University of São Paulo
"Luiz de Queiroz" College of Agriculture

Evaluation of software to construct genetic linkage maps in plants

Jenifer Camila Godoy dos Santos

Dissertation presented to obtain the degree of Master in Science. Area: Genetics and Plant Breeding

Piracicaba
2019

Jenifer Camila Godoy dos Santos
Agronomy Engineer

Evaluation of software to construct genetic linkage maps in plants

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

Dissertation presented to obtain the degree of Master in Science. Area: Genetics and Plant Breeding

Piracicaba
2019

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

Santos, Jenifer Camila Godoy dos
Evaluation of software to construct genetic linkage maps in plants/ Jenifer Camila Godoy dos Santos. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2019 .
46 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Construção de mapas genéticos 2. Pacotes computacionais 3. Espécies diploides 4. OneMap 5. Lep-MAP3 I. Título.

DEDICATORY

To my parents Juscélia and Roberto (*in memoriam*)
To my lovely cousin Amanda,
My angel Bernardo.
I love you.

ACKNOWLEDGMENTS

I'd like to thank God for always being by my side and all those who have supported me to begin this scientific journey:

- My advisor Antonio Augusto Franco Garcia who taught me what science is.
- All the friends of the Statistical-Genetics Lab at ESALQ/USP, specially, Cristiane Taniguti, Gabriel Gesteira, Getúlio Ferreira and Letícia Lara.
- The Department of Genetics at ESALQ/USP.
- The professors of the courses Genetics and Plant Breeding and Agronomic Experimentation at ESALQ/USP.
- My graduation advisor Larissa Barbosa.
- My friends Ana Letycia Basso, Camila Haddad, Fernanda Pazin, Igor Araújo, Nathália Salgado, Victor Pessoa and Walleska Torsian.
- My boyfriend Maycon Fernandes.
- My family Douglas, Laisa, Juscélio and Roberta.
- My parents Juscélia and Roberto (*in memoriam*).

The Lord is my shepherd; I shall not want.

Psalm 23:1

SUMMARY

RESUMO

Avaliação de *softwares* para construção de mapas genéticos em plantas

Mapas genéticos são ferramentas úteis em programas de melhoramento e em estudos evolutivos. Desde a publicação do primeiro mapa, vários conceitos foram propostos e implementados em vários *softwares* de mapeamento. Cada *software* apresenta diferentes características para construção de mapas. Por exemplo, aspectos como disponibilidade do código fonte, licenças, tutoriais e sistema operacional precisam ser considerados. Há ainda pontos importantes relacionados aos métodos estatísticos empregados. Assim sendo, nem sempre a escolha pelos usuários é uma tarefa simples. Os objetivos aqui foram: i) apresentar os principais *software* com licenças gratuitas desenvolvidos nos últimos anos; ii) construir mapas de ligação utilizando esses programas e iii) avaliá-los do ponto de vista dos usuários. Os softwares considerados foram: OneMap, Lep-MAP, HighMap, Lep-MAP2, Flipper, Lep-MAP3, ASMap e GUSMap. Este trabalho poderá orientar os pesquisadores quanto às ferramentas gratuitas disponíveis para construção de mapas genéticos.

Palavras-chave: Construção de mapas genéticos; Pacotes computacionais; Espécies diploides; OneMap; Lep-MAP3

ABSTRACT

Evaluation of software to construct genetic linkage maps in plants

Genetic maps are useful tools in breeding programs and evolutionary studies. Since the publication of the first map, several concepts have been proposed and implemented in various mapping software. Each software presents different characteristics for the construction of maps. For example, aspects such as availability of source code, licenses, tutorials, and operating system need to be considered. There are also important points related to the statistical methods employed. In this context, the users' choice can often be a complicated task. The objectives here were: i) to present the main software with free licenses developed in recent years; ii) construct linkage maps using these software and, iii) evaluate them from the point of view of users. The software considered were: OneMap, Lep-MAP, HighMap, Lep-MAP2, Flipper, Lep-MAP3, ASMap, and GUSMap. This work can guide researchers about the free tools available to construct genetic maps.

Keywords: Construction of genetic maps; Computer packages; Diploid species; OneMap; Lep-MAP3

1 INTRODUCTION

The first genetic linkage map was constructed by A. H. Sturtevant. He, who was still an undergraduate student, suggested to use the frequency of recombinants as a quantitative indicator of the linear distance between two genes (Sturtevant, 1913). From this idea many methods have been developed, and consequently, genetic mapping has become a powerful tool for genomic studies in other organisms, including several plant species (Edwards, 2005). Evolutionary studies, genome assembly, and quantitative trait loci (QTL) mapping are some studies which benefit from having a good genetic map estimate (Lefebvre *et al.*, 1995; Harushima *et al.*, 1998; Oliveira *et al.*, 2007, 2008; Sierro *et al.*, 2014; Fierst, 2015; Zhigunov *et al.*, 2017; Fenton *et al.*, 2018).

Currently, genetic linkage maps are estimated using several mapping software (Cheema and Dicks, 2009). In general, these software have different statistical methods for grouping and ordering molecular markers, they can cope with differing types of mapping populations and several data sets of different sizes, and the executables are provided for different operating systems. In addition, they have several friendly approaches, such as tutorials and graphs, which help users during the construction of maps (Margarido *et al.*, 2007; Rastas *et al.*, 2013; Liu *et al.*, 2014; Rastas *et al.*, 2015; Crane, 2017; Rastas, 2017; Taylor and Butler, 2017; Bilton *et al.*, 2018).

In relation to licenses and availability of source code, mapping software can be divided into three groups: i) free software, ii) open source software, and iii) proprietary software. Free software is a social movement that strives for a balance between the developer and the user (freedom of the user). Software will be considered free if users have the freedom to run the program (freedom 0), to study and change the program (freedom 1), to redistribute exact copies (freedom 2), and to distribute modified versions (freedom 3). In order for freedoms 1 and 3 to be meaningful, the source code must be opened. Open source software is another similar movement but not identical to the free software movement. This movement strives to develop technically superior software tools through the collaborative improvement of source code (software quality). Proprietary software is a non-free software in which the developer or company retains intellectual copyright of the source code (source code is closed). It is important to mention that the three groups have free and paid licenses (Crowston and Howison, 2005).

These different characteristics can make it difficult for users to choose among them. Thus, the present study can guide them in the tools available for the construction of genetic linkage maps.

The work consists of one chapter organized into two parts. In the first, the aim is to present a description considering several aspects of some software with free licences developed in recent years. In the second, the aim is to construct linkage maps using some of these mapping software. For this purpose, several scenarios with different numbers and types of markers were simulated and based on these simulations the genetic maps were estimated.

References

Bilton, T. P., M. R. Schofield, M. A. Black, D. Chagné, P. L. Wilcox, and K. G.

Dodds, 2018 Accounting for errors in low coverage high-throughput sequencing data when constructing genetic maps using biparental outcrossed populations. Genetics 209: 65–76.

Cheema, J. and J. Dicks, 2009 Computational approaches and software tools for genetic linkage map estimation in plants. Briefings in bioinformatics **10**: 595–608.

Crane, C. F., 2017 P0092 mapping 1000000 markers with flipper. In *Plant and Animal Genome XXVI Conference (PAG)*, pp. https://plan.core–apps.com/pag_2018/abstract/fff6e3855de940b33de03bc4eff5c2b4, Plant and Animal Genome XXVI Conference (PAG).

Crowston, K. and J. Howison, 2005 The social structure of free and open source software development. First Monday **10**.

Edwards, A. W. F., 2005 Linkage methods in human genetics before the computer. Human genetics **1922**: 515–530.

Fenton, M. E., B. F. Owens, A. E. Lipka, D. Ortiz, T. Tiede, M. Mateos-hernandez, M. G. Ferruzzi, and T. Rocheford, 2018 High-density linkage mapping of vitamin E content in maize grain. Mol Breeding **38**: 31.

Fierst, J. L., 2015 Using linkage maps to correct and scaffold de novo genome assemblies: Methods, challenges, and computational tools.

Harushima, Y., M. Yano, A. Shomura, M. Sato, T. Shimano, Y. Kuboki, T. Yamamoto, S. Y. Lin, B. A. Antonio, A. Parco, *et al.*, 1998 A high-density rice genetic linkage map with 2275 markers using a single f2 population. Genetics **148**: 479–494.

Lefebvre, V., A. Palloix, C. Caranta, and E. Pochard, 1995 Construction of an intraspecific integrated linkage map of pepper using molecular markers and doubled-haploid progenies. Genome **38**: 112–121.

Liu, D., W. Hong, L. Huang, M. Liu, H. Liu, H. Zeng, D. Deng, H. Xin, J. Song, C. Xu, X. Sun, X. Hou, X. Wang, and H. Zheng, 2014 Construction and Analysis of High-Density Linkage Map Using High-Throughput Sequencing Data. Plos one **9**: e98855.

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.

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, U. D. Sa, M. I. Zucchi, M. H. P. Fungaro, and C. D. Cie, 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 Science **133**: 35–41.

Oliveira, K. M., L. R. Pinto, T. G. Marconi, G. R. A. Margarido, M. M. Pastina, L. H. M. Teixeira, A. F. Garcia, and A. P. Souza, 2007 Functional integrated genetic linkage map based on EST- markers for a sugarcane (*Saccharum* spp.) commercial cross. Mol Breeding **20**: 189–208.

Rastas, P., 2017 Lep-map3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics **33**: 3726–3732.

Rastas, P., F. C. Calboli, B. Guo, T. Shikano, and J. Merilä, 2015 Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example. Genome biology and evolution .

Rastas, P., L. Paulin, I. Hanski, R. Lehtonen, and P. Auvinen, 2013 Lep-map: fast and accurate linkage map construction for large snp datasets. Bioinformatics **29**: 3128–3134.

Sierro, N., J. N. D. Battey, S. Ouadi, N. Bakaher, L. Bovet, A. Willig, S. Goepfert, M. C. Peitsch, and N. V. Ivanov, 2014 The tobacco genome sequence and its comparison with those of tomato and potato. Nature Communications **5**: 1–9.

Sturtevant, A. H., 1913 The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. Journal of experimental zoology **14**: 43–59.

Taylor, J. and D. Butler, 2017 R package asmap: efficient genetic linkage map construction and diagnosis. arXiv preprint arXiv:1705.06916 .

Zhigunov, A. V., P. S. Ulianich, M. V. Lebedeva, P. L. Chang, S. V. Nuzhdin, and E. K. Potokina, 2017 Development of f1 hybrid population and the high-density linkage map for european aspen (*Populus tremula* l.) using radseq technology. BMC plant biology **17**: 180.

2 EVALUATION OF SOFTWARE TO CONSTRUCT GENETIC LINKAGE MAPS IN PLANTS

2.1 Abstract

In recent years, many concepts and methods have been developed to aid the construction of genetic linkage maps. These methods were implemented in a number of software. There are some differences between these software, such as the statistical method employed, the operating system, the availability of the source code, the type of license and friendly approaches. Therefore, the users' choice can often be a complicated task. The objective of this work was to present the software tools OneMap, Lep-MAP, Lep-MAP2, HighMap, Flipper, Lep-MAP3, ASMap and GUSMap, discussing many of the main concepts involved, and to construct linkage maps using OneMap and Lep-MAP3. For doing so, several scenarios were simulated considering a diploid plant species. Although these software have particularities, many share the same algorithmic approaches. Results showed that the software OneMap and Lep-MAP3 had efficient performance in all simulated scenarios. OneMap stood out for the help tools available to construct the maps, while Lep-MAP3 stood out for the speed in the ordering of the markers.

2.2 Introduction

Genetic linkage maps represent the order and distance of the genes or markers along the chromosome. They are useful in evolutionary genetic studies (Gaur *et al.*, 2015; Butler *et al.*, 2017; Wu *et al.*, 2019) and in genome assembly and validation (Fierst, 2015). Furthermore, maps are very important for quantitative trait loci (QTL) mapping. QTLs with strong effects on phenotypic variation have been discovered in several species and, in many cases, were successfully exploited in breeding programs via marker-assisted selection (MAS) (Xu *et al.*, 2000; Sibov *et al.*, 2003; Steele *et al.*, 2006; Carter *et al.*, 2009; Sun *et al.*, 2012; Luo *et al.*, 2017; Ren *et al.*, 2019).

Given the importance of the genetic linkage maps, several mapping software are available. The review of the Cheema and Dicks (2009) makes a brief description of 11 software tools for genetic linkage map estimation, such as MAPMAKER/EXP (Lander and Green, 1987), JoinMap (Stam, 1993), and MSTmap (Wu *et al.*, 2008). However, throughout these 10 years and even before this review, other software have been proposed, including OneMap (Margarido *et al.*, 2007), Lep-MAP (Rastas *et al.*, 2013), HighMap (Liu *et al.*, 2014); Lep-MAP2 (Rastas *et al.*, 2015), ASMap (Taylor and Butler, 2017), Flipper (Crane, 2017), Lep-MAP3 (Rastas, 2017), and GUSMap (Bilton *et al.*, 2018). They have been developed with the purpose of trying to solve problems typical of the construction of maps. For example, the OneMap was the first software capable to estimate integrated maps in outcrossing populations. Lep-MAP, HighMap, Lep-MAP2, and ASMap were developed to construct maps using large data sets. While LepMap3 and GUSMap were developed to construct maps using low coverage and depth sequencing data, respectively.

Beyond these specific characteristics, the software tools have different statistical methods to group and order the molecular markers and they can analyze different data sets. There

are also important points such as licenses, availability of source code, operating system, and friendly approaches that need to be considered (Cheema and Dicks, 2009).

For grouping, software such as MAPMAKER/EXP and OneMap use the nearest neighbour locus strategy (Cheema and Dicks, 2009), while MSTmap and AsMap use probabilistic graphical models (Wu *et al.*, 2008). For ordering, several criteria and algorithms are available and implemented in the mapping software, including the product of adjacent recombination fractions (PARF) (Wilson and Rao, 1988), the sum of adjacent recombination fractions (SARF) (Falk, 1989), the sum of adjacent LOD scores (SALOD) (Weeks and Lange, 1987), hidden Markov model (HMM) (Lander and Green, 1987), seriation (SER) (Buetow, Kenneth H. and Chakravarti., 1987), rapid chain delineation (RCD) (Doerge and Weir, 1994), and multidimensional scaling (MDS) (Preedy and Hackett, 2016).

In general, the software tools can cope with inbred-based populations and outcrossing populations or be restricted to one of two types of population. In addition, they able to analyze various types of markers with different mixing patterns (dominant and codominant) (Collard *et al.*, 2005; Edwards, 2005; Cheema and Dicks, 2009).

In relation to licenses and availability of source code, they can be classified as free software, open source software, and proprietary software. Free software ensures the freedom of the user while open source software ensures the quality of the program, both of which have open source codes. In contrast proprietary software has closed source code. It is important to say that all three groups have free and paid licenses and the the executables are provided for different operating systems such as Mac, Linux, and Windows (Crowston and Howison, 2005).

Finally, these software may contain several tools such as tutorials, graphics, heatmaps, and research forum that will help users during the construction of maps. These tools are user-friendly approaches and are very important because even if a program has many advanced features, users have to be able to use them (Margarido *et al.*, 2007; Rastas *et al.*, 2013; Liu *et al.*, 2014; Rastas *et al.*, 2015; Crane, 2017; Rastas, 2017; Taylor and Butler, 2017; Bilton *et al.*, 2018).

2.2.1 Genetic Linkage Maps

After the rediscovery of Mendel's work several researchers have observed that some genes did not segregate independently. Morgan found similar results while studying some genes in *Drosophila melanogaster*. He realized that the frequency of recombinants progenies varied considerably and suggested that the two genes were located on the same pair of homologous chromosomes (Morgan and Cattell, 1912). Sturtevant (1913) associated this frequency of recombination between the loci with the linear distance between them, giving origin to genetic linkage maps.

The first genetic linkage maps in plants were constructed with morphological and cytological markers (Anderson and Randolph, 1945; Butler, 1952; Dhulappanavar, 1977). Latter, the isoenzymes were discovered (Markert and Møller, 1959) and began to be explored as biochemical markers. However, the number of isoenzymatic loci detected was still small, especially when it was intended to obtain of higher genome coverage.

With the advent of PCR (Polymerase Chain Reaction) and restriction enzymes, markers more abundant and based on polymorphic DNA sequence (molecular marker) began to be used for the construction of genetic linkage maps. These markers can be dominant, such as RAPDs (Randomly Amplified Polymorphic DNA) (Williams *et al.*, 1990), AFLP (Amplified Fragment Length Polymorphism) (Vos *et al.*, 1995), and DArT (Jaccoud *et al.*, 2001), or they can be codominant, such as RFLPs (Restriction Fragment Length Polymorphism) (Botstein, 1980), SSRs (Simple Sequence Repeats) (Tautz, 1989), EST-SSRs (Expressed Sequence Tags containing SSRs) (Cato *et al.*, 2001), and SNPs (Single Nucleotide Polymorphism) (Syvänen, 2001).

SNPs are abundant in plant genomes and their utility as genetic markers has been well established in the last decade. Recently, with next-generation sequencing (NGS) technology, thousands of SNPs are available allowing for the construction of high-density and high-resolution genetic linkage maps in several plant species (Ward *et al.*, 2013; Rabbi *et al.*, 2014; Ott *et al.*, 2015; Pootakham *et al.*, 2015; Peng *et al.*, 2016; Geyer, 2018; Somta *et al.*, 2019).

2.2.2 Applications of the Linkage Genetic Maps

Genetic linkage maps are important in many types of genetic studies because they provide the sequential distribution of the genes along the chromosomes of a species. In evolutionary genetics, they allow the identification of synteny, rearrangements, duplications, and deletions of chromosomal segments among species, such as cotton and *Arabidopsis* (Rong *et al.*, 2005), coffee and tomato (Lefebvre-Pautigny *et al.*, 2010), banana and monocotyledonous plants (D'hont *et al.*, 2012), tobacco and potato (Sierro *et al.*, 2014), cranberry and diploid blueberry (Schlautman *et al.*, 2018).

For species whose genome is not yet sequenced, genetic linkage maps provide a powerful tool to understand the grouping and order of markers through the comparison of genomic sequences to other plant species. For those species that have already been sequenced, genetic linkage maps allow assembly and validation these genomes (Fierst, 2015).

A very important application of genetic linkage maps is the localization of genes that control important agronomic traits such as yield, height of the plant, protein content, and some forms of disease resistance (Zhang *et al.*, 2004; Chutimanitsakun *et al.*, 2011; Cui *et al.*, 2014; Ma *et al.*, 2017). These traits are controlled by many genes and are known as quantitative traits and the genome regions that contribute to variation of quantitative traits are known as quantitative trait loci (QTL). The first QTL identification was made by Sax (1923) in a cross between bean inbred lines using morphological markers. Subsequently, with the development of molecular markers, the number of associations between these markers and quantitative traits has significantly increased.

When comparing the segregation of markers with phenotypic mean values for each individual progeny, it is possible to estimate the number, position, effects, interactions among the QTL, and interactions between QTL and the environments. This approach has enabled the identification of QTL in several species, such as upland cotton (Zhang *et al.*, 2009), tropical maize (Sibov *et al.*, 2003), rubber tree (Souza *et al.*, 2013), wheat (Cui *et al.*, 2014), mayze (Azevedo *et al.*, 2015), millet (Ambawat *et al.*, 2016), sugarcane (Balsalobre *et al.*, 2017),

and rice (Bhatia *et al.*, 2018). QTL mapping is an effective tool in breeding programs, because with this information breeders can define strategies and obtain superior genotypes via marker assisted selection (MAS).

Another promising technique using QTL mapping is the study of the quantitative traits correlation that can be caused by pleiotropy or linkage. This information is also important for the definition of breeding strategies, as the correlation caused by linkage is transient and the correlation caused by pleiotropy will remain in the next generations (Falconer and Mackay, 1996).

2.2.3 Construction of the Genetic Linkage Maps

The main steps to construct linkage maps are: (i) identification of polymorphism, (ii) segregation tests, (iii) linkage analysis of markers and assigning markers to linkage groups, (iv) ordering of the markers within linkage groups, and (v) estimation of genetic distances (Collard *et al.*, 2005; Cheema and Dicks, 2009).

Most mapping populations are obtained from the segregation of plants derived from crosses between inbred lines, such as F_2, backcross (BC_1 and BC_2), RILs (recombinant inbred lines), and DH (doubled haploid) (Tanksley, 1993; Lynch *et al.*, 1998). A number of genome projects with inbred lines are available and have been instrumental in unraveling genetic variation in economical traits (Akond *et al.*, 2013; Lee *et al.*, 2015; Li *et al.*, 2017). There is, however, a large group of important species that are outbreeding (*i.e.* map is done in a progeny of a cross between two non-inbred plants). In these case, it is possible to evaluate segregation in full-sib family or half-sib family (Grattapaglia and Sederoff, 1994; Yin *et al.*, 2003; Bartholomé *et al.*, 2015; Pierro *et al.*, 2016). Regardless of the population used, polymorphisms in the genitors are necessary and the loci must be in linkage disequilibrium.

The identification of polymorphisms is done through molecular markers, and several types of molecular markers, as mentioned above, can be used. They are identified and screened across the entire mapping population, including the parents. Between them, some markers that segregate differently than expected by Mendel's law can be found. It is recommended not to use these markers to build the genetic linkage map, because generally, the statistical models for the map building process follows Mendel's law as an assumption. However, they can be used to saturate the map (Vogl and Xu, 2000).

Linkage analysis of molecular markers is based on the recombination fraction and odds ratios expressed in logarithm (logarithm of odds - LOD) (Fisher, 1922; Morton, 1955; Staub *et al.*, 1996). The recombination fraction is calculated between all pairs of markers (two-point tests) and values less of 0.5 are used to declare the linkage. The LOD score is the statistic used to evaluate the significance of the test and its value takes into consideration the number of tests performed that depend of the number of markers. For example, a LOD score of 3 between two markers, indicates that linkage is 1000 times more likely than no linkage (Morton, 1955).

Based on the values of recombination fraction and LOD, markers are assigned to linkage groups, which represent chromosomal segments or entire chromosomes. For the grouping several types of solutions have been proposed such as nearest neighbour locus and probabilistic graphical models. In nearest neighbour locus, the linkage groups are formed by sequentially adding the

marker that has the lowest recombination fraction in comparison to the others in the cluster. In probabilistic graphical models, a graph is structured in which the nodes are represented by the markers. These markers are connected to each other by weighted edges as estimated by two of recombination fraction. Subsequently, all edges with a recombination fraction above a certain threshold are eliminated, leaving only the subgraphs that correspond to the linkage groups (Cheema and Dicks, 2009).

In inbreeding species, all loci segregate to two alleles and all alleles from the same parent are in the coupling phase in the F_1, therefore, linkage analysis is straightforward (Lynch *et al.*, 1998). Contrarily, a cross between heterozygous parents may present 18 patterns of segregation (there may be up to four different alleles per locus) and the linkage phases are usually unknown and must be estimated (Table A.1). Initially, methods for determining linkage phases have been proposed but these analyses have not simultaneously estimated linkage and linkage phases for a mixed data set of markers (Ritter *et al.*, 1990; Arús *et al.*, 1994; Ritter and Salamini, 1996; Maliepaard *et al.*, 1997; Ridout *et al.*, 1998). Later, Wu (2002) developed a general maximum likelihood algorithm that simultaneously estimates the recombination fractions and linkage phases of all patterns of segregation.

The final step of the construction of a genetic linkage map involves the estimation of the markers' order and distance within linkage groups. The ordering process of the markers is considered an optimization problem (Traveling Salesman Problem - TSP) which seeks to find the best solution among all possible solutions (Mester *et al.*, 2003; Tan and Fu, 2006). For m markers, there are $m!/2$ possible orders to be investigated and in TSP it is impossible to obtain an exact solution when the number of markers is large. Thus, to solve the ordering problem there are two aspects: the choice of a criteria to optimize and the choice of an optimization algorithm.

The optimization criteria is used to quantify a given marker order in light of the data. There are several criteria to evaluate and compare the orders. Some are based on the recombination fraction between pairs of markers (two-point), such as the product of adjacent recombination fractions (PARF) (Wilson and Rao, 1988), the sum of adjacent recombination fractions (SARF) (Falk, 1989), the sum of adjacent LOD scores (SALOD) (Weeks and Lange, 1987), the SALOD-polymorphism information content (SALPIC) (Botstein, 1980; Olson and Boehnke, 1990), the least squares (Weeks and Lange, 1987), and the weighted least squares (Stam, 1993). Another criterion that can be used after the map's construction is the hidden Markov model (HMM) (Lander and Green, 1987). Through HMM, there is possible to obtain the maximum likelihood estimator (MLE) of the values using information from all the markers simultaneously (multipoint). These multipoint estimates can be obtained by the EM (expectation and maximization) algorithm (Dempster *et al.*, 1977). Good mapping software implements this multipoint approach because it allows for the retrieval of information from the entire linkage map, improving estimates mostly when there are dominant markers and missing data.

Several optimization algorithms have also been proposed, including simulated annealing (Thompson, 1984; Weeks and Lange, 1987), stepwise likelihood (Lathrop *et al.*, 1985), branch and bound (Lathrop *et al.*, 1985), try and ripple (Lander and Green, 1987), seri-

ation (SER) (Buetow, Kenneth H. and Chakravarti., 1987), rapid chain delineation (RCD) (Doerge and Weir, 1994), genetic and evolutionary algorithm (Mester *et al.*, 2003), recombination counting and ordering (RECORD) (Van Os *et al.*, 2005), and unidirectional growth (UG) (Tan and Fu, 2006). Recently, Preedy and Hackett (2016) proposed and validated a robust and fast ordering method for constructing high-density linkage maps using multidimensional scaling (MDS). In the face of such a wide variety of methodologies, some comparative research has been done (Olson and Boehnke, 1990; Hackett and Broadfoot, 2003; Mollinari *et al.*, 2009).

The distance in the map is measured in centiMorgans (cM), which is a unit that describes a in percentage (Collard *et al.*, 2005; Cheema and Dicks, 2009). The recombination fraction is not additive and, therefore, mapping functions are required. Several functions have been proposed but the most used are the mapping functions of Haldane and Kosambi. The first assumes no interference between crossover events (Haldane, 1919) and the second assumes interference between crossover events (Kosambi, 1944).

2.2.4 Software Tools Used to Construct Genetic Linkage Maps

With the widespread use of computers, these methodologies to construct linkage maps have been implemented into several genetic mapping software. The review of the Cheema and Dicks (2009) makes a brief description of 11 software tools, but throughout these 10 years and even before this review, other software were built and will be explained in more detail below:

- **OneMap**

  OneMap was the first software developed to simultaneously estimate the linkage and linkage phases (integrated genetic map) between markers in outcrossing species (Margarido *et al.*, 2007). This software analyzes different marker types containing several segregation patterns in outcrossing species (F_1) (Wu, 2002). In addition, it is capable of analyzing populations that have homozygous parental lines in the genealogy, such as F_2, backcross, and RILs. For marker grouping, there are functions to perform the basic two-point analysis between markers via the EM algorithm (Dempster *et al.*, 1977). Marker ordering in linkage groups is done using several two-point algorithms, such as SER (Buetow, Kenneth H. and Chakravarti., 1987), RCD (Doerge and Weir, 1994), RECORD (Van Os *et al.*, 2005), UG (Tan and Fu, 2006), and MDS (Preedy and Hackett, 2016). To finalize the map, the multipoint analysis can be used to check the local order of markers and also to refine the map distance between adjacent markers.

- **Lep-MAP and Lep-MAP2**

  Lep-MAP (Rastas *et al.*, 2013) was developed to construct genetic linkage maps using large data sets in a fast and accurate way. It is memory efficient and it can analyze multiple outcrossing and inbred based families. The grouping and ordering of markers is done using the SeparateChromosomes and OrderMarkers modules, respectively. The OrderMarkers module combines approximate solutions of TSP and multipoint analysis (HMM and EM) (Dempster *et al.*, 1977; Lander and Green, 1987). Although it is capable of creating linkage maps using large data sets, above 2,000 markers per chromosome, the

computational time required for mapping is unfeasible. Hence, a substantially improved version of Lep-MAP (Lep-MAP2) was built (Rastas *et al.*, 2015). Lep-MAP2 is based on Lep-MAP with the following improvements: it can handle all types of genetic marker data and it is capable of creating ultra-high-density genetic linkage maps.

- **HighMap**

  HighMap was developed to construct high-density and high-quality maps from NGS data in outcrossing populations (Liu *et al.*, 2014). Here, the map is built in three steps: grouping, ordering, and genotyping correction. In grouping, recombination fractions and LOD scores are calculated by a two-point analysis and the linkage phases are inferred using Wu (2002), as in OneMap. The ordering and error genotyping corrections are done through an iterative process. This process increases order accuracy and allows for the estimation of map distances in the presence of missing data and genotyping errors. A combination of statistical techniques, such as the SARF criterion (Falk, 1989), the annealing algorithm (Thompson, 1984; Weeks and Lange, 1987), and the enhanced algorithm of Gibbs sampling (Geman and Geman, 1987) are used to obtain the order and genetic distances. The Gibbs sampler is a Markov Chain Monte Carlo algorithm and, in HighMap, it is used in the multipoint analysis. The error correction strategy of SMOOTH (van Os *et al.*, 2005) and the k-nearest neighbor algorithm are used to correct genotyping errors and impute missing genotypes.

- **ASMap**

  ASMap (Taylor and Butler, 2017) is restricted to linkage map construction in F_2, backcross, DH, and RILs. It uses the efficient MSTmap algorithm (Wu *et al.*, 2008) to group and order large sets of markers. In grouping, MSTmap uses the minimum spanning tree of a graph (Cheriton and Tarjan, 1976). In ordering, several strategies such as K-opt, node-relocation, block-optimization (Wu *et al.*, 2008), and EM algorithm (Dempster *et al.*, 1977) are used.

- **Flipper**

  Flipper (Crane, 2017) was presented at the Plant and Animal Genome XXVI (PAG) Conference. According to the author, Kruskal's algorithm (Kruskal, 1956) was implemented in the software to produce a minimal spanning tree of genetic markers. Furthermore, additional heuristics allow for the correction of misplaced markers and has a rudimentary QTL finder.

- **Lep-MAP3**

  Lep-MAP3 (Rastas, 2017) constructs maps with low coverage sequencing data. It also reduces mapping errors by modelling recombination interference and is capable of creating maps using large data sets. It is memory efficient and it can simultaneously use data on multiple full-sib families, F_2, backcross, DH, and RILs. It is similar to earlier versions (Lep-MAP and Lep-MAP2) but there are some differences. The main difference is that the input genotype likelihoods are used in each step of the map's construction. The main

modules of the Lep-MAP3 workflow are: ParentCall2, Filtering2, SeparateChromosomes2, JoinSingles2All, and OrderMarkers2. Grouping and ordering is done using the SeparateChromosomes2 and OrderMarkers2 modules. The SeparateChromosomes2 uses the method of maximum likelihood to estimate the recombination fraction between markers. The marker ordering is an interactive process and is up to $450x - 2000x$ faster than previous versions. The OrderMarkers2 uses HMM (Lander and Green, 1987), ORDER, MERGE, and POLISH algorithms.

- **GUSMap**

  Sequencing data with low depth can lead to two kinds of missing data. The first is a missing genotype when no alleles are found and the second consists of a heterozygous classified as homozygous when only one of the parental allele is sequenced at a specific locus (Dodds *et al.*, 2015; Fragoso *et al.*, 2016). The latter is considered as a genotyping error. This error increases the frequency of inferred recombinants and consequently results in inflated linkage maps (Lincoln and Lander, 1992; Cartwright *et al.*, 2007; Cheema and Dicks, 2009). Several algorithms have been developed for imputing missing genotypes and correcting erroneous genotypes (Spindel *et al.*, 2013; Huang *et al.*, 2014; Fragoso *et al.*, 2016). However, these algorithms have been developed for inbreeding populations and are not applicable to outcrossing populations (full-sib family). Recently, Bilton *et al.* (2018) developed a new statistical method based on HMM (Lander and Green, 1987) for modeling low depth sequencing data in full-sib families in diploid species. This methodology is implemented in GUSMap (Genotyping Uncertainty with Sequencing data and linkage Mapping) software. Linkage groups are formed based on two-point recombination fractions and LOD scores from the EM algorithm (Dempster *et al.*, 1977). Marker ordering is performed using multidimensional scaling (MDS) approach (Preedy and Hackett, 2016).

2.2.5 Objectives

Considering the importance of linkage maps and consequently of the mapping software for the study of genetics, the main objectives here were: i) to provide useful information about software tools with free licences developed in recent years, trying to help users to decide which one to use and ii) to construct genetic linkage maps in different scenarios using some of these software.

2.3 Material and Methods

2.3.1 Software Tools Description

The software with free licences considered were OneMap (Margarido *et al.*, 2007), Lep-MAP (Rastas *et al.*, 2013), HighMap (Liu *et al.*, 2014), Lep-MAP2 (Rastas *et al.*, 2015), ASMap (Taylor and Butler, 2017), Flipper (Crane, 2017), Lep-MAP3 (Rastas, 2017), and GUSMap (Bilton *et al.*, 2018). They were described in chronological order of development according to user-friendliness (tutorial, graphs, forum search) and operating systems. Additional

information such as availability, software status, version, and last update are presented in Table A.2 in the appendix.

2.3.2 Construction of the Genetic Linkage Maps Using OneMap and Lep-MAP3

This study was done using simulations. For this purpose, two mapping populations (inbred-based and outcrossing), with different marker types and missing data, were considered. The simulations were performed using PedigreeSim (Voorrips and Maliepaard, 2012).

2.3.3 Simulation of Genotypes of F_2 Population

In F_2, a diploid species with three metacentric chromosomes of 100 centiMorgans (cM) was considered. This population was constructed from two inbred lines and subsequent self-fertilization (founders of population). The founders genomes were simulated considering $1,000$ biallelic markers (SNP) equally distributed along the three chromosomes. In addition, it was considered a fixed distance between the markers of 0.3 cM. With this information and assuming that no crossover interferences occurred, the gametes were produced and 200 individuals were formed.

2.3.4 Simulation of Genotypes of Outcrossing Population

The outcrossing population (F_1) population was constructed from the crossing of two heterozygous individuals (founder of population). As there may be up to 18 patterns of segregation (Wu, 2002) (Table A.1), the founders genomes were simulated considering three different situations. In the first, 1,000 biallelic markers (B3.7; D1.10; D2.15) (SNP) were considered. In the second, there were 324 markers with 18 patterns of segregation containing fully informative markers (segregating 1:1:1:1) and partially informative markers (missing markers and segregating 1:2:1, 3:1, and 1:1). The latter is a mixture of the first and second scenario, with 1,000 biallelic markers (B3.7; D1.10; D2.15) and 300 markers with the remaining 15 patterns, totaling 1,300 markers. In all situations, the markers were equally distributed along the three chromosomes. In addition, it was considered a fixed distance between the markers of 0.3 cM, 0.93 cM and 0.08 cM in the first, the second and the third scenario, respectively. In relation to the genome structure (ploidy, absence of interference, properties of the chromosomes) and population size, the same information from of the F_2 population was considered.

2.3.5 Software Tools

In short, four scenarios were obtained: i) an F_2 population with $1,000$ SNP markers, ii) an F_1 population with $1,000$ SNP markers, iii) an F_1 population with 300 markers with 18 patterns of segregation, and iv) an F_1 population with $1,300$ markers ($1,000$ biallelic markers and 300 markers with 15 patterns of segregation). Each scenario was simulated 100 times considering 10% of missing data. Some software tools presented in this study were not used due to some limitations that will be discussed later. Therefore, from these simulations, genetic linkage maps were constructed using the software OneMap and Lep-MAP3:

- **OneMap**

OneMap was evaluated in all scenarios totaling 400 linkage maps. The maps were constructed in the R environment (R Development Core Team, 2017) using several functions implemented in version 2.1.1 of the the OneMap package. The simulated data was converted to .raw format and imported using the pedsim2raw and read_onemap functions, respectively. Before the map construction, two steps of filtration were performed. In the first, the markers with redundant information were removed using the find_bins and create_data_bins functions. In the second, the markers with Mendelian segregation deviations were identified using the test_segregation function. This function performs the chi-square test using global alpha (α) of Bonferroni correction for multiple tests. With the select_segreg function, markers without segregation deviations were selected. From these results, the maps were constructed. The grouping of the markers was done using the suggest_lod, twopts, make_seq, and group functions. The markers were grouped with a maximum recombination fraction of 0.5 and a minimum LOD recommended by the suggest_lod function. The ordering of the markers within each linkage group was done using the MDS method (Preedy and Hackett, 2016) using the mds_onemap function. The multipoint analysis based on HMM was also implemented in the mds_onemap function to check the local order of the markers and to refine the map distance between adjacent markers. As the data was simulated considering the absence of interference, the recombination fractions were transformed into distances using the Haldane mapping function (Haldane, 1919). The same functions were used in all scenarios, however, in F_1 populations the twopts function simultaneously estimated the recombination fractions and linkage phases using the methodology proposed by Wu (2002).

- **Lep-MAP3**

Lep-MAP3 was evaluated in scenarios i and ii totaling 200 linkage maps. In both scenarios, the simulated data was converted to the Variant Call Format (VCF) using the pedsim2vcf function and imported using the ParentCall2 module. Markers with Mendelian segregation deviations or redundant markers were removed using Filtering2 and JoinIdenticalLGs modules, respectively. The linkage groups were identified using SeparateChromosomes2 module and considering a maximum recombination fraction of 0.5 and minimum LOD of 6. Singular markers were assigned to existing linkage groups using JoinSingles2All module. The markers were ordered within each linkage group using OrderMarkers module. The recombination fractions were transformed into distances using Haldane mapping function (Haldane, 1919).

2.3.6 Evaluation of results

The grouping, ordering, and lengths of the linkage maps were analyzed for all simulations and all scenarios. In addition, the filtering, grouping, and ordering times were counted. Barplots and boxplots were constructed to visualize the results. All these analyses were performed in the R environment (R Development Core Team, 2017).

2.4 Results and Discussion

2.4.1 Software Tools Description

2.4.1.1 OneMap

OneMap has two tutorials that teach step by step how to construct a genetic linkage map in inbreeding and outcrossing populations. It has graphs that allow the visualization of raw data, of the segregation test and of the markers' order (heat maps). These graphs are useful because they facilitate the analysis and some diagnostics about the maps. For example, criteria and ordering algorithms provide approximate solutions and may contain errors. Thus, the heat maps can be useful tools in the final diagnosis of the marker ordering in a linkage group. The stable version is available on CRAN and the version under development is available on GitHub. Both versions are provided for Mac, Linux and, Windows. It has been widely used to construct genetic linkage maps in several species of plants and animals (*e.g.* Pereira *et al.* (2013); Palaiokostas *et al.* (2015); Hoshino *et al.* (2016); Nugent *et al.* (2017); Murube *et al.* (2019).

2.4.1.2 Lep-MAP, Lep-MAP2, and Lep-MAP3

Lep-MAP (Rastas *et al.*, 2013), Lep-MAP2 (Rastas *et al.*, 2015), and Lep-MAP2 (Rastas, 2017) have a small tutorial and a discussion forum open to the scientific community. The forum is a useful tool because it allows an interaction between the developer and the users. In addition, already answered questions can help other users. It is noted that doubts are answered in a relatively short time. Java source code and documentation are publicly available from sourceforge and the executables are provided for Linux. Note that all versions are widely used in plants and animals (Shearman *et al.*, 2015; Gutierrez *et al.*, 2018; Christensen *et al.*, 2018; Bai *et al.*, 2018; Jeffries *et al.*, 2018; Littrell *et al.*, 2018; Zou *et al.*, 2019).

2.4.1.3 HighMap

HighMap (Liu *et al.*, 2014) provides heat maps and haplotype maps which help users during the construction of maps. In addition, the executables are provided for Mac, Linux and, Windows. High-density genetic maps were constructed in soybean (Qi *et al.*, 2014), cucumber (Xu *et al.*, 2015), watermelon (Shang *et al.*, 2016), cotton (Jia *et al.*, 2016), sorghum (Ji *et al.*, 2017), peas (Zheng *et al.*, 2018), broccoli (Yu *et al.*, 2019), and rubber trees (An *et al.*, 2019).

2.4.1.4 ASMap

ASMap (Taylor and Butler, 2017) contains a suite of tools to assist in the rapid diagnosis and repair of a constructed linkage map. In addition, an overview of the software and an illustrative example showing the complete linkage map construction process are available. It has been available a version on CRAN and the executables are provided for Mac, Linux, and Windows. Although this software development is recent, there are already studies using this tool. Amalraj *et al.* (2019) mapped the resistance of *Phytophthora* root rot and identified inde-

pendent loci from cultivated (*Cicer arietinum* L.) and wild (*Cicer echinospermum* P.H. Davis) chickpea. Ward *et al.* (2019) detected QTLs specific to shoot growth and to leaf elongation.

2.4.1.5 Flipper

In Flipper (Crane, 2017), the source code is hosted on GitHub, but the README file does not have information on the configuration, installation, and operating instructions. According to the author, the executables are provided for MacOSX 10.11.6, Linux, and Windows7.

2.4.1.6 GUSMap

GUSMap (Bilton *et al.*, 2018) is unique to sequencing data because it requires information (number of reads) of the alleles. It contains a an extensive and detailed tutorial and a suite of tools that assists in the construction of the map. As in OneMap, the stable version is available on CRAN and the version under development is available on GitHub. In addition, the executables are provided for Mac, Linux, and Windows.

2.4.1.7 Additional Considerations

The initial idea was to construir genetic linkage maps using all of the software tools with free licenses previously presented in this study. However, Lep-MAP, Lep-MAP2, HighMap, ASMap, Flipper, and GUSMap have not been tested because of some limitations that will be presented below.

The previous two versions of Lep-MAP3 are available to users but we have decided to evaluate only the most current and improved version. Although the HighMap has been used recently, the site hosting the software is temporarily unavailable. According to Taylor and Butler (2017), ASMap is restricted to linkage map construction in inbreeding population (F_2, BC, DH, and RILs). However, the software did not support the simulated F_2 mapping population. The Flipper not have information on the installation and operating instructions. In addition, there is no paper about the software and the only information was presented at the XXVI PAG conference. And finally, the GUSMap, as was said previously, it requires depth information. Depth and coverage information is not within the scope of this study, and consequently, this software can not be included in the evaluations.

2.4.2 Genetic Linkage Maps

The objective of this study was not to compare the efficiency of these software tools, but rather to describe their characteristics and to utilize them to construct maps, when possible. A comparative study would be unfair, because each softwares' tools have their own particularities, which can have a different importance to users with distinct goals. For example, some researchers may prefer faster software, while others are looking for software that has visualization tools or that are easier to use. Furthemore, the scenarios presented here are only a small fraction of the substantial number of possibilities, in which the performance of these software tools may differ.

2.4.2.1 OneMap

In OneMap (Margarido *et al.*, 2007), the genetic linkage maps were constructed using the four simulated scenarios. This software has graphs that allow for the visualization of raw data, the segregation test and the markers' order (heat maps). Therefore, the first simulation of each scenario was used to show these tools throughout the discussion.

Before the construction of the maps, data sets from all scenarios were analyzed. In the first scenario, the markers and missing data are randomly distributed. In addition, all markers are codominant (Figure 2.1). This was already expected because only SNPs were considered.

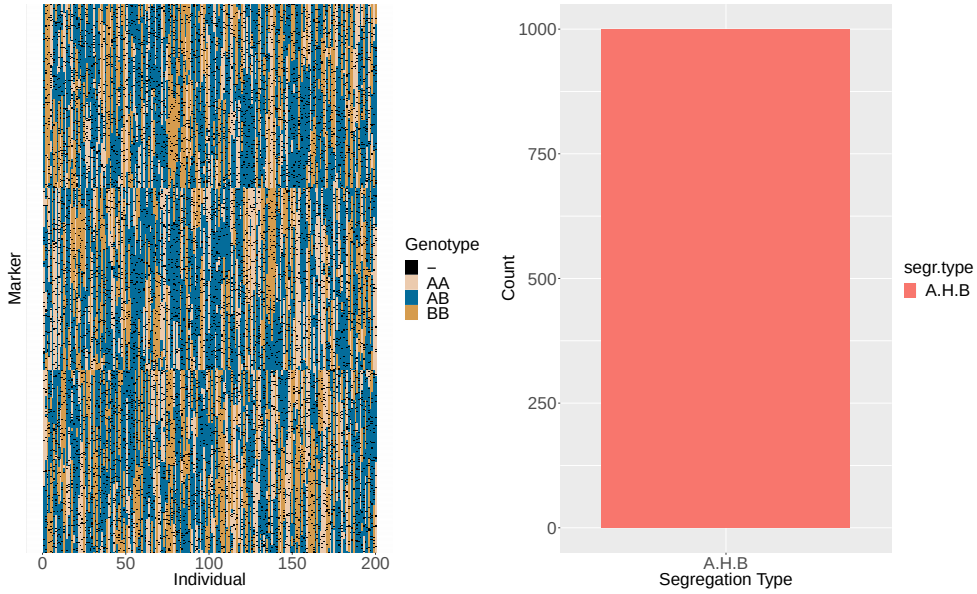

Figure 2.1. Information about raw data (left) and type segregating of marker (right) in F_2 population with 1,000 markers provided by OneMap.

For outcrossing populations, it is possible to make a graphical representation of the types of markers (segregation pattern) and the number of markers by segregation pattern, following the notation proposed by Wu (2002). In the second scenario, the three types of biallelic markers (B3.7, D1.10, and D2.15) are equally frequent in the genome (Figures 2.2). In the third scenario, the 18 types are present and equally frequent in the genome (Figure 2.3). And in the last scenario, the 18 types are also present. However, the biallelic markers are in greater numbers and the remaining types are less numerous (Figure 2.4).

In all scenarios, no redundant markers or markers with segregation distortion were found. Therefore, the maps were constructed considering all simulated markers. OneMap uses the bin strategy to eliminate these redundant markers. First, markers with the same genotypic information are grouped into bins. Subsequently, only the most representative marker of each bin is kept in the data set. Once the map is constructed, they can be added again.

The linkage groups were formed according to estimates of recombination fractions and linkage phases (F_1 only). In four scenarios, three groups were formed by all simulations. In the first scenario, all the markers were assigned to a group. In contrast, in the remaining scenarios some markers were not assigned to any group and it is also possible to observe a

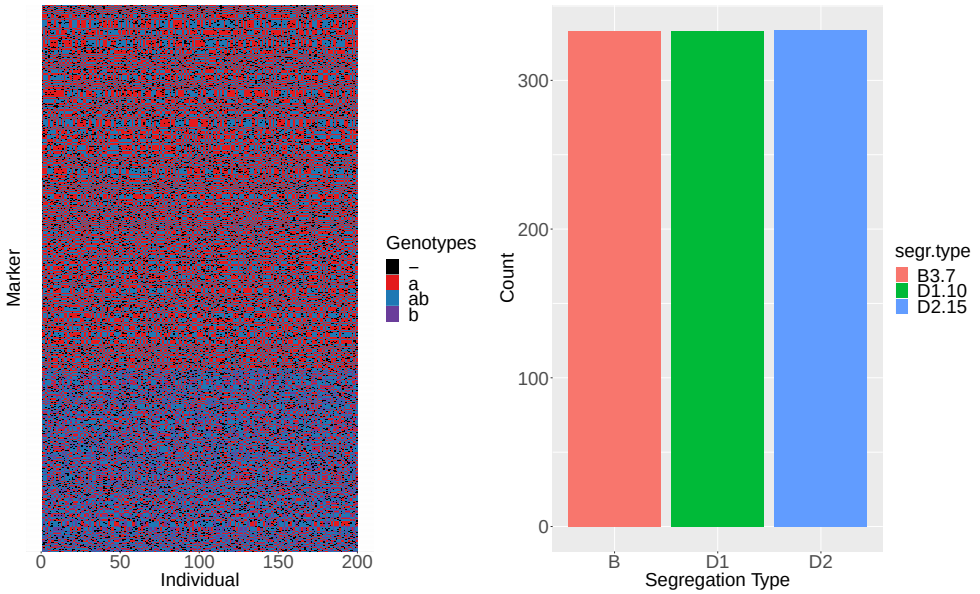

Figure 2.2. Types of segregation pattern (left) and the number of markers by segregation pattern (right) in F_1 with 1,000 markers provided by OneMap.

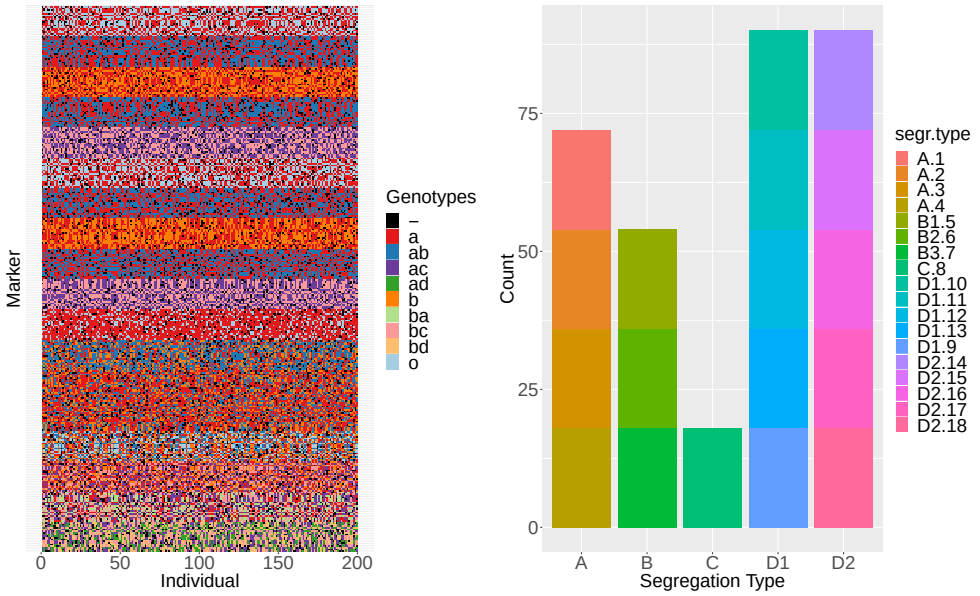

Figure 2.3. Types of segregation pattern (left) and the number of markers by segregation pattern (right) in F_1 with 324 markers provided by OneMap.

higher interquartile amplitude in comparison to first scenario (Figure 2.5). However, OneMap has additional tools and single markers can be reallocated.

The ordering of markers within linkage groups can be diagnosed using heat maps. They are made by plotting the recombination matrix based on a color scale, in which, cool colors (blue areas) represent weak linkage and hot colors (red areas) represent strong linkage. Hot colors are expected to be concentrated along the diagonal while the cold colors fall far from it. In this study, this tool was useful to evaluate the results, because, in all scenarios, some

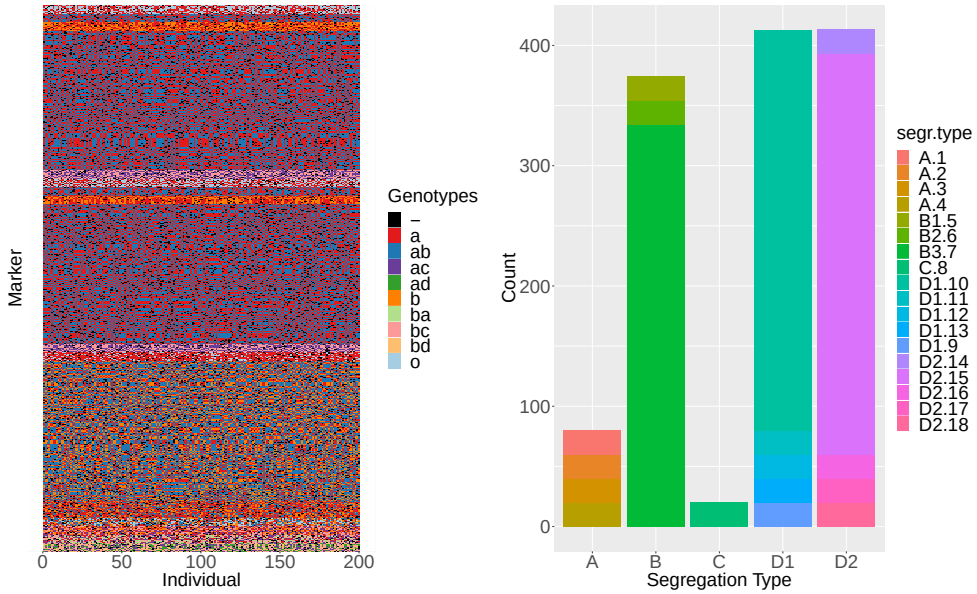

Figure 2.4. Types of segregation pattern (left) and the number of markers by segregation pattern (right) in F_1 with 1,300 markers provided by OneMap.

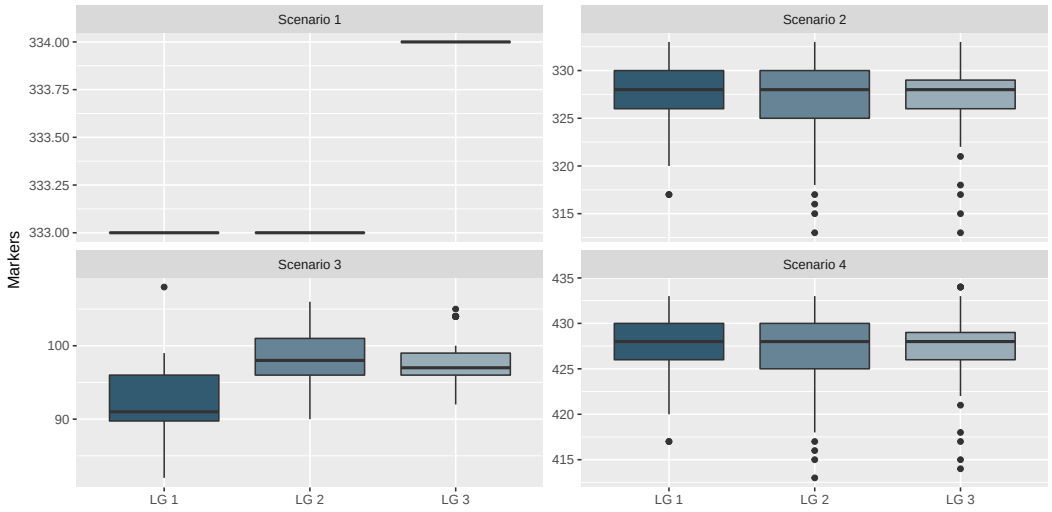

Figure 2.5. Boxplots of the number of markers in each linkage group using OneMap. Scenario 1: F_2 population with 1,000 markers. Scenario 2: F_1 population with 1,000 markers. Scenario 3: F_1 population with 324 markers. Scenario 4: F_1 population with 1,300 markers. The x-axis corresponds to the linkage groups and the y-axis corresponds to the number of markers.

markers are misplaced. These markers were removed and repositioned again to try to improve the ordering, always considering the ones with smaller distances and higher likelihoods (Figures 2.6 and 2.7). The length of the linkage groups varied in all scenarios. However, the estimated values approximated the simulated values (Figure 2.8).

In relation to the filtering and grouping, it is noticed that the larger the number of markers the longer the time needed for the construction of these steps. In relation to the ordering, the first scenario had the shortest time. In the other scenarios with F_1 populations,

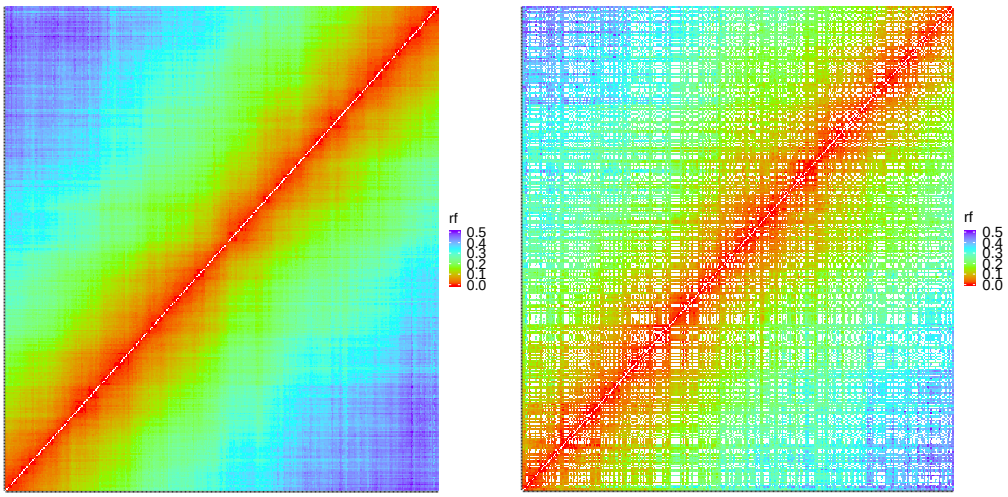

Figure 2.6. Heat Map generated by OneMap to diagnose the order of markers within the linkage group in F_2 population with 1,000 markers (left) and F_1 population with 1,000 markers (right). In x and y axes the markers are plotted according to established order. In the right, the blank cells represent combinations of markers whose recombination fraction can not be estimated based on only two markers (two-point test).

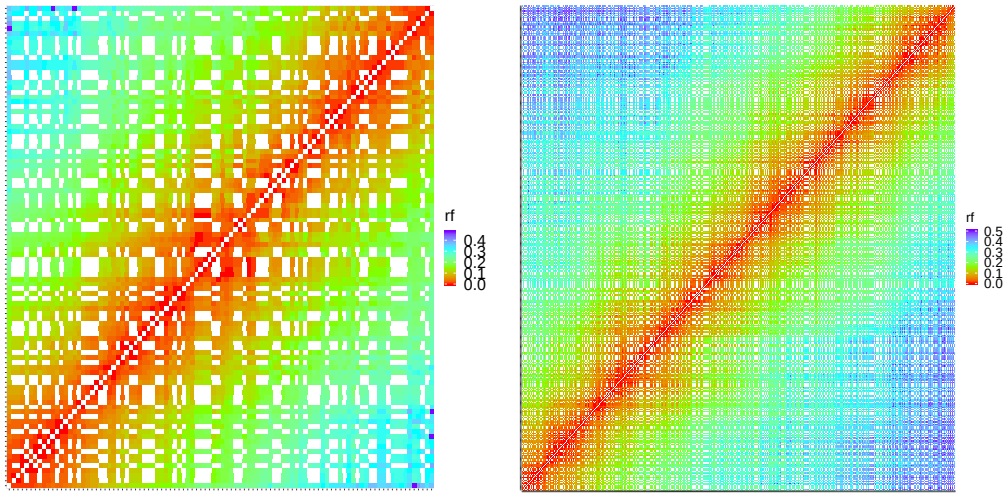

Figure 2.7. Heat Map generated by OneMap to diagnose the order of markers within the linkage group in in F_1 population with 324 markers (left) and F_1 population with 1,300 markers (right). In x and y axes the markers are plotted according to established order. The blank cells represent combinations of markers whose recombination fraction can not be estimated based on only two markers (two-point test).

the time also varied according to the number of markers. For example, the third scenario with 324 markers presented the shortest ordering time while the fourth scenario with 1,300 markers presented the longer ordering time (Figure 2.9). In F_1 populations, OneMap initially estimates the recombination frequencies and the linkage phases of the markers. Subsequently, the order, distances and linkage phases are checked using a multipoint analysis based on HMM (Margarido *et al.*, 2007). The higher the number of individuals, the greater the number of

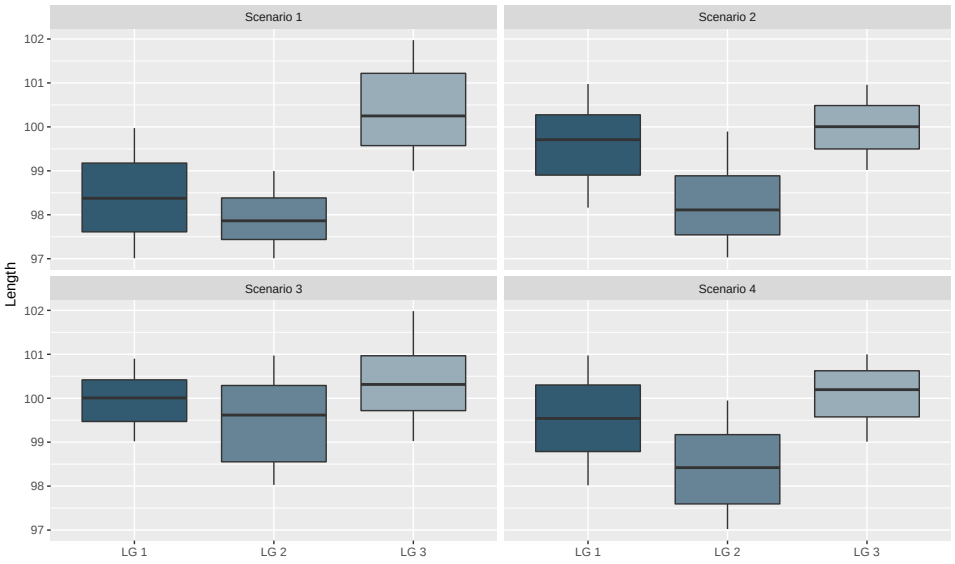

Figure 2.8. Boxplots of the lengths of the linkage groups using OneMap. Scenario 1: F_2 population with 1,000 markers. Scenario 2: F_1 population with 1,000 markers. Scenario 3: F_1 population with 324 markers. Scenario 4: F_1 with 1,300 markers. The x-axis corresponds to the linkage groups and the y-axis corresponds to the lengths of groups.

factors used to calculate the likelihood, and the smaller the amount of information provided by the markers, the greater the number of hidden states that must be inferred. Therefore, depending on the number of individuals and the markers information that steps may take some time (Lander and Green, 1987).

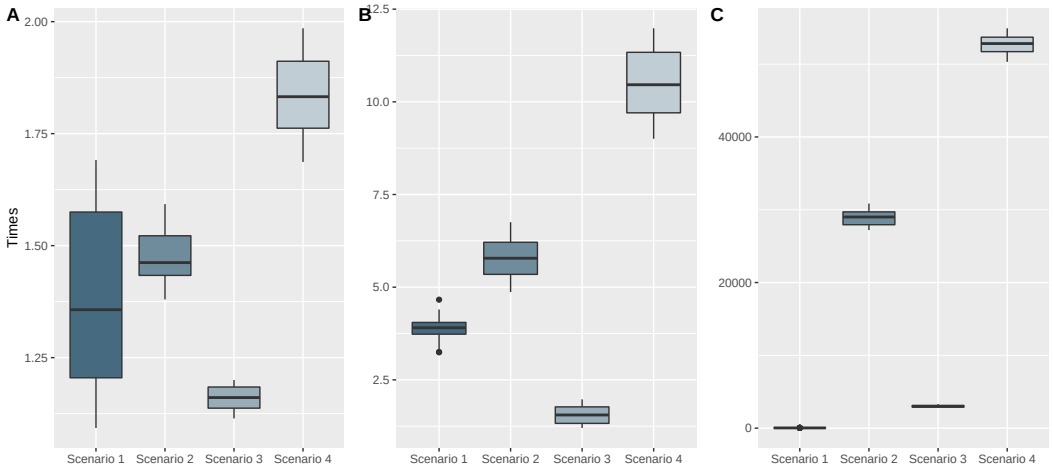

Figure 2.9. Boxplots of the processing time for filtering (A), grouping (B) and ordering (C) using OneMap. Scenario 1: F_2 population with 1,000 markers. Scenario 2: F_1 population with 1,000 markers. Scenario 3: F_1 population with 324 markers. Scenario 4: F_1 with 1,300 markers.The x-axis corresponds to the scenarios and the y-axis corresponds to the times in seconds.

2.4.2.2 Lep-MAP3

Lep-MAP3 (Rastas, 2017) is a software capable of analyzing different populations (inbreeding and outcrossing) and different marker types (SNP and microsatellites). However, only the F_2 and F_1 scenarios with $1,000$ markers were used (first and second scenario).

Like OneMap, Lep-MAP3 also groups the identical markers and leaves only the most representative marker of each group in the data set. Neither redundant markers or markers with segregation distortion were found. Redundant markers have the same genotypic information as other markers. They do not increase the information but increase the computational effort during the construction of the map. Therefore, they can be removed (Liu *et al.*, 2014). Generally, markers with segregation distortion should also be eliminated from the data set because the models for the map construction are based on Mendel's laws. There are several studies about the effect of markers with segregation distortion. In some studies, the presence of these markers affect the order and length of the maps (Lorieux *et al.*, 1995a,b; Xu *et al.*, 1997; Vogl and Xu, 2000). In others, these markers have no significant effect on map quality (Hackett and Broadfoot, 2003). These results may be specific to the data set. Therefore, methods that identify these markers are useful because they give the researcher the opportunity of making their own decision.

In first scenario, 60% of the simulations formed three groups, 34% formed two groups, and 6% formed one group. In the second scenario, 70% of the simulations formed three groups, 24% formed two groups, and 6% formed one group (Figure 2.10). In grouping, the linkage groups represent chromosomal segments or entire chromosomes. In real populations, the markers detected are not always uniformly distributed along the chromosomes. In addition, the frequency of recombination is not equal along chromosomes. In these cases, probably the number of linkage groups and chromosomes will be different (Collard *et al.*, 2005).

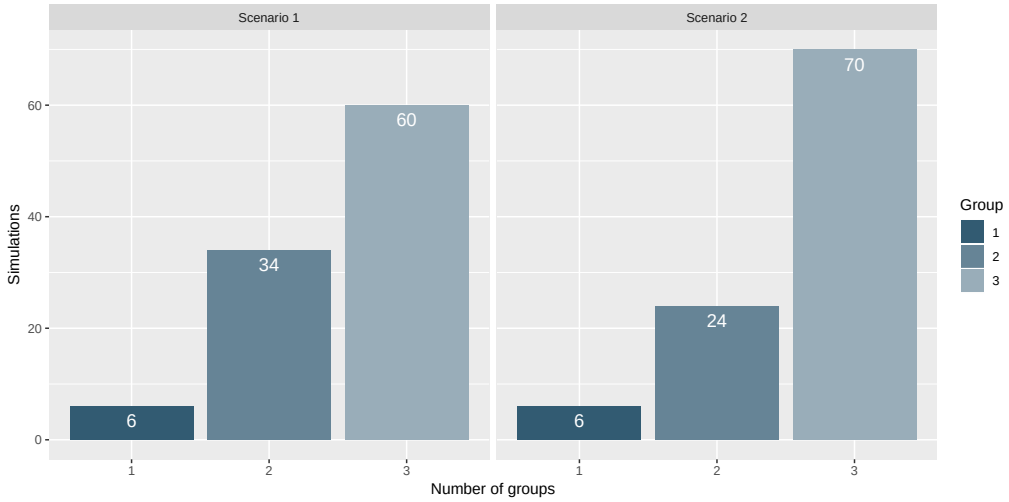

Figure 2.10. The number of linkage groups obtained in each simulation using Lep-MAP3. Scenario 1: F_2 population with 1,000 markers. Scenario 2: F_1 population with 1,000 markers.

The number of markers within each linkage group are shown in (Figure 2.11). In general, the first scenario had a smaller interquartile amplitude and the second scenario had a higher interquartile amplitude. This occurred because some markers were not assigned to any

group. In Lep-MAP3 it is also possible to relocate these markers into a group. This is done using the JoinSingles2Identicals and JoinSingles2All modules.

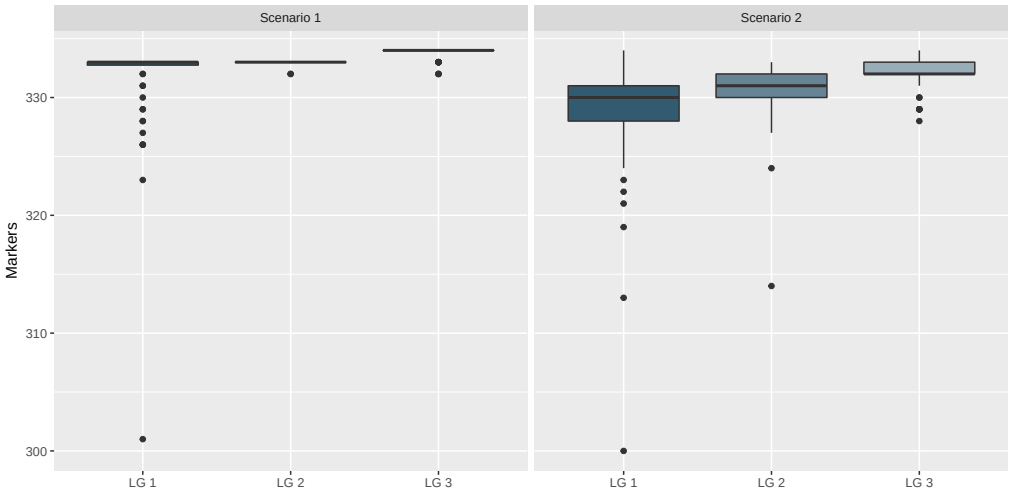

Figure 2.11. Boxplots of the number of markers in each linkage group using Lep-MAP3. Scenario 1: F_2 population with 1,000 markers. Scenario 2: F_1 population with 1,000 markers. The x-axis corresponds to the linkage groups and the y-axis corresponds to the number of markers.

The length of the linkage groups varied in two scenarios. In the first scenario, the variation between the linkage groups is higher. In the second scenario, the variation between the groups is smaller, but is larger when compared to the first scenario. Although there are some atypical values in both scenarios, the estimated groups presented values close to the simulated groups indicating a good ordering (Figure 2.12).

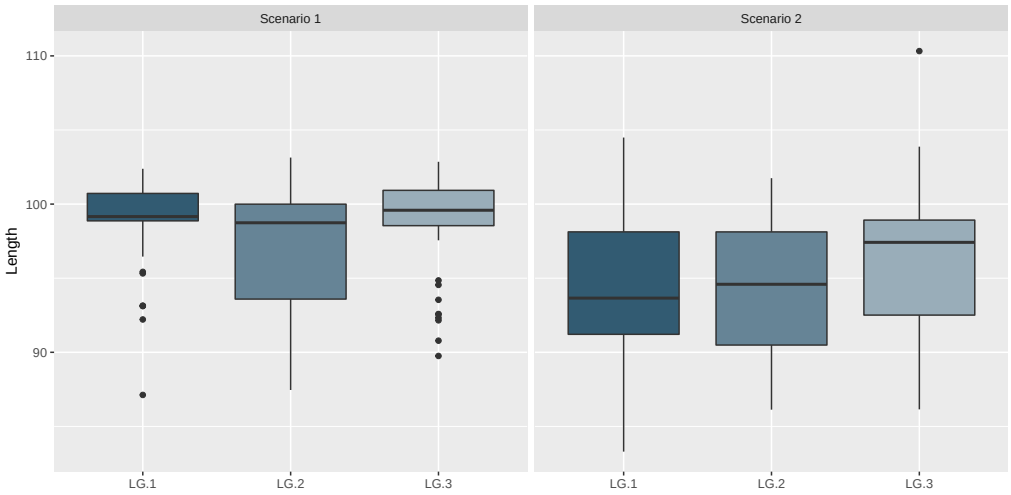

Figure 2.12. Boxplots of the lengths of the linkage groups using Lep-MAP3. Scenario 1: F_2 population with 1,000 markers. Scenario 2: F_1 population with 1,000 markers. The x-axis corresponds to the linkage groups and the y-axis corresponds to the lengths of groups.

Good marker ordering can be associated with multipoint estimates. OneMap and Lep-MAP3 used HMM based multipoint estimates to check the local order of markers and also to

refine the map distance between adjacent markers. According to Lander and Green (1987), multipoint estimates are better than two-point estimates because they avoid the propagation of errors generated by a lack of information between some marker combinations. This advantage is especially important when the data set has a mixture of dominant and codominant markers. In this case, the HMM allows for more informative markers to provide information to the others. In Mollinari *et al.* (2009), methods with approaches of the multipoint type showed better performance in different scenarios.

The processing time of each step was not very different between the scenarios (Figure 2.13). The longest step was the ordering. The ordering time in scenario 1 of Lep-MAP3 and OneMap were similar. However, for scenario 2, the ordering time of OneMap was notably higher. For example, OneMap took approximately 8 hours to order the markers while Lep-MAP3 took approximately 7 minutes. The algorithms implemented in Lep-MAP3 are up to 450 to $2,000$ times faster than previous versions. According to Rastas (2017), these algorithms can order $20,000$ markers in about 1 hour using a single core.

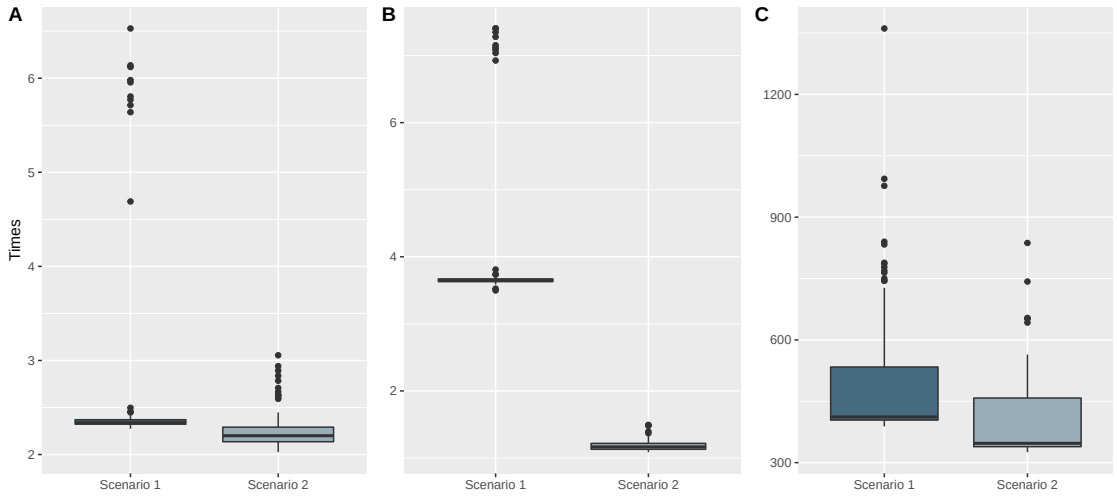

Figure 2.13. Boxplots of the processing time for filtering (A), grouping (B) and ordering (C) using Lep-MAP3. Scenario 1: F_2 population with 1,000 markers. Scenario 2: F_1 population with 1,000 markers. The x-axis corresponds to the scenarios and the y-axis corresponds to the times in seconds.

2.5 Conclusions

- Although software tools have characteristics that make them unique, they also share many methodologies and algorithmic approaches.
- OneMap and Lep-MAP3 were efficient in the construction of the genetic linkage maps considering the different scenarios.
- The results may be different if other scenarios had been considered. In addition, different researchers may obtain different results. Therefore, a comparative study here would be unfair.

Acknowledgments

This work was supported by CNPq ("Conselho Nacional do Desenvolvimento Científico e Tecnológico").

References

Akond, M., S. Liu, L. Schoener, J. A. Anderson, S. K. Kantartzi, Q. Song, D. Wang, Z. Wen, D. A. Lightfoot, and M. Ab, 2013 A SNP-Based Genetic Linkage Map of Soybean Using the SoySNP6K Illumina Infinium BeadChip Genotyping Array. Plant Genetics, Genomics, and Biotechnology **1**: 80–89.

Amalraj, A., J. Taylor, S. Bithell, Y. Li, K. Moore, K. Hobson, and T. Sutton, 2019 Mapping resistance to Phytophthora root rot identifies independent loci from cultivated (*Cicer arietinum* L.) and wild (*Cicer echinospermum* P.H. Davis) chickpea. Theoretical and Applied Genetics **132**: 1017–1033.

Ambawat, S., S. Senthilvel, C. T. Hash, T. Nepolean, V. Rajaram, K. Eshwar, R. Sharma, R. P. Thakur, V. P. Rao, R. C. Yadav, and R. K. Srivastava, 2016 QTL mapping of pearl millet rust resistance using an integrated DArT- and SSR-based linkage map. Euphytica **209**: 461–476.

An, Z., Y. Zhao, X. Zhang, X. Huang, Y. Hu, H. Cheng, X. Li, and H. Huang, 2019 A high-density genetic map and qtl mapping on growth and latex yield-related traits in *Hevea brasiliensis* müll.arg. Industrial Crops and Products **132**: 440–448.

Anderson, E. and L. Randolph, 1945 Location of the centromeres on the linkage maps of maize. Genetics **30**: 518.

Arús, P., C. Olarte, M. Romero, and F. Vargas, 1994 Linkage analysis of ten isozyme genes in f1 segregating almond progenies. Journal of the American Society for Horticultural Science **119**: 339–344.

Azevedo, G. C., A. Cheavegatti-Gianotto, B. F. Negri, B. Hufnagel, L. d. C. e Silva, J. V. Magalhaes, A. A. F. Garcia, U. G. Lana, S. M. de Sousa, and C. T. Guimaraes, 2015 Multiple interval qtl mapping and searching for pstol1 homologs associated with root morphology, biomass accumulation and phosphorus content in maize seedlings under low-p. BMC plant biology **15**: 172.

Bai, B., L. Wang, Y. J. Zhang, M. Lee, R. Rahmadsyah, Y. Alfiko, B. Q. Ye, S. Purwantomo, A. Suwanto, N.-h. Chua, and G. H. Yue, 2018 Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. Scientific Reports **8**: 691.

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, *et al.*, 2017 Gbs-based single dosage markers for linkage and qtl mapping allow gene mining for yield-related traits in sugarcane. BMC genomics **18**: 72.

Bartholomé, J., E. Mandrou, A. Mabiala, J. Jenkins, I. Nabihoudine, C. Klopp, J. Schmutz, C. Plomion, and J.-M. Gion, 2015 High-resolution genetic maps of eucalyptus improve eucalyptus grandis genome assembly. New Phytologist **206**: 1283–1296.

Bhatia, D., R. A. Wing, Y. Yu, K. Chougule, D. Kudrna, S. Lee, A. Rang, and K. Singh, 2018 Genotyping by sequencing of rice interspecific backcross inbred lines identifies QTLs for grain weight and grain length. Euphytica **214**: 1–16.

Bilton, T. P., M. R. Schofield, M. A. Black, D. Chagné, P. L. Wilcox, and K. G. Dodds, 2018 Accounting for errors in low coverage high-throughput sequencing data when constructing genetic maps using biparental outcrossed populations. Genetics **209**: 65–76.

Botstein, D., 1980 Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. American journal of human genetics **32**: 314–331.

Buetow, Kenneth H. and A. Chakravarti., 1987 Multipoint gene mapping using seriation. American journal of human genetics **41**: 180–188.

Butler, J. B., R. E. Vaillancourt, B. M. Potts, D. J. Lee, G. J. King, A. Baten, M. Shepherd, and J. S. Freeman, 2017 Comparative genomics of *Eucalyptus* and *Corymbia* reveals low rates of genome structural rearrangement. BMC genomics **18**: 397.

Butler, L., 1952 The linkage map of the tomato. Journal of Heredity **43**: 25–36.

Carter, A. H., X. Chen, K. Garland-Campbell, and K. K. Kidwell, 2009 Identifying qtl for high-temperature adult-plant resistance to stripe rust (*Puccinia striiformis f. sp. tritici*) in the spring wheat (*Triticum aestivum* l.) cultivar 'louise'. Theoretical and Applied Genetics **119**: 1119–1128.

Cartwright, D. A., M. Troggio, R. Velasco, and A. Gutin, 2007 Genetic Mapping in the Presence of Genotyping Errors. Genetics **176**: 2521–2527.

Cato, S., R. Gardner, J. Kent, and T. Richardson, 2001 A rapid pcr-based method for genetically mapping ests. Theoretical and Applied Genetics **102**: 296–306.

Cheema, J. and J. Dicks, 2009 Computational approaches and software tools for genetic linkage map estimation in plants. Briefings in bioinformatics **10**: 595–608.

Cheriton, D. and R. E. Tarjan, 1976 Finding minimum spanning trees. SIAM Journal on Computing **5**: 724–742.

Christensen, K. A., E. B. Rondeau, D. R. Minkley, J. S. Leong, C. M. Nugent, R. G. Danzmann, M. M. Ferguson, A. Stadnik, R. H. Devlin, R. Muzzerall, M. Edwards, W. S. Davidson, and B. F. Koop, 2018 The Arctic charr (*Salvelinus alpinus*) genome and transcriptome assembly. Plos one **13**: e0204076.

Chutimanitsakun, Y., R. W. Nipper, A. Cuesta-Marcos, L. Cistué, A. Corey, T. Filichkina, E. A. Johnson, and P. M. Hayes, 2011 Construction and application for qtl analysis of a restriction site associated dna (rad) linkage map in barley. BMC genomics **12**: 4.

Collard, B. C., M. Z. Jahufer, J. B. Brouwer, and E. C. Pang, 2005 An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts.

Crane, C. F., 2017 P0092 mapping 1000000 markers with flipper. In *Plant and Animal Genome XXVI Conference (PAG)*, pp. https://plan.core–apps.com/pag_2018/abstract/fff6e3855de940b33de03bc4eff5c2b4, Plant and Animal Genome XXVI Conference (PAG).

Crowston, K. and J. Howison, 2005 The social structure of free and open source software development. First Monday **10**.

Cui, F., C. Zhao, A. Ding, J. Li, L. Wang, X. Li, Y. Bao, J. Li, and H. Wang, 2014 Construction of an integrative linkage map and qtl mapping of grain yield-related traits using three related wheat ril populations. Theoretical and applied genetics **127**: 659–675.

Dempster, A. P., N. M. 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**: 1–22.

Dhulappanavar, C., 1977 A linkage group in rice (*Oryza sativa* l.) involving anti-inhibitory genes. Euphytica **26**: 427–432.

Dodds, K. G., J. C. Mcewan, R. Brauning, R. M. Anderson, T. C. Stijn, T. Kristjánsson, and S. M. Clarke, 2015 Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics **16**: 1047.

Doerge, R. and B. Weir, 1994 Constructing genetic maps by rapid chain delineation .

D'hont, A., F. Denoeud, J.-M. Aury, F.-C. Baurens, F. Carreel, O. Garsmeur, B. Noel, S. Bocs, G. Droc, M. Rouard, *et al.*, 2012 The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature **488**: 213.

Edwards, A. W. F., 2005 Linkage methods in human genetics before the computer. Human genetics **1922**: 515–530.

Falconer, D. and T. Mackay, 1996 Introduction to quantitative genetics. Introduction to quantitative genetics .

Falk, C. T., 1989 A simple scheme for preliminary ordering of multiple loci: application to 45 cf families. Progress in clinical and biological research **329**: 17–22.

Fierst, J. L., 2015 Using linkage maps to correct and scaffold de novo genome assemblies: Methods, challenges, and computational tools.

Fisher, R. A., 1922 On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **222**: 309–368.

Fragoso, C. A., H. Zhao, and S. L. Dellaporta, 2016 Imputing Genotypes in Biallelic Populations from Low-Coverage Sequence Data. Genetics **202**: 487–495.

Gaur, R., G. Jeena, N. Shah, S. Gupta, S. Pradhan, A. K. Tyagi, M. Jain, D. Chattopadhyay, and S. Bhatia, 2015 High density linkage mapping of genomic and transcriptomic SNPs for synteny analysis and anchoring the genome sequence of chickpea. Scientific reports **5**: 13387.

Geman, S. and D. Geman, 1987 Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in computer vision*, pp. 564–584, Elsevier.

Geyer, M., 2018 Exploring the genetics of fertility restoration controlled by Rf1 in common wheat (*Triticum aestivum* L.) using high-density linkage maps. Molecular Genetics and Genomics **293**: 451–462.

Gonzalez-Pena, D., G. Gao, M. Baranski, T. Moen, B. M. Cleveland, P. B. Kenney, R. L. Vallejo, Y. Palti, and T. D. Leeds, 2016 Genome-wide association study for identifying loci that affect fillet yield, carcass, and body weight traits in rainbow trout (*Oncorhynchus mykiss*). Frontiers in genetics **7**: 203.

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 **137**: 1121–1137.

Gutierrez, A. P., T. P. Bean, C. Hooper, C. A. Stenton, M. B. Sanders, R. K. Paley, P. Rastas, M. Bryrom, O. Matika, and R. D. Houston, 2018 A genome-wide association study for host resistance to ostreid herpesvirus in pacific oysters (*Crassostrea gigas*). G3: Genes, Genomes, Genetics **8**: 1273–1280.

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.

Haldane, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. Hournal of Gnetics **8**: 299–309.

Hoshino, A., V. Jayakumar, E. Nitasaka, A. Toyoda, H. Noguchi, T. Itoh, T. Shini, Y. Minakuchi, Y. Koda, A. J. Nagano, M. Yasugi, M. N. Honjo, H. Kudoh, M. Seki, A. Kamiya, T. Shiraki, P. Carninci, E. Asamizu, H. Nishide, S. Tanaka, K.-i. Park, Y. Morita, K. Yokoyama, I. Uchiyama, Y. Tanaka, S. Tabata, K. Shinozaki, Y. Hayashizaki, Y. Kohara, Y. Suzuki, S. Sugano, A. Fujiyama, S. Iida, and Y. Sakakibara, 2016 Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. Nature Publishing Group **7**: 1–10.

Huang, B. E., C. Raghavan, R. Mauleon, K. W. Broman, and H. Leung, 2014 Efficient Imputation of Missing Markers in Low-Coverage Genotyping-by-Sequencing Data from Multiparental Crosses. Genetics **197**: 401–404.

Jaccoud, D., K. Peng, D. Feinstein, and A. Kilian, 2001 Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic acids research **29**: 1–7.

Jeffries, D. L., G. Lavanchy, R. Sermier, M. J. Sredl, I. Miura, A. Borzée, L. N. Barrow, D. Canestrelli, P.-A. Crochet, C. Dufresnes, J. Fu, W.-j. Ma, C. M. Garcia, K. Ghali, A. G. Nicieza, R. P. O. Donnell, N. Rodrigues, A. Romano, I. Martínez-solano, I. Stepanyan, S. Zumbach, A. Brelsford, and N. Perrin, 2018 A rapid rate of sex-chromosome turnover and non-random transitions in true frogs. Nature Communications **9**: 40488.

Ji, G., Q. Zhang, R. Du, P. Lv, X. Ma, S. Fan, S. Li, S. Hou, Y. Han, and G. Liu, 2017 Construction of a high-density genetic map using specific-locus amplified fragments in sorghum. BMC Genomics **18**: 51.

Jia, X., C. Pang, H. Wei, H. Wang, Q. Ma, J. Yang, S. Cheng, J. Su, S. Fan, M. Song, N. Wusiman, and S. Yu, 2016 High-density linkage map construction and QTL analysis for earliness-related traits in *Gossypium hirsutum* L. BMC Genomics **17**: 909.

Kosambi, D. D., 1944 The estimation of map distances from recombination values. volume 12, pp. 172–175, Annals of Eugenics.

Kruskal, J. B., 1956 On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society **7**: 48–50.

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.

Lathrop, G. M., J. Lalouel, C. Julier, and J. Ott, 1985 Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. American journal of human genetics **37**: 482.

Lee, S., K. R. Freewalt, L. K. Mchale, Q. Song, T.-h. Jun, A. P. Michel, A. E. Dorrance, and M. A. R. Mian, 2015 A high-resolution genetic linkage map of soybean based on 357 recombinant inbred lines genotyped with BARCSoySNP6K. Molecular breeding35 **35**: 5–11.

Lefebvre-Pautigny, F., F. Wu, M. Philippot, M. Rigoreau, M. Zouine, P. Frasse, M. Bouzayen, P. Broun, V. Pétiard, S. D. Tanksley, *et al.*, 2010 High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. Tree Genetics & Genomes **6**: 565–577.

Lepais, O., A. Manicki, S. Glise, M. Buoro, and A. Bardonnet, 2017 Genetic architecture of threshold reaction norms for male alternative reproductive tactics in Atlantic salmon (*Salmo salar* L.). Nature Publishing Group **7**: 43552.

Li, L., S. Zhao, J. Su, S. Fan, C. Pang, H. Wei, H. Wang, L. Gu, C. Zhang, G. Liu, D. Yu, Q. Liu, X. Zhang, and S. Yu, 2017 High-density genetic linkage map construction

by F2 populations and QTL analysis of early-maturity traits in upland cotton (*Gossypium hirsutum* L.). Plos **12**: e0182918.

Lincoln, S. E. and E. S. Lander, 1992 Systematic Detection of Errors in Genetic Linkage Data. Genomics **14**: 604–610.

Littrell, J., S.-w. Tsaih, A. Baud, P. Rastas, L. Solberg-woods, and M. J. Flister, 2018 A High-Resolution Genetic Map for the Laboratory Rat. G3: Genes, Genomees, Genetics **8**: 2241–2248.

Liu, D., W. Hong, L. Huang, M. Liu, H. Liu, H. Zeng, D. Deng, H. Xin, J. Song, C. Xu, X. Sun, X. Hou, X. Wang, and H. Zheng, 2014 Construction and Analysis of High-Density Linkage Map Using High-Throughput Sequencing Data. Plos one **9**: e98855.

Lorieux, M., B. Goffinet, X. Perrier, D. G. De Leon, and C. Lanaud, 1995a Maximum-likelihood models for mapping genetic markers showing segregation distortion. 1. backcross populations. Theoretical and Applied Genetics **90**: 73–80.

Lorieux, M., X. Perrier, B. Goffinet, C. Lanaud, and D. G. De León, 1995b Maximum-likelihood models for mapping genetic markers showing segregation distortion. 2. f 2 populations. Theoretical and Applied Genetics **90**: 81–89.

Luo, M., Y. Zhao, R. Zhang, J. Xing, M. Duan, J. Li, N. Wang, W. Wang, S. Zhang, Z. Chen, H. Zhang, Z. Shi, W. Song, and J. Zhao, 2017 Mapping of a major QTL for salt tolerance of mature field-grown maize plants based on SNP markers. BMC plant biology **17**: 140.

Lynch, M., B. Walsh, *et al.*, 1998 *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Ma, Y., C. J. Coyne, M. A. Grusak, M. Mazourek, P. Cheng, D. Main, and R. J. McGee, 2017 Genome-wide snp identification, linkage map construction and qtl mapping for seed mineral concentrations and contents in pea (*Pisum sativum* l.). BMC plant biology **17**: 43.

Maliepaard, C., J. Jansen, and J. 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.

Markert, C. L. and F. Møller, 1959 Multiple forms of enzymes: tissue, ontogenetic, and species specific patterns. Proceedings of the National Academy of Sciences of the United States of America **45**: 753.

Mester, D., Y. Ronin, D. Minkov, E. Nevo, and A. Korol, 2003 Constructing Large-Scale Genetic Maps Using an Evolutionary Strategy Algorithm. Genetics **165**: 2269–2282.

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.

Morgan, T. H. and E. Cattell, 1912 Data for the study of sex-linked inheritance in (*Drosophila*). Journal of Experimental Zoology **13**: 79–101.

Morris, J., N. Navarro, P. Rastas, L. D. Rawlins, J. Sammy, J. Mallet, and K. K. Dasmahapatra, 2019 The genetic architecture of adaptation: convergence and pleiotropy in heliconius wing pattern evolution. Heredity p. 1.

Morton, N. E., 1955 Sequential tests for the detection of linkage. American journal of human genetics **7**: 277.

Murube, E., A. Campa, and J. J. Ferreira, 2019 Integrating genetic and physical positions of the anthracnose resistance genes described in bean chromosomes pv01 and pv04. PloS one **14**: e0212298.

Nugent, C. M., A. A. Easton, J. D. Norman, M. M. Ferguson, and R. G. Danzmann, 2017 A SNP Based Linkage Map of the Arctic Charr (*Salvelinus alpinus*) Genome Provides Insights into the Diploidization Process After Whole Genome Duplication. G3: Genes, Genomees, Genetics **7**: 543–556.

Olson, J. M. and M. Boehnke, 1990 Monte Carlo Comparison of Preliminary Methods for Ordering Multiple Genetic Loci. Am. J. Hum. Genet **47**: 470–482.

Ott, J., J. Wang, and S. M. Leal, 2015 Genetic linkage analysis in the age of whole-genome sequencing. Nature Reviews Genetics **16**: 275.

Palaiokostas, C., M. Bekaert, M. G. Khan, J. B. Taggart, K. Gharbi, B. J. McAndrew, and D. J. Penman, 2015 A novel sex-determining qtl in nile tilapia (*Oreochromis niloticus*). BMC Genomics **16**: 171.

Peng, W., J. Xu, Y. Zhang, J. Feng, C. Dong, L. Jiang, J. Feng, B. Chen, Y. Gong, L. Chen, and P. Xu, 2016 An ultra-high density linkage map and QTL mapping for sex and growth-related traits of common carp (*Cyprinus carpio*). Nature Publishing Group **6**: 1–16.

Pereira, G. S., E. S. Nunes, M. F. Braga, H. A. Penha, A. L. Diniz, C. F. Munhoz, and R. Gazaffi, 2013 Molecular polymorphism and linkage analysis in sweet passion fruit, an outcrossing species. Annals of applied biology, **162**: 347–361.

Pierro, E. A. D., L. Gianfranceschi, M. D. Guardo, J. W. Kruisselbrink, S. Longhi, M. Troggio, L. Bianco, G. Pagliarani, S. Tartarini, T. Letschka, L. L. Luis, L. Garkava-gustavsson, D. Micheletti, R. E. Voorrips, E. Aziz, and R. Velasco, 2016 A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. Horticulture research **3**: 16057.

Pootakham, W., P. Ruang-Areerate, N. Jomchai, C. Sonthirod, D. Sangsrakru, T. Yoocha, K. Theerawattanasuk, K. Nirapathpongporn, P. Romruensukharom,

S. Tragoonrung, *et al.*, 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**: 367.

Preedy, K. F. and C. A. Hackett, 2016 A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. Theoretical and Applied Genetics **129**: 2117–2132.

Qi, Z., L. Huang, R. Zhu, D. Xin, C. Liu, X. Han, H. Jiang, W. Hong, G. Hu, H. Zheng, and Q. Chen, 2014 A High-Density Genetic Map for Soybean Based on Specific Length Amplified Fragment Sequencing. Plos one **9**: e104871.

R Development Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rabbi, I. Y., M. T. Hamblin, P. L. Kumar, M. A. Gedil, A. S. Ikpan, J.-l. Jannink, and P. A. Kulakow, 2014 High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. Virus Research **186**: 87–96.

Rastas, P., 2017 Lep-map3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics **33**: 3726–3732.

Rastas, P., F. C. Calboli, B. Guo, T. Shikano, and J. Merilä, 2015 Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example. Genome biology and evolution .

Rastas, P., L. Paulin, I. Hanski, R. Lehtonen, and P. Auvinen, 2013 Lep-map: fast and accurate linkage map construction for large snp datasets. Bioinformatics **29**: 3128–3134.

Ren, J., Z. Wang, Z. Du, M. Che, Y. Zhang, W. Quan, Y. Wang, X. Jiang, and Z. Zhang, 2019 Detection and validation of a novel major QTL for resistance to Fusarium head blight from Triticum aestivum in the terminal region of chromosome 7DL. Theoretical and Applied Genetics **132**: 241–255.

Ridout, M., S. Tong, C. Vowden, and K. Tobutt, 1998 Three-point linkage analysis in crosses of allogamous plant species. Genetics Research **72**: 111–121.

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

Ritter, E. and F. Salamini, 1996 The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. Genetics Research **67**: 55–65.

Rong, J., J. E. Bowers, S. R. Schulze, V. N. Waghmare, C. J. Rogers, G. J. Pierce, H. Zhang, J. C. Estill, and A. H. Paterson, 2005 Comparative genomics of *Gossypium* and *Arabidopsis*: Unraveling the consequences of both ancient and recent polyploidy. Genome research **15**: 1198–1210.

Sax, K., 1923 The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics **8**: 552.

Schlautman, B., L. Diaz-Garcia, G. Covarrubias-Pazaran, N. Schlautman, N. Vorsa, J. Polashock, E. L. Ogden, A. Brown, Y.-C. Lin, N. Bassil, *et al.*, 2018 Comparative genetic mapping reveals synteny and collinearity between the american cranberry and diploid blueberry genomes. Molecular breeding **38**: 9.

Shang, J., N. Li, N. Li, Y. Xu, and J. Wang, 2016 Construction of a high-density genetic map for watermelon (*Citrullus lanatus* L.) based on large-scale SNP discovery by specific length amplified fragment sequencing (SLAF-seq). Scientia horticulturae **203**: 38–46.

Shearman, J. R., D. Sangsrakru, N. Jomchai, P. Ruang-areerate, C. Sonthirod, C. Naktang, K. Theerawattanasuk, S. Tragoonrung, and S. Tangphatsornruang, 2015 SNP Identification from RNA Sequencing and Linkage Map Construction of Rubber Tree for Anchoring the Draft Genome. Plos one **10**: 0121961.

Sibov, S. T., C. L. De Souza Jr, A. A. F. Garcia, A. R. Silva, A. F. Garcia, C. A. Mangolin, L. L. Benchimol, and A. P. De Souza, 2003 Molecular mapping in tropical maize (*Zea mays* l.) using microsatellite markers. 2. quantitative trait loci (qtl) for grain yield, plant heigth, ear height and grain moisture. Hereditas **139**: 107–115.

Sierro, N., J. N. D. Battey, S. Ouadi, N. Bakaher, L. Bovet, A. Willig, S. Goepfert, M. C. Peitsch, and N. V. Ivanov, 2014 The tobacco genome sequence and its comparison with those of tomato and potato. Nature Communications **5**: 1–9.

Somta, P., J. Chen, C. Yundaeng, X. Yuan, T. Yimram, N. Tomooka, and X. Chen, 2019 Development of an SNP-based high-density linkage map and QTL analysis for bruchid (*Callosobruchus maculatus* F.) resistance in black gram (*Vigna mungo* (L.) Hepper). Scientific Reports **9**: 1–9.

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.

Stam, P., 1993 Construction of integrated genetic linkage maps by means of a new computer package: Join Map. The Plant Journal **4**: 739–744.

Staub, J. E., F. C. Serquen, and M. Gupta, 1996 Genetic markers, map construction, and their application in plant breeding. HortScience **31**: 729–741.

Steele, K., A. H. Price, H. Shashidhar, and J. Witcombe, 2006 Marker-assisted selection to introgress rice qtls controlling root traits into an indian upland rice variety. Theoretical and Applied Genetics **112**: 208–221.

Sturtevant, A. H., 1913 The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. Journal of experimental zoology **14**: 43–59.

Sun, F., P. Liu, J. Ye, L. C. Lo, S. Cao, L. Li, G. H. Yue, and C. M. Wang, 2012 An approach for jatropha improvement using pleiotropic qtls regulating plant growth and seed yield. Biotechnology for biofuels **5**: 42.

Syvänen, A.-C., 2001 Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature Reviews Genetics **2**: 930.

Tan, Y.-D. and Y.-X. Fu, 2006 A novel method for estimating linkage maps. Genetics **173**: 2383–2390.

Tanksley, S. D., 1993 Mapping polygenes. Annual review of genetics **27**: 205–233.

Tautz, D., 1989 Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic acids research **17**: 6463–6471.

Taylor, J. and D. Butler, 2017 R package asmap: efficient genetic linkage map construction and diagnosis. arXiv preprint arXiv:1705.06916 .

Thompson, E., 1984 Information gain in joint linkage analysis. Mathematical Medicine and Biology: A Journal of the IMA **1**: 31–49.

Tsai, H. Y., D. Robledo, N. R. Lowe, M. Bekaert, J. B. Taggart, J. E. Bron, and R. D. Houston, 2016 Construction and Annotation of a High Density SNP Linkage Map of the Atlantic Salmon (*Salmo salar*) Genome. G3: Genes, Genomees, Genetics **6**: 2173–2179.

Van Os, H., P. Stam, R. G. Visser, and H. J. Van Eck, 2005 Record: a novel method for ordering loci on a genetic linkage map. Theoretical and Applied Genetics **112**: 30–40.

van Os, H., P. Stam, R. G. Visser, and H. J. van Eck, 2005 Smooth: a statistical method for successful removal of genotyping errors from high-density genetic linkage data. Theoretical and applied genetics **112**: 187–194.

Vogl, C. and S. Xu, 2000 Multipoint Mapping of Viability and Segregation Distorting Loci Using Molecular Markers. Genetics **155**: 1439–1447.

Voorrips, R. E. and C. A. Maliepaard, 2012 The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC bioinformatics **13**: 248.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. v. d. Lee, M. Hornes, A. Friters, J. Pot, J. Paleman, M. Kuiper, *et al.*, 1995 Aflp: a new technique for dna fingerprinting. Nucleic acids research **23**: 4407–4414.

Ward, B., C. Brien, H. Oakey, A. Pearson, S. Negrão, R. K. Schilling, J. Taylor, D. Jarvis, A. Timmins, S. J. Roy, *et al.*, 2019 High-throughput 3d modelling to dissect the genetic control of leaf elongation in barley (*Hordeum vulgare*). The Plant Journal .

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.

Weeks, D. E. and K. Lange, 1987 Preliminary ranking procedures for multilocus ordering. Genomics **1**: 236–242.

Williams, J. G., A. R. Kubelik, K. J. Livak, J. A. Rafalski, and S. V. Tingey, 1990 Dna polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic acids research **18**: 6531–6535.

Wilson, S. R. and D. C. Rao, 1988 A major simplification in the preliminary ordering of linked loci. Genetic Epidemiology **5**: 75–80.

Wu, D., J. Koch, M. Coggeshall, and J. Carlson, 2019 The first genetic linkage map for Fraxinus pennsylvanica and syntenic relationships with four related species. Plant molecular biology **99**: 251–264.

Wu, R., 2002 Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases in Outcrossing Species. Theoretical population biology **363**: 349–363.

Wu, Y., P. R. Bhat, T. J. Close, and S. Lonardi, 2008 Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. PLoS genetics **4**: e1000212.

Xu, W., P. K. Subudhi, O. R. Crasta, D. T. Rosenow, E. John, and H. T. Nguyen, 2000 Molecular mapping of QTLs conferring stay-green in grain sorghum (*Sorghum bicolor* L. Moench). Genome **43**: 461–469.

Xu, X., R. Xu, B. Zhu, T. Yu, W. Qu, L. Lu, Q. Xu, X. Qi, and X. Chen, 2015 A high-density genetic map of cucumber derived from specific length amplified fragment sequencing (slaf-seq). Frontiers in plant science **5**: 768.

Xu, Y., L. Zhu, J. Xiao, N. Huang, and S. McCouch, 1997 Chromosomal regions associated with segregation distortion of molecular markers in f 2, backcross, doubled haploid, and recombinant inbred populations in rice (*Oryza sativa* l.). Molecular and General Genetics MGG **253**: 535–545.

Yin, M., R. Wang, B. Andersson, and E. Lerceteau, 2003 Nearly complete genetic maps of *Pinus sylvestris* L. (Scots pine) constructed by AFLP marker analysis in a full-sib family. Theoretical and Applied Genetics, v **106**: 1075–1083.

Yu, H., J. Wang, Z. Zhao, X. Sheng, Y. Shen, F. Branca, and H. Gu, 2019 Construction of a high-density genetic map and identification of loci related to hollow stem trait in broccoli (*Brassic oleracea* l. italica). Frontiers in plant science **10**: 45.

Zhang, W.-K., Y.-J. Wang, G.-Z. Luo, J.-S. Zhang, C.-Y. He, X.-L. Wu, J.-Y. Gai, and S.-Y. Chen, 2004 Qtl mapping of ten agronomic traits on the soybean (*Glycine max* l. merr.) genetic map and their association with est markers. Theoretical and Applied Genetics **108**: 1131–1139.

Zhang, Z.-s., M.-c. Hu, J. Zhang, J. Zheng, K. Zhang, W. Wang, and Q. Wan, 2009 Construction of a comprehensive PCR-based marker linkage map and QTL mapping for fiber quality traits in upland cotton (*Gossypium hirsutum* L.). Molecular Breeding **24**: 49–61.

Zheng, Y., F. Xu, Q. Li, G. Wang, N. Liu, Y. Gong, L. Li, Z.-H. Chen, and S. Xu, 2018 Qtl mapping combined with bulked segregant analysis identify snp markers linked to leaf shape traits in pisum sativum using slaf sequencing. Frontiers in genetics **9**.

Zou, C., L. Li, D. Miki, D. Li, Q. Tang, L. Xiao, S. Rajput, P. Deng, L. Peng, W. Jia, R. Huang, M. Zhang, Y. Sun, J. Hu, X. Fu, P. S. Schnable, Y. Chang, F. Li, H. Zhang, B. Feng, X. Zhu, R. Liu, J. C. Schnable, J.-k. Zhu, and H. Zhang, 2019 The genome of broomcorn millet. Nature Communications **10**: 436.

APPENDIX

Table A.1. Wu (2002) Table: Possible marker genotype cross combinations and observed marker band patterns and their offspring.

			Parent			Offspring		
	Cross type		Cross	Observed band	Remark	Observed bands	Segregation	No. phenotypes
A		1	$ab \times cd$	$ab \times cd$	Asymmetry	ac, ad, bc, bd	1 : 1 : 1 : 1	4
		2	$ab \times ac$	$ab \times ac$	Asymmetry	a, ac, ba, bc	1 : 1 : 1 : 1	4
		3	$ab \times co$	$ab \times c$	Asymmetry	ac, a, bc, b	1 : 1 : 1 : 1	4
		4	$ao \times bo$	$a \times b$	Asymmetry	ab, a, b, o	1 : 1 : 1 : 1	4
B	B_1	5	$ab \times ao$	$ab \times a$	Asymmetry	ab, $2a$, b	1 : 2 : 1	3
	B_2	6	$ao \times ab$	$a \times ab$	Asymmetry	ab, $2a$, b	1 : 2 : 1	3
	B_3	7	$ab \times ab$	$ab \times ab$	Symmetry	a, $2ab$, b	1 : 2 : 1	3
C		8	$ao \times ao$	$a \times a$	Symmetry	$3a$, o	3 : 1	2
D	D_1	9	$ab \times cc$	$ab \times c$	Asymmetry	ac, bc	1 : 1	2
		10	$ab \times aa$	$ab \times a$	Asymmetry	a, ab	1 : 1	2
		11	$ab \times oo$	$ab \times o$	Asymmetry	a, b	1 : 1	2
		12	$bo \times aa$	$b \times a$	Asymmetry	ab, a	1 : 1	2
		13	$ao \times oo$	$a \times o$	Asymmetry	a, o	1 : 1	2
D	D_2	14	$cc \times ab$	$c \times ab$	Asymmetry	ac, bc	1 : 1	2
		15	$aa \times ab$	$a \times ab$	Asymmetry	a, ab	1 : 1	2
		16	$oo \times ab$	$o \times ab$	Asymmetry	a, b	1 : 1	2
		17	$aa \times bo$	$a \times b$	Asymmetry	ab, a	1 : 1	2
		18	$oo \times ao$	$o \times a$	Asymmetry	a, o	1 : 1	2

Table A.2. Availability, software status, version and last update of seven tools for genetic map estimation.

Software Tools	Availability	Version	Status	Last Update
Onemap	https://github.com/augusto-garcia/onemap	2.1.1	Free software	May 9, 2019
Lep-MAP	http://sourceforge.net/users/lep-map	1.0	Free software	-
Lep-MAP2	http://sourceforge.net/users/lep-map2	2.0	Free software	October 4, 2017
HighMap	http://highmap.biomarker.com.cn/	-	Free software	-
ASMap	http://CRAN.R-project.org/package=ASMap	1.0-4	Free software	October 27, 2018
Lep-MAP3	http://sourceforge.net/users/lep-map3	3.0	Free software	May 10, 2019
GUSMap	https://github.com/tpbilton/GUSMap	0.1.0	Free software	January 28, 2019