponto. O mapa integrado composto contém 641 marcadores SNP com um comprimento total de 1011 cM. Este mapa representa o melhor ordenamento consenso de marcadores, sendo uma valiosa referência para futuros estudos nesta espécie. O grande número de SNPs identificados permitiu-nos a construção de mapas genéticos de alta densidade, ferramentas moleculares que representam uma contribuição relevante para futuras pesquisas genéticas e avanços no melhoramento genético em *A. sellowiana*.

Palavras-chave: *Acca sellowiana*, Genotipagem por sequenciamento (GBS), Mapa genético, Myrtaceae

ABSTRACT

Construction of a high-density genetic map of *Acca sellowiana* (Berg.) Burret based on two connected mapping populations

Acca sellowiana, known as feijoa or pineapple guava, is a Myrtaceae fruit tree species native to Uruguay and Brazil. The species stand out for its highly aromatic fruits, with nutraceutical and therapeutic value. Despite its agronomically promising valuable, genetics studies on this species are limited. Linkage genetic maps are valuable tools for genetic and genomic studies, and can be employed in breeding programs to support the development of molecular breeding strategies. The lack of a high number of polymorphic markers is one of the main limitation to development saturated genetic maps. Recently, novel genotyping methods based on next generation sequencing technology allow to detect and genotype thousands of markers in mapping populations. This represents a rapid and cost-effective strategy, remarkably useful for minor species with limited genomic resources. In this study, we constructed a high-density integrated genetic linkage map of *A. sellowiana* using two populations, H5 ('TCO $\times$ BR', $n = 160$) and H6 ('TCO $\times$ DP', $n = 184$), which have the same female parent. Genotyping by sequencing (GBS) approach was used to simultaneously discover and genotype single nucleotide polymorphism (SNP) markers in both populations. Two strategies were carried out to identify SNP markers: a reference pipeline using the reference genome of the closely-related species *Eucalyptus grandis*, and a *de-novo* pipeline that do not require a reference genome. After quantitative genotype calling, 5,350 and 4,227 high quality SNP markers were selected for mapping in H5 and H6 populations, respectively. The two resulting maps of populations H5 and H6 comprised 1,236 and 1,302 markers distributed over the expected 11 linkage groups. The H5 and H6 maps spanned a map length of 1,593 cM and 1,572 cM, with an average inter-marker distance of 1.29 cM and 1.21 cM, respectively. A high degree of collinearity was observed between the two maps. In addition, a large proportion of markers were common to both maps and were used to construct the composite genetic linkage map. A novel approach to estimate recombination of two connected populations is described, where the meiosis information of all individuals is captured in a single estimator using a multipoint maximum likelihood estimation. The composite map consisted of 641 SNPs markers with a total map length of 1011 cM. This composite map represent the best consensus ordering of markers, a valuable reference framework for future studies in *A. sellowiana*. The large number of SNPs identified allowed us to construct high-density genetic maps, molecular tools which represent a relevant contribution for future genetic research and breeding efforts in *A. sellowiana*.

Keywords: *Acca sellowiana*, Genotyping by sequencing (GBS), Integrated genetic map, Myrtaceae

LIST OF FIGURES

2.1	Flow diagram summarizing the SNP and genotype calling for GBS data of H5 and H6 mapping populations.	19
3.1	Results of GBS libraries sequencing for each progeny individual of H5 and H6 mapping populations. The y-axis indicates the number of total reads, the x-axes indicates individual progeny.	23
3.2	Distribution of SNP density across the chromosomes of the <i>Eucalyptus grandis</i> genome. The SNP density in the y-axis was estimated in a 1 Mb window. The x-axis indicates the physical position in Mb.	24
3.3	a) Proportion of the final number of SNP markers identified using Discovery and UNEAK pipelines, for H5 and H6 populations. b) Venn diagram showing common and unique SNPs between H5 and H6 populations identified by Discovery and UNEAK pipelines. c) Distribution of marker types: ‘B3.7’, ‘D1.10’ or ‘D2.15’ in H5 and H6 populations. d) Distribution of marker type configurations for common markers between H5 and H6 populations. The joint marker type configuration consider the marker types in H5 and H6 population simultaneously.	26
3.4	Patterns of segregation distortion for H5 and H6 genetic maps. Linkage groups with more than 5% of distorted markers are shown. The x-axis represents the genetic position (cM) on each linkage groups and the y-axis represents the $-\log_{10}(p - values)$ of segregation distortion. The discontinuous lines indicate threshold for significant distortion (Bonferroni-adjusted $p - value = 0.05$).	27
3.5	Synteny of a) H5, b) H6 and c) composite genetic maps of <i>Acca sellowiana</i> and the <i>Eucalyptus grandis</i> genome. The x-axis indicates the genetic position (cM) of SNP markers identified by aligned to the <i>E.grandis</i> reference genome. The y-axis indicates the physical position (Mb) of each SNP marker. For each linkage group, SNPs corresponding to the reference chromosome of <i>E.grandis</i> are labeled in green. The orange points represented SNPs in discordant reference chromosomes.	30
3.6	Comparison of homologous linkage groups from the composite and single population H5 and H6 genetic maps of <i>Acca sellowiana</i> . Map distances (cM) are indicated at the left side of the figure. SNP markers common between LGs of the three genetic maps are connected with blue lines.	31
A.1	Genetic linkage map of the H5 (‘TCO $\times$ BR’) F_1 full-sib mapping population. Heatmaps plots show the pairwise recombination fractions below the diagonal and LOD scores above the diagonal for all marker combinations in each linkage group. The color scale ranges from red (small recombination fraction or large LOD) to dark blue. White cells indicate markers which recombination fraction cannot be estimated with two-point analysis (‘D1’ and ‘D2’).	55

- A.2 Genetic linkage map of the H6 ('TCO $\times$ DP') F_1 full-sib mapping population. Heatmaps plots show the pairwise recombination fractions below the diagonal and LOD scores above the diagonal for all marker combinations in each linkage group. The color scale ranges from red (small recombination fraction or large LOD) to dark blue. White cells indicate markers which recombination fraction cannot be estimated with two-point analysis ('D1' and 'D2'). 56
- A.3 Comparison of homologous linkage groups from H5 genetic maps. The H5 map constructed in this study is displayed in black and the former H5 genetic map developed by QUEZADA *et al.* (2014) is displayed in blue. Map distances (cM) are indicated at the left side of the figure. Marker names are shown for each linkage group. Common markers (green) between the two maps are connected with lines. The marker GGAGA3_7 with a conflict position between maps is show in red. . . . 57

LIST OF TABLES

3.1 Number of SNP markers after each filtering process stage for H5 and H6 populations. 25

3.2 Description of H5, H6 and composite linkages maps 29

1 INTRODUCTION

Acca sellowiana (Berg.) Burret, commonly known as feijoa or pineapple guava, is an agronomically promising Myrtaceae fruit tree species, native to Uruguay and southern Brazil. Its significant potential results from highly aromatic fruits, with nutraceutical value and organoleptic properties. Feijoa fruits are rich in vitamin C (28 mg/100 g), hydrocarbons, minerals, iodine (3 mg/100 g) and bioflavonoids. They also contain polyphenols, terpenes, tannins, steroidal saponins and other volatile compounds, responsible for their characteristic strong aroma (WESTON, 2010). Reports described relevant biological properties of fruits, leaves, seeds and extracts of *A. sellowiana*. An effective antioxidant-activity was found for dry and fresh fruits, leaves, and also the methanol extract showed superior antioxidant effects when compared with commercial or standard antioxidants (BEYHAN *et al.*, 2010). The results of the antimicrobial evaluation suggest *A. sellowiana* as a potential source of antibacterial, antioxidant, and anti-inflammatory agents (VUOTTO *et al.*, 2000). A flavone, present in the acetone extract of the fruits, was identified as the principle responsible for the anti-cancer activity, inducing apoptosis in cancer lines and with low toxicity to normal cells (BONTEMPO *et al.*, 2007). Considering all these properties, *A. sellowiana* stands out as a valuable fruit tree species in the novel market of health-promoting and functional food.

Acca sellowiana is an outcrossing, diploid ($2n = 2x = 22$) species with genome of small size (382 Mb) (DA COSTA *et al.*, 2008). The basic haploid number of $n = 11$ is largely conserved in the Myrtaceae family, where most species are characterized by genome sizes considered small to intermediate (GRATTAPAGLIA *et al.*, 2012). Many economically and ecologically valuable trees belong to the Myrtaceae family, which comprises more than 5,650 species distributed mainly in the Southern Hemisphere. The genus *Acca*, along with *Pisidium*, *Eugenia* and *Syzygium* is composed of fleshy-fruited species. They stand out as taxonomic group with prominent agronomic and economic potential. However, the genus *Eucalyptus*, which comprises woody-fruit species, has a pivotal economic importance, leading the genomics research within the Myrtaceae family (GRATTAPAGLIA *et al.*, 2012). The *Eucalyptus grandis* genome (641 Mb) was the first complete genome sequence available for a myrtle species (MYBURG *et al.*, 2014). Despite the great diversity, ecological adaptation and differences in genome size, the synteny within the family is high, facilitating studies of comparative genetics.

Considering the seasonal presence on the market and the small cultivation area, *A. sellowiana* can be considered a minor crop. The feijoa fruit is well appreciated for fresh consumption, but also in the production of jams, liquors, and juices (THORP and BIELESKI, 2002). In Brazil and Uruguay, the center of origin and genetic diversity of the species, it is cultivated only in small and local scale. However, commercial orchards are found in New Zealand, Colombia and United States, where the production is based on relatively few cultivars. The cultivars commercially grown derived from a reduced number of introduced accessions that had been originally collected in the center of origin, resulting in a narrow genetic base (THORP and BIELESKI, 2002).

The development of new cultivars with superior fruit quality traits and adapted for new environments is needed to accelerate the commercial exploitation of *A. sellowiana*. As with other fruit tree species, the breeding process begins with the selection of superior genotypes

with remarkable agronomic qualities. These selected materials are used as parents for crossing, generating full-sib families, segregating for relevant agronomic traits. Although effective, this breeding process is slow and costly, due to the long juvenile period, extensive phenotyping cost and limited field space (BYRNE, 2012). The development of genetic and genomic resources has been proposed as a suitable strategy to overcome these limitations and speed up the breeding progress in fruit trees (PEACE and NORELLI, 2009).

Despite the potential agronomic value, genomic resources in *A. sellowiana* are limited and not yet useful in breeding programs. High genetic variability of agronomic traits related to fruit quality such as shape, size, color, as the flavor was reported for natural populations located at the center of origin (THORP and BIELESKI, 2002; PUPPO *et al.*, 2014) and also for germplasm accessions (PASQUARIELLO *et al.*, 2015; ALVAREZ *et al.*, 2017). However, few studies document the molecular diversity of the specie using low-throughput molecular markers (DETTORI and PALOMBI, 2000; DOS SANTOS *et al.*, 2007; BACCINO, 2011). The limited number of molecular markers specifically designed for *A. sellowiana* has rarely been applied to genetic analysis or molecular breeding (DOS SANTOS *et al.*, 2008; KLABUNDE *et al.*, 2014). Recently, the complete chloroplast genome was sequenced enabling phylogenomic analysis between *A. sellowiana* and other Myrtaceae species (DE MACHADO *et al.*, 2017). Aiming to initiate genomic studies, help elucidate the genome structure and develop essential tools for genetic improvement, the first genetic map of the species was constructed (QUEZADA *et al.*, 2014). The map consisted of 219 markers, primarily amplified fragment length polymorphisms (AFLPs) and a small number of intersimple sequence repeat (ISSR) markers. The markers were assigned to 10 major linkage groups, with a total map length of 2,875 cM and an average distance between markers of 13.8 cM. This map, along with those developed for the species *Psidium guajava* (PADMAKAR *et al.*, 2015) represents the only ones developed for fruit species within the Myrtaceae family. The development of a high-density genetic map of *A. sellowiana* would provide a valuable resource for molecular breeding in the species and also a powerful tool for comparative genomic analyses in the Myrtaceae family.

Molecular markers are a powerful tool for the study of genetic variability, being widely used in genetic analysis and molecular breeding of plant species (VARSHNEY *et al.*, 2010). The development of molecular markers represents the first effort toward the molecular characterization of a species and is a prerequisite for the construction of genetic linkage maps. Traditionally, genetic maps have been constructed through the scoring of low-throughput genetic markers, such as random amplification of polymorphic DNA (RAPD), restriction fragment length polymorphisms (RFLP) and simple sequence repeat (SSR) markers. As a result, low-density genetic maps have been developed, which are useful tools for comparative analysis, quantitative trait loci (QTL) detection and providing anchor and order scaffolding for physical maps (VELASCO *et al.*, 2010; PEREIRA *et al.*, 2013; SOUZA *et al.*, 2013).

With the advance of new sequencing technologies, single nucleotide polymorphisms (SNPs) have become the most useful type of marker for molecular genetic analysis. SNP markers are the most abundant class of polymorphisms in plant genomes, and the possibility of automating the detection and scoring represents a major advantage (DAVEY *et al.*, 2011). In a first approach, microarray-based technologies were used to analyze simultaneously on a single

multiplexed array large number of SNPs. This methodology allows successfully genotyping in many plant species (VERDE *et al.*, 2012; SILVA-JUNIOR *et al.*, 2015). In order to design SNP genotyping arrays, it is necessary to previously identify a considerable number of SNPs, preferentially with known genomic positions. Relying on genotype diversity used in the marker discovery step, the utility of the arrays for all genetic background is limited. Moreover, the use and transferability of markers between relative species has been shown to be low (VEZZULLI *et al.*, 2008). As a consequence, specific arrays must be developed for the species and population under investigation, which represents a costly strategy for minor or non-model crops.

Currently, next-generation sequencing (NGS) coupled with enzyme-based complexity reduction and DNA barcoding, has been used to simultaneously discover and genotype a large number of SNPs in a single experiment (BAIRD *et al.*, 2008; POLAND *et al.*, 2012b). Genotyping by sequencing (GBS) is one of these novel methods of high-throughput genotyping, where polymorphism identifying and genotyping are completed in a single step (ELSHIRE *et al.*, 2011). The GBS protocol uses methylation-sensitive endonucleases to reduce genome complexity, constructing reduced libraries which are subsequently sequenced on high-throughput platforms. The design of barcodes adapters, a DNA sequence to identify each sample, allows pooling hundreds of samples in a single lane sequence (ELSHIRE *et al.*, 2011). The short sequences generated by NGS are used to identify SNPs by aligning to a reference genome; or alternative, using pairs of nearly identical reads, for species without a reference genome. NGS combined with the reduction genome complexity approach and the multiplexing samples significantly reduces the cost per marker and per sample (DAVEY *et al.*, 2011). This low cost strategy allows the production of thousands SNP markers in a fast and effective way, representing a valuable approach for species such as *A. sellowiana*, with a lack of genomic resources.

GBS technology has demonstrated high efficiency in numerous genetic studies, including molecular marker discovery, population structure and germplasm diagnostics, genetic diversity studies, genome-wide association studies (GWAS) and genomic selection, as well as high-density genetic maps to identify quantitative trait loci (QTL) for example (POLAND *et al.*, 2012a; SPINDEL *et al.*, 2013; HE *et al.*, 2014; HUANG *et al.*, 2014). This technology was first applied in commercial crops, using inbred lines and a sequenced reference genome. For outbred species with highly heterozygous genomes, genotype calling is laborious and challenging (HYMA *et al.*, 2015). The success of GBS strategies relies on the multiplex system, simplified libraries combined with high throughput sequencing, resulting in data with low to moderate sequencing depth (ELSHIRE *et al.*, 2011). However, the low sequencing coverage results in a high proportion of missing data, a variable value of SNP genotyping error, and an under-calling of heterozygous sites in outbred species. Several strategies were reported to overcome these limitations in heterozygous fruit tree species such as raspberry (WARD *et al.*, 2013), apple (GARDNER *et al.*, 2014) and grape (HYMA *et al.*, 2015). Using SNPs from GBS, high-density linkage maps for many fruit tree species have been developed, such as apple (GARDNER *et al.*, 2014), grape (BARBA *et al.*, 2014) and peach (BIELENBERG *et al.*, 2015). Moreover, GBS has successfully been used for genetic linkage map construction in minor crops, such as red raspberry and cranberry, without prior genomic information (WARD *et al.*, 2013; COVARRUBIAS-PAZARAN *et al.*, 2016).

Genetic maps are extremely valuable tools in many genetic and genomic applications.

For this reason a large number of maps has been developed firstly for high commercial value fruit tree species, and, recently, for minor crops. Such maps have been used to investigate the composition and organization of genomes, for comparative genetic mapping analysis and exploration of syntenic and collinearity between species (MATHEW *et al.*, 2014; GUAJARDO *et al.*, 2015). Considering the current increase of complete genome sequences, genetic maps have a renewed importance for genome assembly, by anchoring new scaffolds, correcting scaffold order and orientation, as well as physical and genetic map integration (BARTHOLOMÉ *et al.*, 2014; LI *et al.*, 2017). Genetic maps are a useful tool in breeding programs to establish associations between molecular markers and agronomical traits of interest. The detection of quantitative trait loci leads to a better understanding of the genetic architecture of quantitative traits. In order to identify QTL, genetic maps have been successfully used to study quantitative complex traits of agronomic importance, especially associated with the determination of fruit quality traits (BARBA *et al.*, 2014; BIELENBERG *et al.*, 2015). QTL mapping represents the first step in studies of gene mapping and positional cloning of selection of candidate genes, and towards the development of marker-assisted breeding programs (COLLARD *et al.*, 2005). Especially for fruit tree crops, high-density genetic maps represent a powerful tool to improve breeding efficiency. Genetic maps provide the base for future strategies for selection of desirable QTL alleles and marker-assisted selection (MAS), reducing the time and cost of cultivar development (TROGGIO *et al.*, 2012).

For outcrossing species, like fruit trees and forest species, the inbred depression, long juvenile period as well as self incompatibility prevents the production of inbred lines, and therefore hinder the development of traditional mapping population, such as F_2 , RILs or backcross. For these species, genetic maps are typically based on single full-sib mapping populations derived from crosses of two heterozygous parents. The pseudo-test cross strategy was first developed to construct genetic maps in these populations. An individual map for each parent is obtained, based on dominant markers, heterozygous in only one parent, segregating in a 1:1 ratio (GRATAPAGLIA and SEDEROFF, 1994). Subsequently, integrated maps could be developed using information from all markers simultaneously, estimating the linkage and linkage phases between markers. This approach considers dominant and codominant markers, heterozygous in just one parent or in both parents, with segregation ratios of 1:1:1:1, 3:1, 1:2:1 and 1:1 (MALIEPAARD *et al.*, 1997; WU *et al.*, 2002a). These strategies have been used to construct genetic maps, for several outcrossing species (PALHARES *et al.*, 2012; PEREIRA *et al.*, 2013; QUEZADA *et al.*, 2014; BALSALOBRE *et al.*, 2017).

The construction of accurate genetic maps has faced two main difficulties, the availability of a high number of polymorphic markers and the development of large mapping populations. High-density genetic maps, with dense coverage of markers evenly distributed in the genome, are essential for a comprehensive representation of the genome. Moreover, accurately estimating genetic distances and determining the correct marker order are directly related to the number of recombination events, and therefore to the sample size and type of population (COLLARD *et al.*, 2005). To date, developing a large number of markers is not a limiting factor for map construction, even for species with very few genetic and genomic resources (DAVEY *et al.*, 2011). However, for most fruit tree species, the reported genetic maps are based on a single popula-

tion of medium to small size (PEREIRA *et al.*, 2013; WARD *et al.*, 2013; RUSSELL *et al.*, 2014; GUAJARDO *et al.*, 2015).

The typical mapping populations of fruit tree species are developed from a cross between divergent parents with outstanding fruit quality features and agronomically important traits. In many instances, these populations belong to breeding programs, where they are used for phenotypic evaluations, genetic parameter estimation and cultivar selection. These full-sib progenies present a small to medium size, considering the limited orchard space, and also the high cost to maintain these populations for a long time (PEACE and NORELLI, 2009). Considering the breeding program goals, the same superior genotype can be used in several crosses, generating connected parental progenies. Capitalizing on existing breeding-mapping populations provides an opportunity to develop a composite or a multiparental genetic map, including information of recombination events in multiple populations. This approach allows to overcome the limitations of a reduce number of recombination events captured in single mapping strategies. Composite genetic maps employing mutiparental mapping populations have been constructed for several outcrossing species, such as apple (CLARK *et al.*, 2014), cranberry (SCHLAUTMAN *et al.*, 2017), pinus (PELGAS *et al.*, 2005), oak (BODÉNÈS *et al.*, 2016) and rubber tree (POOTAKHAM *et al.*, 2015). Moreover, a consensus map can be constructed for multiple populations derived from independent crosses (DIAZ *et al.*, 2011; GEORGI *et al.*, 2013).

One of the main advantages of constructing genetic maps based on multiparental populations is to improve the accuracy and resolution of genetic maps, by correcting the order and position of markers. The precise order of markers achieved from a composite map, enabling identify divergences and inconsistencies among maps. The deviations could be a result of genotyping errors or skewed segregations, but can also be related to genome rearrangements or structural variations (KHAN *et al.*, 2012). A composite map with accurate marker order has been used to determine genome duplications and possible misalignments of contigs in the apple reference genome (CLARK *et al.*, 2014). The integration of individual-population genetic maps into a composite map also increases marker density and genome coverage, and fills out genomic regions lacking polymorphic markers. The multiparental map developed in rubber tree presents smaller gaps when compared to individual maps and also achieved the haploid chromosome number of the species when separate small linkage groups of individual maps were connected (POOTAKHAM *et al.*, 2015). Common markers between maps can be used as anchor markers to compare genomic localizations, transfer marker information and facilitate genetics predictions between populations (KUHN *et al.*, 2017).

Composite genetic maps are also a valuable resource for QTL analysis. Currently, the novel QTL approaches are based on the simultaneous analysis of multiple full-sib families derived from connected parents and founders (DI PIERRO *et al.*, 2016). Integrated analysis using a composite map enables a more precise estimation of QTL effects and positions. As a result, narrow QTL confidence interval facilitates the subsequent identification of markers that are useful for MAS or molecular breeding strategies. In addition, this approach increases the power of QTL detection, as well as allows a better characterization of QTL identified in different genetic backgrounds. For example, comparisons of the effect and position of QTL for the same trait in different populations provides deeper insights into the genetic control of phenotypic

traits, allowing the identification of pleiotropic effects (DIAZ *et al.*, 2011).

Construction of composite genetic maps using multiparental populations in outcrossing species is challenging because of genetic factors and different molecular marker informativeness. The analysis of multiparental populations has to address differences in recombination rates due to genomic structural variations between populations. Moreover, nonrandom distribution of recombination along chromosomes in the founders which were used to generate population mapping. Other constraints include the unequal information owing to differences in marker dominance, and the high proportion of missing data resulting from the presence of unique markers scored only in some populations (RONIN *et al.*, 2012).

Two main strategies have been proposed to construct consensus maps applied to outcrossing species. In a first strategy, an approximate consensus recombination fraction is estimated combining all genotypic data sets for multiple populations. In order to construct a composite genetic map in rubber tree, the pairwise recombination fractions of single populations were linearly combined, using weights based on individual population size (VAN OOIJEN, 2006; POOTAKHAM *et al.*, 2015). The approach used in pot azalea was based on phase-known genotypic data for single populations with subsequently merging, assuming that all data represent a single map (DE KEYSER *et al.*, 2010). In this approach, either the same order and distances (recombination fraction) for all populations, or independent distances for each population, can be considered (DE GIVRY *et al.*, 2005). The inability to solve inconsistencies or order conflicts between maps, as well as the disruptive effect of missing data on the mapping process can be considered the main limitations of these approaches (KHAN *et al.*, 2012). A second strategy is based on the order and distances of individual maps, instead of the underlying recombination frequencies. A map integration based on marker position was proposed, with linear programming used to resolve ordering conflicts and minimize the mean absolute errors between the composite map and the individual maps (ENDELMAN and PLOMION, 2014). In another approach, a graph-theoretic was used to construct a framework map according to the shared orders among individual maps. Conflicting order markers are solved using a heuristic algorithm and removing the set of markers with inaccuracies in orders (WU *et al.*, 2011). Both approaches have been applied to genetic map construction in cranberry (SCHLAUTMAN *et al.*, 2017), oak tree (BODÉNÈS *et al.*, 2016) and apple (KHAN *et al.*, 2012).

The ONEMAP R package (MARGARIDO *et al.*, 2007), was initially developed to construct integrated genetic linkage maps in outcrossing species. In the first version, the software implemented the methodology proposed by WU *et al.* (2002a), which uses the maximum likelihood approach to estimate linkage and linkage phases. This approach considers a mixed set of different segregation types of markers, from fully informative markers (segregating 1:1:1:1) to partially informative markers (segregating 1:2:1, 3:1 and 1:1). This method, first applied in the construction of a sugarcane genetic map, overcomes several statistic disadvantages observed in the development of integrated genetic maps (GARCIA *et al.*, 2006). Subsequently, the package was modified to incorporate a multipoint estimation for the recombination fraction between linked markers using a Hidden Markov Model (WU *et al.*, 2002b). The multipoint maximum likelihood linkage maps present a more accurate estimates of map distances and order. This approach was successfully applied in the construction of integrated maps of several outbreeding

species such as yellow passion fruit (OLIVEIRA *et al.*, 2008; PEREIRA *et al.*, 2013), rubber tree (SOUZA *et al.*, 2013) and strawberry (TENNESSEN *et al.*, 2014). This approach was also applied in the construction of the first genetic map of *A. sellowiana* (QUEZADA *et al.*, 2014). The last version, currently under development and available at <https://github.com/augusto-garcia/onemap>, was modified to handle markers obtained from the novel methods of high-throughput genotyping. This version accommodates a large number of molecular markers of different segregation type as well as the genotyping error associated with each marker (TANIGUTI, 2016). The statistical model implemented in this package will be used to develop a novel strategy, allowing the genetic map construction using information about multiparental populations in outcrossing species.

The objective of this study was to develop a reference and composite linkage genetic map for *A. sellowiana*, integrating the genetic information from two connected F_1 populations. For this purpose, two mapping populations segregating for fruit quality and connected to the same female parent were used. For the first time in the species, the high-throughput GBS approach was successfully applied to identify SNP markers distributed throughout the genome. Two high-density individual population genetic maps were constructed, and a novel approach to integrate these maps into a composite genetic map was developed. This new method will be in the future incorporated into ONEMAP (MARGARIDO *et al.*, 2007). The collinearity between common markers in the maps was evaluated, and the final order reflects the most accurate order for the individual and composite maps. The saturated individual genetic maps and the composite map, provide a framework for future genetic and genomic studies in the species. This will be useful for future studies of marker trait association and QTL mapping, in order to speed up the breeding process in *A. sellowiana*.

2 MATERIALS AND METHODS

2.1 Mapping populations

Two full-sib F_1 populations genetically connected by one parental genotype and segregating for relevant fruit quality traits were used to construct an integrated composite map. These populations derived from the crossing of non-inbred clones ‘TCO $\times$ BR’ (H5 population) and ‘TCO $\times$ DP’ (H6 population). The same genotype ‘TCO’ was used as female parent in both crosses. The three parental genotypes were chosen on the basis of their agronomist features: ‘TCO’, collected from the wild, has small fruits with tasty and smooth pulp, as well as a thin green-skin; ‘BR’, collected from a commercial orchard, has large fruits with tasteless pulp and a rough, thick, dark green-skin; ‘DP’, collected from a commercial orchard has a high yield of medium size fruits, and a rough, medium-thick, dark green-skin. The mapping populations developed in 2008 belong to the Native Fruits Breeding Program (Universidad de la República-INIA, Uruguay). Progeny sizes are 160 and 184, for H5 and H6 respectively, where each individual progeny is represented by a single tree. H5 population was planted in the Experimental Station ‘Salto’ of Facultad de Agronomía-UdelaR, Uruguay (31° 19’ S, 57° 41’ W) whereas H6 population was planted in the Experimental Station ‘Salto Grande’ of Instituto Nacional de Investigación Agropecuaria, Uruguay (31° 25’ S, 57° 37’ W). H5 population was used previously in the development of the first genetic map of the species (QUEZADA *et al.*, 2014).

2.2 DNA extraction and library construction

Leaf material was collected from a total of 344 individuals (three parental genotypes, 160 and 184 progenies for H5 and H6 populations, respectively). Genomic DNA was extracted using DNeasy Plant Kit (Qiagen, Germantown, MD, USA). DNA quality was checked by electrophoresis on 0.8% agarose gel and quantified by NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). DNA samples were submitted to the Genomic Diversity Facility, Cornell University (Ithaca, NY, USA) and GBS libraries were prepared according to the original GBS protocol (ELSHIRE *et al.*, 2011). In a previous step of enzymes selection, test libraries were prepared with three restriction enzymes: ApeKI, EcoT22I and PstI. The fragment size distribution among libraries was evaluated on an Experion system (Bio-Rad, Hercules, CA, USA). The six-base-pair site restriction enzyme PstI was chosen because it produced the largest number of fragments minor to 500 bp and amplified little repetitive DNA. DNA samples were prepared in two 192-plex libraries and were sequenced twice in a HiSeq2000 sequencer (Illumina R Inc., San Diego, CA, EUA). To guarantee a satisfactory coverage and confidence in calling heterozygous genotypes, the three parental genotypes were considered as 12 samples each, and all the samples reached a coverage equivalent to 96-plex.

2.3 SNP and genotype calling

First, the quality of the raw sequence GBS data was evaluated using the FASTQC v0.11.5 software (ANDREWS, 2010). For SNP calling, two strategies implemented in the TASSEL GBS software were employed. The first strategy, the Discovery pipeline (TASSEL v4.0), uses a

reference genome to identify SNPs (GLAUBITZ *et al.*, 2014); the second was the UNEAK pipeline (TASSEL v3.0), based on a network approach to identify SNPs without a reference genome (LU *et al.*, 2013).

Raw sequences were filtered in both strategies, discarding reads lacking the barcode, not having the expected PstI cut-site, or containing NA's within the first 64 bp subsequently the barcode. Filtered reads were grouped in tags, and only those with a minimum deep coverage of 5 were retained (parameter -c 5 at the MERGETAXATAGCOUNTPLUGIN step). For the Discovery pipeline, tags were aligned to the *E. grandis* reference genome (v2.0) using BOWTIE v1.2.0 (LANGMEAD and SALZBERG, 2012) with default parameters. The *E. grandis* genome stand out as the reference genome for the Myrtaceae family (GRATTAPAGLIA *et al.*, 2012). In the UNEAK strategy, tags were pairwise aligned, and alignments with a single base pair mismatch were identified. Considering the presence of repeated or paralogous sequences in the genome, many tag pairs form networks. These networks were filtered (error tolerance rate set at 0.03) and only reciprocal tags were maintained representing potential SNPs. The identified SNP markers using both strategies were exported as *vcf* files (information-rich variant call format) where the read depth of the reference and alternative allele is saved (DANECEK *et al.*, 2011). SNP markers identified by the Discovery strategy were denoted as 'Eg' followed by the chromosome number and the position (bp) on the *E. grandis* reference genome. SNPs identified using the UNEAK pipeline were named as 'Un1' plus a number referring to the tags network.

Genotype calling was performed for each population separately using SUPERMASSA software (SERANG *et al.*, 2012). To call genotypes, this software considers the expected allele distribution in a F_1 population, as well as the relative coverage (read counts) of each allele (GARCIA *et al.*, 2013). In a previous step, markers were filtered out for quality; only biallelic markers with less than 25% of missing data were retained. The F_1 segregation model with diploid level was fitted in SUPERMASSA. Each of the 12 replicates of parental genotypes were analyzed separately, to provide additional constraints during estimation. Following the recommendations of MOLLINARI and SERANG (2013), SUPERMASSA naive posterior report threshold was set to zero and individual posterior probability values were estimated. The median of all individual posterior probabilities was used as a quality control, so only SNPs with posterior values higher than 0.8 were selected. The SUPERMASSA output was formatted for mapping purposes using R software (R CORE TEAM, 2017). Following the notation of WU *et al.* (2002a), markers were classified into four groups ('A-D') according to their respective cross type. For each locus, a maximum of four codominant alleles ('a', 'b', 'c' or 'd'), all of them dominant over a null allele ('o') may be present in both parents. For GBS-SNP markers in a F_1 population, only three marker cross types are informative for genetic map construction: 'B3.7' ('ab' $\times$ 'ab'), 'D1.10' ('ab' $\times$ 'aa') and 'D2.15' ('aa' $\times$ 'ab'). The group 'D' comprise markers segregating in a 1:1 ratio, heterozygous in only one parent, also called testcross markers. The 'B3.7' markers are heterozygous and symmetric in both parents, with an expected 1:2:1 segregation ratio. Redundant markers between pipelines and between populations were analysed. A diagram summarizing the work flow of SNP and genotype calling, including softwares and filter parameters used, is shown in Figure 2.1.

Figure 2.1: Flow diagram summarizing the SNP and genotype calling for GBS data of H5 and H6 mapping populations.

2.4 Construction of integrated map for individual populations

For each population, a data set including the SNPs identified by Discovery and UNEAK pipelines was generated. For H5 population, 493 additional markers (ISSR, AFLP and SSR) previously located on a former H5 map were also included in the mapping process (QUEZADA *et al.*, 2014). Integrated genetic maps for each population (H5 and H6) were constructed using ONEMAP (MARGARIDO *et al.*, 2007) R package, the current version is available at <https://github.com/augusto-garcia/onemap>. This software implement the method proposed by (WU *et al.*, 2002a,b) for simultaneous multipoint estimation of the recombination fraction and linkage phases between markers. Considering that H5 and H6 populations have a comparable number of individuals and genotyped markers, the same procedure was applied to construct both maps.

First, the segregation of each marker was evaluated for goodness-of-fit to the expected

Mendelian segregation ratio by a chi-square test, with the Bonferroni correction for multiple testing at a significance level of 0.05. Two-point linkage analysis was carried out between all pairs of markers using the ‘RF_2PTS’ command. Linkage groups (LGs) were established with a LOD score of 7.5 and a maximum recombination fraction of 0.35. Recombination frequency was converted to genetic map distances (cM) using the Kosambi function (KOSAMBI, 1944). To order groups, the ORDER_SEQ function was used, which selects a subset of informative markers (6 – 7 markers), wherein the remaining markers were introduced. A framework map was evaluated, and markers exhibiting a suspect position (ordering problems), producing gaps at the ends of the groups or inflating linkage groups size were removed. Once the framework map presents a reliable initial order, the TRY_SEQ command was used to integrate previously removed markers and also markers that remained unmapped. Incorrect allocation of markers was visually inspected using a heatmap plot (graphical representation of the recombination fraction and LOD score between markers). Considering that the GBS-SNP markers could have many genotyping errors, a higher probability genotyping error value of 0.05 (default= 0.01) was considered to construct the genetic maps. The RIPPLE command was used to verify alternative local orders. LG were numbered according to the chromosome number of *E. grandis* reference genome by the identification of syntenic SNPs. Finally, the linkage maps were drawn using the MAPCHART v2.3 software (VOORRIPS, 2002).

2.5 Construction of a composite integrated genetic map

A new approach to construct a composite genetic map using the information of two connected populations was developed. The proposed model assumes a homogeneous recombination fraction between populations, which is estimated using the segregation data for both populations simultaneously. Therefore, this method will have a higher statistical power, resulting in more accurate estimates of map distances and order.

In a first step, two-point recombination fractions were estimated using the segregation data of two populations. The maximum likelihood estimators for the recombination fraction and LOD score formulas were adapted from MALIEPAARD *et al.* (1997), only for the marker configurations present in our data. For example, we have loci with cross type configurations: ‘B3.7’ (‘ab’ × ‘ab’), ‘D1.10’ (‘ab’ × ‘aa’) and ‘D2.15’ (‘aa’ × ‘ab’). Because the same female parental genotype is shared by populations, the phase linkage of this parent was considered to be equal in both crosses. Subsequently, a multipoint approach was used, enabling more accurate estimation of the recombination fraction. This approach implemented a multipoint likelihood estimation using the methodology of Hidden Markov Models (HMM) (LANDER and GREEN, 1987) implements with the Expectation Maximization (EM) algorithm (DEMPSTER *et al.*, 1977). To estimate the recombination fraction, the expectation step is computed on each population separately and the maximization step is performed considering segregation information for both populations. In this model, the linkage phase between all markers is known, based on the previous individual genetic map of each population. As a result, an accurate recombination fraction is obtained. Details of the statistical modelling are described in the Appendix I.

To construct a composite genetic map, H5 and H6 individual integrated genetic maps were compared to identify common markers. These common markers were identified by name,

considering that SNP calling was performed on the whole data. A new data set for each linkage group was created, containing information for the markers shared between the two populations. An initial framework consensus map with shared colinearity markers was established. Then, markers with conflict order between the two maps were evaluated in all possible positions on the composite map; after, we selected the position with highest likelihood. As in the individual maps, the order was visually verified by inspecting heatmaps. A final local order verification was performed with `RIPPLE` function. Finally, the marker order of the composite map was applied to individual H5 and H6 maps, to correct the position of conflicted markers between maps. To perform such an analysis, functions of the `ONEMAP` package were modified and adapted to handle information for multiple populations. The functions developed here were recorded in an auxiliary R package available on request to the corresponding author; in a near future, they will be integrated into `ONEMAP` package and public released.

3 RESULTS

3.1 GBS libraries

A total of 859,454,459 reads of 100 bp length were obtained from the complete sequencing of the two libraries. The reads number and sequence quality was comparable for the two libraries and the two sequencing processes as determined by FastQC. Also, the quality of the reads was good, with an average base quality score higher than 30 (99.9% base call accuracy). The total number of reads in the progeny ranged from 60 thousand to 6.2 M (Figure 3.1). The 12 fold-repeated samples of parental genotypes resulted in a deeper coverage of parents, with 46,376,113 reads for ‘TCO’, 37,592,034 for ‘BR’ and 28,827,868 for ‘DP’. The average number of reads per progeny was 2.2 M and 3.2 M for the parental genotypes. The samples ‘H5_95’ and ‘H6_156’, presented a relatively low numbers of reads (< 150,000) that represents less of the 10% of the mean reads per sample of the lane on which they were sequenced, so they were removed for subsequent analysis.

Figure 3.1: Results of GBS libraries sequencing for each progeny individual of H5 and H6 mapping populations. The y-axis indicates the number of total reads, the x-axes indicates individual progeny.

3.2 SNP calling

In the first strategy using the Discovery pipeline, GBS reads were aligned to the *E. grandis* reference genome to identify SNPs. After running this pipeline, 3,051,475 high quality tags were identified. Of these, 643,863 (21.10%) aligned exactly once in the reference genome and 280,937 (9.21%) aligned in multiple positions, resulting in an overall alignment rate of 30.31%

(924,800 tags). As a result of this approach, 43,377 putative SNPs were identified. Figure 3.2 presents the SNP markers distribution across the *E. grandis* chromosomes. In the second strategy, 44,889 pairwise alignments running the UNEAK pipeline were obtained. Using the default filter parameters, 13,430 putative SNPs were identified using this approach.

Figure 3.2: Distribution of SNP density across the chromosomes of the *Eucalyptus grandis* genome. The SNP density in the y-axis was estimated in a 1 Mb window. The x-axis indicates the physical position in Mb.

3.3 Genotype calling

A total of 56,807 SNPs, comprised both sets of Discovery and UNEAK raw data, were available for genotype calling (Table 3.1). This number reduced to 53,051 (93.38%) after filtering out markers with variation no identified (missing alternative allele) or triallelic. After removing markers with more than 25% of missing data, a total of 27,595 (48.58%) and 25,670 (45.19%) SNPs for H5 and H6 populations, were retained, respectively. The filtered markers were subjected to quantitative genotyping following a F_1 model in SUPERMASSA. Markers with the median of all individual posterior probabilities greater than 0.8 were selected. This criterion was chosen to ensure a high quality of genotypes as described by MOLLINARI and SERANG (2013). Finally, a set of 5,350 (9.42%) and 4,227 (7.44%) high quality SNPs was used for genetic mapping, for H5 and H6 populations, respectively.

Table 3.1: Number of SNP markers after each filtering process stage for H5 and H6 populations.

Pipeline	H5 Population			H6 Population		
	Discovery	UNEAK	Total	Discovery	UNEAK	Total
Initial	43,377	13,430	56,807	43,377	13,430	56,807
No variant	39,682	13,401	53,083 (93.44%)	39,682	13,041	53,083 (93.44%)
Triallelics	39,650	13,401	53,051 (93.38%)	39,650	13,401	53,051 (93.38%)
< 25% missing	23,113	4,482	27,595 (48.58%)	21,395	4,275	25,670 (45.19%)
Probability genotype > 0.8	2,875	2,475	5,350 (9.42%)	2,046	2,181	4,227 (7.44%)

Of the SNP set for H5 population, 2,875 (53.74%) markers were identified by the Discovery pipeline, so the remaining 2,475 (42.26%) markers by the UNEAK pipeline. For the H6 population, 2,046 (48.40%) markers and 2,181 (51.60%) were identified using the Discovery and UNEAK approaches, respectively (Figure 3.3). The segregation pattern of SNP markers identified by Discovery and UNEAK strategies within each population was evaluated. A low number of overlapping markers between approaches for H5 (174 SNPs) and H6 (88 markers) populations was found. To facilitate map comparison, the redundant markers between pipelines were maintained with their original names.

Considering the segregation pattern, SNP markers were classified into one of the three configuration marker types informative for genetic map construction: ‘B3.7’ (‘ab’ × ‘ab’, heterozygous in both parents); ‘D1.10’ (‘ab’ × ‘aa’, heterozygous only in the maternal genotype); and ‘D2.15’ (‘aa’ × ‘ab’, heterozygous only in paternal genotype) (Wu *et al.*, 2002a). Of the 5,350 segregating markers identified in H5 population, 1,520 (28.41%) were heterozygous in both parents and classified as ‘B3.7’; 1,752 (32.75%) were heterozygous only in maternal genotype and classified as ‘D1.10’ and the remaining 2,088 (39.03%) were heterozygous only in paternal genotype so classified as ‘D2.15’. For H6 population, a set of 4,227 markers were identified, and of those: 993 (23.49%) were heterozygous in both parents (‘B3.7’), 1,520 (35.96%) were heterozygous in the female parent (‘D1’), and 1,714 (40.55%) were heterozygous only in the male parent (‘D2’) (Figure 3.3). Segregation distortion from the Mendelian expected ratio (1:2:1 for ‘B3.7’ markers and 1:1 for ‘D1.10’ and ‘D2.10’ markers) was observed in 1,263 (23.61%) markers of the H5 data set and 1,112 (26.31%) markers of H6 data set.

Comparing SNP data set between H5 and H6 populations, 2,427 SNP markers present in both populations were found. Of these shared markers, 1,204 (49.61%) were identified by Discovery pipeline and 1,223 (50.93%) by the UNEAK pipeline (Figure 3.3). A joint configuration marker type was defined for shared markers, considering the segregation pattern in both populations simultaneously. If the common maternal genotype is heterozygous for a locus

(‘ab’), the paternal genotype could be heterozygous in both populations (‘ab’ × ‘ab’/‘ab’ × ‘ab’, ‘B3.7-B3.7’), only heterozygous in H5 population (‘ab’ × ‘ab’/‘ab’ × ‘aa’, ‘B3.7’-‘D1.10’), only heterozygous in H6 population (‘ab’ × ‘aa’/‘ab’ × ‘ab’, ‘D1.10’-‘B3.7’) or homozygous in both populations (‘ab’ × ‘aa’/‘ab’ × ‘aa’, ‘D1.10’-‘D1.10’). For markers homozygous in the female parent, it is possible only that paternal genotypes were heterozygous in both populations (‘aa’ × ‘ab’/‘aa’ × ‘ab’, ‘D2.15’-‘D2.15’)(Figure 3.3).

Figure 3.3: a) Proportion of the final number of SNP markers identified using Discovery and UNEAK pipelines, for H5 and H6 populations. b) Venn diagram showing common and unique SNPs between H5 and H6 populations identified by Discovery and UNEAK pipelines. c) Distribution of marker types: ‘B3.7’, ‘D1.10’ or ‘D2.15’ in H5 and H6 populations. d) Distribution of marker type configurations for common markers between H5 and H6 populations. The joint marker type configuration consider the marker types in H5 and H6 population simultaneously.

3.4 H5 and H6 genetic map

For H5 population, 5,350 GBS-SNPs were combined with 493 (100 ISSR, 386 AFLP and 7 SSR) makers previously reported by QUEZADA *et al.* (2014). Therefore, 5,843 markers were useful for linkage analysis. The H5 genetic map, derived from the ‘TCO × BR’ cross, comprised 1,256 markers encompassing 11 linkage groups, corresponding to the haploid chromosome number of *A. sellowiana* (Table 3.2, Figure 3.6 and Appendix II). The total length of the map was 1,593.35 cM, with LG4 being the smallest group (97.14 cM) and LG1 the largest (204.43 cM). The number of markers per group ranging from 68 (LG5) to 164 (LG10) with a mean of 112.36 markers per linkage group. The average marker density of the map was 1.29 cM.

For the construction of the H6 genetic map, corresponding to the ‘TCO × DP’ cross, 4,227 GBS-SNPs were available. A total of 1,302 markers were mapped in the H6 map, spanning the expected 11 linkage groups for the species (Table 3.2, Figure 3.6 and Appendix II). The

map covered a total distance of 1,572.10, with the size of linkage groups ranging from 111.49 cM (LG3) to 230.28 cM (LG1). The average number of markers per linkage group was 118.36, with a minimum number of 74 markers in LG5 and a maximum of 172 markers in LG8. The average distances between markers was the 1.21 cM.

A low level of segregation distortion was observed in both H5 and H6 genetic maps (Figure 3.4). Of the 1,256 markers mapped in H5 genetic map, only 47 markers (3.80%) show segregation distortion, and 48 markers of the 1,302 mapped (3.69%) in H6 map presented deviation. However, the pattern of segregation distortion was different between maps. The LG5, LG7 and LG9 of the H5 map, showed more than 5% of the markers with segregation distortion, with 8 (11.76%), 11 (8.21%) and 14 (13.21%) markers distorted, respectively. For H6 map, 9 markers (10.34%) in LG1, 11 markers (11.96%) in LG2 and 15 markers (11.19%) in LG10 showed segregation distortion. Markers in LG1, LG2 and LG3 of H5 map and LG3 of H6 map did not show any segregation distortion.

Figure 3.4: Patterns of segregation distortion for H5 and H6 genetic maps. Linkage groups with more than 5% of distorted markers are shown. The x-axis represents the genetic position (cM) on each linkage groups and the y-axis represents the $-\log_{10}(p - \text{values})$ of segregation distortion. The discontinuous lines indicate threshold for significant distortion (Bonferroni-adjusted $p - \text{value} = 0.05$).

Of the 1,256 mapped markers in H5 genetic map, only 23 markers were previously mapped in the former H5 map (QUEZADA *et al.*, 2014). Although the number of common markers was low, comparison of marker positions revealed collinearity between the two maps. Among these 23 markers, two and six markers were mapped onto LG5 and LG9 in both H5 maps, respectively. The remaining markers were mapped into small linkage groups of the former H5 linkage map. Only one marker showed a conflict position between maps, mapped in LG6 in the former map and in LG5 of the map developed in the present study (Appendix III).

Out of 1,256 markers mapped in H5 map, 301 (24.35%) corresponded to SNPs identified by aligning to the *E. grandis* reference genome (Discovery pipeline), and for H6 map, the same was observed for 335 (25.73%) of the 1,302 SNPs mapped. Analysis of the distribution of the Discovery-identified markers among the LG, showed a high synteny between the genetic maps and the *E. grandis* reference genome (Figure 3.5). It was found that 271 (90.03%) SNP markers in H5 map and 311 (92.84%) markers in H6 map, mapped on the corresponding reference chromosomes of *E. grandis*. Only a few markers, 26 (8.64%) and 20 (5.97%) for H5 and H6 map, respectively, mapped on other linkage groups than the corresponding reference chromosome. The remaining four markers in H5 and H6 genetic map, were assigned to *E. grandis* scaffolds. In both genetic maps, the LG1 presented the higher proportion of syntenic markers (36.96% for H5 and 48.28% for H6), whereas the LG3 showed the lower proportion (12.38% and 12.50%, for H5 and H6 maps, respectively). More than 20% of the total SNPs mapped in LGs 4, 5, 6, 7 and 8 of H5 map and LGs 4, 5, 6, 8, 10 and 11 for H6 map were identified as syntenic markers. Despite the high level of synteny observed in the genetic maps, the marker order (colinearity) was not consistent with the physical position in the *E. grandis* genome.

3.5 Composite genetic map

In order to construct a composite map, 641 common markers between H5 and H6 map were identified. All the shared SNPs were located in the same linkage groups in both maps, and also the overall order was largely consistent between LGs. The composite map developed using 641 common SNP markers, spanned a total length of 1,011.13 cM along 11 LGs (Table 3.2 and Figure 3.6). The average group length was 91.92 cM, with LG3 the smallest group (46.79 cM) and LG1 the largest group (139.14 cM). The number of markers per group ranged from 11 (LG5) to 85 (LG6) with a mean of 58.27 and an average marker density of 1.58 cM.

Only 5 (0.78%) of the markers included in the composite map showed segregation distortion in the H5 population; the same number were distorted in H6 population. The marker Eg15.314547 mapped in the LG5 was the only that showed segregation distortion in both populations. Markers with segregation distortion were mapped in LGs 2, 4, 5, 7, 8, 10.

A high agreement was observed between the genetic positions of the Discovery-identified SNPs mapped in the composite map and the physical positions of these markers on the *E. grandis* genome. This result was expected considering the high synteny observed in the single population maps. Of the total mapped markers, 166 SNPs corresponded to markers identified by the Discovery pipeline, and of these, 149 (89.76%) showed correspondence between the linkage group and the *E. grandis* chromosome. Other 14 markers were mapped in discordant reference chromosome and the remaining three were assigned to *E. grandis* genome scaffolds.

Table 3.2: Description of H5, H6 and composite linkages maps

LG	H5 Map			H6 Map			Composite Map		
	<i>N</i> ^o of Markers	Length (cM)	Average distance	<i>N</i> ^o of Markers	Length (cM)	Average distance	<i>N</i> ^o of Markers	Length (cM)	Average distance
LG1	92	204.43	2.22	87	230.28	2.65	48	139.14	2.89
LG2	84	89.40	1.06	92	124.42	1.35	53	76.88	1.45
LG3	105	117.80	1.12	128	111.49	0.87	70	46.79	0.67
LG4	91	97.14	1.06	106	128.43	1.21	72	85.49	1.19
LG5	68	155.31	2.28	74	135.35	1.83	11	111.51	10.14
LG6	136	110.69	0.81	149	147.86	0.99	85	94.43	1.11
LG7	134	176.24	1.31	137	160.69	1.17	77	121.90	1.58
LG8	143	151.17	1.06	172	130.98	0.76	64	54.38	0.85
LG9	106	196.89	1.86	106	114.94	1.08	54	70.96	1.31
LG10	164	137.70	0.84	134	136.94	1.02	63	81.98	1.30
LG11	113	156.56	1.38	117	150.71	1.29	44	127.66	2.90
Total	1,236	1,593.35	1.29	1,302	1,572.10	1.21	641	1,011.13	1.58

Figure 3.5: Synteny of a) H5, b) H6 and c) composite genetic maps of *Acca sellowiana* and the *Eucalyptus grandis* genome. The x-axis indicates the genetic position (cM) of SNP markers identified by aligned to the *E. grandis* reference genome. The y-axis indicates the physical position (Mb) of each SNP marker. For each linkage group, SNPs corresponding to the reference chromosome of *E. grandis* are labeled in green. The orange points represented SNPs in discordant reference chromosomes.

Figure 3.6: Comparison of homologous linkage groups from the composite and single population H5 and H6 genetic maps of *Acca sellowiana*. Map distances (cM) are indicated at the left side of the figure. SNP markers common between LGs of the three genetic maps are connected with blue lines.

4 DISCUSSION

Developing genome-wide molecular markers and genetic maps is essential for genetic analysis and molecular breeding in crop species. These genomic resources are especially useful in minor crops, to encourage efficient breeding. To initiate genetic studies in *A. sellowiana*, genotyping by sequencing approach was efficiently employed to obtain a large number of SNP markers for the first time in the species. High-density genetic maps were developed for two connected populations, with a high level of collinearity and genome coverage. Integration of the single population maps resulted in a final composite map, a valuable framework for genomic analysis and association studies for desirable agronomic traits in the species.

Applied for the first time in *A. sellowiana*, the GBS approach demonstrated the advantage of coupled next generation sequencing with genome reduction methods to discover and genotype SNP markers for mapping purposes. This approach provides a fast, simple, efficient and cost-effective strategy to obtain a significant number of markers, an invaluable asset for minor or orphan crops, with relatively few genomic and genetic resources. Before conducting a GBS experiment in a new species, a crucial consideration is the choice of the appropriate restriction enzymes. The enzymes used to construct the libraries, reduce the complexity of the genome, but can also be useful to access a specific subset or avoid repetitive regions of the genome (DAVEY *et al.*, 2011). This choice must consider the total number of fragments and their sequencing depth to achieve the marker density required in the study. Therefore, to obtain the highest number of useful markers, the genome characteristics of the species under investigation (size, GC content, methylation patterns), as well as the level of polymorphism in the population studied, should be considered.

Most GBS approaches use restriction enzymes of 5-bp and 6-bp recognition sites, as the ApeKI restriction enzyme used in the original GBS protocol described by ELSHIRE *et al.* (2011). Studies comparing the effects of different restriction enzymes, showed that ApeKI samples a greater proportion of the genome while PstI targets single-copy gene rich regions of the genome (RUSSELL *et al.*, 2014). When the genome sequence is available, the most suitable restriction enzyme can be determined via *in silico* digestion of the reference genome (SONAH *et al.*, 2013). Thus, the best enzyme is selected based on the number of fragments of optimal size produced. For species with large genomes or no reference genome, the use of methylation-sensitive enzymes of less frequent cuts (6-bp or greater restriction sites) was recommended to achieve a higher level of complexity reduction (POLAND and RIFE, 2012). Thus, targeting fewer sites with greater depth at each genomic site results in the identification of fewer SNPs but with greater coverage, reducing missing data.

In our study, PstI was selected between different enzymes evaluated in a previous test, because it produced the largest number of fragments of desired size. Using a rare-cutting enzyme we expected fewer sites but with greater coverage per sample, suitable for a full-sib mapping population of outcrossing species. It is relevant to note, that for full-sib mapping populations with a high degree of linkage disequilibrium (LD), the number of markers required for covering the entire genome is small. The total number of reads and average read per sample was comparable to the observed for other fruit tree species studies (RUSSELL *et al.*, 2014; GUAJARDO *et al.*, 2015). However, these values are influenced by several factors, such as the restriction

enzyme used and the genome features of the specie studied. A high variation was observed in the number of reads per progeny evaluated (ranging from 0.27 to 6.2 M) (Figure 3.1). This result shows the sensitivity of the GBS protocol to the initial DNA quality and quantity, as was observed by GARDNER *et al.* (2014). This observation support our experimental design, where the parents of each mapping population were replicated ensuring the greater sequencing depth needed to establish the correct genotypes for these critical genotypes.

The original GBS protocol introduces the use of a sequenced reference genome to identify SNPs (ELSHIRE *et al.*, 2011). Many commercially fruit trees species as apple, peach, grape and banana, have the completed reference genome available, which may also be use for SNP identification, including closely-related species. *A. sellowiana* is a member of the Myrtaceae family, where *E. grandis* is the only species with a sequenced genome. To map an identified SNP markers, two strategies were carried out. The Discovery strategy used the *E. grandis* genome as reference, whereas the UNEAK approach paired matched sequence tags identify SNPs.

SNP calling using a reference genome for a related species is widely utilized, as in the Rosaceae family, which contains a high number of fruit tree species (GUAJARDO *et al.*, 2015; SALAZAR *et al.*, 2017). The final level of read alignments mainly depends on the quality of the genome model used and also the phylogenetic relationship between the species. HYMA *et al.* (2015) obtained an alignment rates between 66% and 72% studying four grape mapping populations, species with a fully sequenced reference genome available. The percentage of un-aligned reads can be explained by poor quality reads, poor quality genome assembly, as well as the presence of divergent or species-specific sequences not represented in the current assembly. High differences in alignment rates were observed for BALSALOBRE *et al.* (2017) when aligning sugarcane GBS reads against four references. The highest value (87.94%) resulted from the use of methyl-filtered sugarcane genome references, whereas the lowest value (42.29%) was achieved using the complete genome of sorghum, a closely-related species.

Here, we found an acceptable alignment value (30.1%) of *A. sellowiana* reads against the *E. grandis* genome. Using the same reference, *E. urophylla* obtained twice the proportion of aligned reads (69.1%), but these species presented a small phylogenetic distance (BARTHOLOMÉ *et al.*, 2014). It is important to highlight our result considering *A. sellowiana* belongs to the tribe Myrteae, a divergent lineage of tribe Eucalypteae of Eucalyptus species. However, our approach was based on the high synteny within the family, which shares a recent common ancestor resulting in a monophyletic group (GRATTAPAGLIA *et al.*, 2012). In addition, the recent comparison of chloroplast genomes of myrtle species, including *A. sellowiana* and *E. grandis* showed that these species present a highly conserved genome content, gene order and genomic structure (DE MACHADO *et al.*, 2017). Our results remark the high synteny within the family, as also as the utility of *E. grandis* reference genome in genomic studies of minor crops in the Myrtaceae family.

The distribution of reads and SNPs was homogeneous between the *E. grandis* chromosomes (Figure 3.2). Chromosomes 8, 2 and 6 presented the highest number of identified SNPs. This pattern was identical to that reported by BARTHOLOMÉ *et al.* (2014), in the alignment of short reads obtained from the complete genome sequencing of *E. grandis* and *E. urophylla*. Although these chromosomes presented medium to large size, the number of identified markers

has not a linear relationship with chromosome size. The largest chromosomes, 5 and 3, presented the lowest SNP density. For these same chromosomes, BARTHOLOMÉ *et al.* (2014) identified regions less well covered, characterized by a high rate of missing data (gaps) in the genome. In our study, the pattern of SNP markers distribution can be partially explained by unique characteristics of the *E. grandis* assembled genome.

In the absence of a reference genome, some pipelines were developed using a *de-novo* based approach to discover markers. The UNEAK pipeline is the most widely used protocol, with a network-based strategy to identify SNPs (LU *et al.*, 2013). Using this approach, we identified 13,430 SNP markers, ~ 3 fold fewer than the number of markers identified in the first strategy, using the *E. grandis* reference genome. Reference-free approaches are considered less efficient than the reference-based SNP discovery pipelines, and a notable difference in the total number of identified markers was observed when comparing the most widespread GBS pipelines (TORKAMANEH *et al.*, 2016). The low SNP number identified by the UNEAK protocol can be probably explained by the fact that only one mismatch for pair-read was allowed for calling SNPs. This stringent network filter not only reduce the errors in genotype calling, caused by paralogous or repetitive sequences, but also the total number of identified SNPs. In order to overcome the pipeline limitations in the absence of a reference genome, recent studies have suggested complementary approaches, as assembling reads before identifying SNPs or using transcriptome sequence data (RUSSELL *et al.*, 2014). The new strategy proposed by BERTHOULY-SALAZAR *et al.* (2016) interrogates coding regions using a transcriptome data, which facilitates further analysis and annotation of the genomic region, and SNP validation. The two complementary strategies implemented in our study allowed to successfully identify a high number of SNP markers for *A. sellowiana*. With the advance of genomic resources, new approaches may be able to improve the genetic knowledge about the species.

The correct genotype estimation from the NGS data presents unique challenges for highly heterozygous outcrossing species, compared to the study of inbred lines. The low read coverage of GBS data results in a larger proportion of missing data and the under-calling of heterozygous genotypes in full-sib F_1 population samples. Difficulty in identifying heterozygous individuals may be due to both alleles not being equally sampled and to the high sequencing base error rate of NGS reads. The presence of missing values and the genotype error due to the heterozygote under-calling, hinder the linkage analysis. It has a substantial impact in order and phase markers, increasing the linkage map total length. To overcome these limitations, a few approaches were developed to genotype calling in outcrossing species; however, the parameters and algorithms employed affect the final quantity and quality of the SNP data. As an example, KIM *et al.* (2016) proposed the minimum threshold of $6\times$ read coverage at a single locus to call a heterozygous genotype with a confidence greater than 0.95 assuming a binomial distribution in diploid species. However, the trade-off between the number and quality of genotypes needs to be considered, since this theoretical argument can dramatically reduce the number of final SNPs, increasing the missing data. Several mechanisms to impute missing genotypes have been proposed for outcrossing species based on haplotype-phasing algorithms, as well as order or linkage phases of mapped markers (SWARTS *et al.*, 2014; COVARRUBIAS-PAZARAN *et al.*, 2016). But the high heterozygosity and the absence of a reference genome hinder the correct

estimation of missing genotypes. Moreover, it is important to note that the missing data can be of biological significance, evidencing biological divergence, through small insertions, deletions or structural variants in the genome sequence. We used the genotype calling method implemented in the software SuperMASSA, which takes advantages of the relative abundance of each allele (read numbers) and the Mendelian properties of the mapping populations (SERANG *et al.*, 2012). To call genotypes with a high confidence, the software simultaneously considers all the genetic information available, as the parental information, the coverage of each allele and the expected frequencies of individuals for each locus (GARCIA *et al.*, 2013). Despite the fact that it was initially designed for genotyping calling in polyploid species, this quantitative genotyping analysis has proven to be an efficient approach to overcome the main constraints presented by highly heterozygous species.

The SNP and genotyping calling approaches used in the present study allowed us to obtain a large number of high confidence SNP markers useful for mapping purposes. However, the final number of useful SNPs represents a small proportion of the initially identified SNP set. This dramatic reduction has been reported for many fruit tree species, where many SNPs are identified, but genotype is generated only for a small proportion of them (GARDNER *et al.*, 2014). Here we discovered more than 55,000 SNP markers considering the approach with and without a reference genome. Of these, 6.62% were discarded for variant identification errors (variation not identified or triallelic SNPs) and the majority of them due to lower genotype call rate. Considering 25% as threshold value of missing data, 44.81% and 48.20% of the markers in H5 and H6 populations respectively, were excluded. This threshold was chosen considering that the multipoint maximum likelihood algorithm employed for map construction is highly efficient even with missing data (MOLLINARI *et al.*, 2009). Using this filtered process, less than half of the initial markers discovered (48.58% and 45.19% for H5 and H6 populations, respectively) were retained for genotype calling analysis in SuperMASSA software. Subsequently, less than 20% of the filtered markers adjusted to quality control established in this software. Remarkably, the proportion of selected markers significantly differed between the pipelines where SNP markers were identified. Only 12.44% and 9.56% of the markers identified in the Discovery approach for H5 and H6 populations, were selected respectively. This proportion increased to 55.22% for H5 and 51.02% for H6 populations, for marker identified in the UNEAK approach. This result may be due to the phylogenetic divergence between *A. sellowiana* and *E. grandis* genomes. A high number of sequence polymorphism between these genomes was found, but these variations do not represent segregating polymorphisms in the mapping populations. Finally, we obtained a total number of 5,350 and 4,227 SNPs for H5 and H6 populations, respectively, representing the 9.42% and 7.44% of markers initially identified. Our results are within the expected for GBS data in full-sib mapping populations, and are comparable to those found in apple and grape, where only 6.0% and 4.2% of the markers were retained for linkage mapping purposes (GARDNER *et al.*, 2014; HYMA *et al.*, 2015). Despite the high reduction in the number of available markers, the final number of useful markers is fully sufficient for covering the entire genome.

Conservative GBS-SNP sets were used to create two integrated high-density genetic maps in *A. sellowiana*. Maps with 1,256 and 1,302 markers for H5 and H6 populations, respectively were obtained, where only 24 of the markers mapped in H5 map corresponded to PCR

markers previously mapped at the first map of the same cross. Both genetic maps established 11 linkage groups, matching the expected haploid chromosome number ($n = 11$). The total number of mapped markers on both maps was effective to obtain a well saturated linkage map, comparable to high-density maps recently published for tree fruit species (WARD *et al.*, 2013; RUSSELL *et al.*, 2014; GUAJARDO *et al.*, 2015; ?). Nevertheless, the percentage of SNPs mapped from the initial data set was low (23.48% for H5 population and 30.80% for H6 population). This result could be explained by an uneven distribution of GBS-SNPs in the genome, resulting in large under-represented areas that could not be successfully covered in the genetic map. The use of the methylation-sensitive enzyme PstI, can introduce bias in the representation of low-methylated regions, which correspond to repetitive regions. Preliminary studies in *A. sellowiana* showed that repetitive DNA accounts for 40% of the genome, where more than 70% was identified as retrotransposons type LTR (VAIO *et al.*, 2015). Moreover, the lack of markers in large regions could correspond to proportions of the genome that are identical by descent, a result of limited genetic diversity between the two parental genotypes. The absence of polymorphism in particular genome regions affected the total number of markers retained in each map and also the distribution in the genetic maps. Although, the mapped markers showed a uniform distribution across both maps, some gaps were observed in the single population maps. As an example, a gap greater than 30 cM was found in LG11 for H5 map and the largest gap of 46.98 cM was found in LG1 of H6 map. These gaps confirmed the lack of marker detection as well as a low degree of sequence polymorphisms in these regions. The segregation analysis using connected mapping populations and molecular markers with different sequence information might represent a promising strategy to increase the coverage of the genome, resulting in saturated genetic linkage maps. For example, a map region with low marker density can be inferred indirectly by comparing it to the genetic map of a connected population.

The H5 and H6 genetic maps had a comparable number of total map length and average distance between markers. The total H5 map length was 1,593 cM with an average distance between markers of 1.29 cM, whereas H6 map presented a total length of 1,572 cM and an average distance of 1.21 cM. However, H5 map presented a lower number of total markers mapped spanning a greater map length, since it is based on a smaller segregating population (160 individuals). These results were in accordance with previous reports about the influence of population size in genetic linkage maps, specifically in outcrossing species. The number of progeny has a direct effect on the number of recombinants events detected, since smaller populations have fewer recombinants than larger populations. For example, considering a mapping population of $n = 200$, the minimum recombination fraction between markers that can be detected should be 0.5 cM, with one recombinant individual. As a result, mapping resolution (average distance between adjacent markers) of 1 – 2 cM are expected. For H5 and H6 genetic maps, resolution mapping achieved were in accordance with the mapping population sizes ($n = 160$ and $n = 184$ for H5 and H6 populations, respectively). In addition, a small impact was found between the population size and the average map length, but longer map distances were observed for small sample sizes (BARTHOLOMÉ *et al.*, 2014). Moreover, a high correlation was found between population size and the number of markers present in each map, comparing five grape mapping populations (HYMA *et al.*, 2015). Our results provide evidence in agreement to recent reports

(BARTHOLOMÉ *et al.*, 2014; HYMA *et al.*, 2015), which showed that the number of recombination events evaluated, thus the population size, is the current limiting factor to constructing high-density linkage maps in outcrossing species.

The GBS technique enabled the development of thousands of useful SNP markers to construct genetic linkage maps, with improved genome coverage. However, high-density genetic maps reported for outcrossing species showed a remarkable difference in numbers of markers retained in each linkage group. In our study, the linkage groups present a slight variability in the number of markers mapped. Noticeably, the LG5 of H5 and H6 maps present an extremely low number of markers. A similar result was found for the raspberry genetic map, with great differences in total markers between linkage groups, but it was not consistent between the parental and maternal maps (WARD *et al.*, 2013). These results may be due to a nonuniform distribution of SNP markers among chromosomes, consequence of structural variations or genome rearrangements. Moreover, these differences can be explained by the lack of polymorphism in the same regions of the genome. In addition, significant differences between the total length of linkage groups have been found in high-density genetic maps developed for blackberry and sweet cherry (RUSSELL *et al.*, 2014; GUAJARDO *et al.*, 2015). Although, this observation can be due to differences in chromosome size, this relationship was not detected. In the present work, the most peculiar result was the largest total length and average distance of LG1 in H5 and H6 genetic maps. This result can be explained by a high rate of genotyping errors in this LG. Simulated experiments designed to evaluate the effect of genotyping errors in mapping construction, showed an increase of the total length of the linkage groups, due to hindering the ordering process and overestimating the genetic distances (HACKETT and BROADFOOT, 2003). Nevertheless, more genome sequence and cytological information is necessary to understand the main cause of differences observed in markers mapped and total length between linkage groups.

Segregation distortion, a deviation of frequency of genotypes from the expected Mendelian ratio, is a widely spread phenomenon in mapping populations. Distortion of Mendelian segregation ratios may affect the accuracy of linkage map by introducing errors in the recombination frequencies between markers as well marker order estimation, subsequently hindering the QTL detection (WANG *et al.*, 2005). Otherwise, HACKETT and BROADFOOT (2003) showed that segregation distortion does not have a major impact in genetic distances, map length or marker order, and therefore does not affect the construction of genetic linkage maps. Distorted markers were considered in this mapping analysis, in order to investigate the distortion patterns in *A. sellowiana* as well as the effects in genetic map. Here, low proportions of segregation distortion, 3.80% and 3.69% were observed for H5 and H6 maps. For other outcrossing species, as cranberry, apple and grape, higher levels of segregation distortion have been reported (CLARK *et al.*, 2014; HYMA *et al.*, 2015; COVARRUBIAS-PAZARAN *et al.*, 2016). In accordance with these reports, segregation distortion in *A. sellowiana* was not randomly spread across the linkage groups, but distributed as clusters, in segregation distortion regions. The identification of segregation distortion areas along specific linkage groups can be explained by biological factors, such as genome structural variation, chromosome loss, locus duplication, genetic isolation mechanisms, genetic load and gamete competition (WANG *et al.*, 2005). As an example, segregation distortion markers on a cranberry map were associated with areas of possible chromosome inversion

(COVARRUBIAS-PAZARAN *et al.*, 2016). In our maps, a low level of segregation distortion with an inconsistent pattern between H5 and H6 maps, as well as the lack of genomic information hamper a biological interpretation for the segregation distortion observed. Therefore, we can explain our results based on technical problems, as scoring genotyping errors or sampling errors. In order to understand the main factors responsible for segregation distortion in this species, more genetic mapping, genetic and cytogenetic information are required.

The level of synteny between *A. sellowiana* ($n = 11$) and *E. grandis* ($n = 11$) was examined through the distribution of Discovery-identified SNPs, markers which were assigned to a chromosomal position in the *E. grandis* genome. A high degree of synteny was found between the species, considering that more than 90% of the Discovery-identified markers mapped in H5 and H6 genetic maps matched with the correspond chromosome of *E. grandis*. Although this result suggested the conserved localization of markers, colinearity (order of the markers) within each linkage group was not observed. Our result is remarkable considering the phylogenetic divergence and the genome size difference between *A. sellowiana* (382 Mb) and *E. grandis* (641 Mb). Moreover, it is consistent with the observation that chromosome number and synteny is highly conserved in plant families, despite the dramatic differences in genome size (HUDSON *et al.*, 2012a). Comparative mapping studies between species with longer evolutionary distance show a high level of synteny and colinearity, even in the presence of significant genome rearrangements, such as whole genome duplication and polyploidization (TANG *et al.*, 2008). Previous studies mainly between species of the tribe Eucalyptae, showed a high level of sequence conservation within the Myrtaceae family (GRATTAPAGLIA *et al.*, 2012). The eucalypts species with the same a haploid number ($n = 11$) analysed by HUDSON *et al.* (2012a), showed a variable degree of synteny and colinearity according to the phylogenetic relationships and differences in genome size among the species evaluated. However, a more comprehensive analysis is required to quantify the conserved degree within the whole family, including fleshy fruit and dry fruit species. Although our study results are preliminary, this is the first reported of a synteny study between a fresh fruit Myrtacea species with the reference species *E. grandis*. For a small-scale study of synteny, more molecular markers and genome sequences should be evaluated.

GBS-SNP markers common between H5 and H6 maps allowed the construction of the first composite map for *A. sellowiana*. A large number of common markers (51.86% and 49.23% of mapped markers in H5 and H6 map, respectively) was expected considering that both populations shared the same female parent. All the common markers had a complete correspondence with homologous linkage groups in both single population maps. In addition, the markers showed a strong colinearity for each linkage group. As commonly reported in mapping studies, some local inconsistencies in marker order were observed between the H5 and H6 maps. These inconsistencies can be attributed to biological factors, such as chromosomal rearrangements (inversion, translocation), as well as segmental duplications or deletions in parental genotypes. Local rearrangements may serve as hot spots of structural variation, affecting the recombination frequencies in different genotypes, that influenced marker order accuracy (KHAN *et al.*, 2012). Sex-differences in the recombination rates can also result in incorrect orders in the single population maps. Studying multiple mapping populations, local inconsistencies in marker order were attributed to a sampling bias for small mapping populations ($n = 50$) (DOLIGEZ *et al.*, 2006).

The overall agreement in order of loci between the H5 and H6 maps suggests technical errors from genotyping methods as the main cause of the local inconsistencies, rather than major chromosomal rearrangements for the parental genotypes. The small inconsistencies can be a result of missing data, genotyping errors and difference in content information among the markers and populations. Differences in recombination rates between the three parental genotypes were not considered, since they had little effect in incongruence of marker orders. Moreover, markers with segregation distortion can be considered another potential source of error in the ordering process (HUDSON *et al.*, 2012b). But, the small proportion of distorted markers mapped in our genetic maps reduce their effect on marker order. In addition, genetics maps can be presented different distance and order for markers due to sampling bias, a common effect in statistical analysis like linkage mapping studies.

Several mapping procedures for constructing composite genetic maps have been reported and applied to outcrossing species (VAN OOIJEN, 2006; RONIN *et al.*, 2012; ENDELMAN and PLOMION, 2014). Despite differences in the statistical approaches, all of them are based on the combination of genetic distances of single population maps. These approaches are negatively affected by differences in population size, different marker information content, missing data as well as the proportion of common markers. Subsequently, inconsistencies in order map and inflationary map distances reduce the resolution and accuracy of the maps generated through these approaches (DOLIGEZ *et al.*, 2006; DE KEYSER *et al.*, 2010; CLARK *et al.*, 2014).

In order to construct a composite genetic map, the strategy proposed in this work integrated the recombination information from the individuals of two connected populations. The main advantage of our procedure is to capture the meiosis information of all individuals in a single composite map, since the information data of common markers between the two separate populations was merged. Considering the total correspondence between linkage groups and the high colinearity between maps, the integration process was carried out for each linkage group independently. In our strategy, only common markers were considered, with the objective of reducing missing information which hinders the marker ordering process (HACKETT and BROADFOOT, 2003). As a result, the composite map obtained provides the most accurate order markers yet produced for *A. sellowiana*. This reference map allowed the correction of the order inconsistencies in the single population maps H5 and H6, resulting in robust high-density linkage maps. The improved estimation of marker positions can be explained by the larger number of recombination events evaluated in a composite approach, compared to single population approaches. In addition, the efficiency of our strategy is due to the multipoint maximum likelihood estimation using genetic information derived from two populations simultaneously; this is a good approach both for statistical and biological reasons.

Here, the composite genetic map provided a reference framework map ordering for *A. sellowiana*, a valuable tool for future genetic and genomic applications. The composite maps, especially improve the accuracy of order marker and genetic distances, facilitating the comparison between genetic maps. For example, the recent multi-parental genetic map developed in apple allowed to detect discrepancies and modify the orientation of the LG5, to match the orientation of the homologous sequences in LG10 (DI PIERRO *et al.*, 2016). With the development of composite genetic maps, the occurrence of genome structural variation or conserved

synteny can be evaluated across divergent species. To date, the comparative mapping studies in the Myrtaceae family only included dry fruits species of the tribe Eucalypteae (GRATTAPAGLIA *et al.*, 2012). Therefore, the composite map developed in this work can be useful to extend the studies including more divergent species of the family. In addition, the composite maps allowed to align the position of QTLs detected across variable genetic backgrounds, facilitating the transfer of genetic information from molecular markers and gene positions, and accelerating molecular breeding strategies such as marker assisted selection (DIAZ *et al.*, 2011).

In this study, we constructed the first high-density genetic maps for *A. sellowiana* using the genotyping by sequencing approach for the first time in the species. The GBS protocol was an effective strategy for simultaneous SNP discovery and genotyping, identifying thousands of genome-wide polymorphic markers, with the advantage of cost and time reduction. From two full-sib connected mapping populations, we developed the first high-density genetic maps, which a high number of common markers essential for comparative mapping studies as well as for the construction of a framework composite map. The composite linkage genetic map developed using the genetic information for both populations simultaneously provides a better estimation of recombination fraction, resulting in higher accuracy for marker distance and order. The composite map along with the H5 and H6 maps, are the most comprehensive representations of the *A. sellowiana* genome and the main genetic resources available for this minor species. These maps could be useful for future genetic studies, as the detection of QTL of important agronomic traits, comparative genome analysis in the Myrtaceae family and the acceleration of the breeding process in *A. sellowiana*.

REFERENCES

- ALVAREZ, M., G. MACHADO, G. ROSTAGNOL, S. AGUERRE, J. BERNAL, Y. CAZZULO, V. CEJAS, Z. FERNANDA, D. CABRERA, B. VIGNALE, C. PRITSCH, and M. QUEZADA, 2017 Diversidad fenotípica y relaciones entre caracteres de calidad de fruta en dos progenies F1 de guayabos del país [*Acca sellowiana* (Berg) Burret]. In *8to Encuentro Nacional sobre Frutos Nativos*, p. 13.
- ANDREWS, S., 2010 FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> .
- BACCINO, E., 2011 *Estructura genética de cuatro poblaciones silvestres de Acca sellowiana (Berg) Burret situadas en el noreste de Uruguay*. Ph.D. thesis, Universidad de la República, Facultad de Ciencias.
- BAIRD, N. A., P. D. ETTER, T. S. ATWOOD, M. C. CURREY, A. L. SHIVER, Z. A. LEWIS, E. U. SELKER, W. A. CRESKO, and E. A. JOHNSON, 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**: 1–7.
- BALSALOBRE, T. W. A., G. DA SILVA PEREIRA, G. R. A. MARGARIDO, R. GAZAFFI, F. Z. BARRETO, C. O. ANONI, C. B. CARDOSO-SILVA, E. A. COSTA, M. C. MANCINI, H. P. HOFFMANN, A. P. DE SOUZA, A. A. F. GARCIA, and M. S. CARNEIRO, 2017 GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC Genomics* **18**: 72.
- BARBA, P., L. CADLE-DAVIDSON, J. HARRIMAN, J. C. GLAUBITZ, S. BROOKS, K. HYMA, and B. REISCH, 2014 Grapevine powdery mildew resistance and susceptibility loci identified on a high-resolution SNP map. *Theoretical and Applied Genetics* **127**: 73–84.
- BARTHOLOMÉ, J., E. MANDROU, A. MABIALA, J. JENKINS, I. NABIHOUDINE, C. KLOPP, J. SCHMUTZ, C. PLOMION, and J.-M. GION, 2014 High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytologist* **206**: 1283–1296.
- BERTHOULY-SALAZAR, C., C. MARIAC, M. COUDERC, J. POUZADOUX, J.-B. FLOC'H, and Y. VIGOUROUX, 2016 Genotyping-by-Sequencing SNP Identification for Crops without a Reference Genome: Using Transcriptome Based Mapping as an Alternative Strategy. *Frontiers in Plant Science* **7**: 1–11.
- BEYHAN, Ö., M. ELMASTAŞ, and F. GEDIKLI, 2010 Total phenolic compounds and antioxidant capacity of leaf, dry fruit and fresh fruit of feijoa (*Acca sellowiana*, Myrtaceae). *Journal of Medicinal Plants Research* **4**: 1065–1072.
- BIELENBERG, D. G., B. RAUH, S. FAN, K. GASIC, A. G. ABBOTT, G. L. REIGHARD, W. R. OKIE, and C. E. WELLS, 2015 Genotyping by sequencing for SNP-based linkage map construction and QTL analysis of chilling requirement and bloom date in peach [*Prunus persica* (L.) Batsch]. *PLoS ONE* **10**: 1–14.

- BODÉNÈS, C., E. CHANCEREL, F. EHRENMANN, A. KREMER, and C. PLOMION, 2016 High-density linkage mapping and distribution of segregation distortion regions in the oak genome. *DNA Research* **23**: 115–124.
- BONTEMPO, P., L. MITA, M. MICELI, A. DOTO, A. NEBBIOSO, F. D. BELLIS, M. CONTE, A. MINICHIELLO, F. MANZO, V. CARAFA, A. BASILE, D. RIGANO, S. SORBO, R. CASTALDO, E. MARIANO, F. FERRARA, M. D. SIMONE, M. VIETRI, M. CIOFFI, V. SICA, F. BRESCIANI, A. R. D. LERA, L. ALTUCCI, and A. MARIA, 2007 Feijoa selowiana derived natural Flavone exerts anti-cancer action displaying HDAC inhibitory activities. *International Journal of Biochemistry* **39**: 1902–1914.
- BYRNE, D. H., 2012 Fruit Breeding, Handbook of Plant Breeding. In *Trends in Fruit Breeding*, pp. 36–53, Springer US.
- CLARK, M. D., C. A. SCHMITZ, U. R. ROSYARA, J. J. LUBY, and J. M. BRADEEN, 2014 A consensus 'Honeycrisp' apple (*Malus domestica*) genetic linkage map from three full-sib progeny populations. *Tree Genetics and Genomes* **10**: 627–639.
- COLLARD, B. C. Y., M. Z. Z. JAHUFER, J. B. BROUWER, and E. C. K. PANG, 2005 An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142**: 169–196.
- COVARRUBIAS-PAZARAN, G., L. DIAZ-GARCIA, B. SCHLAUTMAN, J. DEUTSCH, W. SALAZAR, M. HERNANDEZ-OCHOA, E. GRYGLESKI, S. STEFFAN, M. IORIZZO, J. POLASHOCK, N. VORSA, and J. ZALAPA, 2016 Exploiting genotyping by sequencing to characterize the genomic structure of the American cranberry through high-density linkage mapping. *BMC Genomics* **17**: 451.
- DA COSTA, I. R., M. C. DORNELAS, and E. R. FORNI-MARTINS, 2008 Nuclear genome size variation in fleshy-fruited Neotropical Myrtaceae. *Plant Systematics and Evolution* **276**: 209–217.
- DANECEK, P., A. AUTON, G. ABECASIS, C. A. ALBERS, E. BANKS, M. A. DEPRISTO, R. E. HANDSAKER, G. LUNTER, G. T. MARTH, S. T. SHERRY, G. MCVEAN, and R. DURBIN, 2011 The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.
- DAVEY, J. W., P. A. HOHENLOHE, P. D. ETTER, J. Q. BOONE, J. M. CATCHEN, and M. L. BLAXTER, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Publishing Group* **12**: 499–510.
- DE GIVRY, S., M. BOUCHEZ, P. CHABRIER, D. MILAN, and T. SCHIEX, 2005 CARTHAGENE: Multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* **21**: 1703–1704.
- DE KEYSER, E., Q. Y. SHU, E. VAN BOCKSTAELE, and J. DE RIEK, 2010 Multipoint-likelihood maximization mapping on 4 segregating populations to achieve an integrated framework map for QTL analysis in pot azalea (*Rhododendron simsii* hybrids). *BMC Molecular Biology* **11**: 1–20.

- DE MACHADO, L. O., L. DO NASCIMENTO VIEIRA, V. M. STEFENON, F. DE OLIVEIRA PEDROSA, E. M. DE SOUZA, M. P. GUERRA, and R. O. NODARI, 2017 Phylogenomic relationship of feijoa (*Acca sellowiana* (O.Berg) Burret) with other Myrtaceae based on complete chloroplast genome sequences. *Genetica* **145**: 1–12.
- DEMPSTER, A., N. LAIRD, and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological* **39**: 1–38.
- DETTORI, M. T. and M. A. PALOMBI, 2000 Identification of Feijoa sellowiana Berg accessions by RAPD markers. *Scientia Horticulturae* **86**: 279–290.
- DI PIERRO, E. A., L. GIANFRANCESCHI, M. DI GUARDO, H. J. KOEHORST-VAN PUTTEN, J. W. KRUISSELBRINK, S. LONGHI, M. TROGGIO, L. BIANCO, H. MURANTY, G. PAGLIARANI, S. TARTARINI, T. LETSCHKA, L. LOZANO LUIS, L. GARKAVA-GUSTAVSSON, D. MICHELETTI, M. C. BINK, R. E. VOORRIPS, E. AZIZ, R. VELASCO, F. LAURENS, and W. E. VAN DE WEG, 2016 A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. *Horticulture Research* **3**: 16057.
- DIAZ, A., M. FERGANY, G. FORMISANO, P. ZIARSOLO, J. BLANCA, Z. FEI, J. E. STAUB, J. E. ZALAPA, H. E. CUEVAS, G. DACE, M. OLIVER, N. BOISSOT, C. DOGIMONT, M. PITRAT, R. HOFSTEDE, P. VAN KOERT, R. HAREL-BEJA, G. TZURI, V. PORTNOY, S. COHEN, A. SCHAFFER, N. KATZIR, Y. XU, H. ZHANG, N. FUKINO, S. MATSUMOTO, J. GARCIA-MAS, and A. J. MONFORTE, 2011 A consensus linkage map for molecular markers and Quantitative Trait Loci associated with economically important traits in melon (*Cucumis melo* L.). *BMC Plant Biology* **11**: 111.
- DOLIGEZ, A., A. ADAM-BLONDON, G. CIPRIANI, G. DI GASPERO, V. LAUCOU, D. MERDINOGLU, C. MEREDITH, S. RIAZ, C. ROUX, and P. THIS, 2006 An integrated SSR map of grapevine based on five mapping populations **113**: 369–382.
- DOS SANTOS, K. L., M. O. SANTOS, P. R. LABORDA, A. P. SOUZA, N. PERONI, and R. O. NODARI, 2008 Isolation and characterization of microsatellite markers in *Acca sellowiana* (Berg) Burret. *Molecular Ecology Resources* **8**: 998–1000.
- DOS SANTOS, K. L. D., L. J. WELTER, A. C. D. M. DANTAS, M. P. GUERRA, J. P. H. J. DUCROQUET, and R. O. NODARI, 2007 Transference of microsatellite markers from *Eucalyptus* spp to *Acca sellowiana* and the successful use of this technique in genetic characterization. *Genetics and Molecular Biology* **30**: 73–79.
- ELSHIRE, R. J., J. C. GLAUBITZ, Q. SUN, J. A. POLAND, K. KAWAMOTO, E. S. BUCKLER, and S. E. MITCHELL, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **6**: e19379.
- ENDELMAN, J. B. and C. PLOMION, 2014 LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* **30**: 1623–1624.

- GARCIA, A. A. F., E. A. KIDO, A. N. MEZA, H. M. B. SOUZA, L. R. PINTO, M. M. PASTINA, C. S. LEITE, J. A. G. D. SILVA, E. C. ULIAN, A. FIGUEIRA, and A. P. SOUZA, 2006 Development of an integrated genetic map of a sugarcane (*Saccharum* spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. *Theoretical and Applied Genetics* **112**: 298–314.
- GARCIA, A. A. F., M. MOLLINARI, T. G. MARCONI, O. R. SERANG, R. R. SILVA, M. L. C. VIEIRA, R. VICENTINI, E. A. COSTA, M. C. MANCINI, M. O. S. GARCIA, M. M. PASTINA, R. GAZAFFI, E. R. F. MARTINS, N. DAHMER, D. A. SFORÇA, C. B. C. SILVA, P. BUNDOCK, R. J. HENRY, G. M. SOUZA, M.-A. VAN SLUYS, M. G. A. LANDELL, M. S. CARNEIRO, M. A. G. VINCENTZ, L. R. PINTO, R. VENCovsky, and A. P. SOUZA, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Scientific reports* **3**.
- GARDNER, K. M., P. J. BROWN, T. F. COOKE, S. CANN, F. COSTA, C. D. BUSTAMANTE, R. VELASCO, M. TROGGIO, and S. MYLES, 2014 Fast and cost-effective genetic mapping in apple using next-generation sequencing. *G3: Genes, Genomes, Genet* **4**: 1681–1687.
- GEORGI, L., J. JOHNSON-CICALESE, J. HONIG, S. P. DAS, V. D. RAJAH, D. BHATTACHARYA, N. BASSIL, L. J. ROWLAND, J. POLASHOCK, and N. VORSA, 2013 The first genetic map of the American cranberry: Exploration of synteny conservation and quantitative trait loci. *Theoretical and Applied Genetics* **126**: 673–692.
- GLAUBITZ, J. C., T. M. CASSTEVENS, F. LU, J. HARRIMAN, R. J. ELSHIRE, Q. SUN, and E. S. BUCKLER, 2014 TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS ONE* **9**: e90346.
- GRATTAPAGLIA, D. and R. SEDEROFF, 1994 Genetic Linkage Maps of *Eucalyptus grandis* and *Eucalyptus urophylla* Using a Pseudo-Testcross: Mapping Strategy and RAPD Markers. *Genetics* **1137**: 1121–1137.
- GRATTAPAGLIA, D., R. E. VAILLANCOURT, M. SHEPHERD, B. R. THUMMA, W. FOLEY, C. KÜLHEIM, B. M. POTTS, and A. A. MYBURG, 2012 Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. *Tree Genetics & Genomes* **8**: 463–508.
- GUAJARDO, V., S. SOLÍS, B. SAGREDO, F. GAINZA, C. MUÑOZ, K. GASIC, and P. HINRICHSEN, 2015 Construction of High Density Sweet Cherry (*Prunus avium* L.) Linkage Maps Using Microsatellite Markers and SNPs Detected by Genotyping-by-Sequencing (GBS). *Plos ONE* **10**: e0127750.
- HACKETT, C. A. and L. B. BROADFOOT, 2003 Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* **90**: 33–38.
- HE, J., X. ZHAO, A. LAROCHE, Z.-X. LU, H. LIU, and Z. LI, 2014 Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science* **5**: 1–8.

- HUANG, Y.-F., J. A. POLAND, C. P. WIGHT, E. W. JACKSON, and N. A. TINKER, 2014 Using Genotyping-By-Sequencing (GBS) for Genomic Discovery in Cultivated Oat. *PLoS ONE* **9**: e102448.
- HUDSON, C. J., J. S. FREEMAN, A. R. KULLAN, C. D. PETROLI, C. P. SANSALONI, A. KILIAN, F. DETERING, D. GRATTAPAGLIA, B. M. POTTS, A. A. MYBURG, and R. E. VAILLANCOURT, 2012a A reference linkage map for Eucalyptus. *BMC Genomics* **13**: 240.
- HUDSON, C. J., A. R. K. KULLAN, J. S. FREEMAN, D. A. FARIA, D. GRATTAPAGLIA, A. KILIAN, A. A. MYBURG, B. M. POTTS, and R. E. VAILLANCOURT, 2012b High synteny and colinearity among Eucalyptus genomes revealed by high-density comparative genetic mapping. *Tree Genetics and Genomes* **8**: 339–352.
- HYMA, K. E., P. BARBA, M. WANG, J. P. LONDO, C. B. ACHARYA, S. E. MITCHELL, Q. SUN, B. REISCH, and L. CADLE-DAVIDSON, 2015 Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: A case study in grapevine. *PLoS ONE* **10**: e0134880.
- KHAN, M. A., Y. HAN, Y. F. ZHAO, M. TROGGIO, and S. S. KORBAN, 2012 A Multi-Population Consensus Genetic Map Reveals Inconsistent Marker Order among Maps Likely Attributed to Structural Variations in the Apple Genome. *PLoS ONE* **7**: e47864.
- KIM, C., H. GUO, W. KONG, R. CHANDNANI, L. S. SHUANG, and A. H. PATERSON, 2016 Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science* **242**: 14–22.
- KLABUNDE, G. H. F., D. OLKOSKI, V. VILPERTE, M. I. ZUCCHI, and R. O. NODARI, 2014 Characterization of 10 New Nuclear Microsatellite Markers in *Acca sellowiana* (Myrtaceae). *Applications in Plant Sciences* **2**: 1400020.
- KOSAMBI, D. D., 1944 The Estimation of Map Distances From Recombination Values. *Annals of Eugenics* **12**: 172–175.
- KUHN, D. N., I. S. E. BALLY, N. L. DILLON, D. INNES, A. M. GROH, J. RAHAMAN, R. OPHIR, Y. COHEN, and A. SHERMAN, 2017 Genetic Map of Mango: A Tool for Mango Breeding. *Frontiers in Plant Science* **8**: 1–11.
- LANDER, E. S. and P. GREEN, 1987 Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences* **84**: 2363–2367.
- LANGMEAD, B. and S. L. SALZBERG, 2012 Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.
- LI, L., C. H. DENG, M. KNÄBEL, D. CHAGNÉ, S. KUMAR, J. SUN, S. ZHANG, and J. WU, 2017 Integrated high-density consensus genetic map of *Pyrus* and anchoring of the ‘Bartlett’ v1.0 (*Pyrus communis*) genome. *DNA Research* p. dsw063.

- LU, F., A. E. LIPKA, J. GLAUBITZ, R. ELSHIRE, J. H. CHERNEY, M. D. CASLER, E. S. BUCKLER, and D. E. COSTICH, 2013 Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS genetics* **9**: e1003215.
- MALIEPAARD, C., J. JANSEN, and J. W. VAN OOIJEN, 1997 Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. *Genetics Research* **70**: 237–250.
- MARGARIDO, G. R. A., A. P. SOUZA, and A. A. F. GARCIA, 2007 OneMap: Software for genetic mapping in outcrossing species. *Hereditas* **144**: 78–79.
- MATHEW, L. S., M. SPANNAGL, A. AL-MALKI, B. GEORGE, M. F. TORRES, E. K. AL-DOUS, E. K. AL-AZWANI, E. HUSSEIN, S. MATHEW, K. F. MAYER, Y. A. MOHAMOUD, K. SUHRE, and J. A. MALEK, 2014 A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms. *BMC Genomics* **15**: 285.
- MOLLINARI, M., G. R. A. MARGARIDO, R. VENCovsky, and A. A. F. GARCIA, 2009 Evaluation of algorithms used to order markers on genetic maps. *Heredity* **103**: 494–502.
- MOLLINARI, M. and O. SERANG, 2013 Plant Genotyping, *Methods in Molecular Biology*. In *Quantitative SNP Genotyping of Polyploids with MassARRAY and Other Platforms*, edited by J. Batley, pp. 215–242, Humana Press.
- MYBURG, A. A., D. GRATTAPAGLIA, G. A. TUSKAN, U. HELLSTEN, R. D. HAYES, J. GRIMWOOD, J. JENKINS, E. LINDQUIST, H. TICE, D. BAUER, D. M. GOODSTEIN, I. DUBCHAK, A. POLIAKOV, E. MIZRACHI, A. R. K. KULLAN, S. G. HUSSEY, D. PINARD, K. VAN DER MERWE, P. SINGH, I. VAN JAARSVELD, O. B. SILVA-JUNIOR, R. C. TOGAWA, M. R. PAPPAS, D. A. FARIA, C. P. SANSALONI, C. D. PETROLI, X. YANG, P. RANJAN, T. J. TSCHAPLINSKI, C.-Y. YE, T. LI, L. STERCK, K. VANNESTE, F. MURAT, M. SOLER, H. S. CLEMENTE, N. SAIDI, H. CASSAN-WANG, C. DUNAND, C. A. HEFER, E. BORNBERG-BAUER, A. R. KERSTING, K. VINING, V. AMARASINGHE, M. RANIK, S. NAITHANI, J. ELSER, A. E. BOYD, A. LISTON, J. W. SPATAFORA, P. DHARMWARDHANA, R. RAJA, C. SULLIVAN, E. ROMANEL, M. ALVES-FERREIRA, C. KÜLHEIM, W. FOLEY, V. CAROCHA, J. PAIVA, D. KUDRNA, S. H. BROMMONSCHENKEL, G. PASQUALI, M. BYRNE, P. RIGAULT, J. TIBBITS, A. SPOKEVICIUS, R. C. JONES, D. A. STEANE, R. E. VAILLANCOURT, B. M. POTTS, F. JOUBERT, K. BARRY, G. J. PAPPAS, S. H. STRAUSS, P. JAISWAL, J. GRIMAPETTENATI, J. SALSE, Y. VAN DE PEER, D. S. ROKHSAR, and J. SCHMUTZ, 2014 The genome of *Eucalyptus grandis*. *Nature* **509**: 356–62.
- OLIVEIRA, E. J., M. L. C. VIEIRA, A. A. F. GARCIA, C. F. MUNHOZ, G. R. A. MARGARIDO, L. CONSOLI, F. P. MATTA, M. C. MORAES, P. O. BOX, M. I. ZUCCHI, and M. H. P. FUNGARO, 2008 An Integrated Molecular Map of Yellow Passion Fruit Based on Simultaneous Maximum-Likelihood Estimation of Linkage and Linkage Phases. *Journal of the American Society for Horticultural Sciences* **133**: 35–41.

- PADMAKAR, B., C. KANUPRIYA, P. M. LATHA, K. PRASHANT, M. DINESH, D. SAILAJA, and C. ASWATH, 2015 Development of SRAP and SSR marker-based genetic linkage maps of guava (*Psidium guajava* L.). *Scientia Horticulturae* **192**: 158–165.
- PALHARES, A. C., T. B. RODRIGUES-MORAIS, M.-A. VAN SLUYS, D. S. DOMINGUES, W. MACCHERONI, H. JORDÃO, A. P. SOUZA, T. G. MARCONI, M. MOLLINARI, R. GAZAFFI, A. A. F. GARCIA, and M. L. C. VIEIRA, 2012 A novel linkage map of sugarcane with evidence for clustering of retrotransposon-based markers. *BMC genetics* **13**: 51.
- PASQUARIELLO, M. S., F. MASTROBUONI, D. DI PATRE, L. ZAMPELLA, L. R. CAPUANO, M. SCORTICHINI, and M. PETRICCIONE, 2015 Agronomic, nutraceutical and molecular variability of feijoa (*Acca sellowiana* (O. Berg) Burret) germplasm. *Scientia Horticulturae* **191**: 1–9.
- PEACE, C. and J. L. NORELLI, 2009 Genomics Approaches to Crop Improvement in the Rosaceae. In *Genetics and Genomics of Rosaceae*, edited by K. M. Folta and S. E. Gardiner, pp. 19–53, Springer New York.
- PELGAS, B., J. BOUSQUET, S. BEAUSEIGLE, and N. ISABEL, 2005 A composite linkage map from two crosses for the species complex *Picea mariana* x *Picea rubens* and analysis of synteny with other Pinaceae. *Theoretical and Applied Genetics* **111**: 1466–1488.
- PEREIRA, G. S., E. S. NUNES, L. D. C. LAPERUTA, M. F. BRAGA, H. A. PENHA, A. L. DINIZ, C. F. MUNHOZ, R. GAZAFFI, A. A. F. GARCIA, and M. L. C. VIEIRA, 2013 Molecular polymorphism and linkage analysis in sweet passion fruit, an outcrossing species. *Annals of Applied Biology* **162**: 347–361.
- POLAND, J., J. ENDELMAN, J. DAWSON, J. RUTKOSKI, S. WU, Y. MANES, S. DREISIGACKER, M. SORRELLS, and J.-L. JANNINK, 2012a Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome* **5**: 103–113.
- POLAND, J. A., P. J. BROWN, M. E. SORRELLS, and J.-L. JANNINK, 2012b Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE* **7**: e32253.
- POLAND, J. A. and T. W. RIFE, 2012 Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome* **5**: 92–102.
- POOTAKHAM, W., P. RUANG-AREERATE, N. JOMCHAI, C. SONTIROD, D. SANGSRAKRU, T. YOOCHA, K. THEERAWATTANASUK, K. NIRAPATHPONGPORN, P. ROMRUENSUKHAROM, S. TRAGOONRUNG, and S. TANGPHATSORNRUANG, 2015 Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). *Frontiers in plant science* **6**.
- PUPPO, M., M. RIVAS, J. FRANCO, and R. L. BARBIERI, 2014 PROPUESTA DE DESCRIPTORES PARA *Acca sellowiana* (Berg.) Burret. *Revista Brasileira de fruticultura* **36**: 957–970.

- QUEZADA, M., M. M. PASTINA, G. RAVEST, P. SILVA, B. VIGNALE, D. CABRERA, P. HINRICHSEN, A. A. F. GARCIA, and C. PRITSCH, 2014 A first genetic map of *Acca sellowiana* based on ISSR, AFLP and SSR markers. *Scientia Horticulturae* **169**: 138–146.
- R CORE TEAM, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- RONIN, Y., D. MESTER, D. MINKOV, R. BELOTSEKOVSKI, B. N. JACKSON, P. S. SCHNABLE, S. ALURU, and A. KOROL, 2012 Two-Phase Analysis in Consensus Genetic Mapping. *G3: Genes, Genomes, Genetics* **2**: 537–549.
- RUSSELL, J., C. HACKETT, P. HEDLEY, H. LIU, L. M. BRENNAN, M. BAYER, D. MARSHALL, L. JORGENSEN, S. GORDON, and R. BRENNAN, 2014 The use of genotyping by sequencing in blackcurrant (*Ribes nigrum*): developing high-resolution linkage maps in species without reference genome sequences. *Molecular Breeding* **33**: 835–849.
- SALAZAR, J. A., I. PACHECO, P. SHINYA, P. ZAPATA, C. SILVA, M. ARADHYA, D. VELASCO, D. RUIZ, P. MARTÍNEZ-GÓMEZ, and R. INFANTE, 2017 Genotyping by Sequencing for SNP-Based Linkage Analysis and Identification of QTLs Linked to Fruit Quality Traits in Japanese Plum (*Prunus salicina* Lindl.). *Frontiers in Plant Science* **8**.
- SCHLAUTMAN, B., G. COVARRUBIAS-PAZARAN, L. DIAZ-GARCIA, M. IORIZZO, J. POLASHOCK, E. GRYGLESKI, N. VORSA, and J. ZALAPA, 2017 Construction of a High-Density American Cranberry (*Vaccinium macrocarpon* Ait.) Composite Map Using Genotyping-by-Sequencing for Multi-pedigree Linkage Mapping. *Genes|Genomes|Genetics* **7**: 1177–1189.
- SERANG, O., M. MOLLINARI, and A. A. F. GARCIA, 2012 Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS ONE* **7**: 1–13.
- SILVA-JUNIOR, O. B., D. A. FARIA, and D. GRATTAPAGLIA, 2015 A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytologist* **206**: 1527–1540.
- SONAH, H., M. BASTIEN, E. IQUIRA, A. TARDIVEL, G. LÉGARE, B. BOYLE, É. NORMANDEAU, J. LAROCHE, S. LAROSE, M. JEAN, and F. BELZILE, 2013 An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLoS ONE* **8**: e54603.
- SOUZA, L. M., R. GAZAFFI, C. C. MANTELLO, C. C. SILVA, D. GARCIA, V. LE GUEN, S. E. A. CARDOSO, A. A. F. GARCIA, and A. P. SOUZA, 2013 QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (*Hevea brasiliensis*) Evaluated in a Sub-Tropical Climate. *PLoS ONE* **8**: e61238.
- SPINDEL, J., M. WRIGHT, C. CHEN, J. COBB, J. GAGE, S. HARRINGTON, M. LORIEUX, N. AHMADI, and S. MCCOUCH, 2013 Bridging the genotyping gap: Using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theoretical and Applied Genetics* **126**: 2699–2716.

- SWARTS, K., H. LI, J. A. ROMERO NAVARRO, D. AN, M. C. ROMAY, S. HEARNE, C. ACHARYA, J. C. GLAUBITZ, S. MITCHELL, R. J. ELSHIRE, E. S. BUCKLER, and P. J. BRADBURY, 2014 Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome* **7**.
- TANG, H., J. E. BOWERS, X. WANG, R. MING, M. ALAM, and A. H. PATERSON, 2008 Synteny and Collinearity in Plant Genomes. *Science* **320**: 486–488.
- TANIGUTI, C. H., 2016 *Construção de mapa genético integrado em uma progênie de irmãos-completos proveniente do cruzamento entre Eucalyptus grandis e Eucalyptus urophylla*. Ph.D. thesis, Universidade de São Paulo, Escola Superior de Agricultura “Luiz de Queiroz”.
- TENNESSEN, J. A., R. GOVINDARAJULU, T. L. ASHMAN, and A. LISTON, 2014 Evolutionary origins and dynamics of octoploid strawberry subgenomes revealed by dense targeted capture linkage maps. *Genome Biology and Evolution* **6**: 3295–3313.
- THORP, G. and R. L. BIELESKI, 2002 *Feijoa: Origins, Cultivation and Uses*. David Bateman Limited, New Zealand.
- TORKAMANEH, D., J. LAROCHE, and F. BELZILE, 2016 Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. *PLoS ONE* **11**: e0161333.
- TROGGIO, M., A. GLEAVE, S. SALVI, D. CHAGNÉ, A. CESTARO, S. KUMAR, R. N. CROWHURST, and S. E. GARDINER, 2012 Apple, from genome to breeding. *Tree Genetics and Genomes* **8**: 509–529.
- VAIO, M., N. BAZ, M. QUEZADA, and C. PRITSCH, 2015 Guayabo de país: 40% de repetición! In *9as Jornadas de la Sociedad de Bioquímica y Biología Molecular*, p. 139.
- VAN OOIJEN, J. W., 2006 JoinMap 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations. Kyazma BV: Wageningen, The Netherlands .
- VARSHNEY, R. K., J. C. GLASZMANN, H. LEUNG, and J. M. RIBAUT, 2010 More genomic resources for less-studied crops. *Trends in Biotechnology* **28**: 452–460.
- VELASCO, R., A. ZHARKIKH, J. AFFOURTIT, A. DHINGRA, A. CESTARO, A. KALYANARAMAN, P. FONTANA, S. K. BHATNAGAR, M. TROGGIO, D. PRUSS, S. SALVI, M. PINDO, P. BALDI, S. CASTELLETTI, M. CAVAIUOLO, G. COPPOLA, F. COSTA, V. COVA, A. DAL RI, V. GOREMYKIN, M. KOMJANC, S. LONGHI, P. MAGNAGO, G. MALACARNE, M. MALNOY, D. MICHELETTI, M. MORETTO, M. PERAZZOLLI, A. SI-AMMOUR, S. VEZZULLI, E. ZINI, G. ELDREDGE, L. M. FITZGERALD, N. GUTIN, J. LANCHBURY, T. MACALMA, J. T. MITCHELL, J. REID, B. WARDELL, C. KODIRA, Z. CHEN, B. DESANY, F. NIAZI, M. PALMER, T. KOEPKE, D. JIWAN, S. SCHAEFFER, V. KRISHNAN, C. WU, V. T. CHU, S. T. KING, J. VICK, Q. TAO, A. MRAZ, A. STORMO, K. STORMO, R. BOGDEN, D. EDERLE, A. STELLA, A. VECCHIETTI, M. M. KATER, S. MASIERO, P. LASSERRE, Y. LESPINASSE, A. C. ALLAN, V. BUS, D. CHAGNÉ, R. N. CROWHURST, A. P. GLEAVE,

- E. LAVEZZO, J. A. FAWCETT, S. PROOST, P. ROUZÉ, L. STERCK, S. TOPPO, B. LAZZARI, R. P. HELLENS, C.-E. DUREL, A. GUTIN, R. E. BUMGARNER, S. E. GARDINER, M. SKOLNICK, M. EGHOLM, Y. VAN DE PEER, F. SALAMINI, and R. VIOLA, 2010 The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nature Genetics* **42**: 833–839.
- VERDE, I., N. BASSIL, S. SCALABRIN, B. GILMORE, C. T. LAWLEY, K. GASIC, D. MICHELETTI, U. R. ROSYARA, F. CATTONARO, E. VENDRAMIN, D. MAIN, V. ARAMINI, A. L. BLAS, T. C. MOCKLER, D. W. BRYANT, L. WILHELM, M. TROGGIO, B. SOSINSKI, M. J. ARANZANA, P. ARÚS, A. IEZZONI, M. MORGANTE, and C. PEACE, 2012 Development and evaluation of a 9k SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PLoS ONE* **7**: e35668.
- VEZZULLI, S., D. MICHELETTI, S. RIAZ, M. PINDO, R. VIOLA, P. THIS, M. A. WALKER, M. TROGGIO, and R. VELASCO, 2008 A SNP transferability survey within the genus *Vitis*. *BMC Plant Biology* **8**: 128.
- VOORRIPS, R. E., 2002 MapChart: software for the graphical presentation of linkage maps and QTLs. *The Journal of heredity* **93**: 77–78.
- VUOTTO, M. L., A. BASILE, V. MOSCATIELLO, P. DE SOLE, R. CASTALDO-COBIANCHI, E. LAGHI, and M. T. L. IELPO, 2000 Antimicrobial and antioxidant activities of Feijoa sellowiana fruit. *International Journal of Antimicrobial Agents* **13**: 197–201.
- WANG, C., C. ZHU, H. ZHAI, and J. WAN, 2005 Mapping segregation distortion loci and quantitative trait loci for spikelet sterility in rice (*Oryza sativa* L.). *Genetics Research* **86**: 97–106.
- WARD, J. A., J. BHANGOO, F. FERNÁNDEZ-FERNÁNDEZ, P. MOORE, J. SWANSON, R. VIOLA, R. VELASCO, N. BASSIL, C. A. WEBER, and D. J. SARGENT, 2013 Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. *BMC Genomics* **14**: 2.
- WESTON, R. J., 2010 Bioactive products from fruit of the feijoa (*Feijoa sellowiana*, Myrtaceae): A review. *Food Chemistry* **121**: 923–926.
- WU, R., C.-X. MA, I. PAINTER, and Z.-B. ZENG, 2002a Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. *Theoretical population biology* **61**: 349–363.
- WU, R., C.-X. MA, S. S. WU, and Z.-B. ZENG, 2002b Linkage mapping of sex-specific differences. *Genetics Research* **79**: 85–96.
- WU, Y., T. J. CLOSE, and S. LONARDI, 2011 Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**: 381–394.

APPENDICES

Appendix I

WU *et al.* (2002a) proposed a general maximum-likelihood-based algorithm for simultaneously estimating linkage and linkage phases in outcrossing species. In this approach, the maximum likelihood estimator of the recombination fraction for two markers is estimated in an iterative procedure based on Expectation and Maximization (EM) algorithms (LANDER and GREEN, 1987; DEMPSTER *et al.*, 1977). For outcrossing populations, the linkage phase between markers is commonly unknown. This approach is implemented in ONEMAP R package (MARGARIDO *et al.*, 2007).

In this study, the EM equations presented by WU *et al.* (2002a) were adapted to estimate the recombination fractions of two connected populations. Our model will be described using the same notation, as well as the matrix $\mathbf{H}$, $\mathbf{D}$ and $\mathbf{I}$, defined in the same article. The possible assignment of linkage phases between the two parental genotypes for two adjacent markers were denoted as $\mathbf{A}^\omega$, where $\omega = 1, \dots, 4$ (coupling $\times$ coupling, coupling $\times$ repulsion, repulsion $\times$ coupling, repulsion $\times$ repulsion, respectively).

The statistical model presented here assumed two connected F_1 mapping populations (denoted as N and N' of sample size j and j' , respectively) genotyped with the same marker set. First, the recombination fraction was estimated in the single-population using the approach of WU *et al.* (2002a). This was used as the initial value for the joint recombination fraction estimation using EM algorithm. The adapted equations are given below.

E step: The expectation step is independently computed on each population as follows:

- N population: at step τ , the matrix $\mathbf{H}_{k(k+1)}^{\omega(\tau)}$ based on the current recombination fraction ($\vartheta_{k(k+1)}^{\omega(\tau)}$) is used to calculate the expected number of recombination events between $\mathbf{M}_k$ and $\mathbf{M}_{k+1}$ for offspring j under the assignment $\mathbf{A}_{k+1}^\omega$,

$$c_{i_k i_{k+1} j}^{\omega\{\tau+1\}} = \frac{\mathbf{m}_{i_k j}^T [\mathbf{I}_{p_k}^T (\mathbf{D}_{k(k+1)}^\omega \circ \mathbf{H}_{k(k+1)}^{\omega\{\tau\}}) \mathbf{I}_{p_{k+1}}] \mathbf{m}_{i_{k+1} j}}{\mathbf{m}_{i_k j}^T \mathbf{P}_{k(k+1)}^{\omega\{\tau\}} \mathbf{m}_{i_{k+1} j}}$$

where

$\mathbf{m}_{i_k j}$, is the p_k -dimensional vector of the indicator variable for marker $\mathbf{M}_k$

$\mathbf{I}_{p_k}$, is a $(4 \times p_k)$ incidence matrix relating the parental chromosome pairings genotypes to phenotypes

$\mathbf{D}_{k(k+1)}^\omega$, is a (4×4) matrix of the number of recombination events between the two markers $\mathbf{M}_k$ and $\mathbf{M}_{k+1}$

$\mathbf{H}_{k(k+1)}^\omega$, is the (4×4) matrix of the transition probability of the parental chromosome pairings from markers $\mathbf{M}_k$ to $\mathbf{M}_{k+1}$

$\mathbf{P}_{k(k+1)}^\omega = \mathbf{I}_{p_k}^T \mathbf{H}_{k(k+1)}^\omega \mathbf{I}_{p_{k+1}}$, is a $(p_k \times p_{k+1})$ matrix of the transition probability from markers $\mathbf{M}_k$ to $\mathbf{M}_{k+1}$

- N' population: at step τ , the matrix $\hat{\mathbf{H}}_{k(k+1)}^{\hat{\omega}(\tau)}$ based on the current recombination fraction ($\hat{\vartheta}_{k(k+1)}^{\hat{\omega}(\tau)}$) was used to calculate the expected number of recombination events between M_k and M_{k+1} for offspring j' under the assignment $A_{k+1}^{\hat{\omega}}$,

$$\hat{c}_{i_k i_{k+1} j'}^{\hat{\omega}\{\tau+1\}} = \frac{\hat{\mathbf{m}}_{i_k j'}^T [\hat{\mathbf{I}}_{\hat{p}_k}^T (\hat{\mathbf{D}}_{k(k+1)}^{\hat{\omega}} \circ \hat{\mathbf{H}}_{k(k+1)}^{\hat{\omega}\{\tau\}}) \hat{\mathbf{I}}_{\hat{p}_{k+1}}] \hat{\mathbf{m}}_{i_{k+1} j'}}{\hat{\mathbf{m}}_{i_k j'}^T \hat{\mathbf{P}}_{k(k+1)}^{\hat{\omega}\{\tau\}} \hat{\mathbf{m}}_{i_{k+1} j'}}$$

where

$\hat{\mathbf{m}}_{i_k j'}$, is the $\hat{p}_k$ -dimensional vector of the indicator variable for marker M_k

$\hat{\mathbf{I}}_{\hat{p}_k}$, is a $(4 \times \hat{p}_k)$ incidence matrix relating the parental chromosome pairings genotypes to phenotypes

$\hat{\mathbf{D}}_{k(k+1)}^{\hat{\omega}}$, is a (4×4) matrix of the number of recombination events between the two markers M_k and M_{k+1}

$\hat{\mathbf{H}}_{k(k+1)}^{\hat{\omega}}$, is the (4×4) matrix of the transition probability of the parental chromosome pairings from markers M_k to M_{k+1}

$\hat{\mathbf{P}}_{k(k+1)}^{\hat{\omega}} = \hat{\mathbf{I}}_{\hat{p}_k}^T \hat{\mathbf{H}}_{k(k+1)}^{\hat{\omega}} \hat{\mathbf{I}}_{\hat{p}_{k+1}}$, is a $(\hat{p}_k \times \hat{p}_{k+1})$ matrix of the transition probability from markers M_k to M_{k+1}

M step: The maximization step is computed jointly for both populations. A single recombination value is estimated using the information of both populations simultaneously as follow:

$$\vartheta_{k(k+1)}^{\omega\{\tau+1\}} = \hat{\vartheta}_{k(k+1)}^{\hat{\omega}\{\tau+1\}} = \frac{1}{2N} \sum_{j=1}^N \sum_{i_k=1}^{p_k} \sum_{i_{k+1}=1}^{p_{k+1}} c_{i_k i_{k+1} j}^{\omega\{\tau+1\}} + \frac{1}{2N'} \sum_{j'=1}^{N'} \sum_{i_k=1}^{\hat{p}_k} \sum_{i_{k+1}=1}^{\hat{p}_{k+1}} \hat{c}_{i_k i_{k+1} j'}^{\hat{\omega}\{\tau+1\}}$$

The iterative procedure is repeated between the E and M step until the recombination fraction value converged to a stable value.

Appendix II

Single-population parental maps for H5 and H6 populations constructed using the software ONEMAP R package (MARGARIDO *et al.*, 2007). Linkages groups are showed as heatmap plots, containing the pairwise recombination fractions and LOD scores.

Appendix III

Comparison between genetic maps developed for the H5 ('TCO $\times$ BR') F_1 full-sib mapping population. Linkage groups of the single population H5 map constructed in this study were aligned with linkage groups of the previous map developed for the same population (QUEZADA *et al.*, 2014)

Figure A.1: Genetic linkage map of the H5 ('TCO × BR') F_1 full-sib mapping population. Heatmaps plots show the pairwise recombination fractions below the diagonal and LOD scores above the diagonal for all marker combinations in each linkage group. The color scale ranges from red (small recombination fraction or large LOD) to dark blue. White cells indicate markers which recombination fraction cannot be estimated with two-point analysis ('D1' and 'D2').

Figure A.2: Genetic linkage map of the H6 ('TCO × DP') F_1 full-sib mapping population. Heatmaps plots show the pairwise recombination fractions below the diagonal and LOD scores above the diagonal for all marker combinations in each linkage group. The color scale ranges from red (small recombination fraction or large LOD) to dark blue. White cells indicate markers which recombination fraction cannot be estimated with two-point analysis ('D1' and 'D2').

Figure A.3: Comparison of homologous linkage groups from H5 genetic maps. The H5 map constructed in this study is displayed in black and the former H5 genetic map developed by QUEZADA *et al.* (2014) is displayed in blue. Map distances (cM) are indicated at the left side of the figure. Marker names are shown for each linkage group. Common markers (green) between the two maps are connected with lines. The marker GGAGA3.7 with a conflict position between maps is shown in red.

Figure A.3: Continued

Figure A.3: Continued