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

Building highly saturated genetic maps with OneMap 3.0: new approaches using workflows

Cristiane Hayumi Taniguti

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

Piracicaba
2021

Cristiane Hayumi Taniguti
Bachelor in Biotechnology

Building highly saturated genetic maps with OneMap 3.0: new approaches using workflows

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2021

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

Taniguti , Cristiane Hayumi

Building highly saturated genetic maps with OneMap 3.0: new approaches using workflows / Cristiane Hayumi Taniguti . – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2021 .

154 p.

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

1. Mapa de ligação 2. Haplótipo 3. Reprodutibilidade 4. Erro de genotipagem I. Título.

To my parents Nanci and Marco Taniguti,
with all my love.

In honor of all women scientists and educators who inspire me.
In particular, my grandmother, the math teacher Edna Donabela Taniguti.

ACKNOWLEDGEMENTS

I am thankful to CNPq ("Conselho Nacional de Desenvolvimento Científico e Tecnológico") for the financial support, which made possible my professional qualification, the research, and the results here presented. I also thank the University of São Paulo, especially the "Luiz de Queiroz" College of Agriculture (ESALQ) for providing the structure needed for all these years of work and for joining so many amazing scientists, professors, and students, who I had the opportunity to meet. I am also grateful to other Brazilian institutions such as the Federal University of São Carlos (UFSCar), Federal University of Lavras (UFLA), and EMPRAPA (Empresa Brasileira de Pesquisa AgropecuáriA) for providing me environments of important scientific discussions.

I especially thank my advisor, Augusto Garcia, who gave support to all my ideas, precisely oriented me, and became a dear friend. He is for me an example of a scientist and optimistic person. I admire his capacity as a professor and the brilliant scientific team he has been qualifying, with significant impact in Brazillian and worldwide context. I do not have enough words to describe the impact of Augusto's influence on my personal and professional development. I also thank him for introducing me to his wife, Luciana, and to his talented son, João.

I am grateful to all the co-workers in the Statistical Genetics Lab (ESALQ/USP) for all the scientific discussions and friendship. In this pandemic time, I miss our coffee breaks. Since my master's, I have known some brilliant and kind people in this lab. I missed a lot the ones who defended the thesis first and left: João Feijó, Guilherme Pereira, Felipe Ferrão, Marianella Quezada, Letícia Lara, Danilo Cursi, Rodrigo Amadeu, Matheus Krause, Jhonathan Pedroso, Rafael Nalin, Pedro Jaloto, Mayara Oliveira, Mariana Niederheitmann, and Jéssica. The ones who just stayed for a while: Elaine Batista and Raíssa. I also thank the ones who are currently members of the lab: Gabriel Gesteira, Wellingson Araújo, Getúlio Ferreira, Camila Godoy, Vitória Bizão, Rafaela, Kaio Olímpio, and Laura. I especially thank Guilherme Pereira for helping me with ideas about the genotyping errors, multiallelic markers, and the design of the workflows for both empirical and simulated data; Marianella Quezada for providing datasets to tests; Jhonathan Pedroso and Getúlio Ferreira for coming with the idea of using splines to simulate the genetic map; Gabriel Gesteira for several works together and for helping me test the workflows in other systems; and Kaio Olímpio for the help with statistical analysis. Also, Rodrigo Amadeu and Letícia Lara for trusting me as their maid of honor. I spent really good times with all these people.

I am grateful to those who first developed OneMap: Gabriel Margarido, Marcelo Mollinari, and Augusto Garcia. And also to all those who contributed somehow: Guilherme Pereira, Rodrigo Amadeu, Getúlio Ferreira, Karl Broman, IdoBar, and all users that sent me an email with questions or suggestions. I especially thank Marcelo Mollinari for also developing MAPpoly, from which several of OneMap updates were based.

I also thank researchers from other groups in ESALQ, with whom I interacted a lot: Bioinformatics Laboratory Applied to Bioenergy, Conservation Genetics and Genomics group, Laboratory of Genetics of Microorganisms, Plant Ecological Genetics Laboratory, Allogamous Plant Breeding Laboratory, and our neighbors, the Cytogenomics and Epigenetics Laboratory. I especially thank the members of Bioinformatics Laboratory Applied to Bioenergy: Gabriel Margarido for the great classes, the discussions, and the help with OneMap updates; Fernando Correr for the friendship, for being a great partner in several activities, and for also helping me with workflow tests in other systems and with ideas about genotyping errors; Lorena Batista for the productive discussions and the trust in my potential; Victor Mello, Amanda Avelar, Ana Letycia, and Guilherme Hosaka. Evelyn Couto and Tamylin for the friendship, and Igor Araújo (Conservation Genetics and Genomics group) for the friendship and the help with evolution and population genetics. Allison and Carol (Plant Ecological Genetics Laboratory) for the company in lunch and sports activities. Maria Letícia (Microorganisms, Plant Ecological Genetics Lab-

oratory) for the talks about scientific communications. Júlia and Fernando (Allogamous Plant Breeding Laboratory) for all activities together. And Thiago Oliveira from Statistical Department (currently in Roslin Institute), for the friendship, very nice conversations, and for helping me with this work statistical analysis.

I acknowledge all the staff and professors of the Genetics Department, specially Prof. Maria Lúcia, Prof. Gabriel Margarido, Prof. Cláudia Vitorello, Prof. Maria Carolina, Prof. Giancarlo, Berdan, Fernandinho, Valdir, Léia, Macedônio, and Rafaelle.

Also, I am grateful for the opportunity to participate in the extension groups GVENCK and GENt, and interact with all members.

I would like to thank Fausto Silva and his team for providing me the joy of being called a "genetic scientist" on national open television. Empowering me even more in this profession. And the friends from Life Circo Piracicaba, PLPs Piracicaba and Cheerleading São Carlos, for the good times together.

I would like to thank my brother, Lucas Taniguti, who came with the idea of building the workflows and helped me to learn WDL syntax for the work here presented. He also had a great influence on my professional carrier choice, once since a kid, he is enthusiastic about technologies and motivated me to know new things and play video-games. I also thank my close friends Aline Nakamura, Bianca Coré, Luiza Kame, Guira Luz, Emeline Boni, Thaís Milanez, Laura Damada, Meenakshi Kannan, Fernanda Rossin, Maria Fernanda Trientini, Caio Boscariol, Rafael Pinto, Ariane Henrique, Daniel Sbravatti, Silvia de Oliveira, Nathalia Taniguti, Nathalia Alves, and Victor Rezende for the support in difficult times and for celebrating together the good times. I am grateful to Veri Firmino for being a great partner in this difficult last year and give me love and support to face all challenges that it brought.

I am mainly grateful to my parents, for making everything possible.

"I am among those who think that science has great beauty.
A scientist in his laboratory is not only a technician,
he is also a child place before natural phenomenon,
which impress him like a fairy tale."
-Marie Curie

"Nothing happens in contradiction to nature,
only in contradiction to what we know of it."
-Dana Scully

"Education is education.
We should learn everything and then choose which path to follow."
-Malala Yousafzai

SUMMARY

RESUMO

Construção de mapas genéticos altamente saturados com OneMap 3.0: novas abordagens usando workflows

OneMap é um pacote do R desenvolvido por membros do Laboratório de Genética Estatística da ESALQ/USP (Brasil) lançado em 2008. Ele ganhou atenção da comunidade científica por ser um dos primeiros programas capazes de construir mapas genéticos integrados para populações F_1 segregantes. Ele é hoje muito usado mundialmente. Entretanto, ele requer aprimoramentos para lidar com novos e abundantes marcadores provindos de técnicas de genotipagem baseada em sequenciamento. Neste trabalho, foi feito um aprimoramento significativo no OneMap para a versão 3.0, o qual inclui: maior velocidade na estimativa das distâncias genéticas; novos métodos de agrupamento e ordenamento dos marcadores; novas ferramentas gráficas para diagnóstico da qualidade dos mapas; novos recursos para realização de simulações; recursos para conversão de arquivos VCF com marcadores bialélicos e multialélicos para os arquivos de entrada do OneMap; possibilidade de incluir probabilidade de erro ou de genótipos para estimar as distâncias genéticas. Uma vez que o OneMap foi atualizado, também foram explorados passos anteriores à construção do mapa, os quais têm impacto na qualidade do mapa resultante. Para isso, foram desenvolvidos os *workflows* **Reads2Map** que realizam análises desde leituras de sequenciamento de dados empíricos ou simulados até mapas genéticos. Por ser escrito em *Workflow Description Language* (`WDL`), os workflows **Reads2Map** disponibilizam aos usuários códigos localizáveis, acessíveis, interoperáveis e reutilizáveis para a construção de mapas genéticos. Os workflows desenvolvidos são capazes de comparar o desempenho dos programas na construção de mapas genéticos: freebayes, GATK como identificadores de SNPs e genotipadores; updog, polyRAD e SuperMASSA como genotipadores; OneMap 3.0 e GUSMap para construção de mapas. Além disso, foi desenvolvido o aplicativo shiny Reads2MapApp para avaliação gráfica dos resultados dos workflows. No caso particular do conjunto de dados de *Populus tremula*, o freebayes foi selecionado como identificador de SNPs e genótipos, e uma probabilidade de erro global de 5%, resultando em um mapa com 6936 marcadores e 3299.96 cM. Em seguida, também utilizando os workflows, foi testado o impacto de duas das maiores melhorias do OneMap 3.0: o uso de probabilidades genotípicas para estimativa das distâncias genéticas; e o uso de marcadores multialélicos baseados em haplótipos provindos de identificadores de SNPs. Usando sequências de leituras simuladas foi possível medir a eficiência de cada identificador de SNP e genótipo e suas influências na construção do mapa. O impacto das probabilidades dos genótipos foi variável entre os programas de acordo com o cenário simulado. Os resultados mostraram que o OneMap 3.0 é capaz de construir mapas genéticos de alta qualidade se i) os genotipadores não cometerem muitos erros e a probabilidade de erro for de 5% para todos os genótipos ou ii) se o genotipador cometer mais erros de genotipagem e atribuir probabilidades menores para os genótipos errados. Além disso, o uso the marcadores multialélicos baseados em haplótipos revelou um aumento na qualidade de ordenamento e estimativa de distância genética. Uma vez que os processos anteriores à construção dos mapas têm grande impacto na sua qualidade, o uso combinado do OneMap 3.0, **Reads2Map** e Reads2MapApp, disponibiliza para os usuários ferramentas para construção de mapas genéticos desde leituras de sequenciamento, e também gráficos diagnóstico para auxílio na escolha da melhor combinação de programas e parâmetros.

Palavras-chave: Mapa de ligação, Haplótipo, Reprodutibilidade, Erro de genotipagem

ABSTRACT

Building highly saturated genetic maps with OneMap 3.0: new approaches using workflows

OneMap is an R package developed by members of Statistical Genetics Laboratory at ESALQ/USP (Brazil) released in 2008. It gained the attention of the scientific community for being one of the first software for building integrated genetics maps for outcrossing species. It is now highly used worldwide. However, it requires updates to deal with the new and abundant markers generated by high-throughput genotyping techniques. In this work, we made a major update of OneMap to version 3.0, which includes: higher speed of the genetic distance estimation; new methods for group and ordering markers; new quality diagnostic graphics tools; new features for making simulations; features to the conversion of VCF file with biallelic and multiallelic to OneMap input file; possibility of include error or genotype probability to estimate the genetic distances. Once OneMap was updated, we explored the steps upstream of the map building process, which has an impact on the resulted map quality. For that, we developed the **Reads2Map** workflows that perform the analysis, starting with empirical or simulated sequencing reads until the final linkage maps. Because the presented workflows are written with `Workflow Description Language` (`WDL`), they provide to users a findable, accessible, interoperable, and reusable code to build maps. The workflows compare the performance of software in the linkage map building: freebayes, GATK as SNP and genotype callers; updog, polyRAD, SuperMASSA as genotype caller; OneMap 3.0 and GUSMap as linkage map builders. We also developed the shiny Reads2MapApp app to evaluate graphically the workflow's results. In the particular case of an example dataset from *Populus tremula*, we select the freebayes as SNP and genotype caller, and a global error probability of 5%, resulting in a map with 6936 markers and 3299.961 cM. After also using the workflows, we tested the impact of two of the major OneMap 3.0 updates in the linkage maps: the usage of genotype probabilities to estimate the genetic distances and the haplotype-based multiallelic markers from assembly-based SNP caller. Using simulated sequence reads data we could measure each SNP and genotype caller efficiency and its influences in the resulted map. The impact of the genotype probabilities was variable between software according to each simulated scenario. The results showed that OneMap 3.0 can build high-quality genetic maps if i) the genotype callers do not estimate wrongly many genotypes and a global error rate of 5% is applied for all genotypes or ii) if the genotype caller estimate more genotypes wrongly it also gives lower genotype probabilities for the wrong genotypes. Furthermore, the usage of haplotype-based markers reveals to increase the order and genetic distance quality. Once the procedures upstream the genetic map building have a strong influence in its quality, the combined usage of OneMap 3.0, **Reads2Map** and Reads2MapApp provide to users tools to build linkage maps since the sequencing reads, and also diagnostic graphics and measures to help them to choose the best combination of software and parameters.

Keywords: Linkage map, Haplotype, Reproducibility, Genotyping error

1 INTRODUCTION

The OneMap package (Margarido *et al.*, 2007) was release in CRAN repository in 2008 with the novelty of building integrated genetic maps for outcrossing species. OneMap can build maps with all types of markers (Table A.2) of outcrossing or inbred (RIL, F_2 intercross, and F_2 backcross) mapping populations. At the time, OneMap was one of the software which opened the opportunity to better understand the genetic architecture of outcrossing species such as yellow passion fruit, loblolly pine, sugarcane, rubber tree, oil palm, eucalyptus, and salmon (Oliveira *et al.*, 2008a; Xiong *et al.*, 2016; Garcia *et al.*, 2006a; Souza *et al.*, 2013; Jeennor and Volkaert, 2014; Bartholomé *et al.*, 2015c; Gonen *et al.*, 2014).

OneMap is written in R, a free software environment for statistical computing and graphics. Users need to have a least a little knowledge of the R language to be able to use the package. According to TIOBE programming community index (the software quality company TIOBE, 2021), R had low (rate of 0.06%) popularity in 2007. In 2010, its popularity started to increase because more people, especially in the statistical science field, started to migrate from commercial statistical languages to R (Figure A.1). The accessibility of OneMap increased together with the R popularity (Figure 1.1).

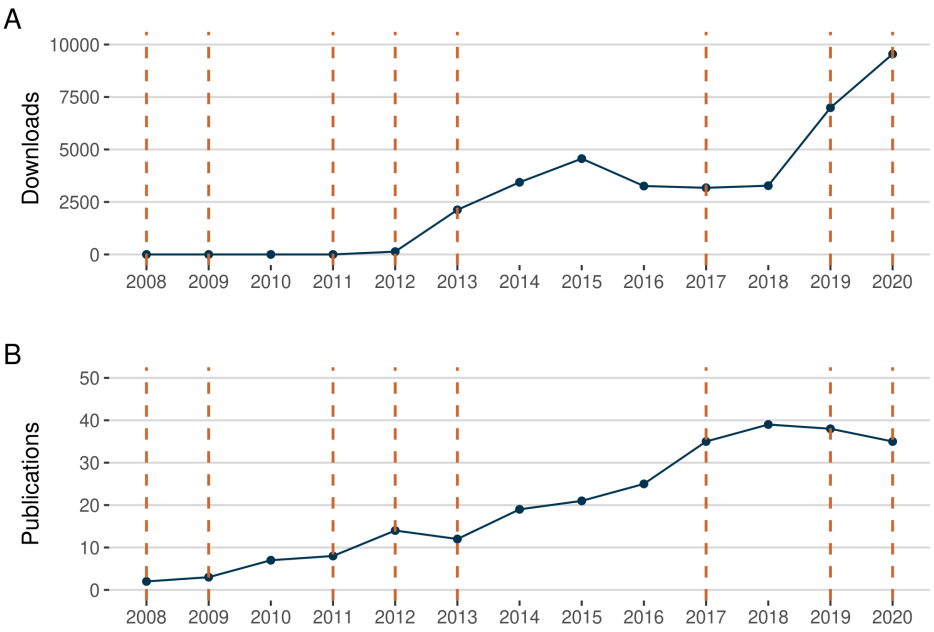

Figure 1.1. OneMap popularity since its release in 2008. A: Total number of OneMap downloads from CRAN by year until 2020. Data obtained using cranlogs package (Csárdi, 2019). B: Total number of publications citing OneMap by year from 2008 to 2020 according to Dimensions (2021) using the words 'onemap', 'linkage map' and ' genetic' as research criterias. The orange dashed lines represents OneMap version updates in CRAN

Members of Statistical Genetics Laboratory in ESALQ/USP (Brazil) make constant improvements and maintenance in OneMap algorithms and documentations, which also contribute to its success. In 2013, we created an GitHub (Chacon and Straub, 2014) repository to store the OneMap development version (`https://github.com/augusto-garcia/onemap`). The platform gives several advantages to code development, including the track of all changes, teamwork optimization, and allows OneMap to receives contributions from anyone around the world through pull requests or messages.

The improvements were made following novelties in statistical genetics research and according to users' feedbacks. Most of the users' demands came together with high-throughput genotyping platforms availability. Most of these demands focused on the conversion of VCF file format (Danecek *et al.*,

2011b), the time-consuming analysis using Hidden Markov Model (HMM), the wrong group and ordering of markers, and the inflated map sizes. In this work, we made updates to attend to each one of these demands. All the updates compose a major modification in the OneMap version. Therefore, here we will refer to the OneMap before this work updates as OneMap 2.0 and after this work updates as OneMap 3.0.

The HMM (Baum *et al.*, 1970; Lander and Green, 1987) combined with the expectation-maximization (EM) (Dempster *et al.*, 1977) algorithm implemented in OneMap is a very robust method to estimate the phases and genetic distance. It calculates iteratively the likelihoods for each possible phase for each marker adding them sequentially and also considering the previous information. At the end of the process, the HMM returns the most likely haplotype for each individual in the population. Thus, the recombination fraction estimated between markers are based on entire sequence information (multipoint approach). This is the best model possible to estimate the haplotypes, but demands an exhaustive method. As the total number of markers in the sequence increases, it also increases the computational resources and time needed (Wu *et al.*, 2002d; Garcia *et al.*, 2006b; Margarido *et al.*, 2007).

OneMap version 2.0 already received updates focusing in speed up the HMM and EM algorithm. The code was rewritten with a lower-level programming language, the `C++`, mainly by the developers M. Mollinari and G. Margarido. The R package Rcpp (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2017) allowed an easy integration of `C++` with the remaining `R` code. Nevertheless, users still demanded more speed.

Another solution was proposed in Schiffthaler *et al.* (2017), which includes calculating the linkage map in overlapping batches. The idea is that we do not need to perform the search for the maximum likelihood for each marker considering all previous markers in the sequence when there are many markers available. The previous information is indeed necessary to obtain accurate calculations of phase likelihoods, but the return saturates after few markers are evaluated. Keeping the search will only overload the model while offering no additional accuracy. In this case, we can limit this search using batches. The batches still keep part of the information from the previous batches using the previously estimated phases in overlapping markers. The Schiffthaler *et al.* (2017) simulations revealed that, once the batch and overlap size are optimized, the method keeps the same accuracy of haplotype estimation as the original one. We also made our simulation to confirm that (Attachment B).

The Schiffthaler *et al.* (2017) method also proposed the parallelization of the analysis in different computer cores. The batches can not be considered as independent processes to compute in separated cores, because of their overlapping markers. However, the parallelization can be made through the four possible phases for outcrossing marker combinations. Thus, the analysis can be divided into a maximum of four different computer cores. We also examined the possibility of parallelization through the batches to increase the possible number of cores to be used. We tested the accuracy of the estimated haplotypes by comparing the estimations in overlap markers among the batches in simulations. Despite it compromises accuracy for some marker combinations, it can be useful to obtain fast estimations (see Attachment B).

The Schiffthaler *et al.* (2017) modification was implemented in BatchMap package, a separated fork of OneMap, released in CRAN in March 2017. However, BatchMap did not receive maintenance and was removed from CRAN in December 2019. There is still the GitHub and Docker Hub versions available. As mentioned before, GitHub platform offers several advantages to code development, however does not impose restrictions on unfunctional codes. Docker hub is a repository for container images. The containers images are lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings (Merkel, 2014a). Thus, with Docker containers users can reproduce the exact computer environment where developers made the code functional. However, the Docker hub and the GitHub versions of BatchMap became outdated once OneMap received new updates. Here, we made the needed updates and merged

the BatchMap modifications with the most recent version of OneMap.

The ordering step of linkage map building in OneMap also needed improvements to adjust to high-throughput markers. All ordering algorithms implemented try to solve the "traveling salesman problem". With high-throughput markers, the traveling salesman would have "more cities to visit" and the problem became more complex. In 2016, Preedy and Hackett (2016) proposed the application of multidimensional scaling for ordering markers in linkage maps as a rapid and efficient method for a large number of markers. We also implemented the method in OneMap and made simulations to compare it with the other ordering algorithms implemented (see Attachment C). It reveals to be an excellent method for a global ordering of markers, despite keeping local misplacement.

Besides the increase in the number of markers available, the high-throughput sequencing genotyping also presented more errors. The users' demands about the wrong grouping and the inflated map sizes are consequences of these errors.

The grouping of markers in separated linkage groups is made in OneMap through the two-points recombination fraction estimations. The logarithm of odds (LOD) score statistics can be applied to test for linkage. LOD score is calculated by the log of the linkage probability of the observed data divided by the probability of the loci be unlinked.

Equation 1.1 shows a simple example of the probability density function used to estimate the described probabilities. This particular function is used for a backcross population.

$$l(r) = (n_1 + n_4)log(\frac{1-r}{2}) + (n_2 + n_3)log(\frac{r}{2}) \tag{1.1}$$

In the equation 1.1, n_1, n_2, n_3 and n_4 refers to the four possible combination of two genotypes for backcross population structure (AB|AB, AB|Ab, AB|aB and AB|ab) and r is the two-point recombination fraction value. Under H_0 hipothesis (unlinked loci) the recombination fraction parameter r is 0.5.

OneMap suggests a LOD threshold defined considering all two-points tests that will be performed for all markers in the data set. It uses a global alpha controlling type I error with Bonferroni's correction. From this global alpha, the corresponding quantile from the chi-square distribution is taken and then converted to LOD score.

Genotyping errors can distort the two-points estimations for some markers, which makes tradicional OneMap grouping algorithm returns wrong linkage groups. The result usually presents a total number of groups different of expected. It is tempting to increase the LOD threshold until the number of groups reaches the expected for the species. This is a common mistake made by users. However, notice that, because it refers to a log function, every slight increase or decrease in LOD value turns the analysis very permissive or rigorous, causing error types I and II, respectively.

A fast, but still not the best solution for group issues was to make use of previous information from the genome, drafts, or past linkage maps. We made an algorithm that separates sequences according to the previous information and, after, tries to group the remaining markers using the group algorithm. The OneMap group algorithm tests the linkage of each marker with markers already in the sequences to decide if they are together.

With the presence of genotyping errors, every step of map building turns more difficult, because the genotype frequencies used to estimate two-points recombination fractions are not trustful. The two-points recombination fraction estimation does not allow to consider an error probability for each genotype, but the HMM does. Once the groups are defined, the HMM can consider an error probability for each genotype in its emission function to estimate the genetic distances. In other words, the HMM can consider the genotypes, not as discrete values such 0 ("aa"), 1("ab"), and 2 ("bb"), but continuous probabilities referring to the chance of each possible genotype to be the true genotype. However, the success of the

approach depends on accurate genotype probabilities estimations in steps upstream OneMap (Taniguti, 2017; Bilton *et al.*, 2018; Mollinari and Garcia, 2019; Mollinari *et al.*, 2020).

Another problem surrounding the high-throughput markers is the low-informativeness of the SNP markers. Their biallelic nature makes available only markers of types B3.7, D1.10 and D2.15 (see A.1). Depending on the population, marker types D1.10 and D2.15 are more frequent than B3.7. In this scenario, it is harder to find information to integrate parents' meiosis information. As consequence, the HMM and EM algorithms need to iterate in more phases possibilities until it reaches the best possible solution, which demands more computer efforts and includes more uncertainties in the analysis.

Some SNP calling software like GATK (Poplin *et al.*, 2017) and freebayes (Garrison and Marth, 2012) uses an assembly-based haplotyping method to search for the polymorphisms in the data sequences. As consequence, they provide phased markers in specific regions of the genomes where they could define a local haplotype. This information can be useful to increase markers informativity in the map building process. Also, it can reduce the possible phases to be estimated and consequently reducing the HMM computational efforts. Therefore, to solve the issues about genotyping errors and low-informativeness, we need to explore the bioinformatics steps upstream of the OneMap analysis.

With our own experience and with users' feedback we already saw that the genetic map itself can be an interesting tool to validate the upstream process because errors observed in the maps point to a dissociation of the data from the genetic concepts. OneMap provides tools to measure the quality of the built map. The heatmap color graphics of the recombination fraction can highlight outlier markers breaking the expected recombination pattern. We implemented new graphical tools to draw the estimated parents and progeny haplotypes and count the number of recombination breakpoints estimated. This new tool demanded a major modification in F_2 intercross algorithms for phase estimation (Attachment C). The new tool highlights the excessive recombination breakpoints estimated when something wrong happened in an upstream process, as contaminant individuals or genotyping errors. However, we need more tools for diagnostics in the entire pipeline (from read sequence to built linkage map) which are the steps affecting the map quality, or which are the ones that can bring solutions.

Even with diagnostic tools available for this upstream process, the dataset context, the software and parameters to test are too many and we can not make a single solution or recommendation. Mostly about the parameters, each software can have dozens of them and changing a single one can produce different results. Therefore, we developed in Chapter 2 the **Reads2Map** workflows. They perform the linkage map building from empirical or simulated sequencing reads datasets. The workflows are written in `Workflow Description Language` (WDL) (Voss *et al.*, 2017), which provides an organized, user-friendly and reproducible structure to all the analysis. With them, users can perform the simultaneous analysis using freebayes, GATK as SNP and genotype callers; updog, polyRAD, SuperMASSA as genotype callers; OneMap 3.0 and GUSMap as linkage map builders. The results can be vizualized in the shiny app Reads2MapApp, which contains graphical tools for diagnostics of the entire procedure to help users the select the best combination of software and parameters for their particular case.

In Chapter 3, we used the workflows developed in Chapter 2 to measure the impact of two of the major modifications of OneMap 3.0 in empirical and simulated data. One is the usage of genotype or error probabilities in the HMM to estimate the genetic distances. We compared the genotype probabilities profiles of the genotype callers and related them to the quality of the linkage map resulted. The second is the usage of haplotype-based multiallelic markers from the assembly-based SNP callers. We observed the capacity of each SNP caller to call these markers and compared the maps including them or not.

References

Bartholomé, J., E. Mandrou, A. Mabiala, J. Jenkins, I. Nabihoudine, C. Klopp, J. Schmutz,

C. Plomion, and J.-M. Gion, 2015c High-resolution genetic maps of Eucalyptus improve Eucalyptus grandis genome assembly. New Phytologist **206**: 1283–1296.

Baum, E., T. Petrie, S. G., and W. N., 1970 A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. The Annals of Mathematical Statistics **41**: 164–171.

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.

Chacon, S. and B. Straub, 2014 *Pro Git*. Apress, USA, second edition.

Csárdi, G., 2019 *cranlogs: Download Logs from the 'RStudio' 'CRAN' Mirror*.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and . G. P. A. Group, 2011b The variant call format and VCFtools. Bioinformatics (Oxford, England) **27**: 2156–2158.

Dempster, A. P., N. M. Laid, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39**: 1–38.

Dimensions, 2021 OneMap publications analytical views.

Eddelbuettel, D., 2013 *Seamless {R} and {C++} Integration with {Rcpp}*. Springer, New York.

Eddelbuettel, D. and J. J. Balamuta, 2017 Extending extit{R} with extit{C++}: A Brief Introduction to extit{Rcpp}. PeerJ Preprints **5**: e3188v1.

Eddelbuettel, D. and R. François, 2011 {Rcpp}: Seamless {R} and {C++} Integration. Journal of Statistical Software **40**: 1–18.

Garcia, A. A. F., E. A. Kido, A. N. Meza, H. M. B. Souza, L. R. Pinto, M. M. Pastina, C. S. Leite, J. A. G. da Silva, E. C. Ulian, A. Figueira, and A. P. Souza, 2006a Development of an integrated genetic map of a sugarcane (Saccharum spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. Theoretical and Applied Genetics **112**: 298–314.

Garcia, A. a. F., E. A. Kido, A. N. Meza, H. M. B. Souza, L. R. Pinto, M. M. Pastina, C. S. Leite, J. A. G. da Silva, E. C. Ulian, A. V. Figueira, A. P. Souza, J. A. G. da Silva, E. C. Ulian, A. V. Figueira, and A. P. Souza, 2006b Development of an integrated genetic map of a sugarcane (Saccharum spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. TAG. Theoretical and applied genetics **112**: 298–314.

Garrison, E. and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. ArXiv e-prints p. 9.

Gonen, S., N. R. Lowe, T. Cezard, K. Gharbi, S. C. Bishop, and R. D. Houston, 2014 Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. BMC Genomics **15**: 166.

Jeennor, S. and H. Volkaert, 2014 Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (Elaeis guineensis Jacq.). Tree Genetics & Genomes **10**: 1–14.

Lander, E. S. and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci USA **84**: 2363–2367.

Margarido, G. R. A., A. P. Souza, and A. A. F. Garcia, 2007 OneMap: software for genetic mapping in outcrossing species. Hereditas **144**: 78–9.

Merkel, D., 2014a Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker : a Little Background Under the Hood. Linux Journal **2014**: 2–7.

Mollinari, M. and A. A. F. Garcia, 2019 Linkage Analysis and Haplotype Phasing in Experimental Autopolyploid Populations with High Ploidy Level Using Hidden Markov Models. G3: Genes|Genomes|Genetics **9**: 3297–3314.

Mollinari, M., B. A. Olukolu, G. d. S. Pereira, A. Khan, D. Gemenet, G. C. Yencho, and Z.-B. Zeng, 2020 Unraveling the Hexaploid Sweetpotato Inheritance Using Ultra-Dense Multilocus Mapping. G3: Genes|Genomes|Genetics **10**: 281–292.

Oliveira, E. J., M. L. C. Vieira, A. A. F. Garcia, C. F. Munhoz, G. R. Margarido, L. Consoli, F. P. Matta, M. C. Moraes, M. I. Zucchi, and M. H. P. Fungaro, 2008a 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.

Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. V. der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks, 2017 Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv p. 201178.

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 .

Schiffthaler, B., C. Bernhardsson, P. K. Ingvarsson, and N. R. Street, 2017 BatchMap: A parallel implementation of the OneMap R package for fast computation of F1 linkage maps in outcrossing species. PLoS ONE **12**: 1–12.

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.

Taniguti, C. H., 2017 *Construção do mapa genético integrado em uma progênie de irmãos-completos proveniente do cruzamento entre Eucalyptus grandis e Eucalyptus urophylla*. Ph.D. thesis, Universidade de São Paulo, Piracicaba.

the software quality company TIOBE, 2021 TIOBE Index for January 2021.

Voss, K., J. Gentry, and G. V. D. Auwera, 2017 Full-stack genomics pipelining with GATK4+ WDL+ Cromwell [version 1; not peer reviewed]. F1000Research p. 4.

Wu, R., C.-X. C.-X. Ma, I. Painter, and Z.-B. Z.-B. Zeng, 2002d Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theoretical population biology **61**: 349–363.

Xiong, J. S., S. E. McKeand, F. Isik, J. Wegrzyn, D. B. Neale, Z.-B. Zeng, L. da Costa E Silva, and R. W. Whetten, 2016 Quantitative trait loci influencing forking defects in an outbred pedigree of loblolly pine. BMC genetics **17**: 138.

2 READS2MAP: PRACTICAL AND REPRODUCIBLE WORKFLOWS TO BUILD LINKAGE MAPS FROM SEQUENCING DATA

Abstract

The high-throughput genotyping methods make available millions of read sequences for genetics research. To build genetic maps from read sequences, researchers need to perform diverse bioinformatics and statistical analysis. For each step of the analysis there are several tools available, all with different methods and parameters to be selected by users. In this context, users struggle to find the best combination of software and parameters to be used and also to create reproducible pipelines for that. Workflows systems such as `Workflow Description Language (WDL)` offers a structure to organize the entire pipeline, producing a fixed structure and metadata of each step. It also guarantees reproducibility by making interface with containers environments, such as Docker. Here we present two workflows: **EmpiricalReads2Map** and **SimulatedReads2Map** to build linkage maps with empirical and simulated sequencing data. **SimulatedReads2Map** simulates sequencing data using SimusCoP for whole-genome sequencing (WGS) and exome DNA libraries; and RADinitio for restriction site associated DNA sequencing RADseq library. In both workflows, the analyses are performed using GATK, freebayes as SNP callers and genotype callers; updog, polyRAD, SuperMASSA as genotype callers; and OneMap and GUSMap as linkage map builders. We also present the shiny app Reads2Map to evaluate graphically the results of workflows. We exemplify their usage running the workflows and selecting the method to re-build a linkage map of *Populus tremula*. We obtained a linkage map with 6936 markers and 3299.961 cM selecting freebayes as SNP and genotype caller and OneMap as map builder.

Key words: Genetic maps; Workflow; Container; Reproducibility.

2.1 Background

The advances in sequencing technologies and the development of different library protocols make available millions of genetic markers able to genotype hundreds of samples in a single sequencing run. For building linkage maps, having more markers and larger sample size better is the capacity of locating where in the genome the recombination events occur (higher genetic map resolution). Also, markers from sequencing technologies overcome efforts with molecular laboratory methods and can be considered more affordable (Heather and Chain, 2016). However, it increases significantly the computer analysis efforts because of the need of using multiple analytic tools and their application to hundreds of experimental samples.

Once the samples are sequenced, many steps follow before being able to build a map. These steps can be divided as the preprocessing of reads, with demultiplexing and cleaning; the alignment to a reference genome; the SNP calling; the genotype calling; the filtering of markers; the grouping; ordering; phases and genetic distances estimations. For each of these steps, there are multiple software available and each software has specific parameters for each feature.

Users that are not familiar with bioinformatic pipelines may find difficult to run their analysis and find the best pipeline for their specific dataset. Performing separately each step can generate several independent scripts and unconnected intermediate files compromising its FAIR Data Principles (Findable, Accessible, Interoperable and Reusable - (Wilkinson *et al.*, 2016)).

Metascience studies highlight a big lack of reproducibility in science, sometimes even called as "reproducibility crises" (Munafò *et al.*, 2017), which have motivated research to adopt new approaches to overcome this concerning issue. For computational methods the lack of reproducibility can be measured by trying to run codes from five or ten years ago. The possibility of success is very low. In 2019, the

Nature journal (NATURE, 2019) even promoted the "Ten Years Reproducibility Challenge" launched by ReScience C journal (RESCIENCE, 2019) to researchers try to do that with their own codes.

During all the map building procedure there are some important parts with decisions that need to be made carefully according to the study of biological aspects, such as the choosen software and their specific parameters. However, other parts are only standard procedures as file conversions, software installation, and computational resources optimization, that can be a lot time-consuming and reduce the research quality if made in a wrong way (REITER *ET AL.*, 2020).

Workflow systems are a good solution to overcome technical issues, allowing users to focus on the important decision and also reach FAIR practices. It integrates the analysis in a single structure connecting each step by inputs and outputs. The steps, here called tasks, can also be combined in sub-workflows to users who want to run only specific parts of the analysis. Intermediate files are stored and organized with a standardized structure together with metadata and reports of parameters used in each task. Furthermore, conditional structures can be used to manage exceptions and options. The workflow organized structure makes it possible to generate flowcharts automatically given a complete overview of the workflow (REITER *ET AL.*, 2020).

Workflow systems also provide built-in tools to monitor and manage resource usage. The Cromwell Execution Engine can execute workflows on any computing platform (local, High Performance Computing - HPC or cloud) (VOSS *ET AL.*, 2017). Furthermore, the integration with containers like Docker (MERKEL, 2014b) and singularity (KURTZER *ET AL.*, 2017) overcome the need for software installation and offers specific software versions, making it significantly more reproducible.

`Snakemake` (GRÜNING *ET AL.*, 2018), `Nextflow` (DI TOMMASO *ET AL.*, 2017), `CWL` (AMSTUTZ *ET AL.*, 2016) and `WDL` (VOSS *ET AL.*, 2017) can be cited as four of the most widely used bioinformatics workflows system. The `Workflow Description Language` (`WDL`) presents advantages for production-level pipelines because of its capacity to deal with hundreds or thousands of samples (REITER *ET AL.*, 2020). It is successfully used by Genome Analysis Toolkit (GATK) team (VAN DER AUWERA *ET AL.*, 2013) to provide read-to-results **Best Practices** workflows that can be executed in the Terra platform (TERRA.BIO, 2020). With these optimized and tested workflows available, researchers do not need to learn the workflow system syntax to make use of its benefits.

In this work, we developed workflows to perform the analysis from read sequencing to linkage maps including some of the most used software for the SNP and genotype calling and map build: freebayes, GATK (POPLIN *ET AL.*, 2017; MCKENNA *ET AL.*, 2010) as SNP and genotype callers; updog (GERARD *ET AL.*, 2018), polyRAD (CLARK *ET AL.*, 2019), SuperMASSA (SERANG *ET AL.*, 2012) as genotype caller; OneMap 3.0 and GUSMap (BILTON *ET AL.*, 2018) as linkage map builders. The provided `WDL` workflow performs sequencing reads simulations based on user empirical library protocol and dataset. With this cross-platform tool, users can make optimized usage of their time and available computer resources to focus on the important decisions about the best software, algorithms, and parameters for their dataset. The simulation studies can validate the analytical methods applied and indicate which requires adaptations or improvements.

The genetic map itself can also be a powerful tool to validate upstream methods. Wrong decisions in any one of the upstream steps can be identified in the outputted map, once errors make the map proprieties dissociate of biological concepts. For example, genotyping errors can generate an inflated map size showing an excessive number of recombinations during the meiosis. Since the first genetic map studies by Sturtevant in 1915, it is observed that is unlikely that crossing-overs happen too close to each other, a phenomenon described as interference (STURTEVANT, 1915). Recent studies also described meiosis molecular mechanisms confirming the low expected number of recombination events during the meiosis (SMITH and NAMBIAR, 2020). The OneMap package to build linkage maps for inbred (RILs, F_2 intercross, and backcross) and outcrossing mapping populations (MARGARIDO *ET AL.*, 2007) have the

biological concepts of meiosis as assumptions and provides graphical tools to diagnose the dissociation between them.

Besides the proposed workflows, we also integrated them with interactive visualizations in a shiny app to visualize generated maps and intermediary results. The R package shiny (Chang *et al.*, 2020) allows to build interactive web pages, where the results can be visualized in real-time according to user-specifiable parameters.

2.2 Conclusions

The HMM approach implemented in software to build linkage maps is robust and able to return the best estimation possible of genetic distances. However, it is a lot of computer resources and time-consuming. Also, many different tools can be applied in the upstream processing of genetic map building with sequencing markers, such as the SNP and genotype calling. Testing every possible scenario to select the best pipeline for the specific dataset can be very difficult.

The workflows here provided offer a tool to test and select different combinations of software and parameters to build linkage maps from sequencing reads. The entire procedure is facilitated by the available configurations, Docker images, and graphical interface. Thus, users can focus on statistical and genetic aspects evolving the linkage map building instead of technical issues.

We consider the **SimulatedReads2Map** workflow more useful for developers because it provides an overview of the entire procedure and facilitates the search for logical errors in codes. The **EmpiricalReas2Map** is useful for users who want to select the best combination of software and parameters to build their genetic maps. To avoid users spend too much time with this pipeline selection, we suggest running the **map_emp** sub-workflow with a subset of data (e.g. a single chromosome). Once selected the pipeline, users can apply it for the entire dataset using genotyping4onemap, OneMap or GUSMap in R environment.

The shiny app Reads2MapApp guides the users through several quality criteria for each approach built genetic map. The main criteria to select the approach are the right color pattern in the recombination fraction heatmaps and the number of recombination breakpoints identified. The heatmaps highlight the grouping and ordering aspects of the map and the number of recombination breakpoints highlight the presence/absence of genotyping errors in the dataset.

We can validate all upstream procedures of obtention of molecular markers and their usage to genotype individuals if they can reproduce known genetic proprieties of linkage maps. The tools developed in this work provide an easy way to test protocols and software for molecular marker application studies.

References

Amstutz, P., M. R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, and E. al., 2016 Common Workflow Language, v1.0.

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.

Chang, W., J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson, 2020 *shiny: Web Application Framework for R.*

Clark, L. V., A. E. Lipka, and E. J. Sacks, 2019 polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. G3: Genes|Genomes|Genetics **9**: g3.200913.2018.

Di Tommaso, P., M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, 2017 Nextflow enables reproducible computational workflows. Nature Biotechnology **35**: 316–319.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping Polyploids from Messy Sequencing Data. Genetics **210**: 789–807.

Grüning, B., J. Chilton, J. Köster, R. Dale, N. Soranzo, M. van den Beek, J. Goecks, R. Backofen, A. Nekrutenko, and J. Taylor, 2018 Practical Computational Reproducibility in the Life Sciences. Cell Systems **6**: 631–635.

Heather, J. M. and B. Chain, 2016 The sequence of sequencers: The history of sequencing DNA. Genomics **107**: 1–8.

Kurtzer, G. M., V. Sochat, and M. W. Bauer, 2017 Singularity: Scientific containers for mobility of compute. PLOS ONE **12**: e0177459.

Margarido, G. R. A., A. P. Souza, and A. A. F. Garcia, 2007 OneMap: software for genetic mapping in outcrossing species. Hereditas **144**: 78–9.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**: 1297–1303.

Merkel, D., 2014b Docker: lightweight linux containers for consistent development and deployment. Linux journal **2014**: 2.

Munafò, M. R., B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie Du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis, 2017 A manifesto for reproducible science. Nature Human Behaviour **1**: 1–9.

Nature, 2019 Challenge to test reproducibility of old computer code.

Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. V. der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks, 2017 Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv p. 201178.

Reiter, T., P. Brooks, L. Irber, S. Joslin, C. Reid, C. Scott, C. T. Brown, and N. T. Pierce, 2020 Streamlining Data-Intensive Biology With Workflow Systems pp. 1–19.

Rescience, 2019 Ten years of reproducibility challenge.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS ONE **7**: 1–13.

Smith, G. R. and M. Nambiar, 2020 New Solutions to Old Problems: Molecular Mechanisms of Meiotic Crossover Control. Trends in Genetics **36**: 337–346.

Sturtevant, A. H., 1915 The behavior of the chromosomes as studied through linkage. Zeitschrift für induktive Abstammungs- und Vererbungslehre **13**: 234–287.

Terra.bio, 2020 Terra: Focus on your science. Available online at: https://app.terra.bio/ .

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, 2013 From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics **43**: 11.10.1–11.10.33.

Voss, K., J. Gentry, and G. V. D. Auwera, 2017 Full-stack genomics pipelining with GATK4+ WDL+ Cromwell [version 1; not peer reviewed]. F1000Research p. 4.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, 2016 The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**: 160018.

3 THE EFFECT OF CONSIDERING GENOTYPE PROBABILITIES AND HAPLOTYPE-BASED MULTIALLELIC MARKERS IN BUILDING LINKAGE MAPS WITH HIGH-THROUGHPUT GENOTYPING

Abstract

Linkage maps are important tools for genetic studies. Nowadays, with high-throughput genotyping, there are millions of genetic markers available to build a map. The markers coming from this technique have different characteristics compared with those commonly used, such as SSRs. The software used to estimate the linkage map needed adaptations to deal with this new scenario and overcome new issues. One of these issues is the genotyping errors generated by different sources such as sequencing errors, alignment, and PCR errors. Without proper analyses, it causes inflated genetic maps. One possible solution for that is to consider the genotype probabilities coming from the genotype caller in the Hidden Markov Model (HMM) used to estimate the genetic distances. However, to the best of our knowledge, there is not yet available a clear recommendation about which genotype probability would give the best results. Here we not only implemented this feature in OneMap 3.0 package but also used the **Reads2Map** workflows to test the impact of freebayes, GATK, updog, polyRAD and SuperMASSA genotype probabilities in the map build process, with empirical and simulated RADseq data. Each approach presented different results depending on the choice of the parameters. The results showed that OneMap 3.0 can build high-quality maps if i) the genotype caller did not make many mistakes and we use a global error rate of 5% or ii) if the genotype caller make mistakes but it also provides lower genotype probabilities for wrong genotypes. In our study, the scenarios that showed to be the most advantageous in terms of genotype probabilities used in the HMM were the freebayes as SNP caller, read counts from VCF, and a single error rate of 5% or with genotype probabilities from SuperMASSA and PolyRAD. They produced a denser map, closer to the expected recombination fraction matrix values and smaller size. The other issue found in building linkage maps with high-throughput markers is the low informativeness of the SNPs. Because of their biallelic nature, SNPs bring a lack of information for the ordering of the markers and the phase estimation in oucrossing species. The solution presented here is the usage of haplotype-based markers identified by the assembly-based haplotyping SNP callers freebayes and GATK. We also tested the differences of using or not these markers with the **Reads2Map** workflows. We observed improvements in the genetic distances and ordering in the presence of multiallelic markers, but also higher map inflation caused by these markers if they contain genotyping errors. The final approach selected was the freebayes as SNP and genotype caller, a global error probability of 5% and the presence of haplotype-based multiallelic markers. We re-built the *Populus tremula* chromosome 10 linkage group with 107 haplotypes-based multiallelic markers and 440 SNPs totaling 216.99 cM.

Key words: Linkage map; Genotyping error; Haplotype-based marker; RADseq.

3.1 Introduction

Since the first genetic map was built by Sturtevant in 1913, methodologies for building genetic maps have been improved to deal with different and more complex genomes, like those from outcrossing and polyploids species, and with different mapping populations. All efforts that have been done are motivated by the importance of this tool for genetic researches, since genetic maps provide valuable information.

Linkage maps are commonly applied to quantitative trait loci (QTL) studies. The QTL mapping estimates the genetic architecture of traits, which include mapping the molecular polymorphisms

responsible for variation in complex traits; determining their gene frequencies and their homozygous, heterozygous, epistatic, and pleiotropic effects in multiple environments (Mackay, 2001b).

Nowadays, with Breeding 4.0, QTL have also a great application for gene editing technologies (Wallace *et al.*, 2018). QTLs allow the identification of genes responsible for important phenotypes and candidates to be edited. The gene edition allows breeders to add new alleles from germplasm in breeding programs without lose the alleles of the well-established lines. As an example, Zsögön *et al.* (2018) applied gene editing of six important genes of tomato to give wild lines the yield and productivity similar to cultivated ones. Previous QTL studies needed to be made to identify these six genes effects (Ashrafi *et al.*, 2012; Frary *et al.*, 2000).

Furthermore, it has becoming more usual to apply linkage maps to solve genome assembly issues (Fierst, 2015). For many species, especially non-model organisms with large genome size, it is still unaffordable to obtain reliable reference genome assemblies. For many species, only draft genomes are available with thousands of sequenced segments (contigs) and very limited information on how these can be assembled into chromosome sequences. Linkage maps can give content for ordering, orienting, positioning and phasing linked sequences (Pengelly and Collins, 2018).

The growing accessibility of sequencing technologies is providing genome information for great understandings about its structures, and how the genome diverges between individuals. The high-throughput genotyping technology to obtain markers can provide ultra-dense genetic maps, which are promising for advances for association studies. However, it also brings challenges for genetic map building procedures. Markers now are identified on large scale automatically, without the need for handwork, which provides clear advantages to produce low-cost markers but they also have more genotyping errors. Especially when high-throughput sequencing technology is applied to reduced representation libraries (RADseq), which have an excess of duplicated sequenced sequences starting and finishing at the same point of the genome. In RADseq data, the most of duplicated sequences are products from PCR, which include errors difficult to be identified by bioinformatics tools (der Auwera *et al.*, 2020; Rivera-Colón *et al.*, 2020).

The excess of genotyping errors in datasets have been generated inflated genetic maps, with unrealistic genetic distances (Smith and Nambiar, 2020), once each genotype error is considered an extra recombination event. Studies have been made to search the sources of errors and correct them in bioinformatics and statistical methods for SNP and genotype calling procedures (Hackett and Broadfoot, 2003; Rivera-Colón *et al.*, 2020; Gerard *et al.*, 2018; Clark *et al.*, 2019). The sources of errors can also be considered in map building algorithms to provide reliable ultra-dense genetic maps (Bilton *et al.*, 2018; Mollinari and Garcia, 2019).

Another characteristic of high-throughput genotyping is the low-informativeness that comes with the biallelic codominant nature of SNPs. In populations coming from inbred lines, this characteristic does not have a significant impact, but, in outcrossing species, this brings as consequence difficulties to integrate recombination from both parents and to ordering the markers. The low-informativeness of non-integrated genetic maps brings limitations in further QTL analysis of multiallelic traits (Gazaffi *et al.*, 2020).

One possible solution to solve low-informativeness is to use assembly-based haplotyping SNP callers in previous steps of map building, such as PolyBayes (Marth *et al.*, 1999), samtools (Li, 2011), freebayes (Garrison and Marth, 2012), GATK (Poplin *et al.*, 2017; McKenna *et al.*, 2010) and Platypus (Rimmer *et al.*, 2014), which combine close located biallelic markers into haplotypes and output them as phased markers or multiple nucleotide markers (MNP).

The available algorithms in OneMap (Margarido *et al.*, 2007) can build integrated genetic maps and generate the linkage group haplotypes of each individual in the mapping population. To estimate linkage phases, OneMap uses a multipoint approach with a Hidden Markov Model (HMM)

method, but it requires high computational resources, time to process, and it is not efficient if there are many genotyping errors and not at least a few informative markers. In this work, we adapted OneMap algorithms to deal with high-throughput markers. Using the **Reads2Map** workflows (chapter 2), we tested the consequences of considering genotype error probabilities and haplotype-based multiallelic markers to build linkage maps from sequencing reads. The workflows allowed the comparision using freebayes, GATK (Poplin *et al.*, 2017; McKenna *et al.*, 2010) as SNP and genotype callers; updog (Gerard *et al.*, 2018), polyRAD (Clark *et al.*, 2019), SuperMASSA (Serang *et al.*, 2012) as genotype caller; OneMap 3.0 and GUSMap (Bilton *et al.*, 2018) as linkage map builder.

3.2 Conclusion

OneMap 3.0 offer to users the possibility to read and consider in the genetic map building the genotype probabilities and haplotype-based multiallelic markers information from the input files (OneMap format or VCF file). The success of genetic map building will be proportional to the quality of the information provided by upstream procedures such as library preparation, SNP and genotype calling, genotype probabilities estimation, and the combination of SNPs into haplotype-based makers. As the upstream procedures for genotyping and identification of haplotype-based multiallelic markers are improved, updates can be easily made in **Reads2Map** workflows, and no modification is required in OneMap 3.0 algorithms, once only the values in the standard VCF file changes.

Only recently, it was developed a software able to simulate RADseq reads, however it still presented a higher number of duplicated sequences and probably more genotyping errors than empirical datasets. Also, it could not reproduce the indels and outlier markers. These differences can limit the possible relation between empirical and simulated results in this study. We can also update **Reads2Map** workflows as RADinitio or other simulation software are improved.

The relation between genotyping errors and genetic map quality was highlighted by our simulation results. We considered high-quality maps those with the expected pattern of heatmap graphics of recombination fraction matrix and number of recombination breakpoints (the consequence of the genetic distance between markers). We could get high quality genetic maps if they were built with datasets with few genotyping errors and single error rate as observed in the genetic map built with freebayes as SNP and genotype caller with a single error rate of 5%. Also, it is possible to obtain high-quality genetic maps with a higher number of genotyping errors if the provided genotype probabilities correct differentiate wrong and right genotypes, as presented by SuperMASSA software in the scenario with freebayes as SNP caller, or by polyRAD when GATK is the SNP caller and VCF is the read counts source.

The simulations also showed consequences of some technical issues, that could be easily changed to significantly increase the final genetic map built by the approach. The OneMap 3.0 new features implemented were validated using the simulation results. OneMap 3.0 is still not able to consider the parents' genotype probabilities in the HMM, thus, it is important to only add in the map building procedure high-quality parents' genotypes.

The usage of haplotype-based multiallelic markers showed to improve significantly the ordering and map distance estimation. They are also important for the integration of both parents' genetic maps. However, they can also add more genotyping errors to the analysis making editions needed for the outlier markers removal.

References

Ashrafi, H., M. P. Kinkade, H. L. Merk, and M. R. Foolad, 2012 Identification of novel quantitative trait loci for increased lycopene content and other fruit quality traits in a tomato recombinant

inbred line population. Molecular Breeding **30**: 549–567.

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.

Clark, L. V., A. E. Lipka, and E. J. Sacks, 2019 polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. G3: Genes|Genomes|Genetics **9**: g3.200913.2018.

der Auwera, G. A., G. van der Auwera, and B. D. O'Connor, 2020 *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Incorporated.

Fierst, J. L., 2015 Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Frontiers in Genetics **6**: 1–8.

Frary, A., T. C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, J. Liu, J. Meller, R. Elber, K. B. Alpert, and S. D. Tanksley, 2000 fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science (New York, N.Y.) **289**: 85–88.

Garrison, E. and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. ArXiv e-prints p. 9.

Gazaffi, R., R. R. Amadeu, M. Mollinari, J. R. B. F. Rosa, C. H. Taniguti, G. R. A. Margarido, and A. A. F. Garcia, 2020 fullsibQTL: an R package for QTL mapping in biparental populations of outcrossing species. bioRxiv .

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping Polyploids from Messy Sequencing Data. Genetics **210**: 789–807.

Hackett, C. a. and L. B. Broadfoot, 2003 Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity **90**: 33–38.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**: 2987–2993.

Mackay, T. F. C., 2001b The Genetic Architecture of Quantitative Traits. Annual Review of Genetics **35**: 303–339.

Margarido, G. R. A., A. P. Souza, and A. A. F. Garcia, 2007 OneMap: software for genetic mapping in outcrossing species. Hereditas **144**: 78–9.

Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. D. Hillier, P. Y. Kwok, and W. R. Gish, 1999 A general approach to single-nucleotide polymorphism discovery. Nature Genetics **23**: 452–456.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**: 1297–1303.

Mollinari, M. and A. A. F. Garcia, 2019 Linkage Analysis and Haplotype Phasing in Experimental Autopolyploid Populations with High Ploidy Level Using Hidden Markov Models. G3: Genes|Genomes|Genetics **9**: 3297–3314.

Pengelly, R. J. and A. Collins, 2018 Linkage disequilibrium maps to guide contig ordering for genome assembly. Bioinformatics pp. 1–5.

Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. V. der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks, 2017 Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv p. 201178.

Rimmer, A., H. Phan, I. Mathieson, Z. Iqbal, S. R. Twigg, A. O. Wilkie, G. Mcvean, and G. Lunter, 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics **46**: 912–918.

Rivera-Colón, A. G., N. C. Rochette, and J. M. Catchen, 2020 Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. Molecular Ecology Resources pp. 1–16.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS ONE **7**: 1–13.

Smith, G. R. and M. Nambiar, 2020 New Solutions to Old Problems: Molecular Mechanisms of Meiotic Crossover Control. Trends in Genetics **36**: 337–346.

Wallace, J. G., E. Rodgers-Melnick, and E. S. Buckler, 2018 On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. Annual Review of Genetics **52**: 421–444.

Zsögön, A., T. Čermák, E. R. Naves, M. M. Notini, K. H. Edel, S. Weinl, L. Freschi, D. F. Voytas, J. Kudla, and L. E. P. Peres, 2018 De novo domestication of wild tomato using genome editing. Nature Biotechnology **36**: 1211–1216.

4 FINAL CONSIDERATIONS

Here we updated the OneMap package and presented user-friendly and reproducible workflows to build linkage maps from read sequencing. The procedure in the **Reads2Map** workflows starts with the filtering of the reads, then the alignment, SNP and genotype calling, and the map building. Several software are considered. The shiny app Reads2MapApp helps users to select the software and parameter combinations that result in the best linkage maps for their specific dataset.

OneMap 3.0 have updates to improve the speed and quality of the built maps, and also quality diagnostic graphic tools. The OneMap 3.0 major modifications were the implementation of variable genotype probabilities in the HMM to estimate the genetic distances and the possibility of include haplotype-based multiallelic markers generated from sequencing technologies. All OneMap 3.0 updates were validated by application in empirical and simulated datasets in **Reads2Map** workflows.

Using genotype probabilities from different software in the HMM for estimating the genetic distances are efficient depending on the dataset characteristics and upstream methods applied. We could not make a single suggestion for all possible dataset profiles, but with the workflows **Reads2Map**, users can try easily all possibilities and select the best for each dataset. The **Reads2Map** flexibility and an organized structure make easy its adaptation if some of the software considered receives updates or to include other software.

The presence of haplotype-based multiallelic markers improves the map quality and helps the building map procedure. However, it can also bring more errors. It is necessary an edition of the maps built by the workflows to remove the outlier markers, which deviates from the expected pattern of recombination fraction and segregation.

Because the quality of the genetic maps depends on all upstream applied approaches, the workflows **Reads2Map** combined with OneMap 3.0 are powerful tools to build linkage maps with markers from sequencing technologies.