(This page has been intentionally left blank for print formatting purposes.)

Marche Polytechnic University
Department of Agricultural, Food and Environmental Sciences
Scientific field: AGR/07 - plant genetics

PhD School of Agricultural, Food and Environmental Sciences
XVI cycle (2014-2017)

The application of reduced-representation sequencing techniques for studying the structure of plant populations: A case study in common bean (*Phaseolus vulgaris* L.)

PhD supervisor:
Prof. **Roberto Papa**

PhD school director:
Prof. **Bruno Mezzetti**

PhD candidate:
Debora Santo

(This page has been intentionally left blank for print formatting purposes.)

This dissertation is dedicated
to those who have supported me in pursuing my aspirations
no matter how unattainable they sometimes might have appeared to be;
to those who have always met me with love, kindness, and understanding;
to my family, friends, and colleagues, who helped me grow as a person.

"To laugh often and much;
to win the respect of intelligent people and the affection of children;
to earn the appreciation of honest critics
and to endure the betrayal of false friends.
To appreciate beauty; to find the best in others; to leave the world a bit better
whether by a healthy child, a garden patch, or a redeemed social condition;
to know that even one life has breathed easier because you have lived.
This is to have succeeded."
- Ralph Waldo Emerson -

"Even if I understand all mysteries and all knowledge,
and if I have all faith so that I can move mountains,
but do not have love, I am nothing."
- 1 Cor 13:2 -

(This page has been intentionally left blank for print formatting purposes.)

Acknowledgments

First of all, I would like to thank all the members of Marche Polytechnic University's plant genetics lab, who I have spent the past three years working with and learning from during the production of this dissertation, including Prof. Roberto Papa, Prof. Laura Nanni, Dr. Elisa Bellucci, Dr. Elena Bitocchi, Dr. Barbara Cerquetti, Ana Velimirović, Antonia Mores, Valerio Di Vittori and Eda Bozkır. I am grateful for the PhD program that I could be a part of, as well as all the colleagues who I was fortunate to meet and share experiences with during my studies. I especially thank Prof. Roberto Papa for his supervision, guidance, and help in navigating through the educational opportunities and collaborations that were established through trainings, conferences and study visits abroad.

Next, I want to thank Prof. Giovanna Attene, Dr. Domenico Rau, Dr. Monica Rodriguez and Dr. Maria Leonarda Murgia from the University of Sassari, for their work on the common bean population development presented in this dissertation, as well as Prof. Paul Gepts from UC-Davis for providing the parental line that enabled the forming of the plant population for this study.

I am also grateful to Prof. Jacques David and the members of his lab at the SupAgro school of the French National Institute for Agricultural Research (INRA) in Montpellier, France, for welcoming me for a three-months-long internship within their team, sharing their knowledge and tools for producing the genotyping data presented in this dissertation. I owe special thanks to Dr. Sylvain Santoni, as the lab manager, Audrey Weber, the lab technician and Dr. Jean-François Martin for his help in starting the bioinformatic analyses.

I am deeply thankful to Prof. Mario Enrico Pè and Dr. Matteo Dell'Acqua from the Sant'Anna School of Advanced Studies in Pisa, Italy, for having their help in the major part of the data analysis and the chance to learn valuable

programming skills during a one-week training in Pisa. Without the collaboration with Matteo, this dissertation would not have the same quality.

Furthermore, I would like to thank Prof. Scott Jackson for enabling an eight-months-long stay in his lab at The Center for Applied Genetic Technologies of University of Georgia in Athens, GA, Jennifer Leverett for her kindness and excellent lab management, Dr. Chunming Xu for advice on bioinformatics analysis, Dr. Jin Hee Shin for help in lab work, Dr. Dongying Gao for moral support, and professors Soraya and David Bertioli, as well as Brian Nadon, Carolina Ballen Taborda and Carolina Chavarro for their feedback on my research, and all staff and students at UGA for building a supportive learning environment for the growth of young scientists there.

I want to also thank for the assistance in GBS genotype calling provided by Dr. Alberto Ferrarini and Dr. Salvatore Benfatto from the University of Verona, Italy, and the feedback on my research and writing provided by the two external dissertation referees, Dr. Maria Carlota Vaz Patto from The António Xavier Institute of Chemical and Biological Technology (ITQB NOVA) based in Oeiras, Portugal, and Prof. Vladimir Meglič from the Agricultural Institute of Slovenia (KIS), Ljubljana, Slovenia.

I would like to use this opportunity to thank my family for enabling my pursuit of education, especially as the first academic within my family. I must thank my mother and father for their love, understanding, and support in the decisions that I needed to make to get this far in my education and also so far from home in order to reach it; my sister for being my cheerful childhood companion and creative inspiration in the family; and my grandparents and aunt for their wise advice. Along with the rest of my family, I am grateful to them for teaching me to always be kind, hardworking and stay courageous.

I cannot forget to thank all my friends for helping me maintain my sanity when life was hard, for their listening ears, hugs, and advice, without which I might

have had given up along the way. My school friends: Marija, Ivana, and Andrea; my university friends: Mira, Mlađo, Marija, and Ilija; the friends I've met in Italy: Katarina, Emel, Musab, Marta, Vittorio, Bud and Marouane; the ones that made my stay in France more joyful: Asya, Vladimir and Stefania; the ones I've met in the US: Cecilia, Caro, Liz, Silas, Ivy, Brian, Sarah, Badie, Leticia and Dave. I consider my friends my extended family and see them as amazing people who enriched my life beyond my imagination. My life would not be nearly as happy and eventful without them. Thank you all so much!

15.02.2018.
Kikinda, Serbia

(This page has been intentionally left blank for print formatting purposes.)

Dissertation abstract and structure

The dissertation is comprised of two chapters. The first chapter is a literature review on the application of the two of the most frequently used reduced-representation sequencing techniques: RADseq (RAD sequencing; Restriction-site Associated DNA-sequencing) and GBS (Genotyping-by-Sequencing). Both techniques are based on the use of restriction enzymes and were developed primarily for genome-wide discovery of single-nucleotide polymorphism (SNP) markers. These methods have already been used in hundreds of studies to investigate the genetic diversity of large numbers of individuals under lowered cost and have shown to perform well in a wide range of research applications. Here, the focus is on their utility for assessing genetic diversity in plants, as the use of these methods in plants can have specific challenges due to genome size, complexity, polyploidy and amount of repetitive sequences, but also provides an advantage of studying big numbers of samples with large genomes at a relatively low price.

The second chapter presents experimental research performed on a common bean (*Phaseolus vulgaris* L.) population developed to segregate for the traits of the domestication syndrome. The population's genomic composition is assessed and described, and its value for research and breeding discussed. The population was genotyped using GBS and the results analyzed to show the structure and diversity of the population, linkage decay, the observed heterozygosity, introgressions, and recombination breakpoints. Its utility for QTL mapping purposes is shown for the traits of flower color and flowering time, while the results for pod related traits of the domestication syndrome. for which the population was developed, are studied and presented separately.

(This page has been intentionally left blank for print formatting purposes.)

Astratto e la strattura dalla tesi

La tesi è composta da due capitoli. Il primo capitolo è una revisione della letteratura sulle due tecniche di sequenziamento a rappresentazione ridotta con: il RAD sequencing (sequenziamento del DNA associato al sito di restrizione) e il GBS (Genotyping by Sequencing). Entrambe sono basate sull'uso di enzimi di restrizione e sviluppate principalmente per la scoperta di marcatori associati a polimorfismi a singolo nucleotide (SNP) su genoma. Queste technique sono già state utilizzati in centinaia di studi per analizzare la diversità genetica di individui numerosi a costo ridotto, e hanno dimostrato di funzionare bene in una vasta gamma di applicazioni di ricerca. Qui, l'attenzione è sulla loro utilità nella valutazione della diversità genetica nelle piante, in quanto l'uso può essere soggetto a sfide specifiche, dovute a dimensioni del genoma, complessità, poliploidia e quantità di sequenze ripetitive, ma offre nello stesso tempo il vantaggio di poter studiare un grande numero di campioni con genomi di grandi dimensioni ad un prezzo relativamente basso.

Il secondo capitolo è un lavoro di ricerca su una popolazione di fagiolo (*Phaseolus vulgaris* L.) sviluppata per segregare per i tratti della sindrome della domesticazione. Abbiamo valutato, descritto e discusso la composizione genomica della popolazione e il suo valore per scopi di ricerca e di breeding. La popolazione è stata genotipizzata utilizzando GBS ed i risultati sono stati analizzati per determinare la struttura e la diversità della popolazione, il decadimento LD, l'eterozigosi osservata, leintrogressioni e i punti di rottura della ricombinazione. E' mostrata l'utilità della popolazione per mappare QTL relativi ai caratteri colore del fiore e tempo di fioritura, mentre i risultati relativi ai caratteri legati agli effetti della domesticazione sul baccelo, per i quali la popolazione è stata sviluppata, sono studiati e presentati separatamente.

(This page has been intentionally left blank for print formatting purposes.)

Table of Contents

Acknowledgments VI

Dissertation abstract and structure X

Astratto e la strattura dalla tesi XII

Table of Contents XIV

Dissertation introduction 18

CHAPTER 1: Review of the application of reduced representation sequencing techniques in plant genetic diversity studies 20

Abstract 22

Introduction 24

Domestication 24
Conservation of plant genetic resources in genebanks 26
Facing agricultural challenges with next-generation sequencing 27
The journey towards molecular breeding in agriculture 28
Molecular markers 29

Reduced representation sequencing 30

Classification of genome resequencing methods 30
The value of genotyping large numbers of individuals with GBS 32
Specific strengths of GBS in plants 33
General steps of the GBS protocol 35
Protocol modifications 35
Reviews on RRS and GBS 39

Conclusion and prospects 46

CHAPTER 2: Genomic characterization of a biparental common bean (*Phaseolus vulgaris* L.) population segregating for the traits of the domestication syndrome 50

Abstract 52

Introduction 54

Study aims 54
The origin and domestication of crop plants 55
The common bean (*Phaseolus vulgaris* L.) and its domestication 57
Plant study population 60

Material and methods 61

Population development 61
Phenotyping 64
Genotyping 65
GBS library sequencing 67
SNP marker discovery and population genotyping 67
SNP dataset quality and filtering 68
SNP density 70
Heterozygosity 70
Population structure 71
Linkage decay analysis 71
Genome composition and introgression detection 72
QTL mapping 73

Results 75

Sequencing quality 75
SNP marker yield 76
Dataset assessment before and after filtering 77
SNP density 79

Gene density 82
Heterozygosity 84
Population structure 89
Linkage decay analysis 92
Genome composition and introgression detection 93
QTL mapping 100

Discussion 103

The population design 103
The GBS genotyping design 104
Sequencing quality 105
SNP markers: yield, density, filtering 106
SNP marker heterozygosity 107
Gene density 108
Population structure 108
Linkage decay analysis 108
Genome composition and introgression detection 109
The power of the population for QTL mapping 112

Conclusions 114

References 116

Appendices 138

Appendix 1: GBS library nested multiplexing primers 138
Appendix 2: GBS NGS library preparation protocol with *ApeKI* 140
Appendix 3. Missing data and heterozygosity plots, filtered dataset 146
Appendix 4. Estimated introgression segment length 150
Appendix 5. Phylogenetic tree 151
Appendix 6. LD heatmaps and LD evolution over chromosomes 152

(This page has been intentionally left blank for print formatting purposes.)

Dissertation introduction

The research presented in this dissertation describes the study outcomes of a three-year-long doctorate program at Marche Polytechnic University in Ancona, Italy. The doctorate encompassed a bibliography review on genotyping-by-sequencing (GBS) as a method for inexpensive genotyping of large numbers of plant samples for assessment of genetic diversity through the discovery of genome-wide single-polymorphism (SNP) markers, as well as on the common bean (*Phaseolus vulgaris* L.), as the plant of choice for the experimental section of the dissertation. The occurrence of dual domestication events makes it an excellent choice for evolutionary studies of domestication syndrome traits.

During the first year, the candidate has spent a three-month long study internship period abroad at SupAgro, INRA in Montpellier, France. The GBS library preparation described in the second chapter was conducted during that stay.

The second year was dedicated to methodology development, conducting research and data analysis. During this period, the candidate has attended a one-week training in bioinformatic data analysis at the Sant'Anna School of Advanced Studies at Pisa, Italy, where collaboration was provided for analyzing the GBS data presented in the second chapter.

The final year was dedicated to data analysis and dissertation writing. During this year, a review article on the common bean was published with the doctoral candidate as a coauthor (Bitocchi et al., 2017) and an eight-month-long study period abroad at the University of Georgia in Athens, GA, US as a visiting research scholar. There a 3'-tag RNA-sequencing (RNA-seq) approach was compared to the classic RNA-seq library preparation method, with a training in bioinformatic data analysis.

(This page has been intentionally left blank for print formatting purposes.)

CHAPTER 1
Review of the application of reduced representation sequencing techniques in plant genetic diversity studies

(This page has been intentionally left blank for print formatting purposes.)

Abstract

This aim of this review is to give an overview of restriction enzyme (RE) based reduced representation sequencing (RRS) techniques and their use for genome-wide single-nucleotide polymorphism (SNP) discovery for the purposes of studying the genetic diversity of diverse plant material collections. Genotyping is performed for assessing the genetic diversity of conserved plant accessions so that knowledge can be used for making well-informed decision on how to manage their storage, remove duplicates, add more accessions of diverse origin and especially to determine which might be of use for specific breeding program purposes, based on some specific allele variants they might carry. Without a detailed knowledge of the diversity of what we have at hand, we might miss out on the potential impact these varieties can have on agricultural crop improvement.

The focus here is on the two most widely used approaches, Restriction-site Associated DNA-sequencing (RADseq; Baird et al., 2008) and Genotyping-by-Sequencing (GBS; Elshire et al., 2011), as well as the possible modifications for optimizing their use for diverse study organisms and aims. This review gives an overview of all RRS approaches developed so far, outlines the palette of those utilizing REs for genome reduction, shows the different study aims for which RADseq and GBS have been used so far, establishes what potential these methods have brought to research and which biases we need to take care of eliminating from the studies, and finally, summarizes the opinions on the application and potential of these techniques as given in both research articles and extensive simulations and reviews published so far.

(This page has been intentionally left blank for print formatting purposes.)

Introduction

The developments in sequencing technologies paired up with our growing knowledge in functional genetics enable us to explore the genome content of crop plants in more detail and with a better understanding than ever before, and further advancements are still being achieved. To be fully aware of the significance this has for human-kind, we must acknowledge the crucial role agricultural plants have in feeding the world, the challenges agriculture is facing and how we can best address them, how domestication has affected the genetic diversity of our crop plants and why it is so important for humans to find ways to find effective ways to preserve plant biodiversity, especially in the form of plant genetic resources (PGRs) conserved and maintained *ex-situ* in genebanks worldwide.

Domestication

The close interdependent relationship of humans and crop plants started when humans began transitioning from the hunter-gatherer lifestyle to forming settlements and starting to grow plants to meet their needs, mainly for having more stable sources of food. The "Origin of Agriculture" or the "Neolithic Revolution" is usually said to have started around 10,000 years ago, but according to more recent findings, it may have started even 30,000 years ago (Allaby et al., 2017). As only material from a limited number of plants was used in this endeavor, the founder effect limited the genetic variation captured from the wild crop ancestors (Doebley et al., 2006; Smith, 2006). Not knowing the mechanisms by which the characteristics of plants were controlled, farmers have long continued selecting plants based only on observable traits, saving the seeds of plants that they preferred over others and planting them in the next sowing season. Over time, this has led to local adaptations of these plants with the anthropogenic selective pressure being added to the ones imposed by the environment. Among the most important traits that were

selectively improved by humans in most crop plants were seeds and fruit size, plant form with determined apical growth, loss of seed dormancy, loss of photoperiod sensitivity, etc. (Doebley et al., 2006). Sometimes, the human need was in conflict with characters developed and maintained during the evolution of the wild crop relatives. A good example for this is the loss of seed dispersal mechanisms, which enabled the seeds to be harvested from plants and prevented yield loss. Rice is an example where the loss of seed shattering comes from a new mutation, not present in the wild progenitor, but instead, it happened after domestication and got selected for in the domesticated genetic pool (Li, 2006). The presence of these traits characteristic for the domesticated crops that result from human selection was first coined as "adaptation syndrome" by Harlan et al. (1973), but the term "domestication syndrome" was adopted for further use, as more intuitive by Hammer (1984).

As crops were mostly grown on small farms, additional genetic loss followed, due of the small effective population size of these crops in fields (Eyre-Walker et al., 1998). Transitioning from traditional to modern agriculture during the "Green Revolution" in the late 1900's, even though the aim of increasing the yield for certain crops was a noble one, overall, it has led to a loss of genetic diversity of the targeted crops, as the few newly created improved cultivars gained popularity and wide use in uniform monocultural fields (Pingali, 2012). Now, the aims have shifted towards improving the crops' diversity through the inclusion of alleles that were left behind during domestication, but are still preserved in old landraces and wild crop relatives (CWRs; Castañeda-Álvarez et al., 2016). Creating crosses of elite cultivars with multiple different wild relatives or landraces, we will be able to select for improved traits among the progeny while trying not to lose the existing good qualities and adaptation of the elite lines to agricultural ecosystems. This approach aims to alleviate the negative effects of the several bottlenecks our crops have gone through in the past, by using the natural variation still preserved in unadapted germplasm.

Conservation of plant genetic resources in genebanks

If we want to understand the genetic diversity and use the full potential of these wild relatives and landraces that contain valuable allele forms that were left behind during domestication, we need to make sure that we preserve as many of them as possible, especially under the pressure of modern agriculture, where to achieve high yield, often a small number of elite cultivars is chosen for being grown as monocultures. As old varieties often do not have the ideal combinations of as many traits as the newer ones, they are becoming less frequently used and propagated, and as *in situ* preservation is becoming more difficult to achieve and catalogize, so active collecting and dynamic storage and propagation of these plant accessions is needed (Dempewolf et al., 2017; Khoury et al., 2010; Kilian & Graner, 2012).

The first person to have noticed the value that CWRs can have in agriculture improvement and who actively started collecting them was Vavilov (Dvorak et al., 2011; Harlan, 1992; Loskutov, 1999; Vavilov, 1926, 1992). This was the motivation for establishing the first genebanks for *ex situ* conservation of "exotic germplasm" in countries worldwide (Dempewolf et al., 2017), while *in situ* conservation of plant diversity also has high significance, whenever applicable. There is a movement now in uniting the information across different genebanks, so IPK-Gatersleben, for an example, hosts and maintains the European Cooperative Programme for Plant Genetic Resources' (ECPGR) Eurisco catalogue, one of the biggest of its kind that unifies information about plant material currently stored in and available from gene and seed banks throughout Europe, while the Agricultural Research Service (ARS), Bioversity International and the Global Crop Diversity Trust works on developing an online information management system for plant genebanks world-wide, based on the already existing Germplasm Resources Information Network's (GRIN) National Plant Germplasm System (NPGS). Good sharing,

information and plant material storage practices will aid better decision making in conservation and provide more efficient practices.

The plant materials stored in genebanks or seedbanks are called plant genetic resources (PGRs) or plant genetic resources for food and agriculture (PGRFA), after the main motivations for their conservation. They were defined at the International Undertaking on Plant Genetic Resources conference (FAO, 1983) of The United Nations' Food and Agriculture Organization (FAO). According to FAO, PGRs are reproductive or vegetative propagating material of cultivars, landraces, CWRs and special genetic stocks (which include elite lines and mutants). As stated by the United Nation's Convention for Biological Diversity (CBD, 1992), PGRs are "any living material of present and potential value for humans". Sometimes also, genes or DNA and RNA fragments can be considered and stored as genetic resources.

Of course, *ex-situ* conservation is not the only way for preserving diversity, but by providing systematic and thorough information about stored accessions, it can be the most useful in agricultural improvement purposes.

Facing agricultural challenges with next-generation sequencing

Facing the estimated growth of the human population and upcoming climate changes, concerns about the necessary yield increase to feed the world is becoming primary for farmers, the seed industry and agricultural scientists in this century (Gerland et al., 2014; McCouch et al., 2013; World Population Prospects, 2015). The limit of land that can be utilized for crop cultivation is being reached and is endangering the preservation of naturally occurring habitats, with now about 38% of all terrestrial surface being under agricultural use (Foley et al., 2005). As we cannot keep extending the agricultural lands, there is a crucial need for innovative approaches in plant cultivation methods and cultivar improvement for increasing yield and lowering losses due to pests and disease.

As already mentioned, a large untapped genetic potential already tested throughout evolutionary time lays in PGRs (Fernie et al., 2006; Gur & Zamir, 2004; Zamir, 2008). The technological advancement in next-generation sequencing (NGS; Wetterstrand, 2012; Mardis, 2011) and development of reduced representation sequencing (RRS) methods led to a great reduction in costs of genotyping and made genomic screening affordable for large numbers of individuals. This has the biggest value for plant breeders who can now more easily produce genotype data for their breeding populations and for researchers wanting to produce genome-wide data on PGRs they are storing, to be more aware of the genetic diversity stored in genebanks. The knowledge gained from finding out the genomic composition and functions of sequences within genomes of our crop species has a large potential for speeding up aimed crop improvement in modern agriculture in order to produce higher yielding and more resilient plants with products of higher nutritional value.

The journey towards molecular breeding in agriculture

Ever since the initiation of domestication of crop species, the consequences of the founder-effect bottleneck and the anthropogenic selective pressure were added to the natural evolutionary forces, shaping and mostly narrowing the genetic diversity of plants cultivated by humans. While in classical plant breeding, the decisions of selection were limited to the observable phenotypic characters of plants, as described earlier, with relatively recent advances in science, the possibility of using molecular markers in breeding has significantly increased the efficiency and precision of screening crop genes and genomes and selecting better performing cultivars via marker-assisted selection (MAS He et al., 2014).

However, in MAS, usually only a few markers are used and that limits research possibilities, especially for complex traits, of which the yield is one of the most often targeted traits for improvement. However, the yield is also sometimes the hardest to achieve stability for due to the dependency on

epistatic interactions with other elements of the genome and environmental interactions. In general, we see that MAS has not fulfilled the initial expectations researchers had for it, but also that some of the limitations that were not accounted for can be overcome and the outcome improved when using newer technologies and including newer knowledge from recent genomic research (Cossio et al., 2010).

While several markers might be sufficient to study the inheritance of qualitative or quantitative traits strongly governed by a major gene, we see that the inheritance of many agronomically important traits, like yield and resistance to abiotic and biotic stress factors, is more complex and is usually the sum of the effects of many minor effect quantitative trait loci (QTL; Collins et al., 2008; Varshney et al., 2011). Combining this awareness with the fact that our knowledge of genomic constitution and genetic variability within many crop species is still rather limited, we can conclude that the possibility of observing genome-wide sequence-based markers seems like the best approach for many future plant breeding applications, especially in the manner of GBS methods, where marker discovery and population genotyping are done simultaneously (Elshire et al., 2011). The use of genome-wide SNP markers for genomics-assisted breeding (GAB), where more markers are assessed than in MAS, or genomic selection (GS), which when paired with phenotypic and environmental data can provide a basis for advanced studies like genome-wide association (GWA) or QTL analyses (Heffner et al., 2009; Jannink et al., 2010; Varshney et al., 2015; Heslot et al., 2015; Jonas & De Koning, 2013).

Molecular markers

Molecular markers can be grouped according to their throughput and method of acquirement, as follows: (1) low-throughput markers, that are hybridization-based; (2) medium-throughput markers, that are polymerase chain reaction (PCR)-based; and (3) high-throughput markers, that are

sequence-based (Mammadov et al., 2012). Among sequence-based markers, single nucleotide polymorphism (SNPs) markers are the most often used in modern genetic and genomic studies, as they are abundant and informative, fast and easy to discover and analyze, and flexible and cost-effective (Mammadov et al. 2012; Vignal et al., 2002; Kim et al., 2015). Currently, the most accessible methods for SNP genotyping are through SNP-arrays (or SNP chips) and (re)sequencing, where the sequencing approaches have a significant advantage over arrays, for they provide simultaneous discovery and genotyping (Bajgain et al., 2016). The rapid decrease in costs of sequencing enabled for genome-wide SNP discovery to become more easily available for breeders and researchers than ever before (Wetterstrand, 2016). Also, as arrays are developed on a specific population, in which the SNPs were discovered, they cannot be used outside that population without a risk of ascertainment bias (Ganal et al., 2011; Lechance & Tishkoff, 2013; Metzker, 2010; Moragues et al., 2010). That means that if in the researched population new SNPs exist which were not present in the population in which the array was developed, these polymorphisms will not be scored. This leads to researchers missing out on detecting this variation and thus underestimating the value of the potentially most unique individuals. Besides this, even when using SNP-arrays might be cheaper than resequencing, developing them is expensive, time-consuming and laborious, and while their production cost has not changed significantly, the costs of sequencing technologies keep dropping.

Reduced representation sequencing

Classification of genome resequencing methods

There are two major genome resequencing approaches: whole-genome sequencing (WGS), which aims to acquire the full genome sequence; and reduced representation sequencing (RRS), where reduced genome sampling is applied (Nielsen et al., 2011; Heffelfinger et al., 2014). While WGS can be

divided into high or low-coverage sequencing, which determines the quality and confidence with which the base calls are made, even with lower coverage, this method restricts the number of samples that can be studied due to the costs of running a whole sequencing lane for one sample. Even though NGS sequencing platforms are being further developed with innovative technology being employed in each new system in order to lower the costs of sequencing, a large number of RRS approaches have emerged, with the aim to lowering the per-sample cost by sequencing only a part of the genome, instead of the whole, and multiplexing multiple barcoded samples to sequence in the same run (Baird et al., 2008; Elshire et al., 2011; Kim et al., 2015). The aim is to reduce genome complexity by either targeting a part of the genome that is already known to harbor useful variation or by separating or eliminating a portion of the genome prior to sequencing, especially by trying to avoid repetitive sequences and focus on SNP variation. These approaches lower sequencing costs for studies that need to take into account large numbers of individuals while still sampling and enabling the observation of genetic diversity across the whole genome.

While the exome-sequencing (capture), which is hybridization based (Bamshad et al., 2011; Dapprich et al., 2016; Yoshihara et al., 2016) and RNA-sequencing (RNA-seq, cDNA sequencing, transcriptomic GBS; David et al., 2014; Nagalakshmi et al., 2008; Wang et al., 2009), that also targets the exome (genic) regions only, a number of RRS methods apply restriction enzymes (REs) to reduce genome complexity and asses SNPs genome-wide, while trying to avoid the repetitive segments. The focus of this study is on this group of techniques.

The first time sequencing was applied to DNA fragments flanking RE cut sites was in the RAD sequencing (RADseq) approach of sequencing RAD tags that were already in use for genomic diversity assessments in microarrays (Baird et al., 2008; Davey et al., 2011; Miller et al., 2007). Since then, a few modifications have been introduced and derived methods have been produced, such as the double digestion RAD method (ddRAD; Peterson et al., 2012), 2b-RAD (Wang

et al., 2012) utilizing REs that make two cuts to produce fixed-size double-stranded DNA (dsDNA) fragments and genotyping-by-sequencing (GBS; Elshire et al., 2011; Poland & Rife, 2012) which is gaining popularity in use in crop diversity research and has a few modified derived protocols of its own (more about this in the Innovative protocol modification section).

The value of genotyping large numbers of individuals with GBS

The main value that RADseq and GBS bring to SNP genotyping are lowered per-sample costs compared to whole-genome sequencing, as only a portion of the genome gets sequenced and several samples are multiplexed and pooled together before sequencing (Elshire et al., 2011; Nielsen et al., 2011). This is significant, as it enables the screening of much higher numbers of individuals through genome-wide SNP marker discovery and genotyping for a fraction of the cost than before. Due to barring polymorphism (Heffelfinger et al., 2014), most RE recognition sites are preserved within a species, so the discovery of DNA polymorphisms within the sequenced fragments can provide sufficient information of genetic diversity for a number of study applications.

GBS was developed as a further improvement to RADseq, with a reduction in price and complexity in mind. In GBS, a simpler protocol with fewer purification steps and no size selection results in reduced sample handling and lowers chances of contamination. Coupled with a more straightforward generation of restriction fragments and not using the more expensive biotinylated adapters as in RADseq, the genotyping expenses are additionally lowered (Elshire et al., 2011). Compared to DArT-sequencing (Jaccoud, Peng, Feinstein, & Kilian, 2001; Sansaloni et al., 2011), even though GBS can have more missing data, it does show higher genomic prediction accuracies in genomic selection (Poland, Endelman, et al., 2012), and as already mentioned, compared to non-sequencing based SNP genotyping methods, it has less ascertainment bias as SNP discovery and population genotyping are performed simultaneously.

Specific strengths of GBS in plants

There are certain features of the plant genomes that make GBS a particularly good fit for genotyping (Elshire et al., 2011). Firstly, plant genomes tend to be large and complex (Cornille et al., 2016; Jiao et al., 2011). One thing that leads to this is the plants' much bigger success in surviving whole-genome duplications that lead to polyploidy compared to many other groups of organisms. Secondly, additional size increase comes from the expansion of repetitive sequences, and while these sequences have been shown to have an effect on preserving the order of genes, they are uninformative when mining for SNP variation (Bevan et al., 2017). This means that sequencing the whole genome would have high costs, but that much of it is not that valuable in data analyses, so sequencing only a subset of the genome while avoiding the repetitive sequences is a logical solution. The ability to use GBS to solve this was demonstrated in the first GBS paper by Elshire et al. (2011). Maize is known to have high genetic diversity, as the average mutation rate in the maize genome is more than one substitution per one-hundred nucleotides, while there is also extensive presence-absence variation (PAV) that results from transposon-mediated rearrangements encompassing genic regions. By using the *ApeKI* RE, the aim was to avoid the repetitive sequences and acquire fragments from the more informative regions of the maize genome (Elshire et al., 2011). The advantage of GBS is that it allows for genotyping larger numbers of SNPs, indels and structural variations without any initial investment compared to hybridization-based strategies (Harfouche et al., 2012).

Even though exome-targeting RRS methods could have achieved a genome complexity reduction, the elements that control gene expression and affect agronomically important traits are most often located outside the protein coding regions, and would not be included in the analyses. Species with high genetic diversity are also hard to examine using single base extension assays, as finding invariant primer binding regions is difficult, or by scoring fixed

positions as in SNP arrays. Also, in some plant species, especially those with high PAV, if the individuals which the research is done for differ significantly from the individual that was used for creating the reference genome, *de novo* fragment assembly can result in a bigger dataset than the one created from the alignment to an ill-fitting reference genome, to which many of the fragments have nothing to align with. All this is, however, can be rather advantageous for any sequencing-based approach, as sequencing efficiency is in direct correlation with genetic diversity.

Despite low coverage, GBS is still suitable for use in segregating populations with strong LD, especially bi-parental populations, where missing values can be additionally lowered by using multiple barcoded tags for sequencing the parents multiple times and using this information for subsequent imputation (Kim et al., 2015). The genotyping information from breeding populations can be used for accurate genomic selection (Heffner et al., 2009; Jannink et al., 2010; VanRaden, 2008)

General advantages of GBS in comparison to other RRS approaches are that it is a less technically challenging, less time consuming, very reproducible, multiplexed and inexpensive high-throughput method (Elshire et al., 2011). This makes it easy to apply it on large numbers of samples and the production of libraries do not require any specific equipment, while now both the library production and the sequencing can be, and often is, outsourced. Using methylation-sensitive enzymes, we can assess important non-genic regions of the genome (as opposed to capture) while avoiding highly repetitive and targeting low copy sequences, which helps avoid computationally challenging alignment.

Many plants of interest in agriculture currently are wild crop relatives or orphan crop species whose genomes are quite different from those crops whose reference genomes are assembled, and GBS can still be applied in their research, as *de novo* reference maps can be assembled from the GBS fragments themselves that encompass the recognition site of the restriction enzyme.

As many restriction sites are conserved within a species, the sequenced portion of the genome is consistent within a population and therefore GBS is ideal for use in QTL mapping, breeding and natural population genomics for experiments that need to survey many markers across a large number of individuals (Heffelfinger et al., 2014).

General steps of the GBS protocol

The main steps in any GBS protocol can be divided into three segments: performing study-specific choices and protocol adaptation, preparing the GBS library itself and processing the bioinformatic data, so that the genetic interpretation may take place.

The study-specific choices include sample choice (sampling the natural variation or developing a study population), RE selection adapted to the studied organism of choice and adapter construction for barcoding and multiplexing the samples. The laboratory part consists of DNA extraction and GBS library preparation, following the protocol as described in Elshire et al. (2011) with optional modifications followed by DNA sequencing on the sequencing platform of choice. The bioinformatic processing of the raw sequencing reads includes demultiplexing the samples (where the barcodes and RE recognition sites in the sequence are identified and determining which reads belong to which sample), quality filtering, aligning the reads to the reference genome (if it is available) and using genotyping pipelines for SNP discovery and genotype construction. Imputation can be also applied, to reduce the amount of missing data, after which a genetic and statistical description of the acquired data may be done.

Protocol modifications

The papers that first published applications of both RADseq (Baird et al., 2008) and GBS (Elshire et al., 2011; Kim et al., 2015) utilize a single RE for cutting DNA fragments in the genome. For both, double RE methods were

developed (Poland, Brown, et al., 2012; Peterson et al., 2012) with the aim to sample an even smaller portion of the genome, where a smaller number of fragments would be selected and fewer markers discovered, but with an expected decrease in missing data under the same level of sample multiplexing and coverage when sequencing.

There is a number of similar protocols with different modifications that were applied in studies, a list of those based on RADseq is given in Table 1, modified GBS protocols in Table 2, and other RRS methods that apply RE for genome reduction can be seen in Table 3. Among these protocols, a few can be highlighted for introducing innovative changes that significantly impact the use of the method or further improve price reduction.

Rife et al. (2015) introduce spiked-GBS, similar to the approach used by Wells et al. (2013), which seeks to combine both targeting known genes of interest for use in MAS and whole-genome marker SNP discovery for use in GS. By using primers developed for KASP assays with the selective base removed to introduce the barcode and sequence the genic regions KASP would target, and using the rest of the sequencing capacity for GBS genotyping, an economical combination of the two is established.

Heffelfinger et al. (2014) demonstrate an adaptable protocol for use in population studies. Optimization and price reduction are achieved through a few modifications. First, by choosing REs that introduce blunt-end cuts combined with universal Illumina blunt-end Y-adaptors for multiplexing samples, savings were achieved in adapter compatibility, as there is no need to design different adaptors for each enzyme applied, but also in cheaper adaptor incorporation (Lamble et al., 2013). Second, a solid phase bead-based in-solution fragment selection and reversible immobilization is applied, which makes it available for the process to be done in smaller volumes in microliter plates which reduces handling and can enable automatization (Hawkins et al., 1994; Fisher et al., 2011).

Table 1. List of RADseq methods and modifications

Method name	Reference
RADseq, restriction-site associated DNA-sequencing (single-end)	Baird et al., 2008 Miller et al., 2007
PE RADseq, paired-end RADseq	Etter et al., 2011
ddRADseq, double-digested RADseq	Peterson et al., 2012
2b-RADseq	Wang et al., 2012
ezRAD	Toonen et al., 2013
I2b-RAD, improved 2b-RADseq	Guo et al., 2014
nextRAD	Russello et al., 2015
ddRADseq-ion	Recknagel et al., 2015

Table 2. List of GBS methods and modifications

Method name	Reference
GBS, genotyping-by-sequencing	Elshire et al., 2011
Double-digested GBS	Poland et al., 2012
Ion-torrent GBS	Mascher et al., 2013
GBS with selective primers	Sonah et al., 2013
Modified GBS for population studies	Heffelfinger et al., 2014
AFSM sequencing, amplified-fragment SNP and methylation sequencing	Xia et al., 2014
Spiked-GBS	Wells et al., 2013 Rife et al., 2015
545 pyro GBS	Rocher et al., 2015
GT-seq, genotyping in thousands by sequencing	Campbell et al., 2015
msGBS, methylation sensitive GBS	Kitimu et al., 2015
rtGBS, random tagging GBS	Hilario et al., 2015
epiGBS, reference-free reduced representation bisulfite sequencing	van Gurp et al., 2016

Table 3. List of other non-exclusive genic region targeting RRS methods

Method name	Reference
Reduced representation **shotgun sequencing**	Altshuler et al., 2000
CroPS, complexity reduction of polymorphic sequencing	Orsouw et al., 2007
MSG, multiplex shotgun genotyping	Andolfatto et al., 2011
DArT-seq, diversity array technology	Jaccoud et al., 2001 Sansaloni et al., 2011
SBG, sequence-based genotyping	Truong et al., 2012
RESCAN, RE sequence comparative analysis, RE-phased sequencing	Monson-Miller et al., 2012
DG, digital genotyping	Evans et al., 2013 Morishige et al., 2013
GWAFF, genome-wide allele frequency fingerprints	Byrne et al., 2013
GGRS, genotyping by genome reducing and sequencing	Chen et al., 2013
SLAF-seq, specific-locus amplified fragment sequencing	Sun et al., 2013 Zhang et al., 2013
REST-seq, restriction fragment sequencing	Stolle & Moritz, 2013
iRRL, improved reduced representation sequencing	Greminger et al., 2014
RAPiDseq, randomly amplified polymorphic DNA sequencing	Carletti et al., 2016

Reviews on RRS and GBS

Several review articles were published which consider different aspects of using RRS, RADseq or GBS genotyping in plants and for a variety of study aims. As RADseq is most often applied in animal population genomic or phylogenetic studies, most reviews on this method tend to focus on how good these markers are for this particular use, while GBS is more often used for genotyping plant populations developed for various breeding purposes. Some reviews, as the one by Deschamps et al. (2012) and Kumar et al. (2012) assess genotyping by sequencing in its widest sense, discovering SNP markers using NGS technologies, not specifically just the GBS method developed by Elshire et al. (2011).

In the first review on RADseq, Davey & Blaxter (2011) have outlined the RADseq library preparation process and highlighted the power RADseq had in population genomics of the three-spine sticklebacks (*Gasterosteus aculeatus* L.; Baird et al., 2008; Hohenlohe et al., 2010), a well-known model organism for studying evolutionary mechanisms, and in phylogeographic research of a mosquito species (*Wyeomyia smithii*, Coquillett; Emerson et al., 2010), which is an example of research done on a species without a reference genome. They provide ideas on how to maximize the utility of RADseq by choosing longer pair-end sequencing and taking into account the general sequencing issues that exist in next-generation sequencing (NGS), like sequencing errors and GC bias (Benjamini & Speed, 2012; Y. C. Chen, Liu, Yu, Chiang, & Hwang, 2013; Harismendy et al., 2009). In the end, they envision the potential that RADseq can have for diverse genetic analyses by opening the door for more affordable sequencing of increasingly larger numbers of individuals as the costs of sequencing keep decreasing (Wetterstrand, 2016). Similarly, a RADseq symposium meeting review by Rowe et al. (2011) presents results of several research investigations which used RADseq for genotyping purposes, along with a description of the basis

on which the method functions. Two years later, a review of special features of RADseq data that need to be considered as they affect the genotyping was published by Davey et al. (2013). This encompasses the specific biases in RADseq data which result from the nature of how REs and polymerase chain reaction (PCR) amplification function, mainly that there is a significant variation in read depth resulting from restriction fragment length bias and GC content bias in PCR. Restriction site heterozygosity causes problems leading to calling presence/absence sites homozygous when the other allele is not sampled at a heterozygous site. Some of these problems can be addressed by applying conservative filtering approaches, but as those lead to a loss of some data that is informative, the best solution is developing more sophisticated statistical methods to incorporate in genotyping tools' RAD contig assembly.

Li et al. (2011) provide an assessment of datasets of large numbers of individuals sequenced at different coverage, ranging from 2x to 30x, showing how imputation can improve the informational content of these datasets known for the strong downside of having a high amount of missing data. They show how these perform in studying the genetics of complex traits. Finally, they provide a guideline for researchers who need to combine data from sequenced, genotyped and imputed samples, in order to choose the best compromises for the most informative results. Similarly, Torkamaneh & Belzile (2015) explore the extent to which missing data can be tolerated in SNP datasets and how successful imputation can be in filling in the missing genotypes in GBS data. Further discussion on similar topic is continued in the article by Lowry et al., (2016), who present how the amount of missing data and genome coverage of RADseq markers effects genome scans for adaptation in populations in comparison with other approaches, such as gene-targeting exome and transcriptome sequencing, pool sequencing of individuals and whole genome sequencing. Genome size and the extent of LD need to be considered. This is followed by a comment by Mckinney et al. (2017) on the unprecedented utility of RADseq for molecular ecology and evolutionary

genetics, how powerful a tool it is for studying adaptation in nature (Catchen et al., 2017) and which practices to follow for best results in population genomics studies of adaptation (Lowry et al., 2017)..

There are several reviews focusing on specific applications of GBS. The one by Poland & Rife (2012) focuses on the advantages of GBS over other methods in plant breeding and genetics, its adaptability to different research questions, the innovative possibility of its use for genotyping species without having information of their reference genome, for linkage and association mapping, gene mapping, and GS, and shows ways to handle missing data issues. Narum et al. (2013), however, focus on the application of GBS in ecological and conservation genomics. This review shows the potential of RE employing RRS methods through case studies assessing population genomic data for wildlife conservation aims, but also for QTL mapping, and genome-wide SNP discovery, while addressing potential biases, software solutions, and future perspectives. Andrews et al. (2016) assessed the power of RADseq in ecological and evolutionary genomics. They aim to show how to choose a reduced representation sequencing approach based on what scientific question needs to be addressed and types of bias and error inherent to RADseq data, similarly to the discussion seen in Davey et al. (2013). He et al. (2014) offer a view of GBS as the ultimate tool for MAS, that will accelerate plant breeding, which fits the aim of its development for genotyping plant breeding populations. This review gives a general overview of GBS, DNA markers and NGS technologies, showcasing the successful use of GBS as a tool in plant breeding studies in maize, potato, soybean, barley, switchgrass, yellow mustard, Arabidopsis, rice, and bread wheat; as well as species without a reference genome, like rapeseed, lupin, and lettuce; pointing out drawbacks and perspectives in the conclusion. Thomson (2014) points out the advantages of using genome-wide SNP markers for crop improvement and reviews high-throughput genotyping platforms for their acquirement. It shows how different tools fit different needs and goes through examples of the successful use of

fixed and flexible SNP assays and genotyping by sequencing techniques. It draws out the issues to consider when deciding on SNP genotyping options and gives an example of what setting up an in-house genotyping facility looks like.

In Kilian & Graner (2012), NGS-based DNA marker systems for fingerprinting germplasm accessions stored *ex-situ* in genebanks are encompassed in general, not GBS in particular. But as this article has similar goals to this study, including their considerations is crucial. In short, this article considers the exploration of patterns of genetic diversity, mapping quantitative traits and mining novel alleles, describing both advances and bottlenecks that still exist.

Peterson et al. (2014) gives an example of a 2 RE GBS protocol for genetic diversity assessment of flax (*Linum usitatissimum* L.), and introduces npGeno, a custom bioinformatics pipeline for processing raw sequencing reads. Like most other reviews, it reflects on considerations that need to be taken into account when performing this kind of analysis, which can be a useful guideline for experimental design. It demonstrates the time and price ranges expected for conducting this type of research for the period in which the article was written.

The review of Heffelfinger et al. (2014) examines the effect RE selection has on genome reduction in different genomes, using eight different enzymes on F_2 populations of maize and rice. Enzymes are compared based on their mapping quality scores, fragment size effect on coverage, site density, coverage in genic regions, and methylation sensitivity, and the data is used for population genomics and trait mapping. The aim is to show how results can differ and how a careful enzyme selection can improve the power of the resulting data. They also modified the protocol for higher cost-effectiveness and better adaptability to diverse population study research aims.

Varshney et al. (2014) provide a valuable overview of how integrating NGS technologies and genomics knowledge into crop breeding programs can help us use the available natural variation better and improve our crop varieties more efficiently. It lists types of breeding populations used, selection and mapping approaches used, and sequencing methods for acquiring genome-wide molecular marker data. They stress how putting focus onto further improvements in technology, education, collaboration and data sharing is crucial for best results and future advancement.

Kim et al. (2015) show the impact of GBS on crop breeding programs through examples of sorghum, brassica, cotton, and silvergrass. Their study includes GBS pipeline suggestions for data processing, pointing out some potential issues that need to be accounted for, and giving an example of how to calculate the number of sequence reads needed for a discovery of a targeted number of SNPs within a species (using silvergrass as an example), which includes equations for the estimation of the number of restriction fragments expected in the relation of the recognition site length, expected number of SNPs discovered in relation to sequencing length and polymorphism rate in the species. They also discuss how to achieve desired coverage per site, and conclude with the future prospects of the technology.

Patel et al. (2015) predict where plant genotyping is headed in the future by discussing the most popular currently used sequencing approaches for genotyping, encompassing short overviews of RRL, CRoPS, RADseq and GBS..

Zimmer & Wen (2015) lists a broad range of NGS methods that have been successfully used in deciphering phylogenetic relationships in plants, including microsatellite markers, genome skimming, transcriptomics, targeted enrichment, EPIC markers, RADseq and GBS. They address the character evolution in the genomic era, challenges and prospects.

Some reviews provide great resources for experiment planning and decision making in research which plans to use GBS or similar methods. The review by Jiang et al. (2016) introduces a term that encompasses all the reduced representation sequencing methods: Genome-wide sampling sequencing (GWSS), while noting how in most cases it could be synonymously used with reduced genome complexity sequencing, reduced representation genome sequencing and selective genome target sequencing (as they include exome sequencing in this classification as well, even though it does not use restriction enzymes for a random reduction in genome complexity). They have divided the methods they include into 4 main groups: 1) GWSS without size selection (GBS, 2RE GBS, GGRS); 2) GWSS with size selection (RADseq, ddRADseq, PE RADseq, flexible and scalable GBS); 3) GWSS with semi-size selection (RRS, RRLs, paired-end RPLs, 1RE GBS, 2RE GBS, iRRL, 2b-RAD); and 4) GWSS with selective amplification (CRoPS, scalable GBS, WES). Lastly, technical challenges in GWSS, like the inconsistency in the number of reads per sample, number of reads per site, missing data, and number of sites sequenced per sample are outlined and suggestions for future GWSS development are given based on these considerations. They include a list of all the restriction enzymes used in studies so far, as well as a library construct sequence comparison between major approaches, which are useful resources for researchers planning similar research. Torkamaneh et al. (2016) reports on the speed and output dataset quality in terms of accuracy, amount of missing data and heterozygous calls using two different raw datasets (one with and the other without a reference genome) put through seven bioinformatics pipelines, some of which are specifically aimed at addressing SNP calling without the use of a reference genome. This comparison provides help in choosing the pipeline that best suits the data in hand, showing how the assumptions in their algorithms are adapted for different study questions or organisms.

The review paper by Voss-Fels & Snowdon (2016) focuses on genotyping for plant diversity discovery and encompasses all sequence-based genotyping

methods, including WGS or skimGBS, RRS/GBS, transcriptome sequencing and sequence capture approaches, as well as multiple possible applications of this data in breeding (MAS, GS, GWAS, etc.). It shows increased use of genotyping data for breeding purposes in estimating germplasm diversity for use in crop improvement breeding programs and lists many successful example studies for each method and application.

Scheben et al. (2017) reviewed the currently used genotyping by sequencing methods, pointed out the pros and cons in decision making when it comes to choosing the appropriate genotyping method for different application purposes showing study examples, and provided an overview of the bioinformatics software that is currently most popular in handling genotyping by sequencing data. Grover & Sharma (2014) have shown the development of molecular markers over time, also covering RADseq and GBS as methods of marker discovery.

Conclusion and prospects

GBS was developed on the model of the RADseq technique that reduced the representation of the genome by using restriction enzymes, after which only the ends of the cut fragments were sequenced, usually after size-selection applied. This kind of approach enables a relatively random and even genome-wide coverage of the resulting sequence data, but also a possibility of avoiding repetitive sequences and modulating the density genome coverage through careful choice of the restriction enzyme. An enzyme with a longer recognition site would cut in fewer genomic locations, while combining two enzymes and sequencing only fragments with different cuts on their opposite ends would provide further reduction. Using a completely or partially methylation-sensitive enzyme brings the possibility of avoiding repetitive DNA sequences that are less informative for most types of plant trait studies, but carries with itself a need of more careful sample collection and preparation, for an example, to avoid imbalances in coverage due to biological differences in methylation.

With targeted modifications that lead to GBS having a lower per-sample cost the assessment of populations with much larger numbers of individuals became possible, while innovative tweaks are still being thought of to create new methods for special uses, better results and even more affordable costs. This is of particular advantage for both plant breeders, who benefit from genotyping data from as many plants from their test populations as possible, but also for those interested in the diversity of wild populations, where including more samples can bring a more thorough understanding of the evolutionary dynamics of populations in the wild.

However, while genotyping cost and availability used to be the bottleneck in various research before, now we see that this has shifted towards phenotyping and subsequent data storage and analysis. Innovations that are bringing down

the sequencing costs need to be coupled with an investment in high-throughput phenotyping technologies, providing more affordable storage for the ever-increasing data we are producing and interdisciplinary education and collaboration, that will lead to breakthroughs in how we view and process data. The fastest improvements are now probably happening in the bioinformatics field, especially when talented and motivated specialists are brought together to form teams which have sufficient financial freedom to test the boundaries of what's possible. To create new algorithms for data processing for having a clearer apprehension of the results, we need to couple good understanding of the logic behind the experimental design of biological research with the knowledge about currently available technology and the awareness of its current shortcomings. This can be accomplished by supporting the forming of well-integrated teams of diverse backgrounds, where creativity thrives through thorough collaboration.

In literature, we see that both RADseq and GBS have been used in diverse study organisms and for different study purposes, while also many reviews provide critical opinion on the advantages and drawbacks these approaches have in general. We might need to invest more effort for integrating and interpreting the results of these studies with a broader view, and keep comparing how different modifications perform in different organisms under different assumptions. That will help form better-informed decisions in future research endeavors, where we would get the deepest insights into biological systems under the current possibilities.

This literature review so far includes a summary of currently available reviews that cover certain aspects of the presented genotyping methods, but a database of research articles published using these genotyping methods was created and its planned to be added as a resource and easy overview of what research, on which organisms, with what tweaking and for which reasons has been done up until now.

(This page has been intentionally left blank for print formatting purposes.)

(This page has been intentionally left blank for print formatting purposes.)

CHAPTER 2

Genomic characterization of a biparental common bean (*Phaseolus vulgaris* L.) population segregating for the traits of the domestication syndrome

(This page has been intentionally left blank for print formatting purposes.)

Abstract

A thorough understanding of the domestication process in plant crop species is beneficial for two main reasons. From a theoretical perspective, it enables us to study the genetic and molecular mechanisms of evolution, towards the identification of the molecular basis of heritable phenotypic variation among organisms. At the same time, in applied sciences, this knowledge serves as a model for discovering genomic regions governing traits of agronomic importance, which can be used in breeding programs for crop improvement. Scientist have therefore recognized the need to facilitate the mining of the genetic diversity present in wild crop relatives, that was unintentionally left behind during the human selection process of domestication, but can be valuable novel resources in future breeding strategies, that already passed the test and pressures of natural selection.

Here, we present a common bean (*Phaseolus vulgaris* L.) population of introgression lines (ILs) developed to dissect the genetic basis of pod-related traits of the domestication syndrome. The Mesoamerican G12873 accession was used as the wild donor, while the Andean variety Midas was chosen as the recurrent parent for a triple backcross with the MG38 (G12873xMidas, F_9 generation of single-seed descend under self-pollination). The purpose of this population is to enable the study of complex traits related to domestication, and to serve as a tool for the introgression of useful alleles from the wild into a domesticated common bean accession for application in breeding programs.

In this study, we have assessed the single-nucleotide polymorphism (SNP) marker dataset based on genotyping-by-sequencing (GBS) data, its marker density, heterozygosity, linkage decay, the population structure, performed genomic characterization of introgressions in selected Ils and QTL analysis for the traits of flower color and flowering time.

(This page has been intentionally left blank for print formatting purposes.)

Introduction

Study aims

The main aim of this research project was to develop a common bean (*Phaseolus vulgaris* L.) study population using a domesticated and a wild parent to create a population that segregates for the traits of the domestication syndrome. This was done according to the ideas and joint efforts of teams of Prof. Roberto Papa's lab at Marche Polytechnic University (UnivPM, Ancona, Italy) and Dr. Domenico Rau's research group at the University of Sassari (UNISS, Sassari, Italy), and thanks to the accession MG38 provided by Prof. Paul Gepts from the University of California at Davis (UC-DAVIS, Davis, CA, US), which was developed for studying common bean domestication traits (Koinange et al., 1996). This project is a continuation of that endeavor.

The population was developed as a triple backcross between the domesticated Andean variety Midas and wild Mesoamerican accession G12873 with the main focus on studying the pod domestication syndrome traits, specifically, pod-shattering, as it's a desired trait in wild beans and was selected against during domestication in human grown beans. Phenotypic traits were recorded both in greenhouse and field conditions, while precision phenotyping was applied for pod and seed measurements. A subset of Recombinant Inbred Lines from 10 different families in generating F_5 and F_7, were genotyped via genotyping-by-sequencing (GBS) following a modified protocol of Elshire et al. (2011). This dissertation focuses on the genomic characterization of the population and the demonstration of its power for quantitative trait loci (QTL) mapping. The population is planned to be further developed and used in a Near-Isogenic Line approach through the comparison of sister lines of the population, to more precisely map the genetic regions affecting domestication trait phenotypes.

The origin and domestication of crop plants

Domestication represents the evolutionary process of the adaptation of organisms through directed selective breeding to better suit human needs. All crop plants have gone through several bottlenecks during the process of their domestication, starting from the founder effect resulting from limited sample sizes when the first wild individuals were picked out for being grown and harvested by humans. Since then, humans have been applying selective pressures on their crops, choosing to grow those which they've seen perform better or had more appealing fruit or other products meant for our use. This artificial selection has lead to further impoverishment of the genetic diversity in the crop plants, but also, to their specific adaptation to the agroclimatic factors in the region where they were grown as well as loss in seed dispersal, loss of sensitivity to the length of day in relation to the initiation of flowering and fruit baring developmental phases, but also bigger fruit and more yield overall, which was the main focus of many farmers in order to feed their families. Figure 1 shows a graphical representation of the genetic loss during domestication.

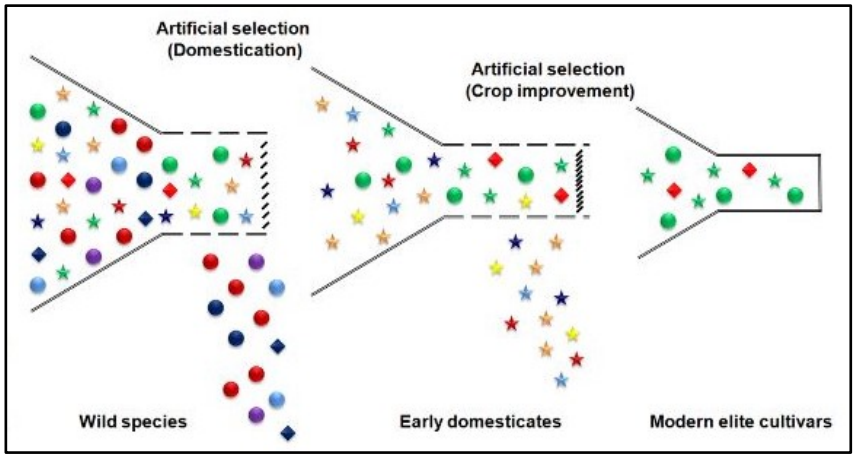


Figure 1. Predomestication and domestication bottlenecks in crop plant genetic diversity, an example from cultivated rice (modified from Gopala et al., 2014).

Vavilov introduced the theory of centers of diversity, where the origin of a particular plant would be deduced based on geographical locations where the highest genetic diversity in natural populations of their wild relatives can still be found (Dvorak, 2011; Vavilov, 1926, 1992). For species where the natural populations of the wild progenitors are well known, geographically defined and relatively preserved in their distribution and genetic diversity, this approach can be used in a straightforward way.

There are two main reasons we are interested in getting to know more about the natural diversity of the wild progenitors, sister species as well as old landraces nowadays. In evolutionary biology, these plants can be excellent models for studying the evolutionary mechanisms of domestication and the effects of selection on genetic and phenotypic diversity and plasticity, while in agricultural applications, there is an interest in collecting, cataloguing and preserving this diversity for using it in genetic improvement of our crops in a try to reintroduce some of the useful variation that has been unintentionally left behind during domestication. With the population in this study, we try to tackle both, by increasing the knowledge of the genetic bases of phenotypic variation in the common bean, that can later be applied for decision making in projects for agricultural improvement.

The presence of specific traits characteristic in the domesticated crops that are a result of artificial anthropogenic selection was first coined as “adaptation syndrome” by Harlan et al. (1973), but the term “domestication syndrome” was adopted for further use, as more intuitive by Hammer (1984). This term is now widely used and encompasses all the phenotypic variants that are useful for the growing, breeding and use of agricultural crops as food, feed, energy or material.

The common bean (Phaseolus vulgaris L.) and its domestication

The bean genus (*Phaseolus* ssp.) counts around 70 recognized species (Bitocchi et al., 2017), of which 5 are domesticated: the common bean (*P. vulgaris*), the year bean (*P. dumosus*), the runner bean (*P. coccineus*), the tepary bean (*P. acutifolius*) and the lima bean (*P. lunatus*). Most have an origin that ranges from Southwestern USA to Northwestern regions of North America, but with the biggest abundance of wild forms in Mesoamerica. Due to their differences in mating systems (predominantly or exclusively allogamous or autogamous), life cycles (annual, perennial or both), adaptation to different agroclimatic regions and especially the dual domestication events in the common bean and possibly in the tepary bean, as well, make this genus an excellent tool for studying evolution, especially under domestication.

The common bean (2n = 22) is an annual predominantly autogamous species with a relatively small genome size (587 Mbp; Schmutz et al., 2014) and a well-documented case of dual domestication. The origin of the common bean is thought to be Mesoamerican, while beside the gene pool present there, there is another large gene pool in the Andean region and a smaller one in Peru. The domesticated beans have originated from both the Mesoamerican and Andean gene pools, which makes this species a rare tool for studying specific mechanisms in evolutionary studies.

The domestication syndrome in the common bean has already been studied and reported on by Koinange et al. (1996). This study builds on those findings and is developing a population from one of the lines created in the study of Koininge et al. The aim is to provide more detailed findings on the genes that are underlying these traits by further developing the population, introducing more recombination points to lower the size of the introgressed segments from the wild parent in the domesticated genomic background, so the trait mapping resolution would be higher, and using the latest technology and bioinformatics approaches available for producing and analyzing the data.

The domestication syndrome in the common bean is most evident in traits related to growth habit, photoperiod sensitivity and pod and seed related traits. The list of domestication syndrome traits that were recorded and tracked in both the original and current study can be seen in Table 4, and the additional phenotypic traits included in this study are reported on in the Material and methods section. Considering the growth habit, the wild beans show climbing tendencies, while the domesticated beans are more compact. The domesticated beans were also selected for reduced photoperiod sensitivity, so its growth would not be inhibited when the days are shorter, which in turn extends the period during the year when it can be grown. That is especially useful during breeding population development, where more generations can be grown in one year, especially in regulated greenhouse conditions. The traits of the common bean that, like in many other crops, have undergone the biggest change under strong selection during domestication are the ones related to pods and seeds, as those are the parts used in our diet, for which the plant was domesticated in the first place. While the wild beans most often have smaller seeds of uniform dark color placed into smaller pods that shatter when they mature, the domesticated ones have larger non-shattering pods and seeds that are bigger and more variable in color (see Figure 2).

Figure 2. The diversity in seed phenotypes in cultivated common bean (modified from Gentry, 1969).

Table 4. Phenotypes of the parental accessions used for RIL population development for studying the domestication syndrome in common bean (taken and modified from Koininage et al., 1996).

Attribute	Trait	G12873	Midas
pod shattering	pod suture fibers	present	absent
	pod wall fibers	present	absent
seed dormancy	germination	70.5%	100%
growth habit	determinancy	indeterminate	determinate
	twining	present	absent
	no. of nodes on main stem	22.5	7.5
	no. of pods	43.2	13.9
	internode length	1.6 cm	2.9 cm
gigantism	pod length	5.7 cm	9.8 cm
	100 seed weight	3.5 g	19.5 g
earliness	days to flowering (12 h day)	69	46
	no. of days to maturity	107	80
photoperiod sensitiv.	delay in flowering (16 h day)	> 60 days	0 days
harvest index	seed yield/biomass	0.42	0.62
seed pigmentation	presence/absence	present	absent

Perhaps the most significant trait for selection during the domestication of the common bean is the pod shattering trait. Pod shattering is a useful seed dispersal strategy in the wild beans, while in the domesticated varieties, it is desired to prevent this happening, as it would cause unnecessary yield loses. This trait has been closely investigated in this population by Murgia et al., (2017), where the trait was described by the mode (dehiscent or indehiscent) and level of shattering (indehiscent, fissured with a slight separation between the pod valves, non-twisting dehiscent and twisting dehiscent). The chemical analysis of the pods, which examined the content of carbon, nitrogen, and hydrogen in the pods, has shown a correlation between the carbon content and shattering trait, where more carbon appeared in pods with observed higher shattering levels. When the fiber content (lignin, hemicellulose, and cellulose) and location were investigated, the pods with higher fiber content, specifically, increase in lignin in ventral sheets and inner fibrous layers of

pods were showing higher levels of shattering. In total, there was an association of this trait with seven genomic locations (genes), one on chromosome 5 which was having the strongest effect, two regulating the level and four regulating the modes of shattering.

Besides its importance for theoretical studies of evolution, the common bean also has a big practical value as food and feed due to its nutritional values; and in crop rotation systems and intercropping agricultural systems due to its nitrogen-fixation capabilities. Nutritionally, the common bean has a great micronutrient content, especially considering iron, and a high protein content (Gepts et al., 2008; Wiesinger et al., 2016), while it is also thought that it can be used to alleviate the lowering nutritional quality of plants under raising CO2 by ballancing N:P plant ratios through providing higher availability of nitrogen compounds in soils (Deng et al., 2015; Loladze, 2002), while it is also proposed to replace beef with beans in our diets in the beans-for-beef against climate change movement (Harwatt et al., 2017).

Plant study population

Even though linkage mapping using bi-parental populations is being less used over the years in favor of association mapping that relies on high genetic diversity and low LD in germplasm to provide better resolution in QTL mapping (Álvarez et al. 2014), linkage mapping can still provide useful results, especially when applying certain breeding designs. Here, by crossing a wild and domesticated parent to create a segregating population for domestication syndrome traits, and selecting for the presence of the domestication traits in a domesticated genomic background during population development, starting from one F_1 plant that have rise to 16 F_2 families and 250 F_3 subfamilies, doubling the population size at each further generation until F_5 or F_7 , we have selected 285 plants for using in a near-isogenic line approach, where we want to utilize the small differences between lines from sister families to achieve finer mapping of the QTL traits, as demonstrated by Stam & Zeven (1981) and Tanksley & Nelson (1996).

Material and methods

Population development

For developing a population that segregates for the traits of the domestication syndrome, the parents were selected to have contrasting characters for these traits. As mentioned, the domestication syndrome was already studied and reported on by Koinange et al. (1996). The objective of this study was to continue in that endeavor with the goal to achieve a finer mapping of the traits, particularly through the exploitation of the near-isogenic lines (NIL) population approach, using the comparison of the sister lines within the families (Stam & Zeven, 1981). The accessions that were used in the initial population development by Koinange et al. (1996) were a domesticated stringless variety from the Andean gene pool, Midas, and a wild accession G12873 from the Mesoamerican common bean gene pool. The domestication syndrome traits in which they differed can be seen in Table 4, and the population development scheme in Figure 3. From a cross between these two accessions, a bi-parental recombinant inbred line (RIL) population was developed, following the approach proposed by Broman, 2005. For increasing homozygosity and decreasing the length of the introgressed segments in the lines, single-seed descent (SSD) self-fertilization was used for population propagation, as suggested by Goulden (1939), later modified by Brim (1966) and evaluated in different crops (Adamski et al., 2014; Haddad & Muehlbauer, 1981; Lalic, Kovacevic, & Novoselovic, 2000; Salas & Friedt, 1995; Tee & Qualset, 1975).

From the F_9 generation of the RIL population, MG38 was selected for having around 55% of the wild genome introgressed (based on AFLP marker data, Prof. Papa, personal communication) and the desired combination of wild and domesticated traits (as described below). The line was provided by Prof. Paul Gepts (UC-DAVIS) for use as the semi-wild parent for this study.

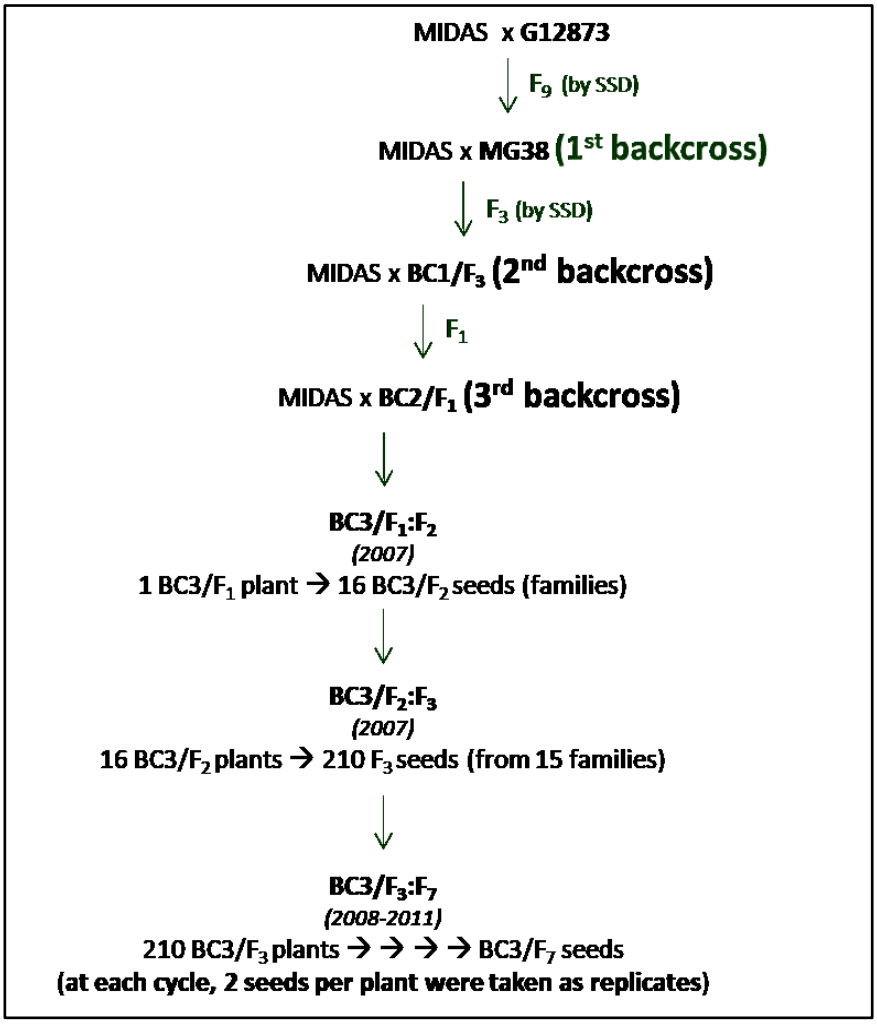


Figure 3. The population development scheme

A triple backcross population design followed, using Midas as the recurrent parent, with the aim to follow the QTL-NIL approach suggested by Tanksley & Nelson (1996). Through this, it was expected to achieve the shortening of the introgression segments from the wild parent and increase the background genome of the domesticated parent, while preserving the desired phenotype characters.

Both when choosing MG38 and during backcrossing, we selected for the wild phenotype for the pod-related traits of the domestication syndrome (e.g. pod shattering, small pod and seed size, wild shape and color of pods and seeds), while selecting for the rest of the traits of the recurrent domesticated parent (e.g. determinate growth habit, photoperiod insensitiveness). The later was done to facilitate population development and ease of further breeding, through maintaining the traits selected for by farmers and breeders during domestication. The contrasting traits for the pod and seed between the two parental accessions can be seen in Figure 4, where the domesticated Midas has much larger, colorless pods and seeds (Fig. 4A), compared to the wild G12873 genotype (Fig. 4B), while we also see a visible difference between the amount of twisting in the pods that leads to pod shattering, a mechanism of seed dispersal in the wild beans, which is not desired in the domesticated bean, as it leads to yield losses during harvest.

Figure 4. A) Pod and seed traits of the domesticated variety, Midas; **B)** Pod and seed traits of the wild G12873 genotype.

A BC3/F_1 plant with the desired traits was chosen for further developing the population, producing 16 seeds which eventually gave rise to 15 BC3/F_2 families. Starting from 250 BC3/F_3 plants, 2 seeds were grown at each following generation, to achieve doubling of the population size at each generation, while theoretically keeping the absolute number of heterozygous loci among generations constant (assuming the absence of drift and selection). The population was designed with the aim to build sets of near-isogenic lines

(NILs) among the sister lines within the families, to exploit the segregation as in the heterogeneous inbred family (HIF) approach (Fletcher et al., 2013; Tuinstra, Ejeta, & Goldsbrough, 1997; Yeri et al., 2014). By comparing the lines, after five and seven generations of SSD, the homozygosity is expected to increase, and most loci across the genome would be isogenic, but it will be possible to find and use lines that are segregating for loci of some of the targeted QTL regions (Tuinstra et al., 1997).

A total set of over 1,600 nested introgression lines (ILs) were produced for use in the family-based association test for QTL detection and fine mapping. The plant material that was selected to be genotyped using genotyping-by-sequencing (GBS; Elshire et al., 2011) and presented in this study consists of a selection of 68 ILs from F_5 generation plants, 217 F_7 lines, and three replicates of each of the parental lines, Midas and MG38, which together come to a total of 291 multiplexed samples.

Phenotyping

Phenotypic data related to the pod traits (e.g. pod dimensions, shattering, level of twisting), general plant morphology (e.g. habitus, branching, cotiledone number, angle and length) and traits of agronomic importance (e.g. flowering time, plant height, number and weight of pods and seeds, germination success rate, fruit setting time) has shown that this population comprises lines presenting a range of phenotypic values as well as showing transgressive phenotypes (Rieseberg, Archer, & Wayne, 1999) as compared to the parents (unpublished data, reported in the dissertation og Murgia M.L., 2016). The phenotypic data was collected over the course of population development in field and greenhouse conditions, while for pod shattering, precision phenotyping was applied, with the pod-shattering investigation reported by (Murgia et al., 2017). In Figure 5A, a pod with high pod shattering (as in wild), but low pigmentation (as in domesticated) can be observed, while Figure 5B shows how the maintenance of the plants in the greenhouse was organized.

A qualitative trait that was investigated in this study to demonstrate the power of the population for QTL mapping application was the flower color, where the G12873 parent has a violet variant, while the flower of Midas is white (showing the absence of pigmentation). A quantitative trait presented for the same purposes was the flowering time.

Figure 5. A) The wild phenotype for the pods and seeds; **B)** Growth of plants in controlled conditions of a greenhouse at the Department of Agricultural, Food and Environmental Sciences of Marche Polytechnic University.

Genotyping

DNA extraction was performed by Maria Leonarda Murgia in the lab of Prof. Giovanna Attene at the University of Sassari (Università degli Studi di Sassari), located in Sardinia in Italy. The genomic DNA was extracted from 100 mg of young leaf tissue per plant. The leaf tissue, frozen in liquid nitrogen, was ground using TissueLyser II (Qiagen) and the DNA was extracted using the DNeasy 50 Mini Plant Kit (Qiagen). The quantity and quality of extracted DNA was examined using spectrophotometry (GeneQuant II, Pharmacia Biotech LTD). The DNA stock was stored at -20°C until it was sent to The French National Institute for Agricultural Research (INRA) in Montpellier, France for GBS library preparation. At INRA, the quality and quantity of the extracted DNA was controlled again using spectrofluorimetry with benzimidazole derivative H33258 (Hoechst) on a Spark 10M multimode microplate reader, while the DNA quality

and possible degradation was examined using gel electrophoresis with ethidium bromide used for DNA visualization.

Genotyping was conducted using genotyping-by-sequencing (GBS) as described by Elshire et al. (2011), with two major modifications applied to the original protocol: (i) a nested multiplexing design and (ii) the application of fragment size selection. The rationale behind the decisions in the genotyping design is explained in detail in the Discussion, and the sequences of the Illumina indices and barcoded adapters can be viewed in Appendix 1.

In short, the GBS library preparation protocol steps were the following: (1) **The enzymatic digestion** of 200 ng of extracted genomic DNA per sample with *ApeKI* restriction enzyme (partially sensitive to CpG methylation, with the G'CWGC recognition motif) in the NEB 3.1 buffer, 2h on 75°C; (2) **The ligation of barcoded adapters** to the digested DNA in a ligase mix consisting of a water solution of the ligase buffer and T4 DNA ligase, 10 mins on 30°C, 4h on 22°C, 8°C overnight; (3) **The enzyme inactivation** was done by holding the samples for 30 mins on 65°C; (4) **The samples were pooled** together into libraries, 24 samples with different barcoded adapters at a time, making up a total of 13 libraries; (5) **The purification** was conducted in 2 cycles on Invitrogen magnetic racks with metal beads in a modified buffer solution; (6) Sizing was confirmed using BluePippin (Sage Science); (7) **The Illumina indices** were added to the library fragments in a PCR amplification step where a PCR mix containing the Taq Phusion HF buffer, Taq Phusion polymerase, dNTPs and primers were added to the pooled GBS libraries (the PCR program held 30 s at 98°C for denaturation, ran 18 cycles of 10 s at 98°C for denaturation, 30 s at 65°C for annealing and 30 s at 70°C for elongation, with 5 mins at 72°C for final elongation, and hold on 4°C after); (8) **A final purification step** was performed the same way as before; (9) The **GBS library dosage** was controlled using Agilent DNA 7500 Kit and qPCR, following the user manuals (available upon request). The full GBS protocol applied in this study can be found in Appendix 2.

GBS library sequencing

The GBS libraries were pair-end sequenced on the Illumina HiSeq 3000 platform in two lanes with 150 bp in sequence length at the INRA facility in Toulouse, France. The first pool was also used in a test run and pair-end sequenced at 150 bp in length on the Illumina MiSeq 2000 platform at Supagro INRA, Montpellier, France. These sequence reads were also included in the SNP discovery and genotyping process, providing better coverage for the chosen samples with contrasting pod-shattering phenotypes.

The sequencing quality of all files was examined with FasQC v0.1.3 (Andrews, 2010). An additional assessment of the batch effect between the two sequencing lanes and among the GBS library pools was performed on the SNP dataset itself, using a custom script in R version 3.4.2 (R Core Team, 2017) used in the RStudio v1.0.143 editor for R (RStudio team, 2015). The script groups the samples based on their Illumina indices and creates a barplot for each pool presenting the distribution of missing data within the pool, while also marking the average of per sequencing lane.

SNP marker discovery and population genotyping

Demultiplexing was conducted using a custom Python script. Filtering of low quality reads and the removal of adapter sequences was done in cutadapt (Martin, 2011; *https://github.com/marcelm/cutadapt*). Sequence alignment to the reference genome was done using the Burrows-Wheeler Aligner, BWA v0.7.16 (Li & Durbin, 2009), using the bwa-mem algorithm. As the reference genome, the second version of the common bean genome was used (*P. vulgaris* v2.1; *https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvulgaris*). The realignment of indels was done in GATK 3.7 (McKenna et al., 2010), but as the caller is not reliable for calling indels in sequencing analyses like this, they were not included in the dataset. The filtering of multiple mappings was performed in SAMtools v1.4 (Li, 2011; H. Li et al., 2009), with the minimum

Q-score cutoff at 13. The SNP discovery and genotyping were performed using GATK, where 537.218.636 x 10^8 sites were imported from the raw, demultiplexed, aligned sequence data for use in the variant discovery. At the end of the calling, a basic filtering of SNPs was performed using the GATK Variant Filtration, setting the parameters as follows: (i) QD < 2.0, quality filtering by depth; (ii) MQ < 40.0, the Root Mean Square of the mapping quality of the reads across all samples; (iii) FS > 60.0, control against false positive calls using the Phred-scaled p-value in Fisher's Exact Test; (iv) SOR > 4.0, control against strand bias; (v) HaplotypeScore > 13.0, to ensure the reads at a site come from at most two haplotypes; (vi) MQRankSum < -12.5 applied to heterozygous calls as the u-based z-approximation of the Mann-Whitney Rank Sum Test for mapping quality. The output of the SNP variant discovery procedure was an SNP dataset with 3.259.191 marker sites in total. Genotype data was produced for each sample, but also, joint calls from all reads from the replicates of each parent were produced, to be used as the reference call for the parental lines, Midas and MG38. This part of the data processing was done by Dr. Alberto Ferrarini and Dr. Salvatore Benfatto at the University of Verona, Italy.

SNP dataset quality and filtering

Subsequent filtering was applied as a quality control, to exclude sequencing and alignment errors, and reduce missing data to have less noise in downstream analyses. VCFtools v0.1.13 was used to filter the dataset (Danecek et al., 2011). The filtering for minimum coverage (depth) per site was performed using a custom Python script applied in VCFtools, keeping only variants that appeared in at least 2 reads. This is seen as a good alternative to the usual read depth filtering set to 3 to 5 reads, as that type of filtering takes into account the sum of mapped reads, not per variant, which provides better control over the data quality, as each variant needs to be observed at least twice to be kept in the dataset.

Subsequently, only biallelic markers that were located on the 11 chromosomes were kept. To reduce the dataset size for further manipulation in the graphical user interface of the TASSEL v5.2.39 software (Glaubitz et al., 2014), markers with more than 80% missing data were removed, which resulted in the core dataset. Then, for increasing dataset quality and excluding potential sequencing errors, markers with minor allele frequency below 5% were removed using TASSEL. Additionally, filtering samples for maximum missing data (80%) and heterozygosity (25%) was done to exclude several samples that had an extremely low coverage or showed excessively high heterozygosity compared to the rest of the samples. Markers were also filtered for maximum missing data (80%) and heterozygosity (20%) to exclude SNPs resulting from misalignment of paralogs. The sample and marker quality filtering was done using a custom script in R, based on the proportion of missing and heterozygous calls that were counted per genotype and marker. Additionally, the SNPs that were not genotyped in both parents or genotyped as heterozygous in any of the two parents were also excluded from downstream analyses. An additional examination of the estimation of the error count was done using the ErrorCount Python script developed by J.B. Puritz for the dDocent pipeline for RADseq data (Puritz et al., 2014; *https://github.com/jpuritz/dDocent/raw/ master/scripts/ErrorCount.sh*).

The proportion of missing and heterozygous calls per marker and per genotype (before and after filtering), as well as the density of the discovered SNP markers along the common bean chromosomes were calculated and plotted using custom scripts in R. Most custom scripts in R that the results are based on were developed in collaboration with Dr. Matteo Dell'Acqua from Sant'Anna School of Advanced Studies in Pisa, Italy.

SNP density

The plots that show the comparison of the density of discovered SNP markers and the ones retained after the filtering procedure (the core dataset and the final filtered dataset for the downstream analyses) were based on TASSEL's Geno summary output for the two datasets. In each case, the markers were grouped in 1 Mb windows and the average across the windows was visualized in the plot, with the centromere ends marked with vertical dotted lines.

We have also compared the SNP density with the gene density across individual chromosomes. The SNP data relied on the filtered dataset's TASSEL Geno summary for the SNP position information, and the common bean genes' start and end positions extracted from the *Phaseolus vulgaris* reference genome gene annotation data. The density was visualized using R's density function under the default Gaussian smoothing kernel (the probability density function of the normal distribution), the centromeres being delimited with vertical dotted lines.

Heterozygosity

The observed allele frequencies and observed genotype frequencies were counted within the core dataset using a custom script in R. Based on this information, the observed proportion of heterozygous genotypes was plotted and the expected proportion of heterozygous genotypes based on allele frequencies and the Hardy-Weinberg equation. The expected proportion of heterozygous genotypes was both plotted without and with having the inbreeding coefficient (F) taken into account. When correcting for the inbreeding within the population, the inbreeding coefficient was set to 0.97 (as we had both lines from the F_5 and F_7 generation, we have decided to use the coefficient estimated for the F_5 lines) and the following equation was applied (Gillespie, 1988):

$$He = 1 - ((p^2 \times (1-F) + pF) + (q^2 \times (1-F) + qF))$$

The observed heterozygous genotype proportion was also plotted alongside the missing data proportion, and it can be viewed in Appendix 3. The data was plotted in a 150 bp long rolling window using the 'rollapply' function of the 'zoo' R package (Zeileis & Grothendieck, 2005) employing the mean function. Other R packages that were used to assist the manipulation of this data were 'diveRsity′ (Keenan et al., 2013), 'pegas′ (Population and Evolutionary Genetics Analysis System; Paradis, 2010), and 'adegenet` (Jombart, 2008; Jombart & Ahmed, 2011).

Population structure

We examined how the lines of the population are genetically related to eachother through: (i) Van Raden's population kinship matrix calculated through the 'GAPIT' R package (Lipka et al., 2012); (ii) the principal component analysis (PCA) of the genetic data calculated and plotted using the 'SNPrelate' (Zheng et al., 2012) and 'maptools' packages (Bivand & Lewin-Koh, 2017); and (iii) a dendrogram representation based on the kinship matrix visualized using the 'gclus' (Hurley et al., 2012) and 'APE' R packages (Paradis et al., 2004).

Linkage decay analysis

LD heatmaps were created using the R package 'LDheatmap' (Shin et al., 2006) coupled with the 'genetics' package (Warnes et al., 2013) for genotype data transformation into the right input format, 'gdata' package (Warnes et al., 2017) and 'reshape' package (Wickham, 2007) for data manipulation, all applied within a custom script in R. After creating the LD heatmap, we transformed the data into pairwise marker LD matrices over individual chromosomes with the melt function and grouped the pairwise distances into 1 Mb window increments in order to plot average LD within the certain distance grouping per chromosome. Using the same data, we have created a visualization of the LD profiles of chromosomes, first calculating the mean LD for each SNP position in relation to all other SNPs within a 10 Mb range, and then creating a plot using the rolling window approach ('rollapply'

function of 'zoo' package; (Zeileis & Grothendieck, 2005)) visualizing the mean LD value per SNPs positioned within a 100 bp window.

Genome composition and introgression detection

We have approached the estimation of the amount of wild genome introgression into the parental line MG38 and the RIL population in two ways. First, we wanted to observe the introgression profile of MG38 and each genotyped lines using only the SNPs that were homozygous and polymorphic between the parental lines, so that we could visualize along the genome each SNP variant in MG38 that differed from Midas and should originate from the wild original parent, G12873. Starting from the core dataset, we have set all heterozygous calls to missing and then removed all SNPs where any of the parents was genotyped as missing. We have created binary flags that marked the polymorphic sites between the parental genotypes (marking the polymorphic sites with 1 and the rest with 0), while the sites which were now set as missing data were also separately flagged. Scanning the entire dataset, we have marked within the population all sites which were genotyped with the same homozygous variant as the G12873 introgression to MG38 (1 marking the introgression). We have then formed a genomic visualization, where we mapped the G12873 flags on the genome representation of MG38 and all the lines. This resulted in a line by line overview of the genomic composition.

These results were also used to visualize the average introgression proportion over a 150 bp rolling window from G12873 into MG38 (using the 'rollapply' function of the 'zoo' R package) and then plot the average introgression per site across all the RILs, visualized also in a 150 bp rolling window. This gives an insight into the proportion of lines in which a recombination occurred during the population development where the introgression was shortened at a particular part of the chromosome in favor of increasing the amount of background Midas genome.

To visualize the introgressions from the semi-wild MG38 instead of the G12873, as we had genotyping data available only from the MG38 parental line in this study, we have repeated a similar process in an imputed haplotype dataset. First, we have recoded the heterozygous calls as missing using TASSEL's Homozygous genotype function. As this has increased the amount of missing data per marker, we have repeated the filtering for allowing a maximum of 80% of missing data per SNP. Then, imputation was performed using the LD-KNNi approach, which is based on a k-nearest neighbor genotype imputation method that takes into account the LD between the markers and can be used also for unordered markers in non-model organisms (Troyanskaya et al., 2001; Money et al., 2015; Money et al., 2017). Next, the TASSEL's ABH genotype caller was used to assign parental haplotypes to the RIL genotypes, and based on this data, the MG38 introgressions were visualized for each chromosome, presenting each line, using a custom script in R and the 'rollapply' rolling window function from the 'zoo' R package (Zeileis & Grothendieck, 2005). The estimated individual introgression lengths within chromosomes within all lines can be seen in the histogram in Appendix 4.

QTL mapping

The QTL analysis was performed using the genome-wide association algorithms in TASSEL (Bradbury et al., 2007; Glaubitz et al., 2014) and GAPIT (Lipka et al., 2012) implemented in R. The results using the general linear model (GLM), mixed linear model (MLM) and weighted linear model (weightedMLM) in TASSEL, and GLM and MLM in GAPIT were compared for the traits of flowering time and flower color. We have decided to present the results for the flower color using the weighted MLM in TASSEL with the Bonferroni correction for multiple testing (Sedgwick, 2014) using the first five PCs as fixed factors and the kinship within the population, as calculated within TASSEL, as a random factor. The results

presented for the flowering time trait were done in GAPIT using the MLM model, the VanRaden's kinship correction (VanRaden, 2008), examining adding the first five PCs as cofactors and applying the Bonferroni correction for multiple testing for the significance values (Sedgwick, 2014). To assist the manipulation of data for use in GAPIT, the following packages were used: 'multtest' (Pollard et al. 2005), 'gplots' (Warnes et al., 2016), 'genetics' (Warnes et al., 2013) and 'scatterplot3d' (Ligges & Mächler, 2003), while the source code for GAPIT and EMMA was acquired from Zhivu Zhang's website (zzlab.net). The remaining recorded phenotypic trait associations were analyzed by our collaborators and will be published and discussed separately.

Results

Sequencing quality

The HiSeq run resulted in 26 compressed fastq datafiles, two for each Illumina index used in the nested multiplexing approach. Sequencing has yielded 86.27 GB of data in fastq.gz compressed file format, with the individual file sizes ranging from 1.83 to 5.32 GB. In Figure 6, a boxplot representation of the batch effect in data yield across libraries and sequencing lanes is presented through the amount of missing data within the core SNP dataset. In the FastQC reports, the main reason for the difference between the yield in the two lanes that were ran at different times can be seen, as in the second lane, that comprised the GBS libraries from seven through 13, had a drop in read quality towards the ends of the reads due to a technical failure in the sequencer. As a consequence of this, there are blocks of SNP marker sites with high missing data at certain parts of the dataset, at location of the ends of fragments sequenced in that lane.

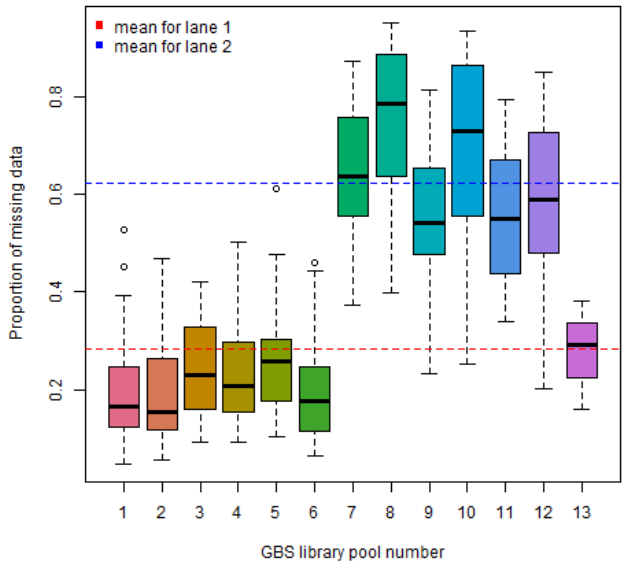


Figure 6. The proportion of missing data across GBS library pools.

SNP marker yield

There were 3,259,191 SNP markers discovered in the initial dataset. After filtering for minimum variant depth of 2 reads, 2,073,044 SNPs (63.6%) remained in the dataset. Of those, 2,048,766 (98.83%) were located on the 11 chromosomes of the common bean genome reference (version 2). From those, 2,038,719 (99.5%) were biallelic, and 78,918 SNPs (3.87%) had less than 80% missing data per site: these markers made up the core dataset. After filtering for 5% MAF, the dataset was reduced to 31,344 SNPs (39.72%). Following the sample-wise and marker-wise filtering based on heterozygosity and missing data, as well as controlling for those markers that were genotyped as homozygous in parents, 18,385 markers (58.66%) were retained. Table 5 represents the chromosome length, centromere position, and the marker numbers per chromosome in the core versus the filtered dataset. After the filtering procedure, chromosome 5, 10 and 11 had the biggest reduction in the number of markers.

Table 5. Chromosome length (in bp), centromere positions (in Mb, as identified by Schmutz et al. (2014) and in the *P. vulgaris* V2 reference genome), the number of SNPs discovered by chromosome (biallelic, with max 80% missing data) and retained after quality filtering (see Methods). Marked with gray are the three chromosomes where the lowest number of SNP markers were discovered.

chr. numb.	centromere start	centromere end	chromosome length	SNPs discovered	SNPs retained
1	12.2	19.9	51,433,939	7,158	2,168
2	5.4	10.0	49,670,989	9,126	2,286
3	14.8	16.9	53,438,756	11,560	3,090
4	15.7	22.2	48,048,378	7,706	1,573
5	15.3	22.7	40,923,498	4,183	385
6	2.6	2.7	31,236,378	5,010	1,973
7	16.7	30.3	40,041,001	5,968	1,327
8	24.3	38.2	63,048,260	9,757	2,220
9	1.5	5.8	38,250,102	6,605	2,725
10	30.6	31.3	44,302,882	6,483	492
11	16.0	17.1	53,580,169	5,362	46
		Total:	**513,974,352**	**78,918**	**18,385**

Dataset assessment before and after filtering

We have compared the core and filtered dataset, to assess how filtering changed certain metrics across markers and samples. In Figure 7 are shown the distributions of missing data and heterozygosity, calculated both marker-wise and sample-wise, as well as minor allele frequencies (MAF) per SNP site. The distributions from the core dataset were used for decision making for the filtering cutoff values, while the ones after filtering were obtained to show the effect of the filtering procedure on the resulting filtered dataset for downstream analyses.

The most notable difference between the dataset is visible through the elimination of SNP sites with MAF under 5% (44,574 markers removed; 56.48%) and heterozygosity higher than 20% (12,959 markers removed; 16.42%) which reduced the filtered dataset to 18,385 markers (23.30%). Based on the heterozygosity cutoff at 25% and maximum allowed missing data at 80% sample-wise, 32 genotypes were removed from the dataset (one Midas replicate and 31 RILs), leaving 261 genotypes within the dataset (including the joint parental calls and replicates which passed the filtering criteria). Based on TASSEL's Geno Summary, the dataset summary statistics were extracted (see Table 6.).

Table 6. Dataset summary statistics

	Core dataset	**Filtered dataset**
Number of taxa	291	261
Number of sites	78,918	18,385
Percentage missing in dataset	42.73%	29.20%
Max missing per marker	81.44%	79.69%
Max missing per sample	95.03%	75.21%
Percentage heterozygous in dataset	3.40%	3.41%
Max heterozygous per marker	94.35%	20.00%
Max heterozygous per sample	15.50%	21.86%
Average MAF	0.1292	0.3690

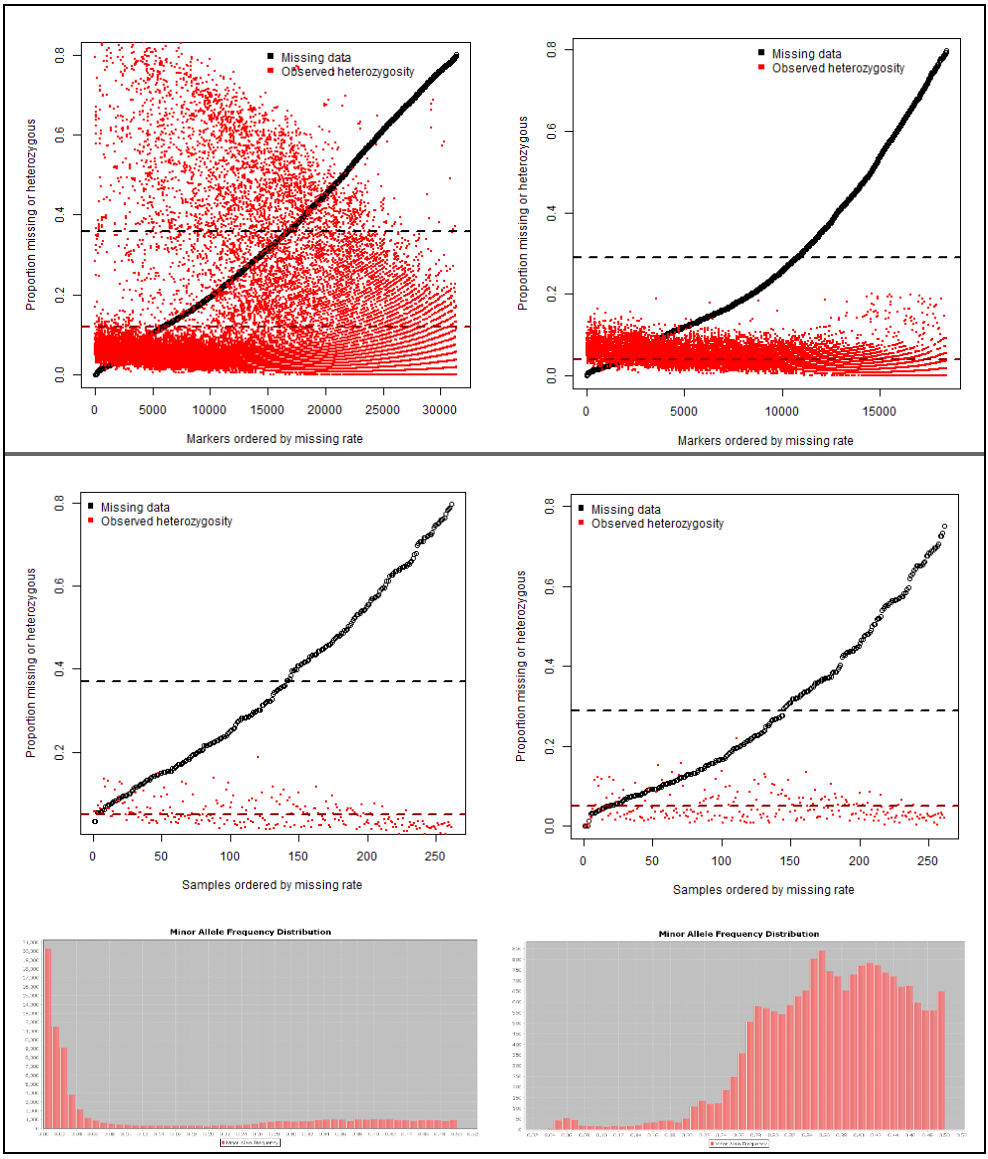


Figure 7. Left, plots for the core dataset; **right**, plots for the filtered dataset. **Top**, marker-wise heterozygosity, and missing data proportions; **middle**, the sample-wise geterozygosity and missing data proportions; **bottom**, minor allele frequency distribution (MAF). The black dotted lines mark the mean value for missing data proportion, while red the mean of heterozygosity proportion.

SNP density

Figure 8 and 9 show the density of SNP markers in the core dataset (black line) versus the filtered dataset (red line) by plotting the average number of markers at SNP sites located within 1 Mb bins. The dashed gray lines mark the start and end of the centromeric region. The SNP reduction between the two datasets is a result of the application of a filter for minimum 5 % MAF value and maximum heterozygosity cutoff at 20%, as described in the Data assessment before and after filtering.

In several chromosomes there is a drastic reduction in SNP number in the telomeric regions, that reduces a high peak with hundreds of markers to a near zero count (see the end of chromosome 1 and both ends of chromosome 3 and 7). There are also high peaks in centromeric and pericentromeric regions that are affected by this filtering (see the region from 22 to 26 Mb in chromosome 1; 19 to 22 Mb and 26 to 28 Mb in chromosome 3 and numerous peaks in chromosome 8). Chromosomes in which the entire central area is affected by this filtering are chromosome 4 and 5, while in chromosome 6 and 10 this affects only the beginning half of the chromosome. Almost all markers get filtered out in chromosome 11, retaining only 46 markers in total.

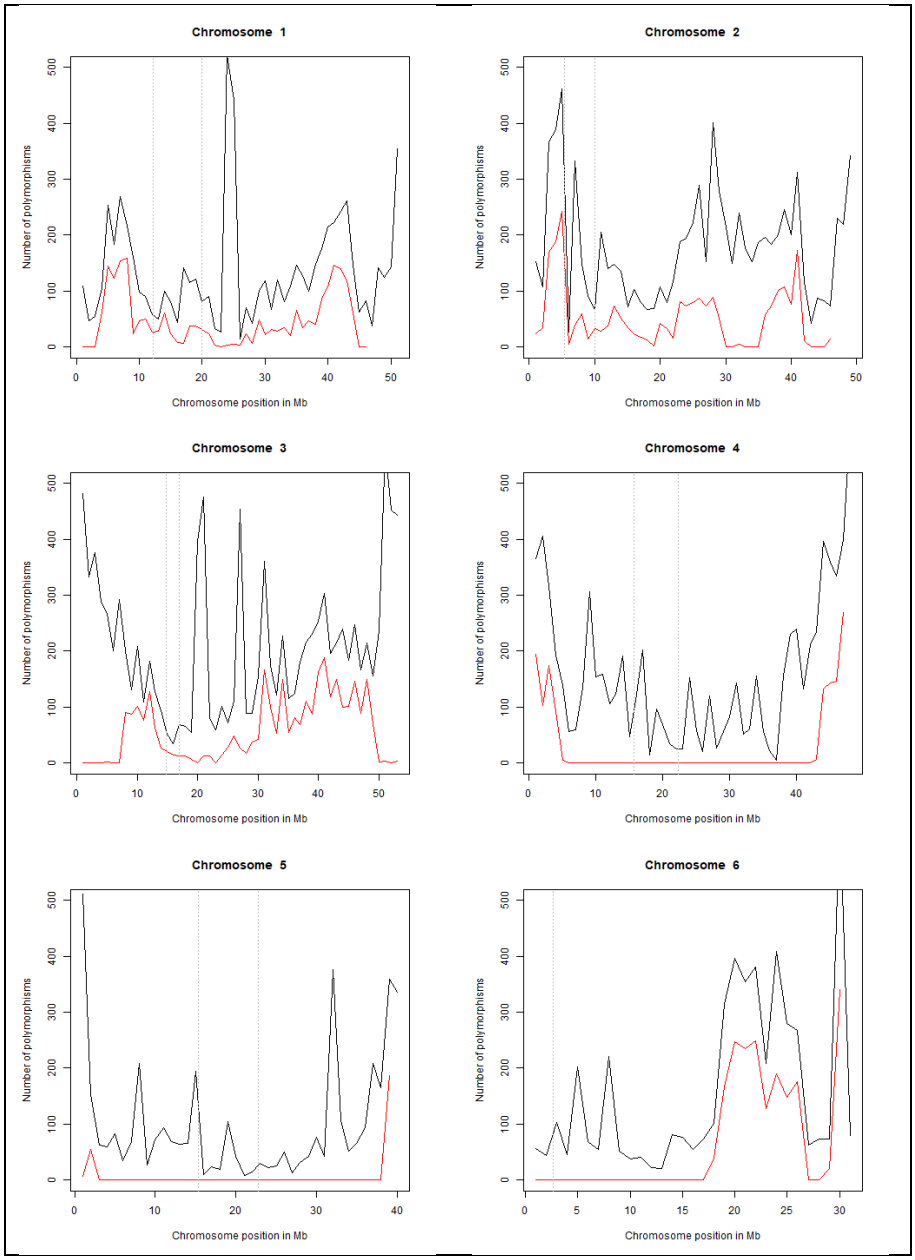


Figure 8. SNP density across chromosomes 1 to 6 in core (black line) and filtered (red line) datasets. Average values within 1 Mb windows shown.

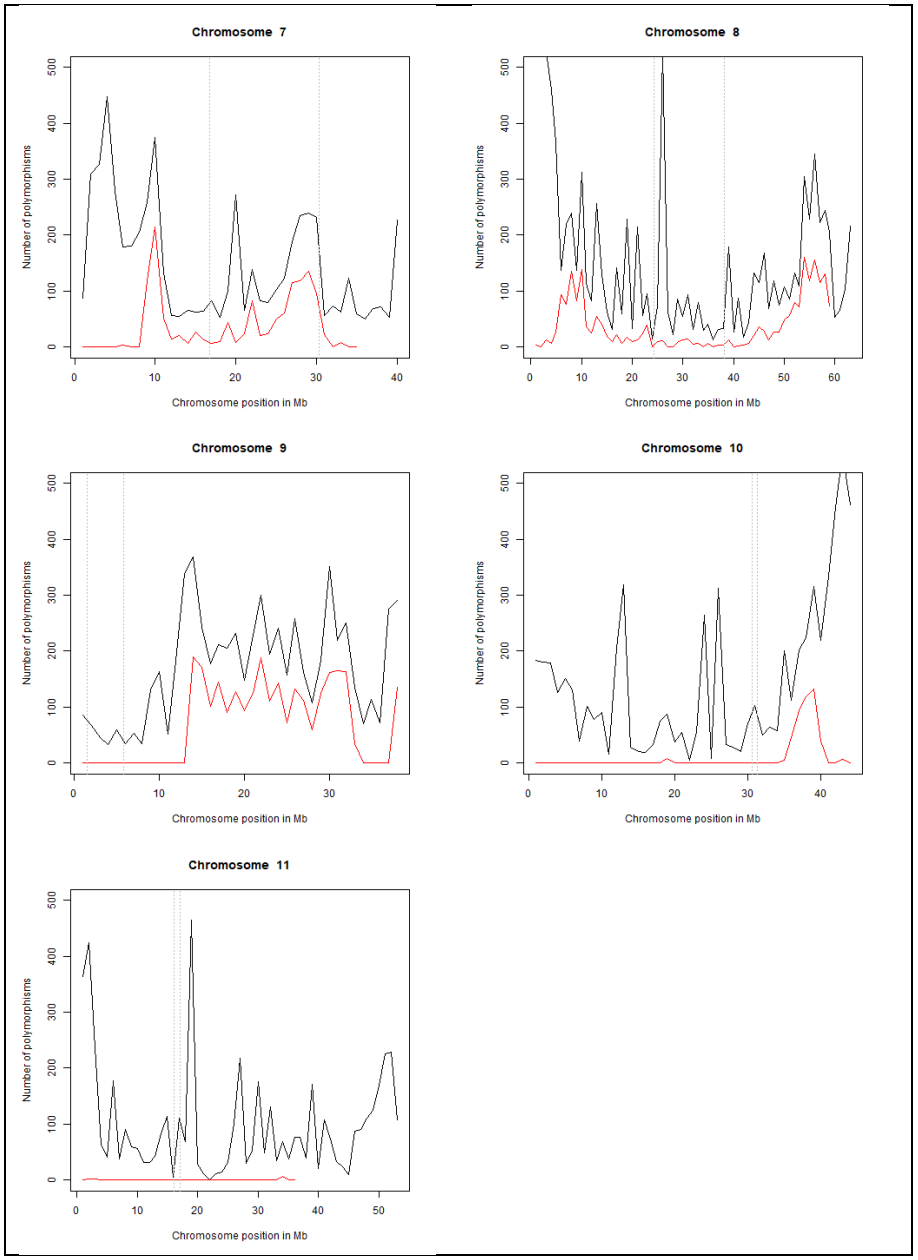


Figure 9. SNP density across chromosomes 7 to 11 in core (black line) and filtered (red line) datasets. Average values within 1 Mb windows shown.

Gene density

Figure 10 shows how the gene density (black line) correlates to the distribution of SNP marker density (red line) in the filtered dataset, with Gaussian smoothening kernel applied, presented per single chromosome. In most chromosomes, high SNP density correlates with high gene density, but as polymorphism detection is only possible in regions which harbor wild introgressions, the SNP marker distribution can also be taken as a proof for introgression.

Higher densities are often observed in the central regions of chromosomal arms, except in chromosomes 2 and 7, where on each chromosome there is an SNP density peak that in part overlaps with the centromeric region. On chromosomes 5 and 10 the SNP density peaks are the highest, as most markers that passed the filtering criteria are distributed over very narrow regions.

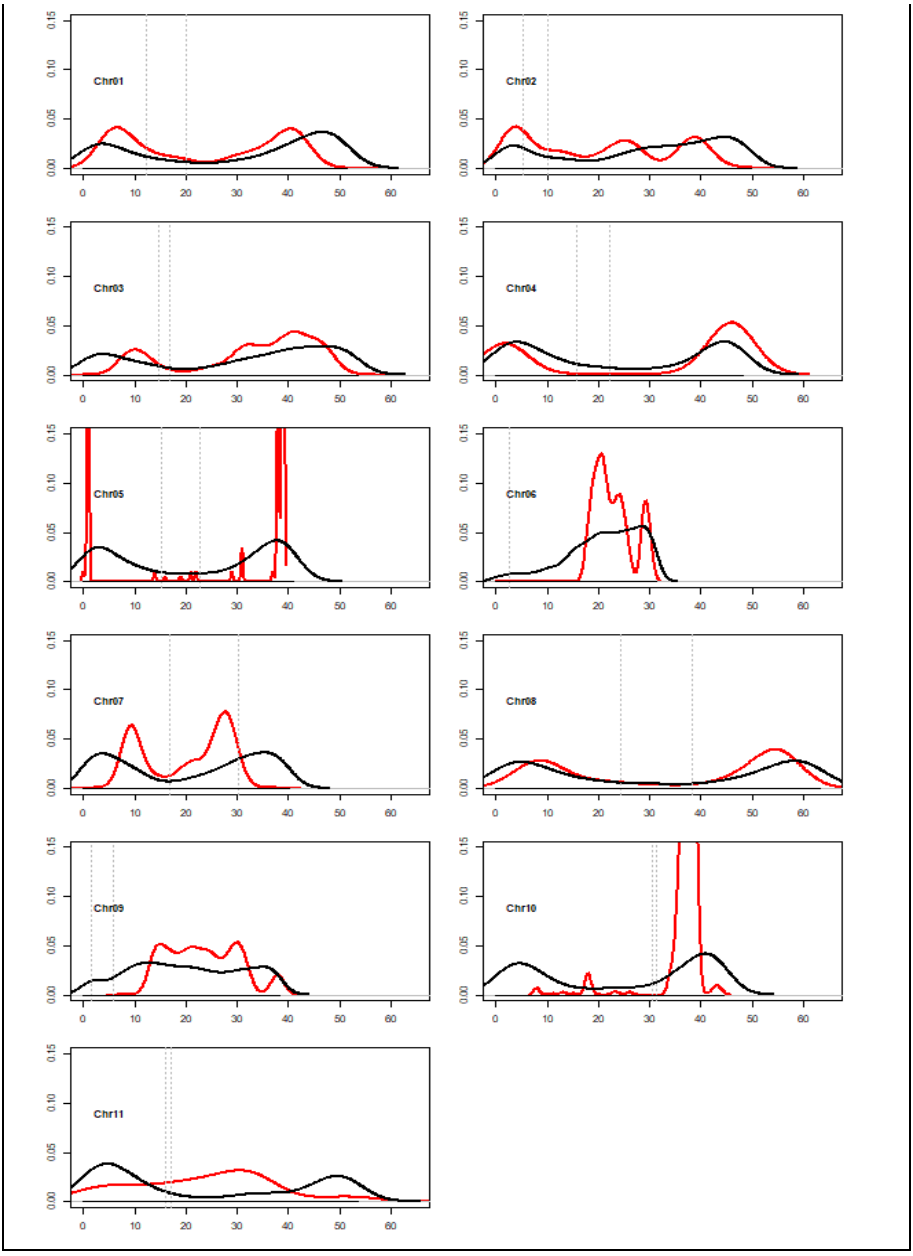


Figure 10. Gene density (black line) and SNP density within the filtered dataset (red line) shown across the 11 chromosomes. The dashed lines mark the location delimitos for the centromeres.

Heterozygosity

SNP marker heterozygosity was examined through the proportion of the observed heterozygous genotypes within the core dataset (red line), the expected proportion of heterozygous genotypes based on the Hardy-Weinberg equation calculated from the allelic proportions of marker variants (gray line), and the expected proportion of heterozygous genotypes corrected with the inbreeding coefficient (F; black line). In order to not have the results affected by the marker-wise heterozygosity filter, the values were calculated on the core (unfiltered) dataset. The values were plotted as the average in a 150 bp rolling window and are presented in Figures 11 to 14.

The uncorrected expected heterozygosity is high where there are introgressions, especially where there is a relatively balanced segregation within the population (allele frequencies having a tendency towards 0.5). As the genotyped individuals belong to generations F_5 and F_7, the inbreeding coefficient of 0.97 was used for the correction, which corresponds to the F_5 generation inbreeding value. The observed heterozygosity peaks were arbitrarily determined and marked with arrows. Black arrows mark the regions which overlap with those that in which the majority of markers was filtered out in the filtered dataset, while the ones marked with red were in regions that were less affected by the heterozygosity and MAF filter. Red rectangles mark regions which show an introgression based on the high expected heterozygosity, but which have a low observed heterozygosity (see chromosomes 2, 4, 6 and 9).

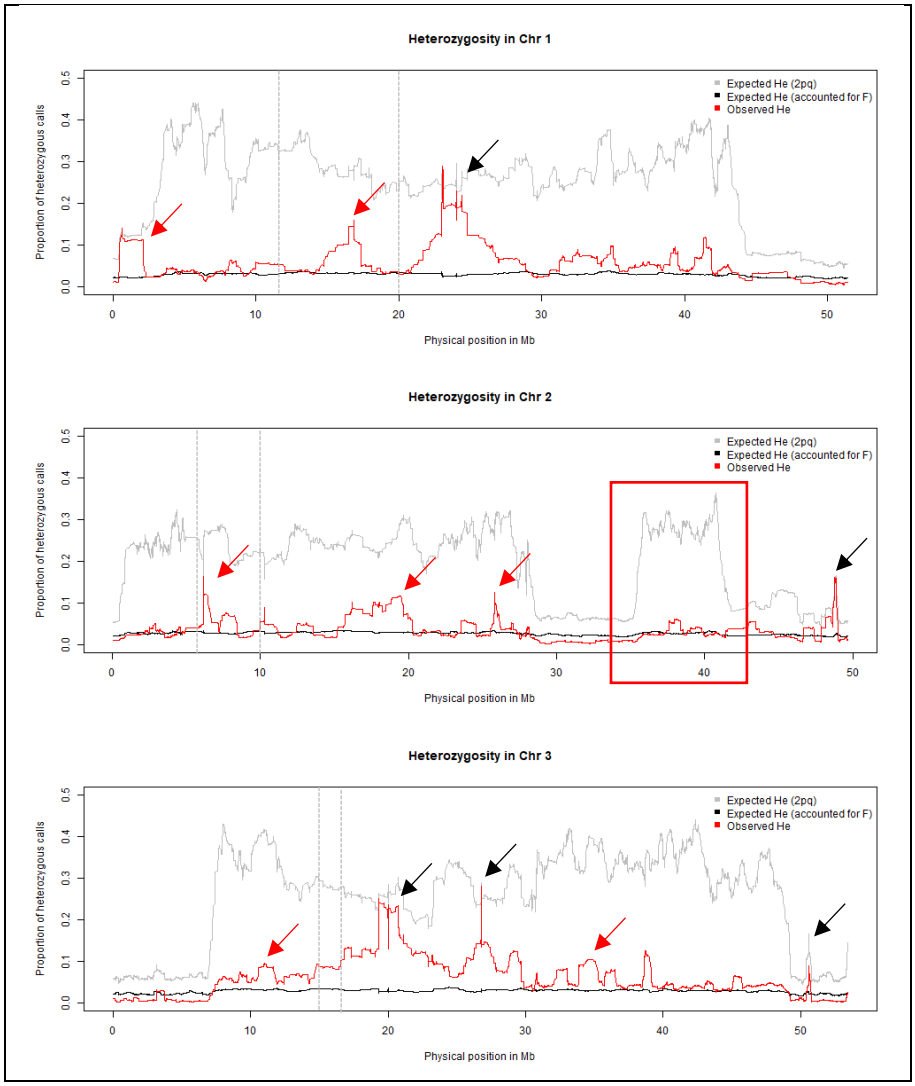


Figure 11. Observed heterozygosity (red), expected heterozygosity (gray) and expected heterozygosity corrected for the inbreeding coefficient (F, black) for the population plotted as the average within a rolling window of 150 bp in length. The dasshed lines mark the centromere location.
Data is shown for chromosomes 1 to 3.

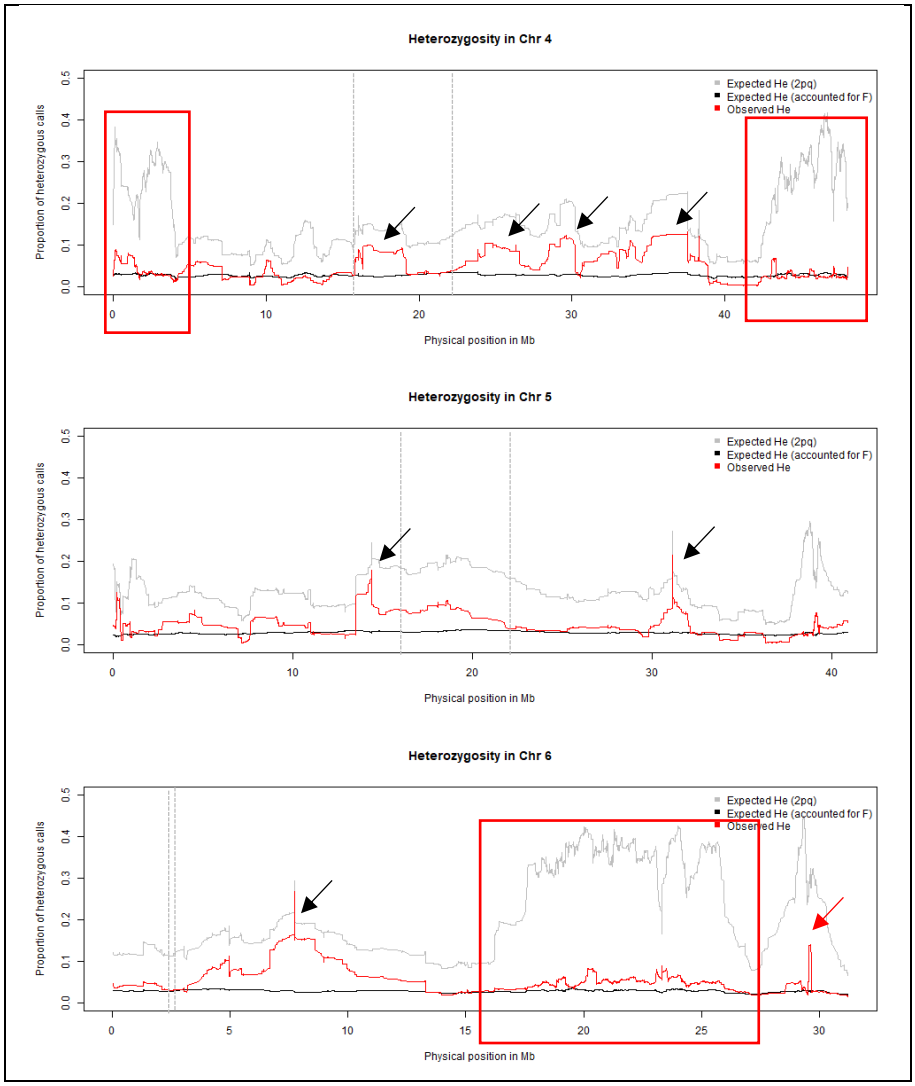


Figure 12. Observed heterozygosity (red), expected heterozygosity (gray) and expected heterozygosity corrected for the inbreeding coefficient (F, black) for the population plotted as the average within a rolling window of 150 bp in length. The dasshed lines mark the centromere location.
Data is shown for chromosomes 4 to 6.

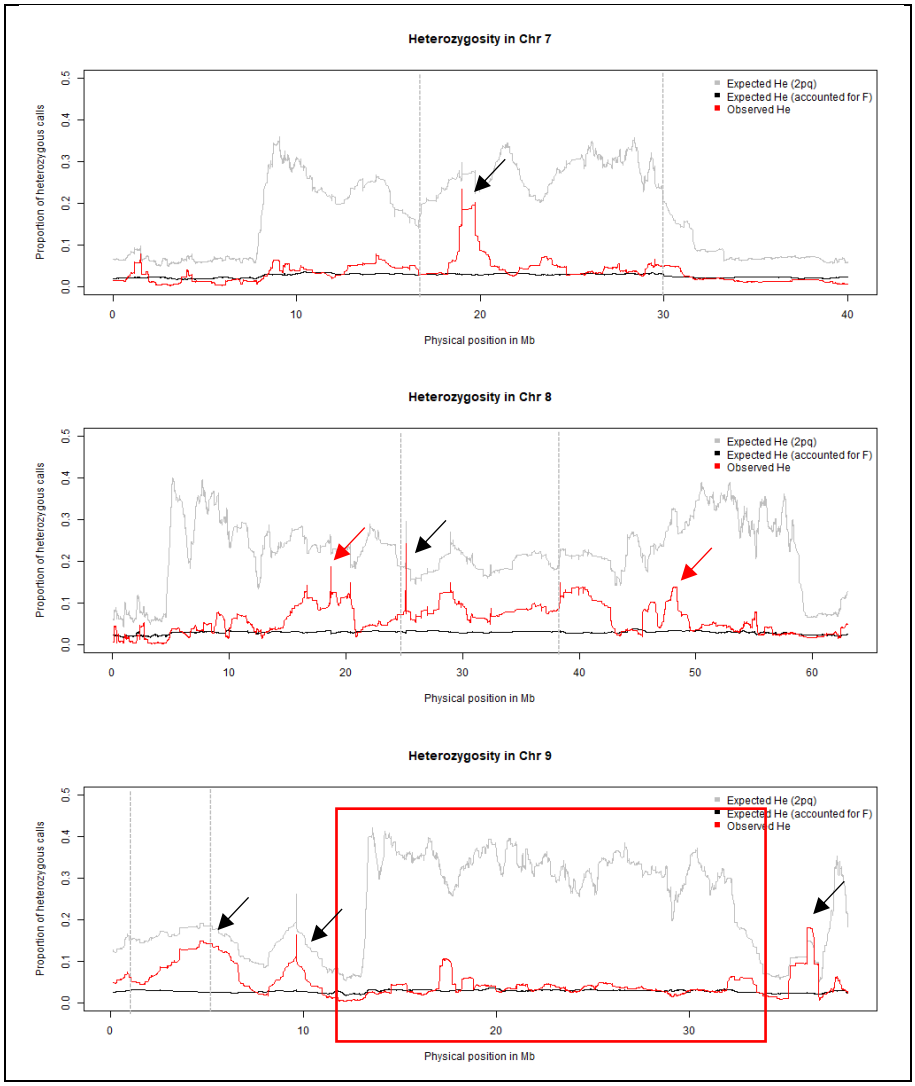


Figure 13. Observed heterozygosity (red), expected heterozygosity (gray) and expected heterozygosity corrected for the inbreeding coefficient (F, black) for the population plotted as the average within a rolling window of 150 bp in length. The dasshed lines mark the centromere location.
Data is shown for chromosomes 7 to 9.

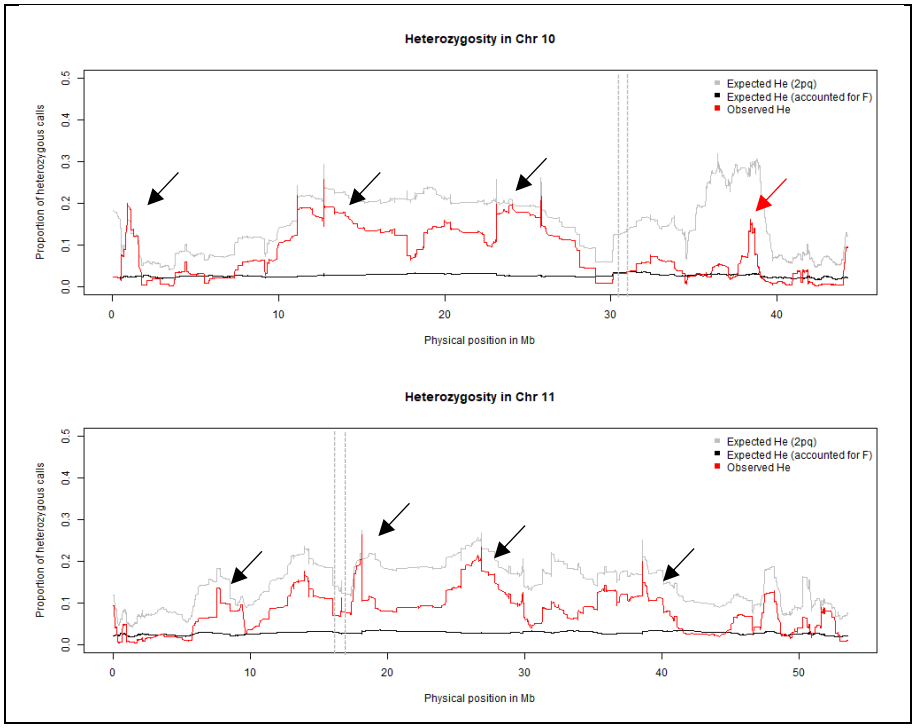


Figure 14. Observed heterozygosity (red), expected heterozygosity (gray) and expected heterozygosity corrected for the inbreeding coefficient (F, black) for the population plotted as the average within a rolling window of 150 bp in length. The dasshed lines mark the centromere location. Data is shown for chromosomes 10 and 11.

Population structure

The relationships between the lines in the population were examined using the principal component analysis (PCA) of genotype data. The population shows intermediate levels of structure, where a few families with a bigger representation in individuals have a clear separation, while being interspersed with individuals from some families with fewer representatives (see Figure 15).

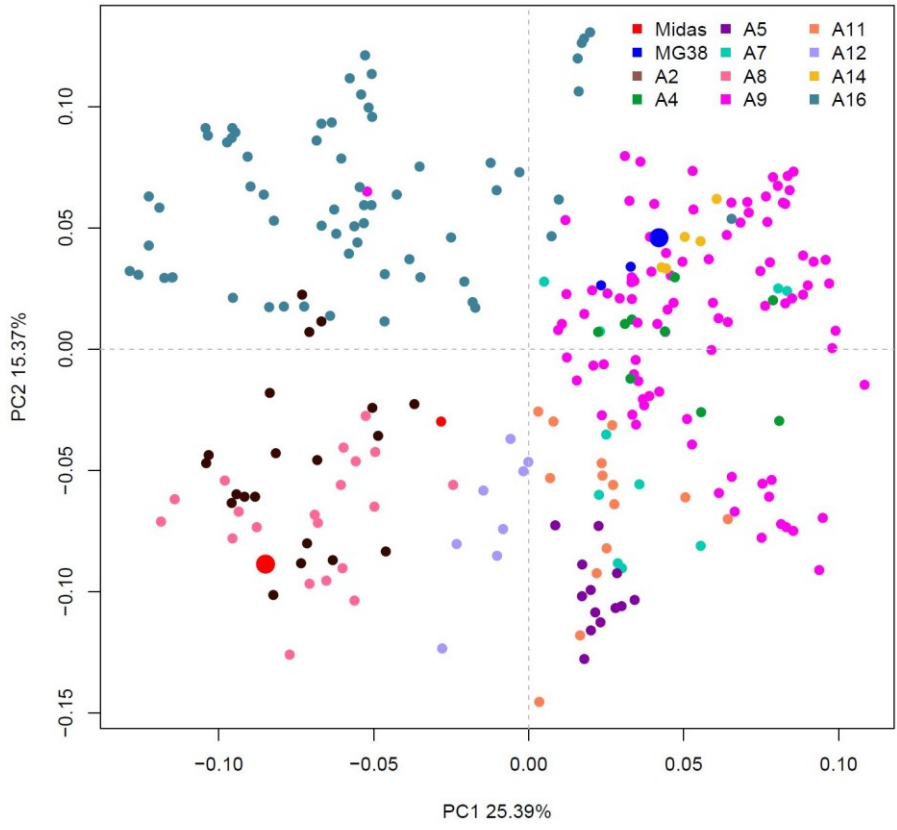


Figure 15. PC1 and PC2 plot

The first 5 principal components explain 25.4%, 15.4%, 10.8%, 7.1% and 5.4% of the genetic variation respectively (see Figures 15 and 16). The RILs show a diffuse distribution with respect to the parental lines, while the parental lines are genetically distant, as expected (Figure 16).

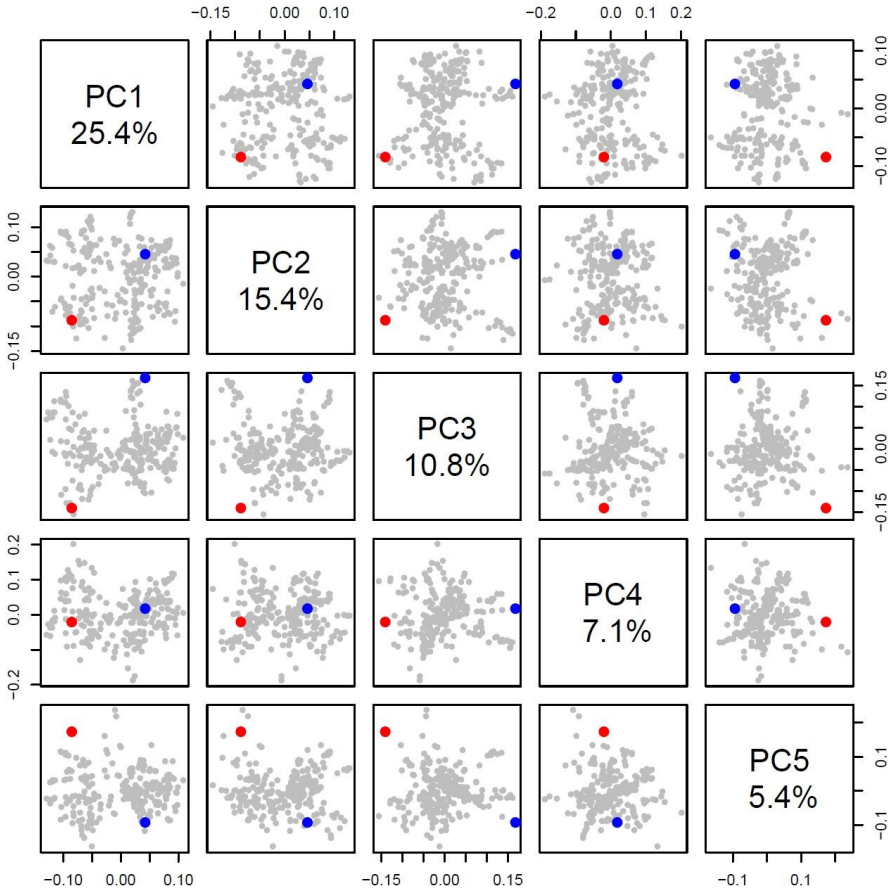


Figure 16. PCA plot showing the first 5 principal genetic components. The Midas parent is marked with red and MG38 with blue.

The kinship between RILs was calculated in GAPIT based on marker relationships as proposed by VanRaden (2008) and can be seen in Figure 17. The population is subdivided into two main groups (and several subgroups), where the first (about 1/3 of lines) is more similar to Midas (top right) and the second (about 2/3 lines) is more similar to MG38 (bottom left). The kinship matrix also doesn't show an overall strong population structure population structure, and the phylogenetic tree does not directly resemble the F_2 family structure of the population (see Appendix 5).

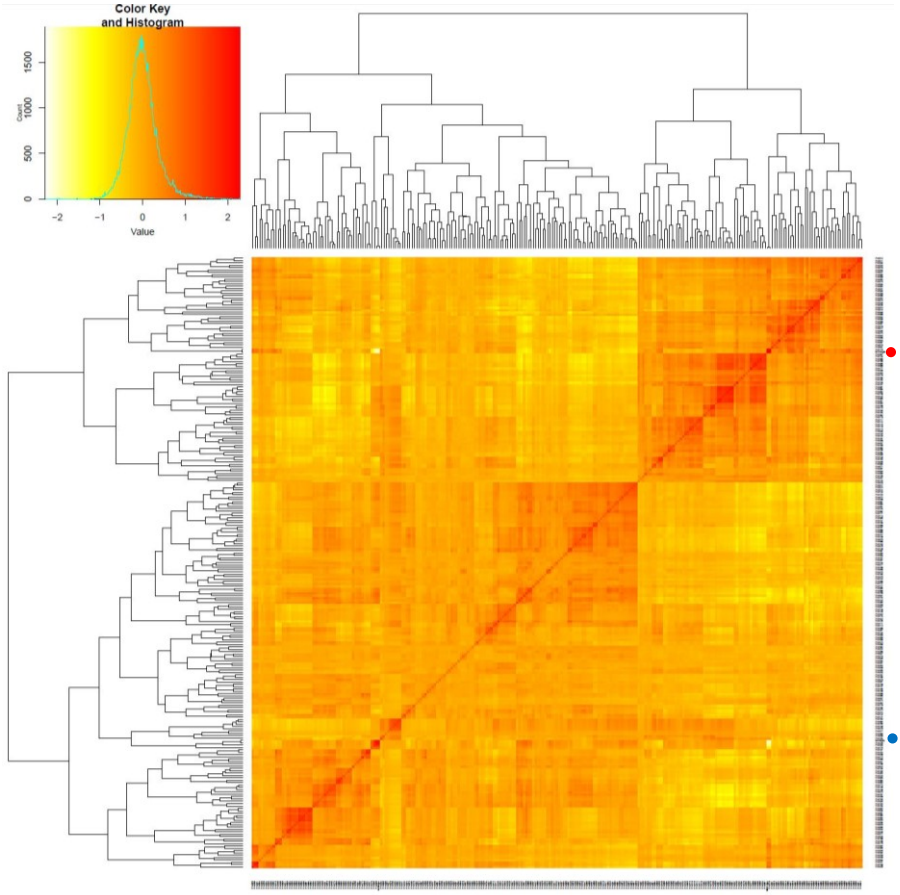


Figure 17. The kinship heatmap of the population based on the Van Raden marker relationship calculation (VanRaden, 2008). The parents are marked with dots, Midas in red and MG38 in blue.

Linkage decay analysis

The linkage decay (LD) is presented in bins that contain pairs of loci at a certain distance, up to 10 Mb apart (Figure 18). At this physical distance, the LD across chromosomes 1,7 and 8 is slow (not lowering past mean $r^2 = 0.5$), in chromosomes 2, 3 and 9 having intermediate decay (between $r^2 = 0.2$ and 0.5), while being the fastest in chromosomes 4, 5, 6 and 10, and being hard to estimate for chromosome 11, due to low marker coverage. LD heatmaps and the LD evolution across single chromosomes can be seen in Appendix 6.

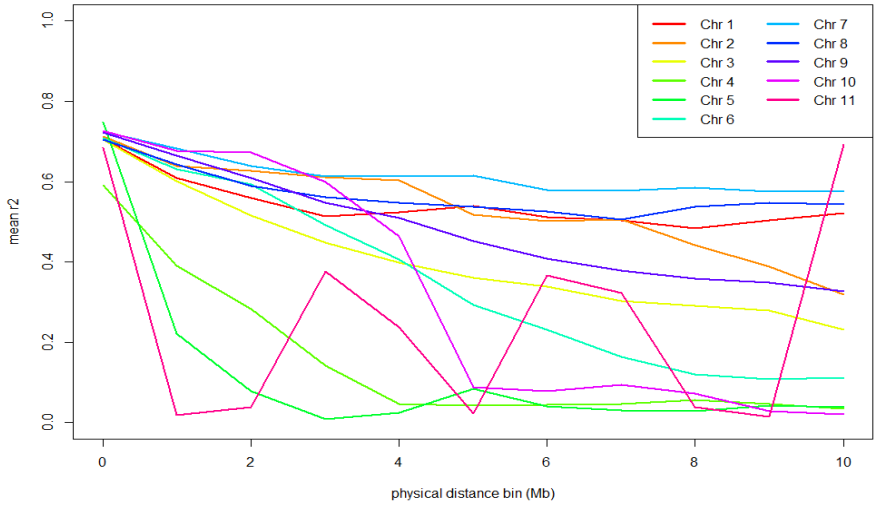


Figure 18. The decay in linkage observed using the mean R^2 value between markers at a certain physical distance, grouped by chromosome location.

Genome composition and introgression detection

The reconstruction of the genome composition and detection of introgressions and parental haplotypes was performed in two ways. The first is a conservative and straightforward method, where the polymorphisms detected in the parental genotypes across the RIL population were mapped, marking the G12873 calls (homozygous variant calls unique to MG38) with blue and the Midas calls (the remaining homozygous variant calls) with yellow, keeping only the markers genotyped in both parents. Plots were created for each RIL, showing its genome composition. In Figure 19 is the mapping of G12873 introgressions for the MG38 parent. This represents the maximum size of introgressions that can be passed on to RILs in the population, while in RIL genomes we can then see how and where the backcrossing and inbreeding have reduced the size of the introgressions. As a summary representation of the results of this approach, the average proportion of G12873 calls detected across the RILs was visualized along the average proportion of G12873 calls in MG38 in 150 bp rolling windows (on the left side in Figures 20 to 22). Red arrows mark stretches of introgression present in MG38, but which were not retained in the RILs. The second approach was applied in a dataset cleaned from heterozygous calls (set to missing calls), followed by imputation (based on k-nearest neighbor and marker LD) and the calling of parental haplotypes (ABH genotypes). The results are shown on the right side in Figures 20 to 22, where introgressions in each line are stacked one atop the other, with the chromosome length shown in blue at the bottom and a centromeric region marked with red. The plots with the frequency of detected haplotype lengths per chromosome are in Appendix 4.

Based on the presence, location and number of introgressions, chromosomes can be grouped into 5 groups: (i) one introgression spanning across the centromeric region (chromosomes 1, 2, 3, 7 and 8); (ii) one introgression, not covering the centromeric region (chromosome 10); (iii) two introgressions of which one spans across the centromeric region (chromosome 4); (iv) two introgressions of which none cover the centromeric region (chromosomes 5, 6 and 9); and (v) no detected introgression (chromosome 11).

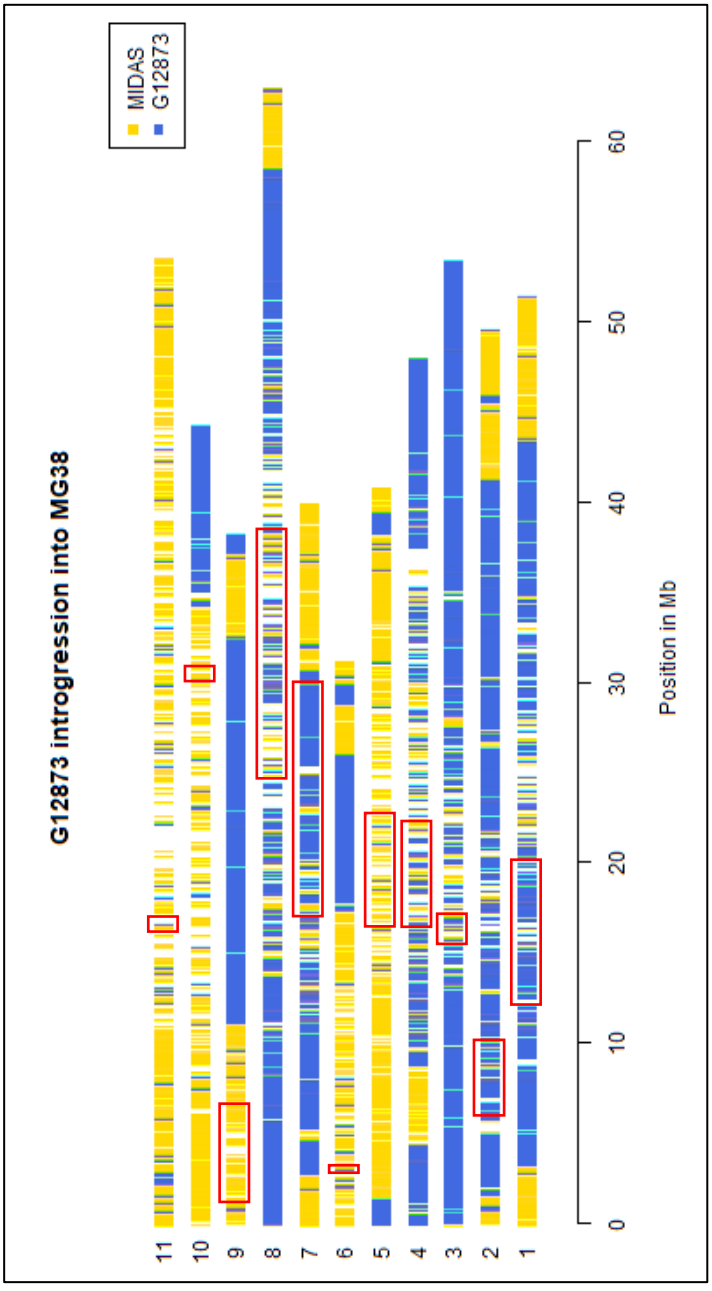


Figure 19. G12873 introgressions (blue) into the Midas genomic background (yellow) in the parental line MG38. Centromeric regions are marked with red rectangles.

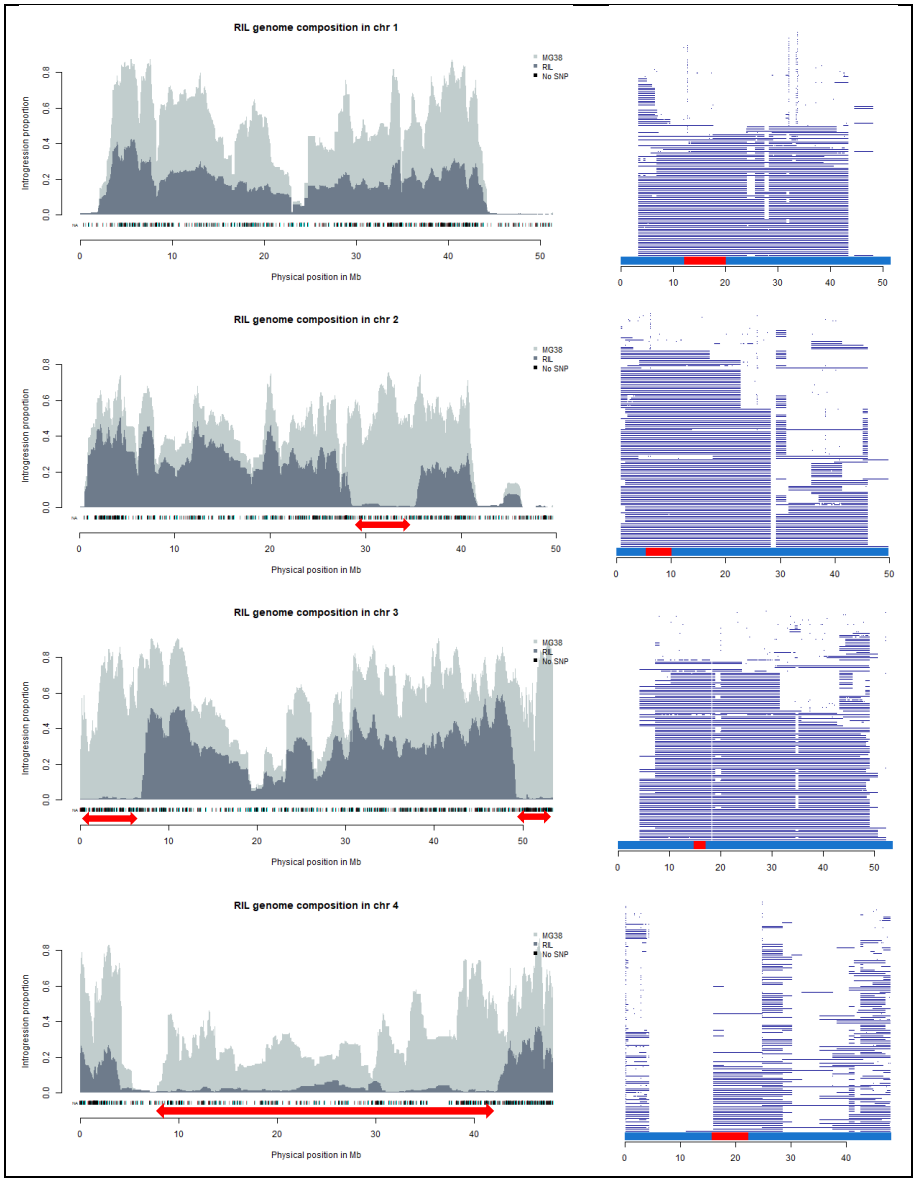


Figure 20. Detection of G12873 introgressions in MG38 and RILs (left) and of wild parent (MG38) haplotypes in the RILs (right). Chromosomes 1 to 4 shown.

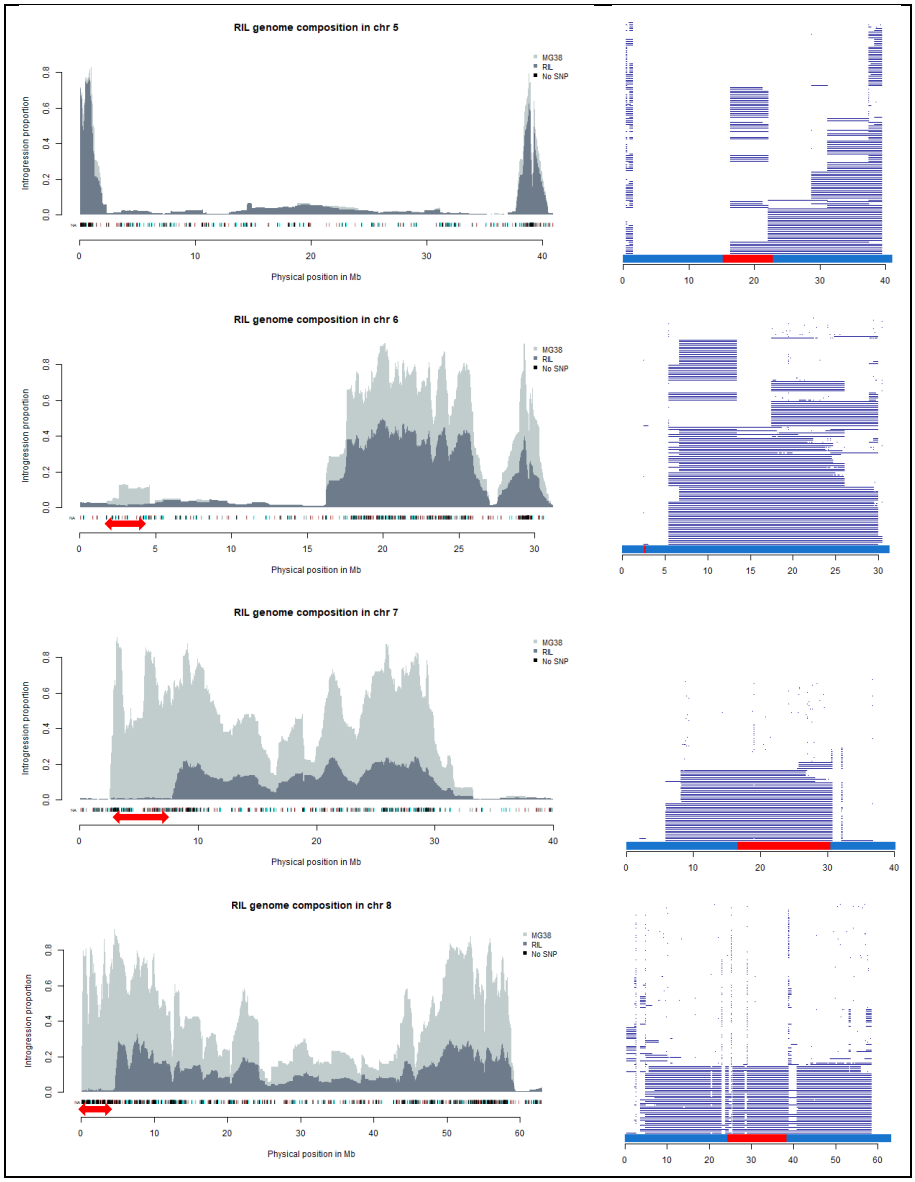


Figure 21. Detection of G12873 introgressions in MG38 and RILs (left) and of wild parent (MG38) haplotypes in the RILs (right). Chromosomes 5 to 8 shown.

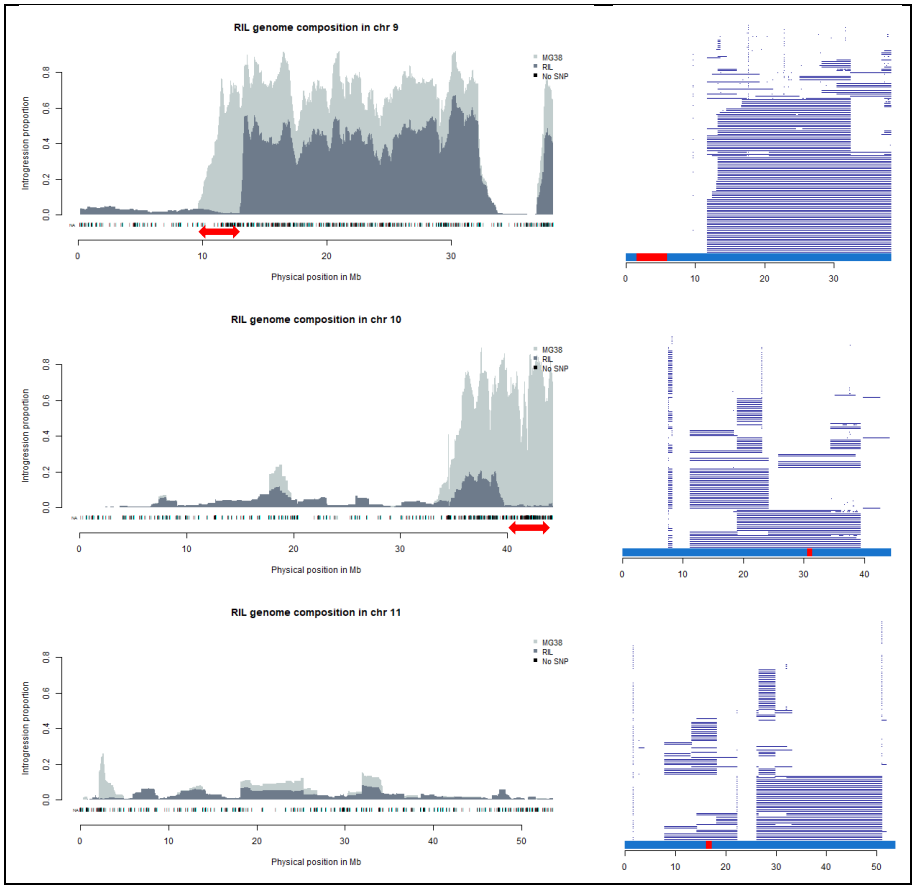


Figure 22. Detection of G12873 introgressions in MG38 and RILs (left) and of wild parent (MG38) haplotypes in the RILs (right). Chromosomes 9 to 11 shown.

The RILs from this population have a potential to be used in the QTL-NIL mapping approach, and in Figure 23 it is demonstrated how the introgression size and location varies in chromosomes 1 and 7 (which had the strongest QTL association signals in QTL mapping for flower color and flowering time, marked with triangles; see 'QTL mapping segment' for more details) to enable this type of trait mapping.

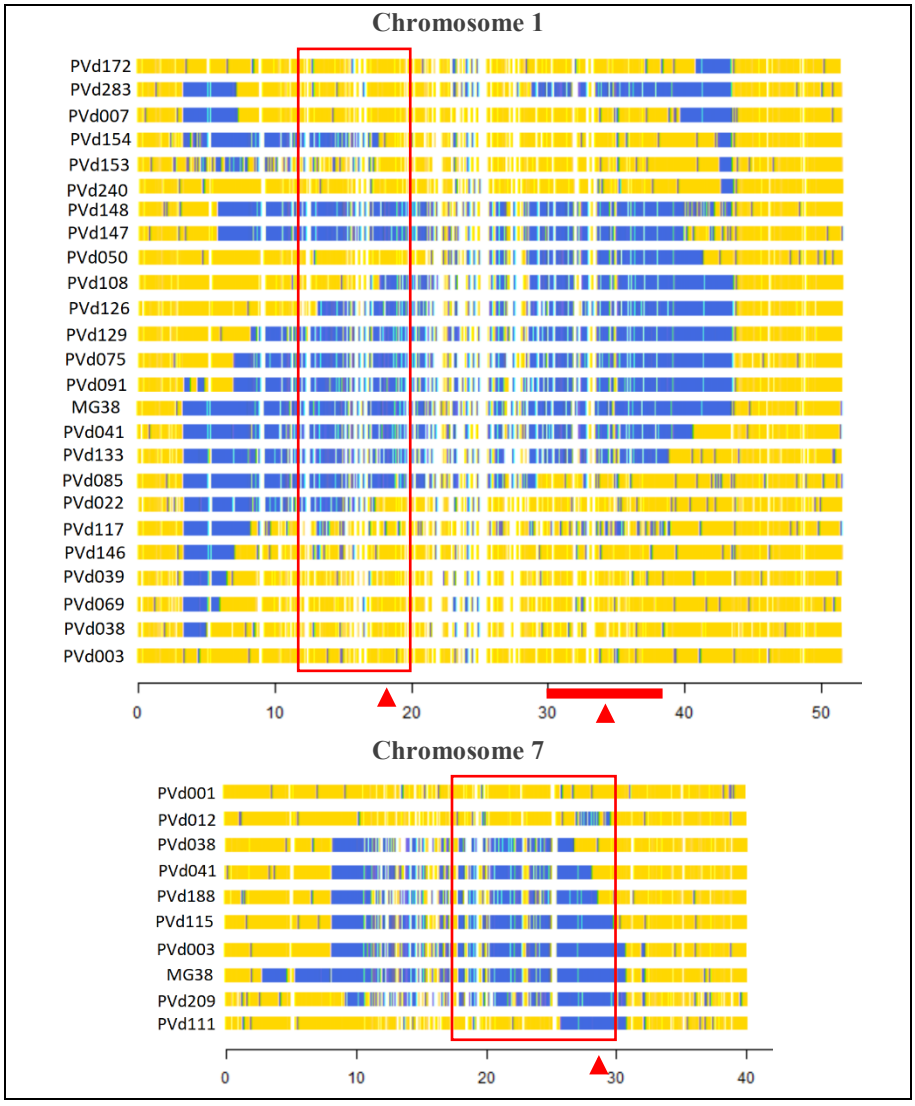


Figure 23. The variation in introgression size and position in chromosome 1 and 7. Red rectangles mark the location of the centromeric region. Triangles mark the chromosome location where QTLs were discovered in this population.

QTL mapping

Two of all the traits that were phenotyped in the population were used to demonstrate the utility that this kind of population can have for QTL mapping: flower color, as a qualitative trait, showing a white, violet or intermediate phenotype; and flowering time, as a quantitative trait, measured as the number of days to flowering from the flowering of the first plant in the population. The distributions of the trait values are presented in Figure 24. Almost 2/3 of the RILs have a white flower phenotype (equal to the Midas parent), and from the remaining RILs, about 2/3 have a violet flower and 1/3 an intermediate violet color (heterozygous genotype). For the trait of flowering time, we see that the majority of the RILs flowered from four to nine days from the first flowering RIL. RILs with more genomic similarity Midas would flower earlier (as it is insensitive to photoperiod length), while the RILs with more genomic similarities to G12873 would have a delay in flowering time, as the wild parent is photoperiod sensitive, which causes a delay in flowering.

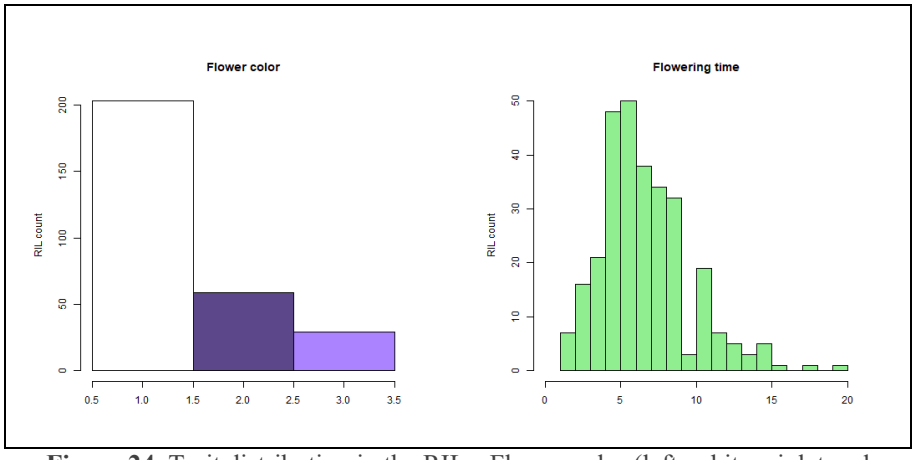


Figure 24. Trait distribution in the RILs. Flower color (left; white, violet and intermediate violet) and days from the flowering of the first RIL (right).

The results for the flower color trait were produced using the weighted MLM model in TASSEL, using the first five PCs as fixed factors and the kinship matrix as a random factor, with the Bonferroni correction for multiple testing. The Manhattan plot of the $\log_{10}$ transformation of corrected p-values for flower color is at the top of Figure 25. One QTL region on chromosome 7 has shown the strongest association. The 13 markers with the strongest signal are in the region from S07_27916108 (29 Mb) to S07_29412376 (29.5 Mb). There were also significant associations on chromosomes 2 (S02_38217847), 6 (S06_29871340) and 8 (near S08_55012146 and S08_18669107). The QQ plot does not show that there is an inflated false positive discovery (Figure 25, bottom left).

The results for the flowering time trait were obtained in GAPIT using the MLM model, VanRaden kinship correction, Bonferroni correction for multiple testing, and the first PC as a cofactor. There are 2 QTL signals (Figure 25, the Manhattan plot in the center), one wide on chromosome 1, where from the 12 markers with most significant association, one is at S01_18496016 (18.5 Mb) and the rest span over the region from S01_30136692 to S01_38241648 (30.1 to 38.2 Mb); and one marker with a strong association on chromosome 8 at S08_38812200 (38.8 Mb). The QQ plot with the p-values for flowering time is inflated due to a large number of associations on chromosome 1 (Figure 25, bottom right).

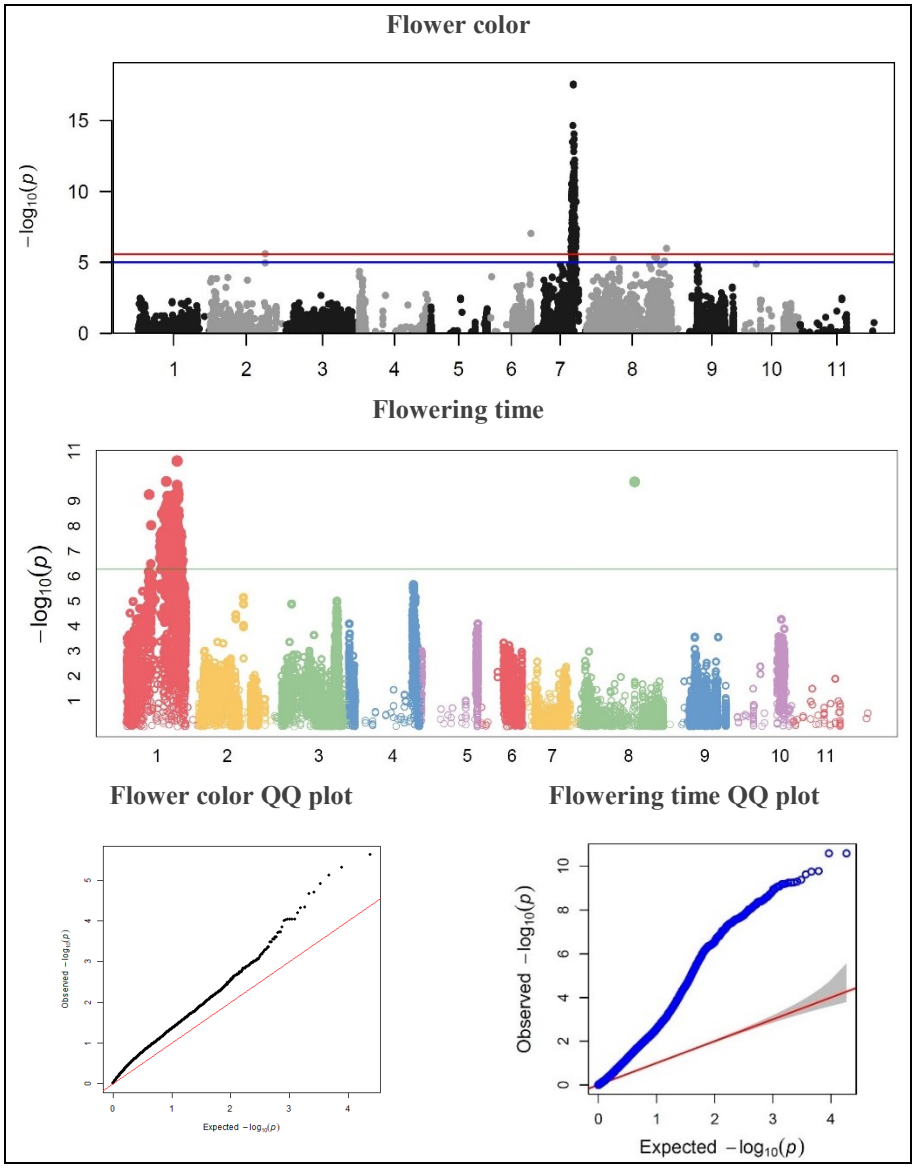


Figure 25. QTL analysis for flower color (up) and flowering time (middle) with QQ plots for the two traits (flower color on bottom left and flowering time on bottom right).

Discussion

The population design

The population was designed to segregate for the traits of the domestication syndrome at the pod level. The parents chosen for developing the population were the domesticated Andean non-shattering variety Midas and the wild Mesoamerican variety G12873 with a high pod-shattering phenotype. After a cross between the parents and 9 generations of SSD, the MG38 line was selected for having the wild phenotype for the pod traits ad the domesticated phenotype for the other recorded traits. Through AFLP analysis it was observed that MG38 had around 55% of the wild genome introgressed into the domesticated genomic background (information attained personal communication).

Even though the linkage mapping bi-parental populations are becoming less used in comparison to the association mapping populations that are gaining interest, since they can provide better resolution in QTL mapping, still, the design of this common bean population is expected to enable us to use it in a QTL-NIL approach, which relies on recombinations occurring in near regions in lines from sister sub-families. By comparing the position of introgressions between these lines and their observed phenotypes, it will be possible to pinpoint the narrow introgression regions that convey the wild phenotype within the population.

Besides the value that this population has for research of domestication and QTL mapping, as it is a cross between a wild and domesticated accession, it also holds a pre-breeding value, where RILs with phenotype combinations that might lead to a certain gain in agricultural use are possible to find. Further backcrossing and population development are in plan and underway.

The GBS genotyping design

For genotyping the selected RILs from 10 F_2 families for this study, a modified version of the original GBS method was used (Elshire et al., 2011), applying a nested multiplexing system which is more inexpensive, and a size-selection, to avoid sequencing fragments that are too short for the 150 bp paired-end sequencing or too long to be efficiently amplified or sequenced.

A nested multiplexing system was applied, similar to the one used by Peterson et al., (2012), where the ddRADseq method was introduced for the first time. There, a two restriction enzyme RADseq modification is employed to achieve larger genome reduction and deeper coverage for a smaller subset of fragments, in comparison to the single restriction enzyme RADseq method (Baird et al., 2008), but a nested multiplexing system is an additional change that is introduced. The main purpose of applying such multiplexing design is the cost reduction, as a lower number of barcoded adapters is required to be designed and synthesized. Creating barcoded adapters for tagging 12 different samples can cost as much as the sequencing cost of an Illumina lane (Peterson et al., 2012). Here, a smaller set of unique barcoded adapters than the total number of samples that need to be sequenced is produced (unlike in Elshire et al., 2011) and samples tagged by one unique set are separately pooled after the addition of barcoded adapters, after which different unique Illumina indices are added to fragments in each pool. After the addition of the indices, all samples can be pooled together as at that stage they are distinguishable by unique pairings of indices and barcodes, multiplexed and prepared for sequencing. Using this approach, each pool tagged with a different Illumina index will be stored as a separate file in the sequencing output data (having two files per index when using pair-end sequencing), which enables simple distinguishing of samples by most demultiplexing pipelines. In this study, 24 barcodes were created and used with 13 different Illumina indices and the lists of their unique sequence regions are available in tables in Appendix 1.

The size-based selection was applied for maintaining only fragments of 300-700 bp in length during the purification steps in the protocol and then confirmed using BluePippin (Sage Science). The decision about the lower cutoff was based on the planned paired-end sequencing of 150 bp in length. Therefore, including fragments shorter than 300 bp would lead to a coverage increase in middle regions of the fragments, where uneven coverage could potentially introduce bias in the downstream analyses, but also waste some of the sequencing efforts by unnecessary increase in coverage in these regions. The upper cutoff was introduced to avoid the bias in fragment representation due to the difference in length. It has been observed in previous studies that longer fragments produced diffuse sequence clusters which lowered the power of SNP calling (Elshire et al., 2011), while there is also a notable PCR amplification bias in favor of shorter fragments that results in underrepresentation of longer reads in the sequence read output (DaCosta & Sorenson, 2014). Besides that, having fragments of more similar length, improves the resulting SNP dataset quality, as having less diverse fragments that are more uniform in length improves coverage per site.

Sequencing quality

It is evident that even though all known precautions were taken to normalize the DNA content and sequencing output within and between GBS libraries for achieving approximately the same read coverage per sample, a variation was observed that cannot be clearly attributed to any of the controllable factors in the protocol design, but results from the technical limitations that currently exist.

Due to a technical failure, the ends of the fragment reads in sequencing lane 2 had a low quality towards the end nucleotides. For this reason, the missing data proportion was increased across SNP markers in certain regions of the dataset, but this was later controlled by the quality filtering.

SNP markers: yield, density, filtering

The SNP markers discovered through GBS that provided sufficient genome-wide coverage and number. As a partially methylation sensitive restriction enzyme was used, we could observe a lower marker coverage in regions of the chromosomes that contain more repetitive sequences and also often have a higher methylation level, as in centromeric and pericentromeric regions (Iwata-Otsubo et al., 2016; Iwata et al., 2013; Mehrotra & Goyal, 2014; Schmutz et al., 2014; Shcherban, 2015). The mapping of the centromeric regions is done exactly by marking and tracking specific repetitive sequences and deducing their location from them (Fonsêca et al., 2010; Gao et al., 2016; Pedrosa-Harand et al., 2009).

It has been observed that the common bean chromosomes have differences in recombination rates across chromosomes and their regions, as well as that in some crosses, especially between wild and domesticated accessions, there seems to be a suppression of recombination, that might be due to structural incompatibilities between the underlying sequence in these chromosomes (Fonsêca et al., 2010; Moscone et al., 1999; Pedrosa-Harand et al., 2009)

A general trend is observed, that in markers that have less missing data, a higher proportion of observed heterozygosity is present. Partially, this could be because, in these regions, there might be some misalignment due to sequence similarity, where the coverage in that region would be increased as multiple paralogous reads would be mapped at the same location, which would result in the discovery of stretches of heterozygous calls that are false.

The marker density plots show us which chromosomic regions are rich in polymorphisms. As this is a biparental population, these regions are expected to correlate with locations where there are wild genome introgressions segregating within the population, as the wild introgressions into the domesticated genomic background are the source of detected variability. By looking at how filtering impacts the patterns of marker density along chromosomes, we can see that near-

centromeric and telomeric regions were most severely affected by filtering most often (best seen in chromosome 3 in Figure 9), and that most narrow peaks contained high numbers of markers that were almost fully filtered out based on our criteria (see the highest peak in chromosome 1 as example).

SNP marker heterozygosity

Filtering out highly heterozygous markers at such a low cutoff value as in this analysis (at maximum 20% heterozygosity) is a double-edged sword. Where narrow large peaks disappear almost completely, there is some chance that there is a problem with paralogous mapping that can cause blocks of heterozygosity to appear, but also, there is always a chance for it to be a result of the preservation of high heterozygosity due to natural or artificial selection during population development. The best approach would be to find a balanced cutoff for filtering or to even try to observe the markers that were filtered out subsequently.

As for the representation of the heterozygosity along the chromosomes, accounting for the inbreeding in the population have brought down the expected heterozygosity genome-wide to near-zero values (Gillespie, 1998; Singh & Singh, 2015), and it shows more realistically the degree of the difference between the observed and the expected heterozygosity. However, the uncorrected expected heterozygosity, as it depends on the balance in the allelic frequencies within the population, can be used to deduce where could introgressions be present along chromosomes.

We see that there are regions where the expected heterozygosity is high, while the observed does not vary much and is relatively low. This could be a result of selective pressures during population development that disrupt the genotype expectancies based on the assumptions of Hardy-Weinberg equilibrium (which include neutral selection). For certain traits, a heterozygous or homozygous genotype can be favored and selected for or against, which then leads to skewed heterozygosity proportions in the genomic region surrounding the causative loci. This could be of interest to be examined in further analyses.

Gene density

The gene and SNP density comparison can be used to get an idea on how useful the SNP dataset will be for the downstream analyses, as having more markers in gene-rich regions can enable a higher QTL mapping power, as those markers have a higher chance to be in high LD with causative loci, and therefore more useful for association analyses later on.

Population structure

The population structure was also examined so that the power of the use of the population for QTL mapping could be understood. It is desired for the population to not have a too strong structure based on the F_2 family division, as less structure in the initial phases of the population development could lead to higher genomic diversity in later population development, and higher resolution in QTL mapping.

Linkage decay analysis

The linkage is the measure of non-random association of pairs of loci (markers) and here it is presented by the r2 value which ranges from zero (no linkage, leading to random occurance of co-inherritance) to one (complete linkage, where a pair of loci is always inherited together). Linkage depends on the distance between the loci and recombination rates between them, but in this population, linkage decay is only possible where G12873 introgressions (focalized allele diversity) exists. So, while linkage is known to be rather high within the common bean chromosomes (Schmutz et al., 2014) and usually due to low recombination rates that are a consequence of structural rearrangements preventing recombination (Feder, Nosil, & Flaxman, 2014; Lowry & Willis, 2010; Ragland et al., 2017), at the same time, it is not expected for the linkage to decay too much within this population either at this stage of its development.

Genome composition and introgression detection

In chromosome 1, there is one large central introgression in MG38, which encompasses the centromeric region and spans over almost the entire chromosome (from 2 to 43 Mb). Based on both LD heatmap data and haplotype reconstruction, a large number of RILs has inherited the entire introgression or none, but there is also a number of lines that have after a recombination breakage inherited only the part of the introgression of different size and location (see Figure 23). Recombination has mostly lead to the shortening of the introgression segment on one side or both, but there are also lines to which the very end parts of the introgression have been passed on, having two introgressions with a segment of the Midas genome in between. There were no recombination events detected in the region from 30 to 38 Mb, and therefore we can not have a higher resolution for the flower color QTL that is in that region on this chromosome.

In chromosome 2, two introgressions were detected in MG38, one spanning from the start of the chromosome until 42 Mb and a smaller introgression from 45 to 47 Mb. Within the RILs, none of them contain the part of the large introgression from 28 to 35 Mb, which can be a result of two recombination points and a selection against the wild genotype in that location.

In chromosome 3, the introgression detected in MG38 spans over the entire chromosome. As in many other chromosomes that have large introgressions in MG38 (chromosomes 1, 2, 3, 4, 7 and 8), there is a lot of missing data in centromeric and pericentromeric regions, that might be due to the partial methylation sensitivity of the *ApeKI* restriction enzyme, which might have been prevented from cutting in highly methylated regions of the genome, such as those in the centromeres. For this reason, it seems hard to deduce which is the real number of introgressions in chromosome 3. The LD heatmap shows three linkage groups, and the haplotype reconstruction is showing a similar pattern, but almost all the three segments are present in each line. However,

as in chromosome 2, here we also see two parts of the introgression regions that were present in MG38, but later absent in all the lines. They are towards the telomeric ends of the introgression.

In chromosome 4, the overall marker coverage is quite low. Similarly to chromosome 3, we have 3 high linkage regions visible in the LD heatmap, but we see that within the lines, by using both genome composition approaches, we can only confidently confirm the existence of two smaller introgressions on the telomeric ends of the chromosome. The one that is located at the beginning of the chromosome (from 1 to 4 Mb) seems to be missing in more lines than the one on the other end.

Considering all approaches, chromosome 5 seems to have two very narrow introgressions in MG38. We see that during population development, there is a strong selective pressure for keeping these introgressions in the population, as they remain in almost all RILs, but there are some that only have the introgression from the beginning of the chromosome, not the other one. Based on the LD analysis, the linkage in the second half of the chromosome is high, so there is a chance that the introgression is larger than detected.

Chromosome 6 has two introgressions of G12873 detected in MG38 (from 16 to 27 Mb, and from 27.5 to 31 Mb), but as we do not have the genomic data of G12873 and can deduce its genomic introgression solely based on the differences between Midas and MG38, it might be that this is one large introgression encompassing both regions, that did not contain polymorphism between the two parents. Within the population, we see that recombinations have occurred at various locations in this region, shortening the introgressions to a different extent. The introgression segment might extend further towards the start of the chromosome as well, but we do not seem to have a sufficient marker coverage in that region to confirm it. In almost the entire region of the chromosome that harbors the G12873 introgression, we can observe a strong selection for homozygosity.

Chromosome 7 shows a strong linkage across the most part of the chromosome (see LD heatmap in Appendix 6). Based on the haplotype reconstruction and introgression detection, we see that in all RILs, the part of the introgression from 3.5 to 8 Mb is missing. There might be a suppression of recombination in the region of the introgression (spanning from 8 to 26 Mb), but with a number of lines having the introgression shortened at the far end to a different extent. As our QTL for flower color is detected in that region, these recombinations allow for higher resolution of mapping in that area.

Chromosome 8 is also having a lower marker coverage in the centromeric and pericentromeric region, and in MG38 it has an introgression covering almost the entire chromosome (except from 59 Mb to the end of the chromosome). The LD heatmap is showing a complex profile, but we can see recombination occurring near the ends of the introgression, where different lengths of the introgression remain in different RILs, usually in the area from 4 to 24 Mb and from 43 to 59 Mb. As in chromosome 7, the area of the introgression from 0 to 4 Mb present in MG38 appears to strongly be selected against in the RILs.

In chromosome 9, we can observe two separate introgressions, one spanning from 10 to 35 Mb and the other from 38 to 39 Mb, near the end of the chromosome. Even though they are rather close, based on the introgression detection and linkage analyses, we see that there is weak linkage between these segments and also recombination occurring in several locations along the larger introgression, breaking linkage along it, but with the segment from 10 to 13 Mb wild introgression being removed during population development and not present in most RILs.

Chromosome 10 has a small introgression in MG38, from 33 Mb to the end of the chromosome (44 Mb), but with only the region from 35 to 40 Mb being retained in some RILs. Chromosome 11 has the lowest coverage of all chromosomes and that could be because there is not enough polymorphism to be detected, as there might not be any introgression present in this chromosome.

The power of the population for QTL mapping

We have approached QTL mapping using both TASSEL and GAPIT implemented in R, as well as by applying the GLM, MLM and weighted MLM models to observe which model can give the best results in our population for these two traits. Not all the models have succeeded in finding an association of the markers with the trait, and this depends on the assumptions that the models are based on, for which they perform with different success and different stringency on different population and traits. Using the GLM model in TASSEL, we have obtained similar results as with MLM and weighted MLM, but with inflated significance values. As GAPIT could not find any significant associations between the markers and the flower color, it might be that the phenotype co-varies with the population structure, and that the model cannot identify a QTL because it corrects for the population structure.

Flower color loss often relates to loss of seed color, as has been shown in several studies (Caldas & Blair, 2009; Johnson, 2002; Hallqvist, 1921). Flower and seed color have been often studied together (McClean et al., 2002), while flower color alone was of interest more rarely (Lamprecht, 1936; Basset et al., 1990; Bassett, 2003; McClean et al., 2002). The first linkage maps in common bean already contained markers associated with flower color traits (Gepts et al., 1993). But there is still not too much known about flower pigment inheritance in common bean, especially as it is a predominantly autogamous species, where flower color does not play a role in plant pollination.

In our population, we have a parent with violet flower color and brown seeds and pod color (MG38; resulting from anthocyanin) and the other parent without coloring in its flowers, seeds, and pods (Midas). In the population, light violet flower color also appears, presumably in the RILs heterozygous at loci for this trait. Based on the comparative QTL map from Galeano et al., (2011) (available at *http://cmap.comparative-legumes.org/cgi-bin/cmap/*

viewer?data_source=LIS;saved_link_id=562;), where the strong association on chromosome 7 is located for flower color, there might be QTLs for tannin pigments. Even though in our population, there is also a loss in tannin pigmentation in pod and seed coloring, further investigation should be performed to understand how the flower color is regulated and if this population can provide further aid in that undertaking.

Flowering time has been shown to be strongly influenced by growth habit (Michelangeli et al., 2013) where we see that photoperiod-insensitive, determinate common bean cultivars usually flower and mature early (Koinange et al., 1996). The flowering time QTL in chromosome 1 corresponds to the major gene affecting flowering time discovered by González et al, (2016) where the Phvul.001G189200 gene is located, homologous to the Arabidopsis thaliana TERMINAL FLOWER1 (TFL1) gene. It was also suggested that Phvul.001G189200 (PvTFL1y) is a candidate gene for determinacy locus.

The flowering time QTL in chromosome 8 falls within the Phvul.008G158112 gene, which is a tRNA-splicing endonuclease subunit, but, two other genes are located in the very proximity of it: (i) Phvul.008G158106 that codes a cis-zeatin O-glucosyltransferase (CISZOG) and (ii) Phvul.008G158118 that codes a protein from the glycoprotein family. From these three genes, the CISZOG has been shown high expression in flower bud tissue samples and has a role in the cytokinin-O-glucosides biosynthesis pathway (data from the gene annotation information on the JGI Phytozome website: *https:// phytozome.jgi.doe.gov*). Cytokinins discovery and role in plant growth regulation (Mok et al., 2000; Veach et al., 2003; Werner et al., 2001; Mok, 1994).

Bigger GWAS and WTL mapping projects on common bean traits of agronomic importance have been reported in the recent years (Kamfwa et al., 2015; Moghaddam et al., 2016; Tock et al., 2017). We expect that there will be further interest in these studies.

Conclusions

The population shows high phenotypic diversity for the targeted traits of the domestication syndrome, especially on the pod level, for which it was developed. However, the results on that data is presented in separate work, while here only the genomics composition and the power of this population in further research appplications is considered.

The population is characterized by genomic differences between the RILs and can, therefore, be used for QTL mapping studies, while some progeny shows transgressive phenotypes for certain traits as well, that are more extreme than those observed in the parental accessions. We see that the contribution of the wild genome donor in different lines in the population varies and partially depends on the linkage drag associated to the phenotypic selection applied for the target traits. A large collection of NILs is present in the sub-families, which will be of use for fine mapping of traits of interest. The population can be developed further by producing following generations, applying another backcross or by introducing different donor material, while more in-depth genomic characterization and QTL mapping for other traits are planned, as well. This population also has a pre-breeding value, and from the developed lines, some could show improved characteristics in comparison to the domesticated parent after additional analyses are conducted.

While some of the traits included in this research ahve been studied before, this study is unique in addressing them within the complex frame of the domestication syndrome and provides a contribution to the understading of the domestication of the common bean. We also observe GBS as a robust, simple and inexpensive method that was used for genotyping of the RILs of this population, and recognize it as a useful tool in developing and examining genome-wide markers in breeding and study populations.

(This page has been intentionally left blank for print formatting purposes.)

References

Adamski, T., Krystkowiak, K., Kuczyńska, A., Mikołajczak, K., Ogrodowicz, P., Ponitka, A., … Ślusarkiewicz-Jarzina, A. (2014). Segregation distortion in homozygous lines obtained via anther culture and maize doubled haploid methods in comparison to single seed descent in wheat (*Triticum aestivum* L.). *Electronic Journal of Biotechnology*, *17*(1), 6–13. http://doi.org/10.1016/j.ejbt.2013.12.002

Allaby, R. G., Lucas, L., Stevens, C., Maeda, O., & Fuller, D. Q. (2017). Geographic mosaics and changing rates of cereal domestication. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. http://doi.org/10.1098/RSTB.2016.0429

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., & Lander, E. S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, *407*, 513–516. http://doi.org/10.1038/35035083

Álvarez, M. F., Mosquera, T. and Blair, M. W. (2014) The Use of Association Genetics Approaches in Plant Breeding, in Plant Breeding Reviews: Volume 38 (ed J. Janick), John Wiley & Sons, Inc., Hoboken, New Jersey. doi: 10.1002/9781118916865.ch02

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-morita, T., & Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, *21*(4), 610–617. http://doi.org/10.1101/gr.115402.110

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/ projects/fastqc

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet*, *advance on*(2), 81–92. http://doi.org/10.1038/nrg.2015.28

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), 1–7. http://doi.org/ 10.1371/journal.pone.0003376

Bajgain, P., Rouse, M. N., & Anderson, J. A. (2016). Comparing genotyping-by-sequencing and single nucleotide polymorphism chip genotyping for quantitative trait loci mapping in wheat. *Crop Science*, *56*(1), 1–17. http://doi.org/10.2135/cropsci2015.06.0389

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, *12*(11), 745–755. http://doi.org/10.1038/nrg3031

Basset, M. J., Bao, X. L., & Hannah, L. C. (1990). Flower colors in common bean produced by interactions of the Sal and V loci and a gametophyte factor Ga Linked to Sal. *J. Amer. Soc. Hort. Sci*, *115*(6), 1029–1033.

Bassett, M. J. (2003). Allelism between the P and Stp genes for seedcoat color and pattern in common bean. *J. Amer. Soc. Hort. Sci.*, *128*(4), 548–551.

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10), 1–14. http://doi.org/10.1093/nar/gks001

Bevan, M. W., Uauy, C., Wulff, B. B. H., Zhou, J., Krasileva, K., & Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature Review*, *543*, 346–354. http://doi.org/10.1038/nature22011

Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M. L., Gioia, T., Santo, D., … Papa, R. (2017). Beans (*Phaseolus* ssp.) as a Model for Understanding Crop Evolution. *Frontiers in Plant Science*, *8*(May), 1–21. http://doi.org/10.3389/fpls.2017.00722

Bivand R. & Lewin-Koh N. (2017). maptools: Tools for Reading and Handling Spatial Objects. R package version 0.9-2. https://CRAN.R-project.org/package=maptools

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

Brim, C. A. (1966). A Modified Pedigree Method of Selection in Soybeans. *Crop Science*, *6*(March-April), 220.

Broman, K. W. (2005). The genomes of recombinant inbred lines. *Genetics*, *169*(2), 1133–1146. http://doi.org/10.1534/genetics.104.035212

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., & Asp, T. (2013). Genome Wide Allele Frequency Fingerprints (GWAFFs) of Populations via Genotyping by Sequencing. *PLoS ONE*, *8*(3). http://doi.org/10.1371/journal.pone.0057438

Caldas, G. V., & Blair, M. W. (2009). Inheritance of seed condensed tannins and their relationship with seed-coat color and pattern genes in common bean (*Phaseolus vulgaris* L.). *Theoretical and Applied Genetics*, *119*(1), 131–142. http://doi.org/10.1007/s00122-009-1023-4

Callaway, E. (2017). New concerns raised over value of genome-wide disease studies. *Nature*, *546*(7659), 463–463. http://doi.org/10.1038/nature.2017.22152

Campbell, N. R., Harmon, S. a., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*(4), 855–867. http://doi.org/10.1111/1755-0998.12357

Carletti, G., Carra, A., Allegro, G., Vietto, L., Desiderio, F., Bagnaresi, P., … Nervo, G. (2016). QTLs for Woolly Poplar aphid (*Phloeomyzus passerinii* L.) resistance detected in an inter-specific Populus deltoides x P. nigra mapping population. *PLoS ONE*, *11*(3), 1–18. http://doi.org/10.1371/journal.pone.0152569

Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., … Toll, J. (2016). Global conservation priorities for crop wild relatives. *Nature Plants*, *2*(4), 16022. http://doi.org/10.1038/nplants.2016.22

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, 362–365. http://doi.org/10.1111/1755-0998.12669

CBD. (1992). Convention on biological diversity (United Nations). *Diversity*, 30. Retrieved from http://www.cbd.int/doc/legal/cbd-en.pdf

Chen, Q., Ma, Y., Yang, Y., Chen, Z., Liao, R., Xie, X., … Pan, Y. (2013). Genotyping by genome reducing and sequencing for outbred animals. *PLoS ONE*, *8*(7), 6–11. http://doi.org/10.1371/journal.pone.0067500

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in Next-Generation-Sequencing data on *de novo* genome assembly. *PLoS ONE*, *8*(4). http://doi.org/10.1371/journal.pone.0062856

Clavijo Michelangeli, J. A., Bhakta, M., Gezan, S. A., Boote, K. J., & Vallejos, C. E. (2013). From flower to seed: Identifying phenological markers and reliable growth functions to model reproductive development in the common bean (*Phaseolus vulgaris* L.). *Plant, Cell and Environment*, *36*(11), 2046–2058. http://doi.org/10.1111/pce.12114

Collins, N. C., Tarieu, F., & Tuberosa, R. (2008). Quantitative Trait Loci and Crop Performance under Abiotic Stress - Where Do We Stand?, *147*(June), 469–486. http://doi.org/10.1104/pp.108.118117

Cornille, A., Salcedo, A., Kryvokhyzha, D., Gl??min, S., Holm, K., Wright, S. I., & Lascoux, M. (2016). Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). *Molecular Ecology*, *25*(2), 616–629. http://doi.org/10.1111/mec.13491

Cossio, M. L. T., Giesen, L. F., Araya, G., Pérez-Cotapos, M. L. S., VERGARA, R. L., Manca, M., … Héritier, F. (2010). Molecular Plant Breeding. (Y. Xu, Ed., 1st edition). London, UK: CAB International. http://doi.org/10.1007/s13398-014-0173-7.2

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. http://doi.org/10.1093/bioinformatics/btr330

Dapprich, J., Ferriola, D., Mackiewicz, K., Clark, P. M., Rappaport, E., D'Arcy, M., … Monos, D. (2016). The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics*, *17*(1), 486. http://doi.org/10.1186/s12864-016-2836-6

Davey, J. W., & Blaxter, M. L. (2011). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, *9*(5–6), 416–423. http://doi.org/10.1093/bfgp/elq031

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*, *22*(11), 3151–3164. http://doi.org/10.1111/mec.12084

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, *12*(7), 499–510. http://doi.org/10.1038/nrg3012

David, J., Holtz, Y., Ranwez, V., Santoni, S., Sarah, G., Ardisson, M., … Tavaud-Pirra, M. (2014). Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population. *Molecular Breeding*, *34*, 1534–1548. http://doi.org/10.1007/s11032-014-0179-z

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., & Guarino, L. (2017). Past and future use of wild relatives in crop breeding. *Crop Science*, *57*(3), 1070–1082. http://doi.org/10.2135/cropsci2016.10.0885

Deschamps, S., Llaca, V., & May, G. D. (2012). Genotyping-by-Sequencing in Plants. *Biology*. http://doi.org/10.3390/biology1030460

Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006). The Molecular Genetics of Crop Domestication. *Cell*, *127*(7), 1309–1321. http://doi.org/10.1016/j.cell.2006.12.006

Dvorak, J., Luo, M. C., & Akhunov, E. D. (2011). N.I. Vavilov's theory of centres of diversity in the light of current understanding of wheat diversity, domestication and evolution. *Czech Journal of Genetics and Plant Breeding*, *47*(SPEC. ISSUE 1).

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5). http://doi.org/10.1371/journal.pone.0019379

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. a, Cresko, W. a, Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 16196–16200. http://doi.org/10.1073/pnas.1006538107

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A., & Johnson, E. A. (2011). Local de novo assembly of rad paired-end contigs using short sequencing reads. *PLoS ONE*, *6*(4). http://doi.org/10.1371/journal.pone.0018561

Evans, J., McCormick, R. F., Morishige, D., Olson, S. N., Weers, B., Hilley, J., … Mullet, J. (2013). Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. *PLoS ONE*, *8*(11). http://doi.org/10.1371/journal.pone.0079192

Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L., & Gaut, B. S. (1998). Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(April), 4441–4446. http://doi.org/10.1073/pnas.95.8.4441

FAO. (1983). International Undertaking on Plant Genetic Resources (with Annexes I, II & III). In *Resolution 8/83*.

Feder, J. L., Nosil, P., & Flaxman, S. M. (2014). Assessing when chromosomal rearrangements affect the dynamics of speciation: Implications from computer simulations. *Frontiers in Genetics*, *5*(AUG), 1–14. http://doi.org/10.3389/fgene.2014.00295

Fernie, A. R., Tadmor, Y., & Zamir, D. (2006). Natural genetic variation for improving crop quality. *Current Opinion in Plant Biology*. http://doi.org/10.1016/j.pbi.2006.01.010

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T. M., … Nusbaum, C. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology*, *12*(1), R1. http://doi.org/10.1186/gb-2011-12-1-r1

Fletcher, R. S., Mullen, J. L., Yoder, S., Bauerle, W. L., Reuning, G., Sen, S., … McKay, J. K. (2013). Development of a next-generation NIL library in *Arabidopsis thaliana* for dissecting complex traits. *BMC Genomics*, *14*(1), 655. http://doi.org/10.1186/1471-2164-14-655

Foley, J. a, Defries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., … Snyder, P. K. (2005). Global consequences of land use. *Science (New York, N.Y.)*, *309*(5734), 570–4. http://doi.org/10.1126/science.1111772

Fonsêca, A., Ferreira, J., dos Santos, T. R. B., Mosiolek, M., Bellucci, E., Kami, J., … Pedrosa-Harand, A. (2010). Cytogenetic map of common bean (*Phaseolus vulgaris* L.). *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *18*(4), 487–502. http://doi.org/10.1007/s10577-010-9129-8

Galeano, C. H., Fernandez, A. C., Franco-Herrera, N., Cichy, K. A., McClean, P. E., Vanderleyden, J., & Blair, M. W. (2011). Saturation of an intra-gene pool linkage map: Towards a unified consensus linkage map for fine mapping and synteny analysis in common bean. *PLoS ONE*, *6*(12). http://doi.org/10.1371/journal.pone.0028135

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., … Falque, M. (2011). A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE*, *6*(12). http://doi.org/10.1371/journal.pone.0028334

Gao, D., Zhao, D., Abernathy, B., Iwata-Otsubo, A., Herrera-Estrella, A., Jiang, N., & Jackson, S. A. (2016). Dynamics of a Novel Highly Repetitive CACTA Family in Common Bean (*Phaseolus vulgaris*). *G3: Genes|Genomes|Genetics*, *6*(7), 2091–2101. http://doi.org/10.1534/g3.116.028761

Gentry, H. S. (1969). Origin of the Common Bean , Phaseolus vulgaris Author (s): Howard Scott Gentry Published by : Springer on behalf of New York Botanical Garden Press Stable URL : http://www.jstor.org/stable/4253014 New York Botanical Garden Press , Springer are collabora, *23*(1), 55–69.

Gepts, P., Nodari, R., Tsai, S. M., Koinange, E. M. K., Llaca, V., Gilbertson, R. L., & Guzman, P. (1993). Linkage mapping in common bean. *Annual Report of the Bean Improvement Cooperative*.

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., … Wilmoth, J. (2014). World population stabilization unlikely this century. *Science*, *346*(6206), 234–237. http://doi.org/10.1126/science.1257469

Gillespie, J. H. (1998). Population genetics: A concise guide. Baltimore, Md: The Johns Hopkins University Press.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, *9*(2), 11. http://doi.org/10.1371/journal.pone.0090346

González, A. M., Yuste-Lisbona, F. J., Saburido, S., Bretones, S., De Ron, A. M., Lozano, R., & Santalla, M. (2016). Major Contribution of Flowering Time and Vegetative Growth to Plant Production in Common Bean As Deduced from a Comparative Genetic Mapping. *Frontiers in Plant Science*, *7*(December). http://doi.org/10.3389/fpls.2016.01940

Gopala, K. S., Waters, D. L. E., & Henry, R. J. (2014). Australian wild rice reveals pre-domestication origin of polymorphism deserts in rice genome. *PLoS ONE*, *9*(6). http://doi.org/10.1371/journal.pone.0098843

Goulden, C. H. (1939). Problems in plant selection. *Proceeding of the Seventh International Genetical Congress.*, (1941), 132–133.

Greminger, M., Stolting, K., Nater, A., Goossens, B., Arora, N., Bruggmann, R., … Krutzen, M. (2014). Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. *BMC Genomics*, *15*(1), 16. http://doi.org/10.1186/1471-2164-15-16

Grover, A., & Sharma, P. C. (2014). Development and use of molecular markers: past and present. *Critical Reviews in Biotechnology*, *8551*(0), 1–13. http://doi.org/10.3109/07388551.2014.959891

Guo, Y., Yuan, H., Fang, D., Song, L., Liu, Y., Liu, Y., … Zhang, H. (2014). An improved 2b-RAD approach (I2b-RAD) offering genotyping tested by a rice (Oryza sativa L.) F2 population. *BMC Genomics*, *15*(956), 1–13. Retrieved from http://www.biomedcentral.com/1471-2164/15/956

Gur, A., & Zamir, D. (2004). Unused natural variation can lift yield barriers in plant breeding. *PLoS Biology*, *2*(10). http://doi.org/10.1371/journal.pbio.0020245

Haddad, N. I., & Muehlbauer, F. J. (1981). Comparison of random bulk population and single-seed-descent methods for lentil breeding. *Euphytica*, *30*(3), 643–651. http://doi.org/10.1007/BF00038792

Hallqvist, C. (1921). The inheritance of the flower color and the seed color in *Lupinus angustifolius*. *Weibullsholm, Landskona*.

Hammer, K. (1984). Das Domestikationssyndrom. *Die Kulturpflanze*, *32*(3), 11–34. http://doi.org/10.1007/BF02098682

Harfouche, A., Meilan, R., Kirst, M., Morgante, M., Boerjan, W., Sabatti, M., & Scarascia Mugnozza, G. (2012). Accelerating the domestication of forest trees in a changing world. *Trends in Plant Science*, *17*(2), 64–72. http://doi.org/10.1016/j.tplants.2011.11.005

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., … Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, *10*(3), R32. http://doi.org/10.1186/gb-2009-10-3-r32

Harlan, J. R. ., Wet, J. . M. . J. . de, & Price, E. . G. (1973). Comparative Evolution of Cereals. *Society for the Study of Evolution*, *27*(2), 311–325.

Harlan J.R., editor, (1992). Crops & Man. ASA, CSSA, Madison, WI. doi:10.2135/1992.cropsandman

Hawkins, T. L., O'Connor-Morin, T., Roy, A., & Santillan, C. (1994). DNA purification and isolation using a solid-phase. *Nucleic Acids Research*, *22*(21), 4543–4544. http://doi.org/10.1093/nar/22.21.4543

He, J., Zhao, X., Laroche, A., Lu, Z., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, *5*(484), 1–8. http://doi.org/10.3389/fpls.2014.00484

Heffelfinger, C., Fragoso, C. A., Moreno, M. A., Overton, J. D., Mottinger, J. P., Zhao, H., … Dellaporta, S. L. (2014). Flexible and scalable genotyping-by-sequencing strategies for population studies. *BMC Genomics*, *15*(979), 1–23. http://doi.org/10.1186/1471-2164-15-979

Heffner, E. L., Sorrells, M. E., & Jannink, J. (2009). Genomic Selection for Crop Improvement. *Crops*, (February), 1–12. http://doi.org/10.2135/cropsci2008.08.0512

Heslot, N., Jannink, J.-L., & Sorrells, M. E. (2015). Perspectives for Genomic Selection Applications and Research in Plants. *Crop Science*, *55*(february), 1–12. http://doi.org/10.2135/cropsci2014.03.0249

Hilario, E., Barron, L., Deng, C. H., Datson, P. M., De Silva, N., Davy, M. W., & Storey, R. D. (2015). Random Tagging Genotyping by Sequencing (rtGBS), an Unbiased Approach to Locate Restriction Enzyme Sites across the Target Genome. *Plos One*, *10*(12), e0143193. http://doi.org/ 10.1371/journal.pone.0143193

Hohenlohe, P. a., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. a., & Cresko, W. a. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, *6*(2). http://doi.org/ 10.1371/journal.pgen.1000862

Hurley C. (2012). gclus: Clustering Graphics. R package version 1.3.1. https://CRAN.R-project.org/package=gclus

Iwata-Otsubo, A., Radke, B., Findley, S., Abernathy, B., Vallejos, C. E., & Jackson, S. A. (2016). Fluorescence in situ Hybridization (FISH)-Based Karyotyping Reveals Rapid Evolution of Centromeric and Subtelomeric Repeats in Common Bean (*Phaseolus vulgaris*) and Relatives. *G3 (Bethesda, Md.)*, *6*(4), 1013–1022. http://doi.org/10.1534/g3.115.024984

Iwata, A., Tek, A. L., Richard, M. M. S., Abernathy, B., Fonsêca, A., Schmutz, J., … Jackson, S. A. (2013). Identification and characterization of functional centromeres of the common bean. *The Plant Journal*, n/a-n/a. http://doi.org/10.1111/tpj.12269

Jaccoud, D., Peng, K., Feinstein, D., & Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research*, *29*(4), E25. http://doi.org/10.1093/nar/29.4.e25

Jannink, J.-L. L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, *9*(2), 166–177. http://doi.org/10.1093/bfgp/elq001

Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, *9*(2), 166–177. http://doi.org/10.1093/bfgp/elq001

Jiang, Z., Wang, H., Michal, J. J., Zhou, X., Liu, B., Woods, L. C. S., & Fuchs, R. A. (2016). Genome wide sampling sequencing for SNP genotyping: Methods, challenges and future development. *International Journal of Biological Sciences*, *12*(1), 100–108. http://doi.org/10.7150/ijbs.13498

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., … dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*(7345), 97–100. http://doi.org/10.1038/nature09916

Johnson, C. S. (2002). TRANSPARENT TESTA GLABRA2, a Trichome and Seed Coat Development Gene of Arabidopsis, Encodes a WRKY Transcription Factor. *The Plant Cell Online*, *14*(6), 1359–1375. http://doi.org/10.1105/tpc.001404

Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403-1405. doi:10.1093/bioinformatics/btn129

Jombart T. and Ahmed I. (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. doi:10.1093/bioinformatics/btr521

Jonas, E., & De Koning, D. J. (2013). Does genomic selection have a future in plant breeding? *Trends in Biotechnology*, *31*(9), 497–504. http://doi.org/10.1016/j.tibtech.2013.06.003

Kamfwa, K., Cichy, K. A., & Kelly, J. D. (2015). Genome-Wide Association Study of Agronomic Traits in Common Bean. *The Plant Genome*, *8*(2), 0. http://doi.org/10.3835/plantgenome2014.09.0059

Keenan, K., McGinnity, P., Cross, T.F., Crozier, W.W., & Prodöhl, P.A. (2013). diveRsity: An R package for the estimation of population genetics parameters and their associated errors, Methods in Ecology and Evolution. doi:10.1111/2041-210X.12067

Khoury, C., Laliberté, B., & Guarino, L. (2010). Trends in ex situ conservation of plant genetic resources: A review of global crop and regional conservation strategies. *Genetic Resources and Crop Evolution*, *57*(4), 625–639. http://doi.org/10.1007/s10722-010-9534-z

Kilian, B., & Graner, A. (2012). NGS technologies for analyzing germplasm diversity in genebanks. *Briefings in Functional Genomics*, *11*(1), 38–50. http://doi.org/10.1093/bfgp/elr046

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., & Paterson, A. H. (2015). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science*, 1–9. http://doi.org/ 10.1016/j.plantsci.2015.04.016

Kitimu, S. R., Taylor, J., March, T. J., Tairo, F., Wilkinson, M. J., & Lopez, C. M. R. (2015). Meristem micropropagation of cassava (*Manihot esculenta*) evokes genome-wide changes in DNA methylation. *Frontiers in Plant Science*, *6*(August), 1–12. http://doi.org/10.3389/fpls.2015.00590

Koinange, E. M. K., Koinange, E. M. K., Singh, S. P., Singh, S. P., Gepts, P., & Gepts, P. (1996). Genetic control of the domestication syndrome in common-bean. *Crop Sci*, *36*, 1037–1045. http://doi.org/10.2135/ cropsci1996.0011183X003600040037x

Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, *2012*. http://doi.org/10.1155/2012/831460

Lalic, A., Kovacevic, J., & Novoselovic, D. (2000). Comparison of Pedigree and Single Seed Descent Method (Ssd) in Early Generation of Barley. *Poljoprivreda*, (1975), 1–6. Retrieved from http://www.pfos.hr/~poljo/sites/default/data/2003_2/5_LALIC.pdf

Lamble, S., Batty, E., Attar, M., Buck, D., Bowden, R., Lunter, G., … Piazza, P. (2013). Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnology*, *13*, 104. http://doi.org/10.1186/1472-6750-13-104

Lamprecht, H. (1936). Zur Genetik von *Phaseolus vulgaris* XII. Uber die Vererbung der Bluten- und Stammfarbe. *Hereditas*, *21*(2–3), 129–166. http://doi.org/10.1111/j.1601-5223.1936.tb03196.x

Lechance, J., & Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays*, *35*(9), 780–786. http://doi.org/10.1007/s12671-013-0269-8.Moving

Li, C. (2006). Rice Domestication by Reducing Shattering. *Science*, *311*(5769), 1936–1939. http://doi.org/10.1126/science.1123604

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. http://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. http://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research*, *21*, 940–951. http://doi.org/10.1101/gr.117259.110.individuals

Ligges, U. & Mächler, M. (2003). Scatterplot3d - an R Package for Visualizing Multivariate Data. Journal of Statistical Software 8(11), 1-20.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., … Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, *28*(18), 2397–2399. http://doi.org/10.1093/bioinformatics/bts444

Loskutov, I. G. (1999). *Vavilov and his institute: A history of the world collection of plant genetic resources in Russia*. Rome, Italy: International Plant Genetic Resources Institute. Retrieved from http://www.vir.nw.ru/files/pdf/books/Vavilov_and_his_institute.pdf

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2016). Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, n/a-n/a. http://doi.org/10.1111/1755-0998.12596

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular Ecology Resources*, (April), 366–369. http://doi.org/10.1111/1755-0998.12677

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, *8*(9). http://doi.org/10.1371/journal.pbio.1000500

Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *International Journal of Plant Genomics*, *2012*. http://doi.org/10.1155/2012/728398

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, *470*(7333), 198–203. http://doi.org/10.1038/nature09796

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. http://doi.org/10.14806/ej.17.1.200

Mascher, M., Wu, S., St. Amand, P., Stein, N., & Poland, J. (2013). Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. *PLoS ONE*, *8*(10), 1–11. http://doi.org/10.1371/journal.pone.0076925

McClean, P. E., Lee, R. K., Otto, C., Gepts, P., & Bassett, M. J. (2002). Molecular and Phenotypic Mapping of Genes Controlling Seed Coat Pattern and Color in Common Bean (*Phaseolus vulgaris* L.). *Journal of Heredity*, *93*(2), 148–152.

McCouch, S., Baute, G. J., Bradeen, J., Bramel, P., Bretting, P. K., Buckler, E., … Zamir, D. (2013). Agriculture: Feeding the future. *Nature*, *499*, 23–24. http://doi.org/10.1038/499023a

Mckinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: Comment on Breaking RAD by Lowry et al. (2016). *Molecular Ecology Resources*, 356–361. http://doi.org/10.1111/1755-0998.12649

Mehrotra, S., & Goyal, V. (2014). Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, Proteomics and Bioinformatics*, *12*(4), 164–171. http://doi.org/10.1016/j.gpb.2014.07.003

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. http://doi.org/10.1038/nrg2626

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 240–248. http://doi.org/10.1101/gr.5681207.high-throughput

Moghaddam, S. M., Mamidi, S., Osorno, J. M., Lee, R., Brick, M., Kelly, J., … McClean, P. E. (2016). Genome-Wide Association Study Identifies Candidate Loci Underlying Agronomic Traits in a Middle American Diversity Panel of Common Bean. *The Plant Genome*, *9*(3), 0. http://doi.org/10.3835/plantgenome2016.02.0012

Mok M.C. (1994). Cytokinins and plant development: an overview. In: Mok DWS, Mok MC, editors. Cytokinins: Chemistry, Activity, and Function. Boca Raton, FL: CRC Press; 1994. pp. 155–166.

Mok, M. C., Martin, R. C., & Mok, D. W. S. (2000). Cytokinins: Biosynthesis metabolism and perception. *In Vitro Cellular & Developmental Biology - Plant*, *36*(2), 102–107. http://doi.org/10.1007/s11627-000-0021-7

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., & Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Non-Model Organisms. *G3: Genes|Genomes|Genetics*, *X*(September), 1–10. http://doi.org/10.1534/g3.115.021667

Money, D., Migicovsky, Z., Gardner, K., & Myles, S. (2017). LinkImputeR: user-guided genotype calling and imputation for non-model organisms. *BMC Genomics*, *18*(1), 523. http://doi.org/10.1186/s12864-017-3873-5

Monson-Miller, J., Sanchez-Mendez, D. C., Fass, J., Henry, I. M., Tai, T. H., & Comai, L. (2012). Reference genome-independent assessment of mutation density using restriction enzyme-phased sequencing. *BMC Genomics*, *13*(1), 72. http://doi.org/10.1186/1471-2164-13-72

Moragues, M., Comadran, J., Waugh, R., Milne, I., Flavell, A. J., & Russell, J. R. (2010). Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theoretical and Applied Genetics*, *120*, 1525–1534. http://doi.org/10.1007/s00122-010-1273-1

Morishige, D. T., Klein, P. E., Hilley, J. L., Sahraeian, S. M., Sharma, A., & Mullet, J. E. (2013). Digital genotyping of sorghum - a diverse plant species with a large repeat-rich genome. *BMC Genomics*, *14*(1), 448. http://doi.org/10.1186/1471-2164-14-448

Moscone, E. a., Klein, F., Lambrou, M., Fuchs, J., & Schweizer, D. (1999). Quantitative karyotyping and dual-color FISH mapping of 5S and 18S-25S rDNA probes in the cultivated *Phaseolus* species (Leguminosae). *Genome*, *42*(6), 1224–1233. http://doi.org/10.1139/gen-42-6-1224

Murgia, M. L., Attene, G., Rodriguez, M., Bitocchi, E., Bellucci, E., Fois, D., … Rau, D. (2017). A Comprehensive Phenotypic Investigation of the "Pod-Shattering Syndrome" in Common Bean. *Frontiers in Plant Science*, *8*(March). http://doi.org/10.3389/fpls.2017.00251

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, *320*(5881), 1344–1349. http://doi.org/10.1126/science.1158441.The

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*. http://doi.org/10.1111/mec.12350

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–51. http://doi.org/10.1038/nrg2986

Paradis E. (2010). pegas: an R package for population genetics with an integrated modular approach. Bioinformatics 26: 419-420.

Paradis E., Claude J. & Strimmer K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

Patel, D. A., Zander, M., Dalton-Morgan, J., & Batley, J. (2015). Advances in Plant Genotyping: Where the Future Will Take Us. In J. Batley (Ed.), *Methods in Molecular Biology* (Vol. 1245, pp. 1–11). New York: Springer Science+Business Media. http://doi.org/10.1007/978-1-4939-1966-6_8

Pedrosa-Harand, A., Kami, J., Gepts, P., Geffroy, V., & Schweizer, D. (2009). Cytogenetic mapping of common bean chromosomes reveals a less compartmentalized small-genome plant species. *Chromosome Research*, *17*(3), 405–417. http://doi.org/10.1007/s10577-009-9031-4

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). http://doi.org/10.1371/journal.pone.0037135

Peterson, G. W., Dong, Y., Horbach, C., & Fu, Y. B. (2014). Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity*, *6*(4), 665–680. http://doi.org/10.3390/d6040665

Pingali, P. L. (2012). Green Revolution: Impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences*, *109*(31), 12302–12308. http://doi.org/10.1073/pnas.0912953109

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, *7*(2). http://doi.org/10.1371/journal.pone.0032253

Poland, J. A., Endelman, J., Dawson, J., Rutkoski, J., Wu, S. Y., Manes, Y., … Jannink, J. L. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *Plant Genome*, *5*(3), 103–113. http://doi.org/Doi10.3835/Plantgenome2012.06.0006

Poland, J., & Rife, T. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome*, *5*(3), 92–102. http://doi.org/10.3835/plantgenome2012.05.0005

Pollard K.S, Dudoit S. & van der Laan M.J. (2005). Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit (Editors). Springer (Statistics for Biology and Health Series), pp. 251-272.

Porch, T. G. (2013). List of genes - Phaseolus vulgaris L. *Bean Improvement Cooperative*, *11*, 1–35. Retrieved from http://beangenes.cws.ndsu.nodak.edu/genes/genlist3.htm

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2*, e431. http://doi.org/10.7717/peerj.431

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ragland, G. J., Doellman, M. M., Meyers, P. J., Hood, G. R., Egan, S. P., Powell, T. H. Q., … Feder, J. L. (2017). A test of genomic modularity among life-history adaptations promoting speciation with gene flow. *Molecular Ecology*, *26*(15), 3926–3942. http://doi.org/10.1111/mec.14178

Recknagel, H., Jacobs, A., Herzyk, P., & Elmer, K. R. (2015). Double-digest RAD sequencing using Ion Proton semiconductor platform (ddRADseq-ion) with nonmodel organisms. *Molecular Ecology Resources*, *15*(6), 1316–1329. http://doi.org/10.1111/1755-0998.12406

Rieseberg, L. H., Archer, M. a, & Wayne, R. K. (1999). Transgressive segregation, adaptation and speciation. *Heredity*, *83 (Pt 4)*(July), 363–372. http://doi.org/10.1038/sj.hdy.6886170

Rife, T. W., Wu, S., Bowden, R., & Poland, J. a. (2015). Spiked GBS: a unified, open platform for single marker genotyping and whole-genome profiling. *BMC Genomics*, *16*(1), 1–7. http://doi.org/10.1186/s12864-015-1404-9

Rocher, S., Jean, M., Castonguay, Y., & Belzile, F. (2015). Validation of Genotyping-By-Sequencing Analysis in Populations of Tetraploid Alfalfa by 454 Sequencing. *Plos One*, *10*(6), e0131918. http://doi.org/10.1371/journal.pone.0131918

Rowe, H. C., Renaut, S., & Guggisberg, A. (2011). RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, *20*(17), 3499–3502. http://doi.org/10.1111/j.1365-294X.2011.05197.x

RStudio Team. (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Russello, M. a., Waterhouse, M. D., Etter, P. D., & Johnson, E. a. (2015). From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ*, *3*, e1106. http://doi.org/10.7717/peerj.1106

Salas, G., & Friedt, W. (1995). Comparison of pedigree selection and single seed descent for oil yield in linseed (*Linum usitatissimum* L.). *Euphytica*, *83*(1), 25–32. http://doi.org/10.1007/BF01677857

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., & Kilian, A. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC Proceedings*, *5*(Suppl 7), P54. http://doi.org/10.1186/1753-6561-5-S7-P54

Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*(2), 149–161. http://doi.org/10.1111/pbi.12645

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., … Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, *46*(7), 707–713. http://doi.org/10.1038/ng.3008

Sedgwick, P. (2014). Multiple hypothesis testing and Bonferroni's correction. *BMJ*, *349*(October 2014), 1–3. http://doi.org/10.1136/bmj.g6284

Shcherban, A. B. (2015). Repetitive DNA sequences in plant genomes. *Russian Journal of Genetics: Applied Research*, *5*(3), 159–167. http://doi.org/10.1134/S2079059715030168

Shin, J.-H., Blay, S., McNeney, B., & Graham, J. (2006). LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria between Single Nucleotide Polymorphisms. *Journal of Statistical Software*, *16*(October), 1–10. http://doi.org/http://dx.doi.org/10.18637/jss.v016.c03

Singh, B. D., & Singh, A. K. (2015). Mapping Populations. In *Marker-Assisted Plant Breeding: Principles and Practices* (p. 514). Springer.

Smith, B. D. (2006). Eastern North America as an independent center of plant domestication. *Proceedings of the National Academy of Sciences*, *103*(33), 12223–12228. http://doi.org/10.1073/pnas.0604335103

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., … Belzile, F. (2013). An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLoS ONE*, *8*(1), 1–9. http://doi.org/10.1371/journal.pone.0054603

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant Journal 3: 739-744.

Stam, P., & Zeven, A. C. (1981). The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica*, *30*(2), 227–238. http://doi.org/10.1007/BF00033982

Stolle, E., & Moritz, R. F. A. (2013). RESTseq - Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing. *PLoS ONE*, *8*(5), 4–8. http://doi.org/10.1371/journal.pone.0063960

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., … Zheng, H. (2013). SLAF-seq: An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing. *PLoS ONE*, *8*(3). http://doi.org/10.1371/journal.pone.0058700

Tanksley, S. D., & Nelson, J. C. (1996). Advanced backcross QTL analysis: A method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theoretical and Applied Genetics*, *92*(2), 191–203. http://doi.org/10.1007/s001220050114

Tee, T. S., & Qualset, C. O. (1975). Bulk populations in wheat breeding: comparison of single-seed descent and random bulk methods. *Euphytica*, *24*(2), 393–405. http://doi.org/10.1007/BF00028206

Thomson, M. J. (2014). High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breeding and Biotechnology*, *2*(3), 195–212. http://doi.org/10.9787/PBB.2014.2.3.195

Tock, A. J., Fourie, D., Walley, P. G., Holub, E. B., Soler, A., Cichy, K. A., … Miklas, P. N. (2017). Genome-Wide Linkage and Association Mapping of Halo Blight Resistance in Common Bean to Race 6 of the Globally Important Bacterial Pathogen. *Frontiers in Plant Science*, *8*(July), 1–17. http://doi.org/10.3389/fpls.2017.01170

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, *1*(February), e203. http://doi.org/10.7717/peerj.203

Torkamaneh, D., Laroche, J., & Belzile, F. (2016). Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. *PLoS ONE*, *11*(8), 1–14. http://doi.org/10.1371/journal.pone.0161333

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. http://doi.org/10.1093/bioinformatics/17.6.520

Truong, H. T., Ramos, A. M., Yalcin, F., de Ruiter, M., van der Poel, H. J. A., Huvenaars, K. H. J., … van Eijk, M. J. T. (2012). Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE*, *7*(5). http://doi.org/10.1371/journal.pone.0037565

Tuinstra, M. R., Ejeta, G., & Goldsbrough, P. B. (1997). Heterogeneous inbred family (HIF) analysis: A method for developing near-isogenic lines that differ at quantitative trait loci. *Theoretical and Applied Genetics*, *95*(5–6), 1005–1011. http://doi.org/10.1007/s001220050654

van Gurp, T. P., Wagemaker, N. C. A. M., Wouters, B., Vergeer, P., Ouborg, J. N. J., & Verhoeven, K. J. F. (2016). epiGBS: reference-free reduced representation bisulfite sequencing. *Nature Methods*, *13*(4), 322–4. http://doi.org/10.1038/nmeth.3763

Van Ooijen, J.W. (2006). JoinMap 4, Software for the calculation of the genetic linkage maps in experimental populations. Kyazma B.V., Wageningen, Netherlands.

Van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., … van Eijk, M. J. T. (2007). Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE*, *2*(11). http://doi.org/10.1371/journal.pone.0001172

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–23. http://doi.org/10.3168/jds.2007-0980

Varshney, R. K., Bansal, K. C., Aggarwal, P. K., Datta, S. K., & Craufurd, P. Q. (2011). Agricultural biotechnology for crop improvement in a variable climate: Hope or hype? *Trends in Plant Science*. http://doi.org/10.1016/j.tplants.2011.03.004

Varshney, R. K., Singh, V. K., Hickey, J. M., Xun, X., Marshall, D. F., Wang, J., … Ribaut, J. M. (2015). Analytical and Decision Support Tools for Genomics-Assisted Breeding. *Trends in Plant Science*, *xx*, 1–10. http://doi.org/10.1016/j.tplants.2015.10.018

Varshney, R. K., Terauchi, R., & McCouch, S. R. (2014). Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. *PLoS Biology*, *12*(6), 1–8. http://doi.org/10.1371/journal.pbio.1001883

Vavilov, N. I. (1926). Studies on the origin of cultivated plants. *(Russian) Bulletin of Applied Botany and Plant-Breeding*, *14*, 1–245.

Vavilov, N. I. (1992). Origin and geography of cultivated plants. *Cambridge University Press, Cambridge*.

Veach, Y. K., Martin, R. C., Mok, D. W. S., Malbeck, J., Vankova, R., & Mok, M. C. (2003). O-glucosylation of cis-zeatin in maize. Characterization of genes, enzymes, and endogenous cytokinins. *Plant Physiology*, *131*(3), 1374–80. http://doi.org/10.1104/pp.017210

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, *34*(2002), 275–305. http://doi.org/10.1051/gse

Visscher, P. M., Andrew, T., & Nyholt, D. R. (2008). Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *European Journal of Human Genetics*, *16*(3), 387–390. http://doi.org/10.1038/sj.ejhg.5201990

Voss-Fels, K., & Snowdon, R. J. (2016). Understanding and utilizing crop genome diversity via high-resolution genotyping. *Plant Biotechnology Journal*, *14*(4), 1086–1094. http://doi.org/10.1111/pbi.12456

Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, *9*(8), 808–810. http://doi.org/10.1038/nmeth.2023

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. http://doi.org/10.1038/nrg2484

Warnes G., Gorjanc G., Leisch F. & Man M. (2013). genetics: Population Genetics. R package version 1.3.8.1. https://CRAN.R-project.org/package=genetics

Warnes G.R., Bolker B., Bonebakker L., Gentleman R., Liaw W.H.A., Lumley T., Maechler M., Magnusson A., Moeller S., Schwartz M. & Venables B. (2016). gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1. https://CRAN.R-project.org/package=gplots

Warnes G.R., Bolker B., Gorjanc G., Grothendieck G., Korosec A., Lumley T., MacQueen D., Magnusson A., Rogers J. & others (2017). gdata: Various R Programming Tools for Data Manipulation. R package version 2.18.0. https://CRAN.R-project.org/package=gdata

Wells, R., Trick, M., Fraser, F., Soumpourou, E., Clissold, L., Morgan, C., … Bancroft, I. (2013). Sequencing-based variant detection in the polyploid crop oilseed rape. *BMC Plant Biology*, *13*(1), 111. http://doi.org/10.1186/1471-2229-13-111

Werner, T., Motyka, V., Strnad, M., & Schmülling, T. (2001). Regulation of plant growth by cytokinin. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(18), 10487–92. http://doi.org/10.1073/pnas.171304098

Wetterstrand, K.A. (2016). DNA sequencing costs: Data from the NHGRI large-scale genome sequencing program. National Human Genome Research Institute, Bethesda, MD. http://www.genome.gov/sequencingcostsdata.

Wickham, H. (2007). Reshaping data with the reshape package. Journal of Statistical Software, 21(12) 2007.

World Population Prospects. (2015). *World Population Prospects 2015 Data Booklet of United Nations*. United Nations. http://doi.org/ST/ESA/ SER.A/377

Xia, Z., Zou, M., Zhang, S., Feng, B., & Wang, W. (2014). AFSM sequencing approach: a simple and rapid method for genome-wide SNP and methylation site discovery and genetic mapping. *Scientific Reports*, *4*, 7300. http://doi.org/10.1038/srep07300

Yeri, S. B., Shirasawa, K., Pandey, M. K., Gowda, M. V. C., Sujay, V., Shriswathi, M., … Bhat, R. S. (2014). Development of NILs from heterogeneous inbred families for validating the rust resistance QTL in peanut (*Arachis hypogaea* L.). *Plant Breeding*, *133*(1), 80–85. http://doi.org/10.1111/pbr.12130

Yoshihara, M., Saito, D., Sato, T., Ohara, O., Kuramoto, T., & Suyama, M. (2016). Design and application of a target capture sequencing of exons and conserved non-coding sequences for the rat. *BMC Genomics*, *17*(1), 593. http://doi.org/10.1186/s12864-016-2975-9

Zamir, D. (2008). Plant breeders go back to nature. *Nature*, *40*(3), 269–270.

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, *30*(6), 1--27. http://doi.org/10.1017/CBO9781107415324.004

Zhang, Y., Wang, L., Xin, H., Li, D., Ma, C., Ding, X., … Zhang, X. (2013). Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. *BMC Plant Biology*, *13*(1), 141. http://doi.org/10.1186/1471-2229-13-141

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data, 1–3.

Zimmer, E. A., & Wen, J. (2015). Using nuclear gene data for plant phylogenetics: Progress and prospects II. Next-gen approaches. *Journal of Systematics and Evolution*, *53*(5), 371–379. http://doi.org/10.1111/jse.12174

(This page has been intentionally left blank for print formatting purposes.)

Appendices

Appendix 1: GBS library nested multiplexing primers

ApeKI barcoded GATA_tag adapter example:

ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAGTA
P-CWGTACTAGAGATCGGAAGAGCACACGTCT

List of unique tag sequences within the custom barcoded adapters:

Tag 101	AACAA	TTGTT
Tag 102	CCACC	GGTGG
Tag 103	TTGTT	AACAA
Tag 104	GGTGA	TCACC
Tag 105	AACAGT	ACTGTT
Tag 106	CCATGA	TCATGG
Tag 107	TTGCCA	TGGCAA
Tag 108	ACTGTT	AACAGT
Tag 109	GGAACGT	ACGTTCC
Tag 110	CACCTGA	TCAGGTG
Tag 111	CTTGAAT	ATTCAAG
Tag 112	TCGTGTA	TACACGA
Tag 113	GGAACGGT	ACCGTTCC
Tag 114	AACCTAGA	TCTAGGTT
Tag 115	CTTGATGA	TCATCAAG
Tag 116	AGGTCGGT	ACCGACCT
Tag 117	TAACGAACA	TGTTCGTTA
Tag 118	GCCAACCAT	ATGGTTGGC
Tag 119	CTTGTGTTA	TAACACAAG
Tag 120	ACGTGTGGT	ACCACACGT
Tag 121	TGAACACAA	TTGTGTTCA
Tag 122	GACCACACT	AGTGTGGTC
Tag 123	CTTGGTTGA	TCAACCAAG
Tag 124	ACGTTGGTT	AACCAACGT

PPI Illumina index example (in 5'-3' orientation for one of the strands):

CAAGCAGAAGACGGCATACGAGAT **CGTGAT** GTGACTGGAGTTC

List of unique recognition sequences within the Illumina indices:

Index 1	ATCACG	CGTGAT
Index 2	CGATGT	ACATCG
Index 3	TTAGGC	GCCTAA
Index 4	TGACCA	TGGTCA
Index 5	ACAGTG	CACTGT
Index 6	GCCAAT	ATTGGC
Index 7	CAGATC	GATCTG
Index 8	ACTTGA	TCAAGT
Index 9	GATCAG	CTGATC
Index 10	TAGCTT	AAGCTA
Index 11	GGCTAC	GTAGCC
Index 12	CTTGTA	TACAAG
Index 13	AGTCAA	TTGACT

Appendix 2: GBS NGS library preparation protocol with *ApeKI*

Translated and adapted protocol
Reference code: PEX-NGS-003, v2
Dates: 21.3.2013 (created), 22.9.2014. (updated)
Responsible editors: Webber Audrey, Latreille Muriel
Approved by: Santoni Sylvain, laboratory manager
Institution: INRA Montpellier
Group: UMR AGAP: Genetic improvement and adaptation of Mediterranean and tropical species
Team: GE²Pop, AMM: Molecular marker workgroup

Aim: Preparation of DNA libraries for next generation broadband sequencing. The choice of a digestion enzyme is based on the restriction sites and the amount of DNA to digest.

Preparation of adapters:

> **Prepare a solution**:

50 mM Tris pH7	500 µl of the 1 M solution
50 mM NaCl	200 µl of the 2.5 M solution
mQ RNase-free water	9300 µl

> Filter the solution with a 0.22µ size filter

> **Adapters** (40 µM) – use strips of 8 PCR tubes to mix

GATA1_ tag 100 µM	20 µl
GATA2_ tag 100 µM	20 µl
Tris/NaCl solution	10 µl

> Launch the PCR program for oligo hybridization: **Adapter**

97°C	2 min
97°C	1 min --> gradual decrease in 1°C per cycle for 72 cycles
25°C	5 min
14°C	hold

> Dilute the adapters in water to 0.5 µM and store at -20°C

* Optional: check dosage by UV

> For 100mL of the Home Magic Solution (HMS), mix:

PEG 8000	20ml
NaCl 2.5M	50ml of 5M NaCl
	100ml water (ultrapure)

1. Enzymatic digestion with the *ApeKI* restriction enzyme (day 1)

> Place **200ng of DNA** (8µL to 25ng/µl) per sample into a 96-well plate

> Prepare a **mix** according to the conditions below:

	volume in µl
Buffer NEB3 10X	2 µl
ApeKI 1U	0.25 µl
water	9.75 µl

> Dispense 12µl of the mix into the wells containing the 8µl of DNA

* Mix up and down with the pipette 10 times

* Seal the plate with the sealer

> Launch the digestion program: **GBS digestion 75**

75°C	2h
4°C	hold

* Adjust the cover temperature for 20-30°C higher than the block

2. Adapter oligonucleotide ligation (day 1)

* Determine unique sample - barcoded adapter - Illumina index associations

> **Incubation**:

Add 5µl of the ligated ds-barcoded adapters at 0.5 µM

(That will make up for 1 pmol of adapter per 200 ng of digested product.

It's possible to asjust according to the amount of PCR product available.)

> Prepare a **mix** with, for each sample:

5X Ligase Buffer	10µl
water	14 µl
T4 DNA Ligase (1U/µl)	1µl

> Distribute 25µl of mix per sample

* Mix up and down 10 times

* Seal the plate to the sealer

> Start the ligation program:

* no heated lid

30°C	10 min
22°C	4 h
8°C	hold

3. Mixing of the DNA tagged with different adapters: (day 2)

> Perform enzymatic inactivation at 65°C (30 min)

* The amount of pooled DNA pool must not exceed 2 µg

> Mix in equiproportion 25µl or 50µL* of each ligation product in one 1.5 ml Eppendorf lowbind tube and add up to 600 µL with water if necessary

* 200 ng were digested (in 50µl) so for 24 ligations we will take 24 x 25 µl

* Save the rest of the individual ligations in the plate and store at -20°C

4. Modified AMpure Bead Purification (Agilent Genomics) on the Invitrogen Magnetic Rack for 1.5 ml tubes

1st purification cycle (for 600 µl of ligated DNA)

> Add 1.25 times the volume of pooled DNA of modified bead solution: 75 µL of AMPure + 675 µL of Home Magic Solution HMS, stored at 4 ° C

* Mix up and down

> Incubate the tubes for 15 min at room temperature

> Place the tubes on the magnetic support 5 min

> Remove and discard the supernatant
> Without removing the tubes from the support, add 1 ml of ethanol 80%
> Wait 30 sec and remove the supernatant
> Repeat the ethanol wash 80%, then remove all the ethanol
> Let the tubes dry for 5 min and remove them from the magnetic holder
> Add 105 µl of ultra pure water and mix up and down
> Leave 2 min at room temperature
> Place the plate on the magnetic support 5 min
> Transfer 100 µl of the supernatant to a new tube

2nd purification cycle

> Add 100 µL of the bead solution:
75 µl of AMPure + 25 µl of Magic Solution HMS, stored at 4 ° C
* Mix up and down
> Incubate the tubes for 15 min at room temperature
> Place the tubes on the magnetic support 5 min
> Remove and discard the supernatant
> Without removing the tubes from the support, add 1 mL of ethanol 80%
> Wait 30 sec and remove the supernatant
> Repeat the ethanol wash 80%, then remove all the ethanol
> Let the tubes dry for 5 min and remove them from the magnetic holder
> Add 35 µl of ultra pure water and mix up and down
> Leave 2 min at room temperature
> Place the plate on the magnetic support 5 min
> Transfer 30 µl of the supernatant to a new tube

5. Sizing on Blue Pippin (Sage Science) * optional

> Perform sizing on Blue Pippin with 30 µl, size range: tight (550 bp)
> After sizing, recover a maximum of 60 µl in the elution chamber
> If elution volume is less than 60 µl, then add water up to a total of 60 µl

6. Amplification of the ligated and sized fragments

* Assign correct RPI (Illumina) indices per DNA pool.

> Distribute the 60µl eluate in two wells (2x30µl)
that will have the same RPI index

For each well containing 30 µl of ligated DNA:
> Add 1µL of the RPI index at 25µM

> Prepare a mix with, for each sample:

dNTP (2.5mM each)	6µl
TP 5X (Phusion HF)	10µl
Taq Phusion HF (2U/µl)	1µl
25 µM dilution of MP1	1 µl
0.5µM dilution of MP2	1µl

> Distribute 19µL of the mix per sample
* Mix up and down with the pipette 10 times (50µL)
* seal the plate to the sealer

> Place the plate in the thermocycler using the following PCR program:

98°C	30 sec	
98°C	10 sec	⎫
65°C	30 sec	⎬ 18 cycles
72°C	30 sec	⎭
72°C	5 min	
4°C	hold	

> Mix 50µl of each of the PCR copies of the same pool together

7. Purified modified XP amide on the Invitrogen Magnetic Rack for Tubes 1.5 ml:

> Add 100 µL of the bead solution (10 µL of AMPure + 90 µL of HMS) to the 100 µL of PCR product.
> Incubate the tubes for 15 min at room temperature
> Place the tubes on the magnetic rack for 5 min

> Remove and discard the supernatant
> Without removing the tubes from the support, add 0.4 ml of 80% ethanol
> Wait 30 sec and remove the supernatant
> Repeat the ethanol, then remove all the ethanol
> Let the tubes dry for 5 min and remove them from the magnetic rack
> Add 30 µl of ultra pure water and mix up and down with the pipette
> Leave 2 min at room temperature
> Place the plate on the magnetic rack for 5 min
> Transfer 25 µl of the supernatant to a new tube

8. Validation, dosing and mixing of indexed libraries

> All pooled libraries can be mixed together or based on the lanes they will be sequenced in (nested multiplexing provides correct recognition)

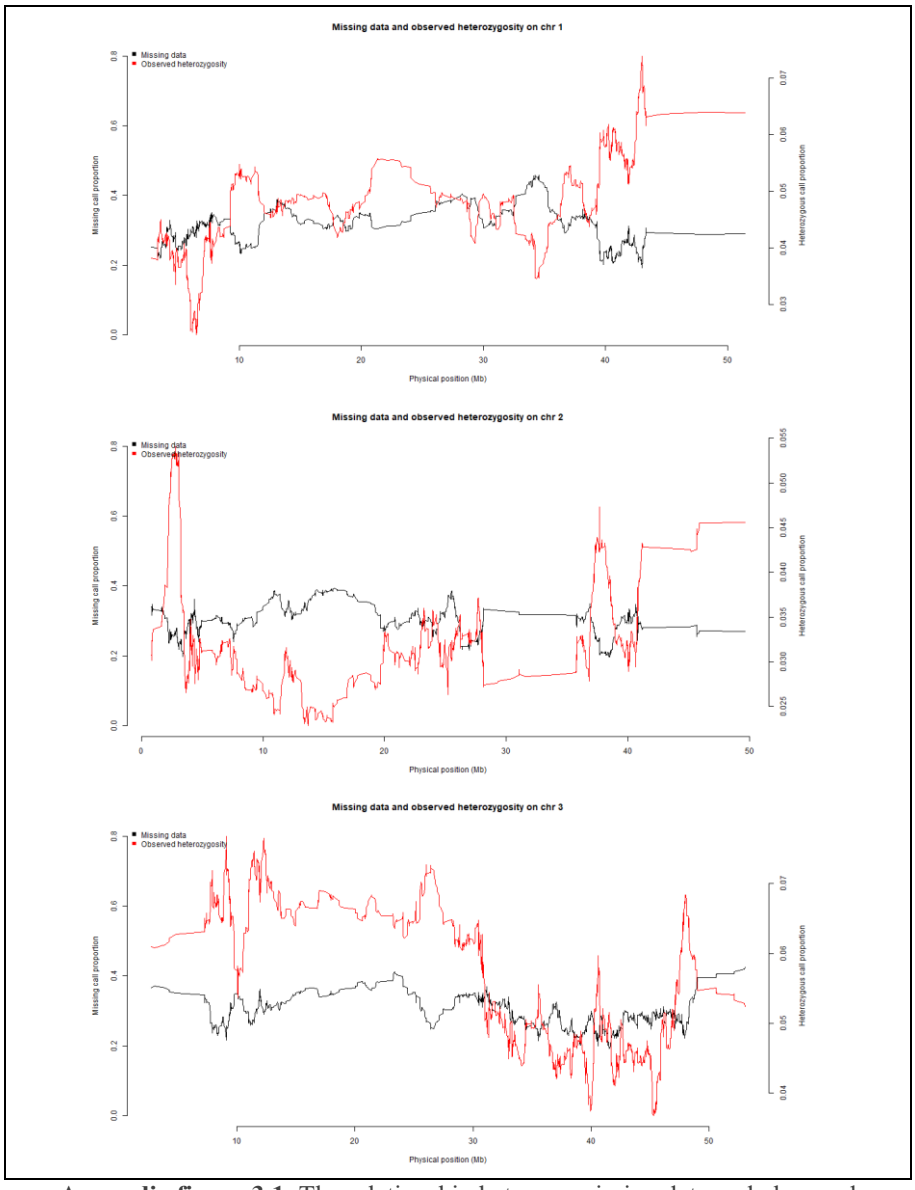
> Deposit 1 and 2 µl of an indexed and purified GBS library on an Agilent DNA 7500 chip to measure the dosage

> If necessary, perform also UV dosing (Nanodrop)

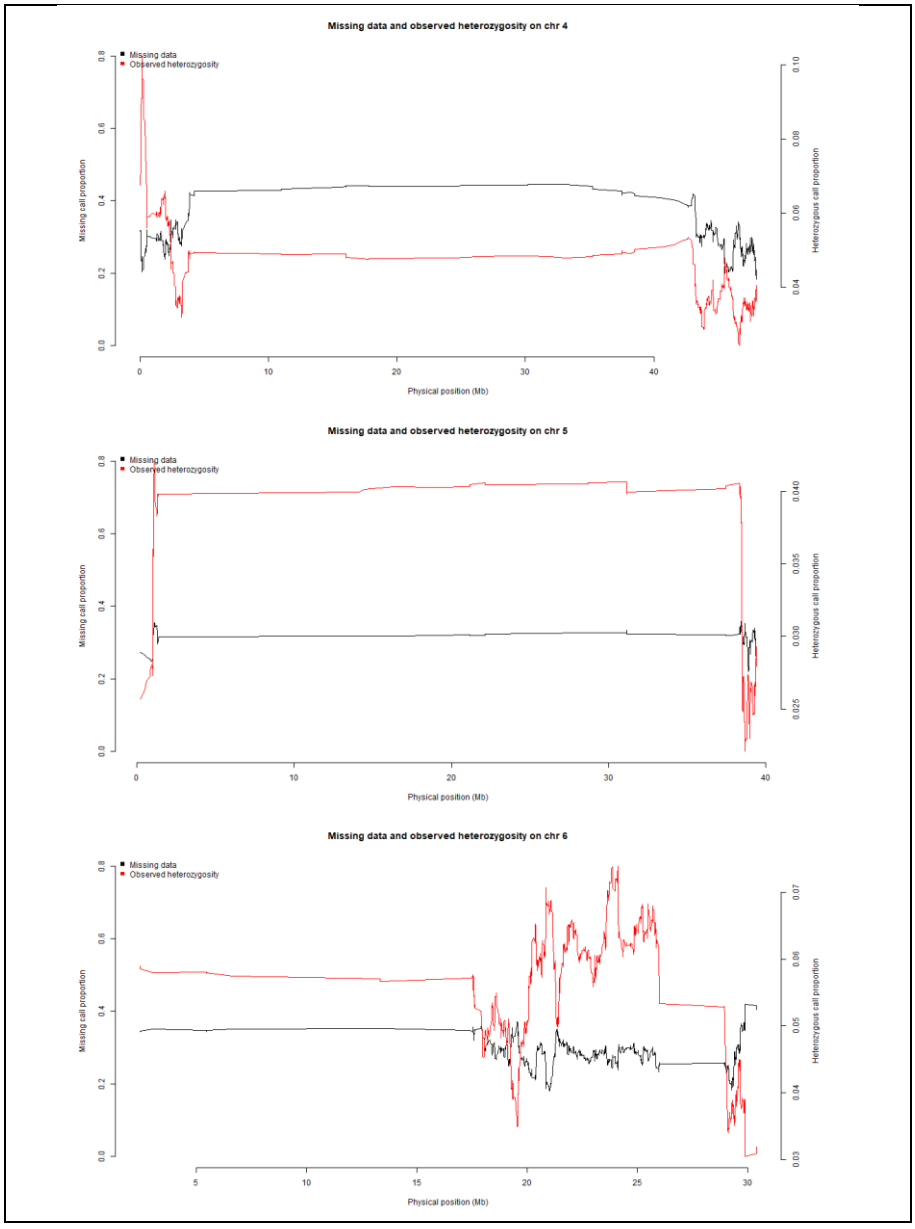
> Perform precise dosage of libraries using qPCR NGS_KAPA

Note: BluePippin Protocol, MiSeq protocol, Agilent DNA 7500 chip and qPCR NGS_KAPA protocols, all are available upon request.

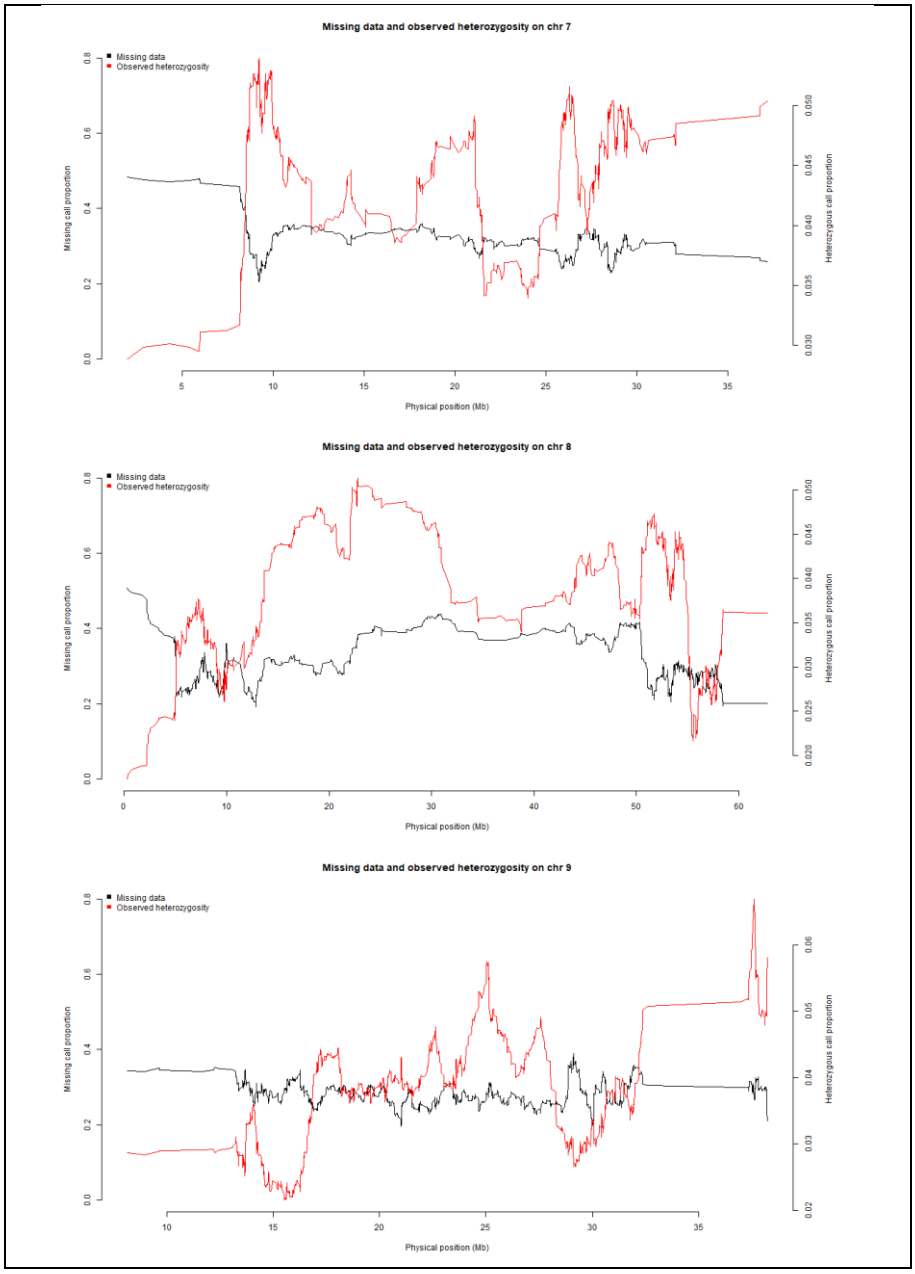
Appendix 3. Missing data and heterozygosity plots, filtered dataset



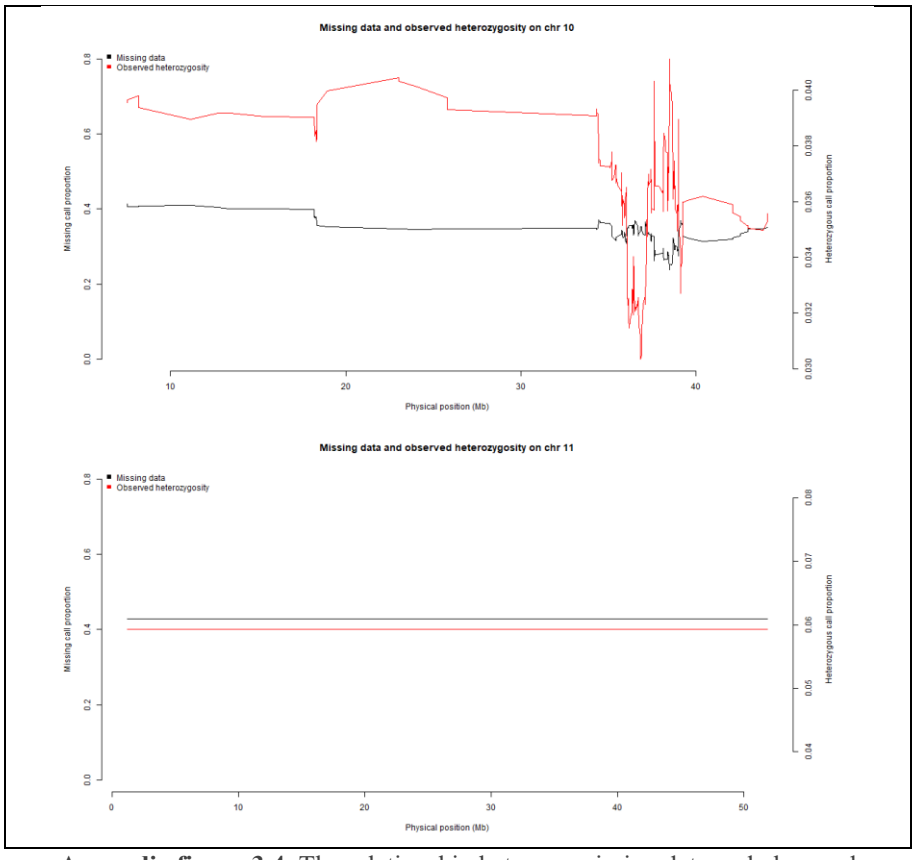
Appendix figure 3.1. The relationship between missing data and observed heterozygosity in the filtered dataset (chromosomes 1 to 3).



Appendix figure 3.2. The relationship between missing data and observed heterozygosity in the filtered dataset (chromosomes 4 to 6).

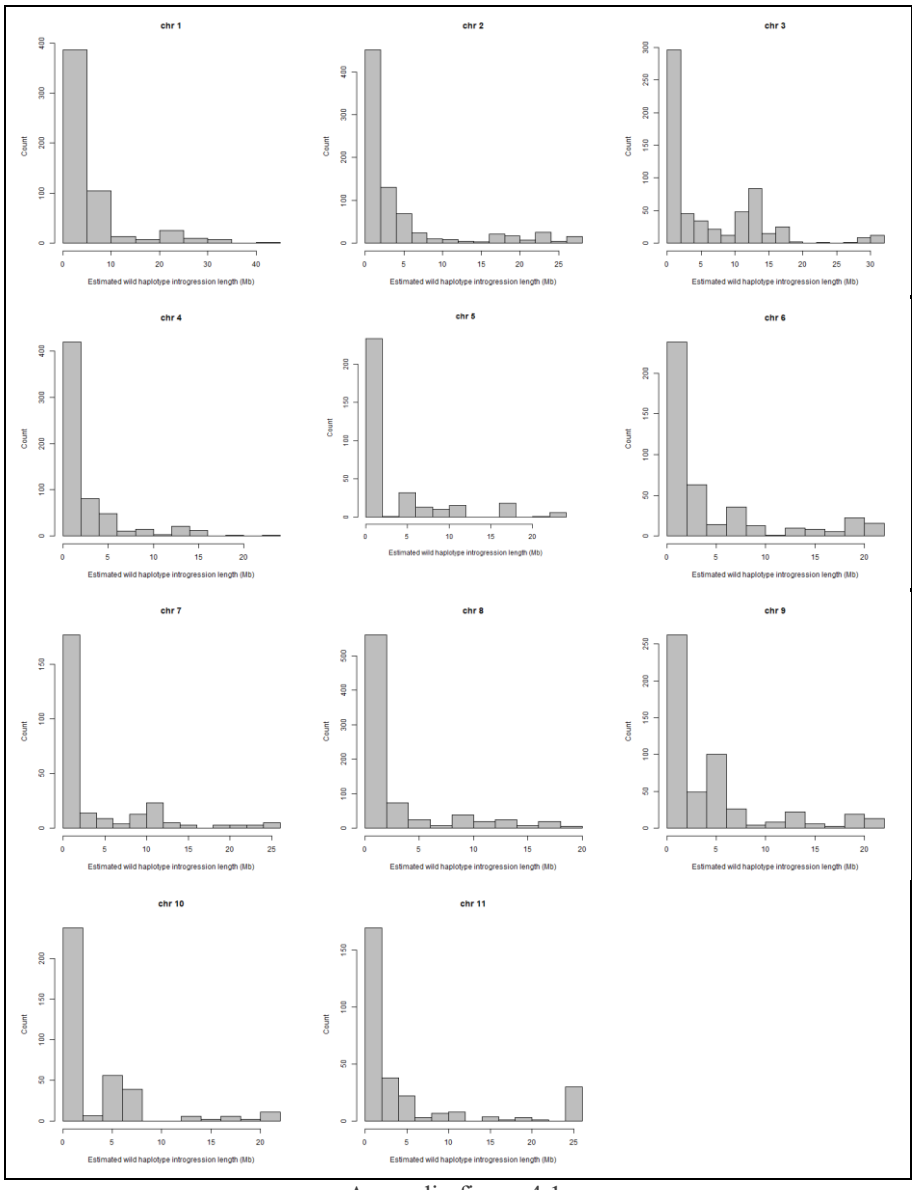


Appendix figure 3.3. The relationship between missing data and observed heterozygosity in the filtered dataset (chromosomes 7 to 9).



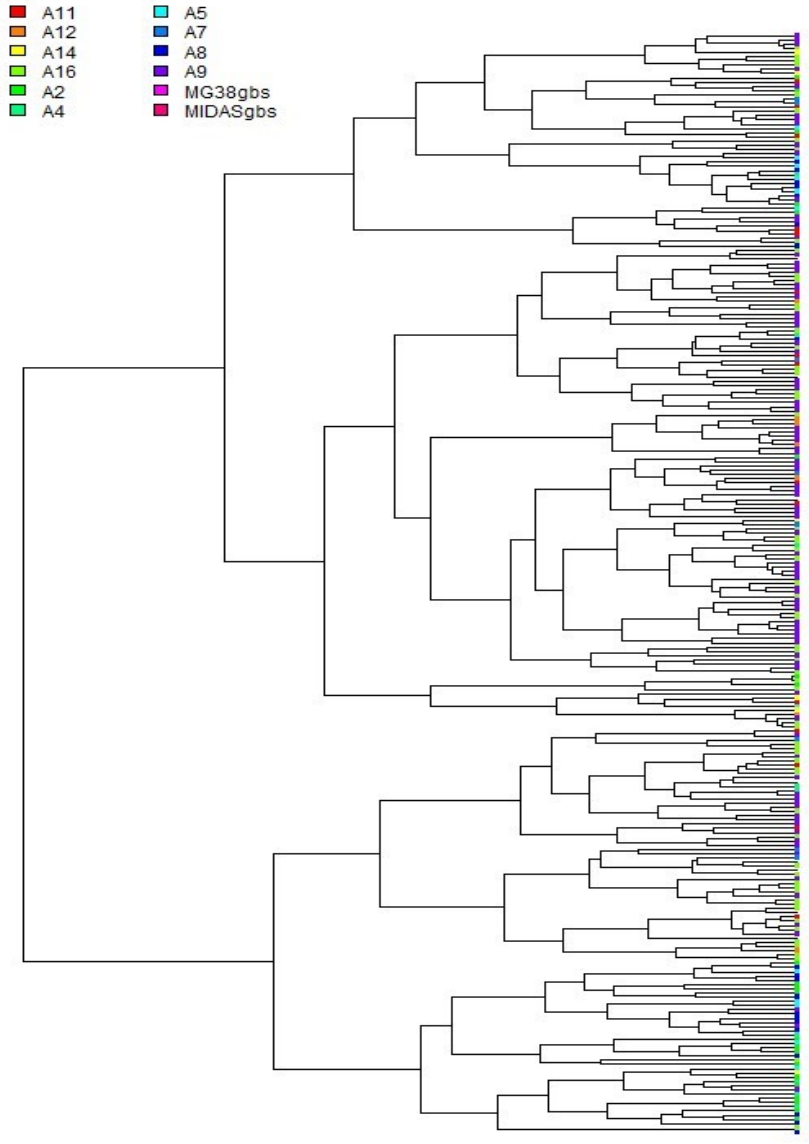
Appendix figure 3.4. The relationship between missing data and observed heterozygosity in the filtered dataset (chromosomes 10 and 11).

Appendix 4. Estimated introgression segment length



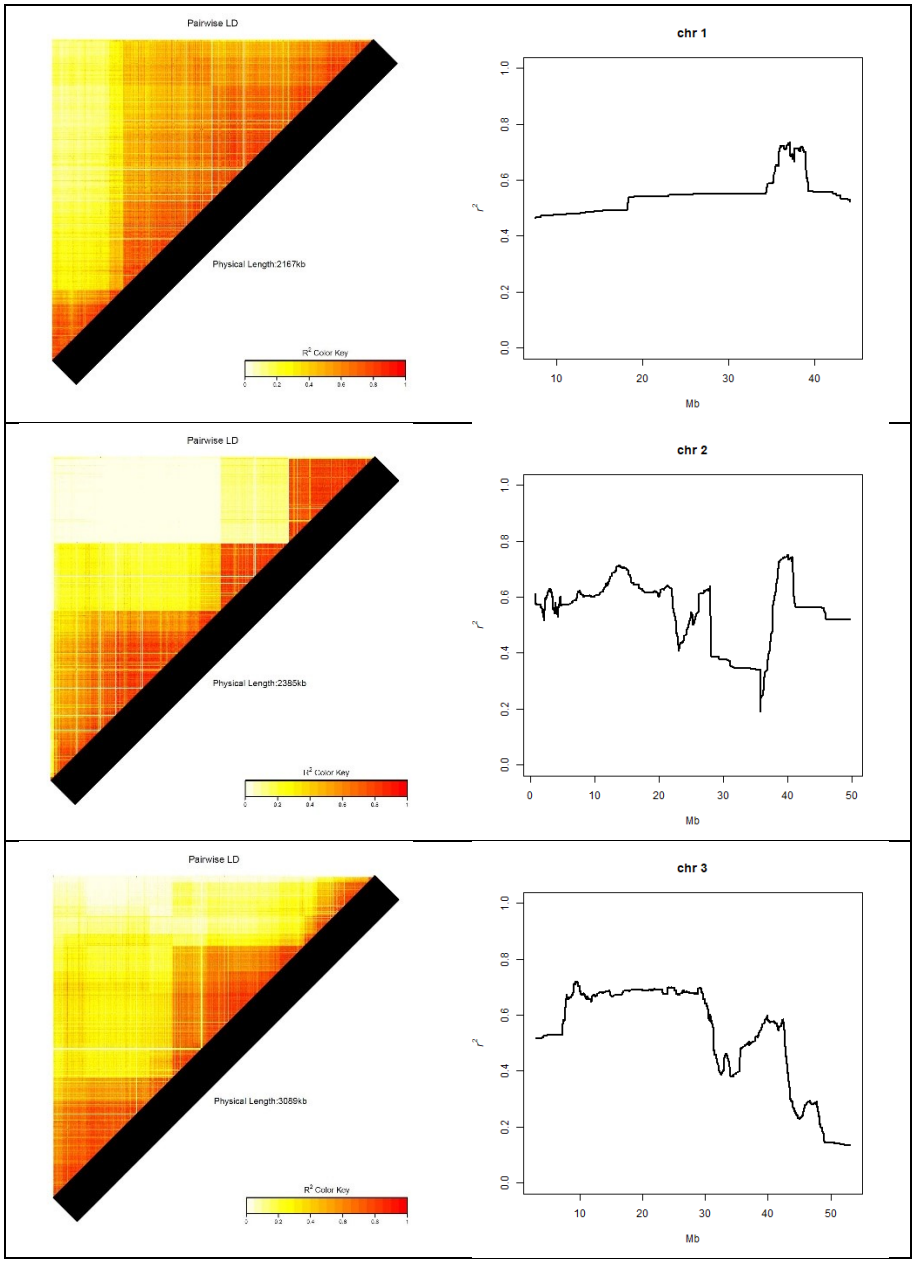
Appendix figure 4.1.

Appendix 5. Phylogenetic tree

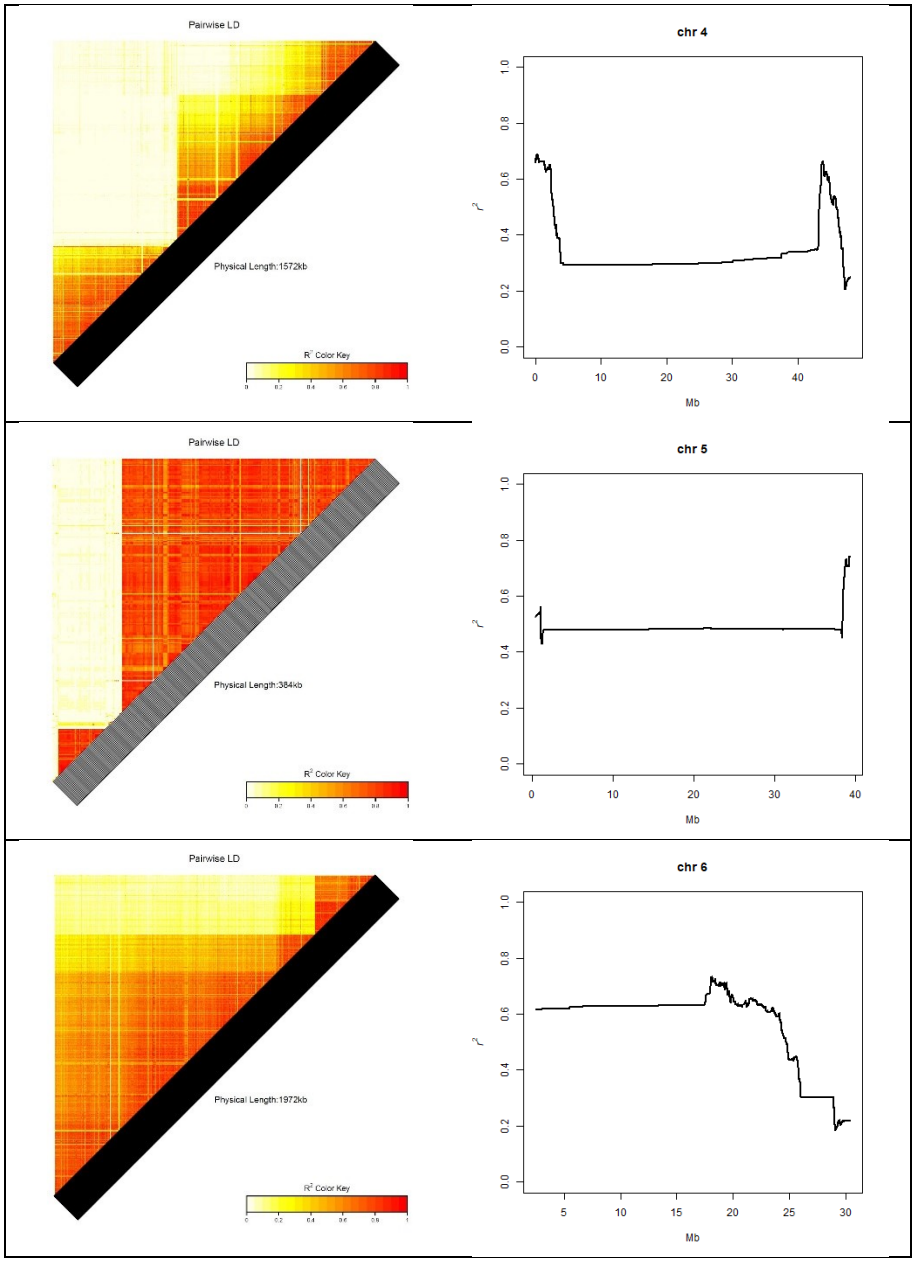


Appendix figure 5.1. Dendrogram based on the kiship heatmap produced by GAPIT.

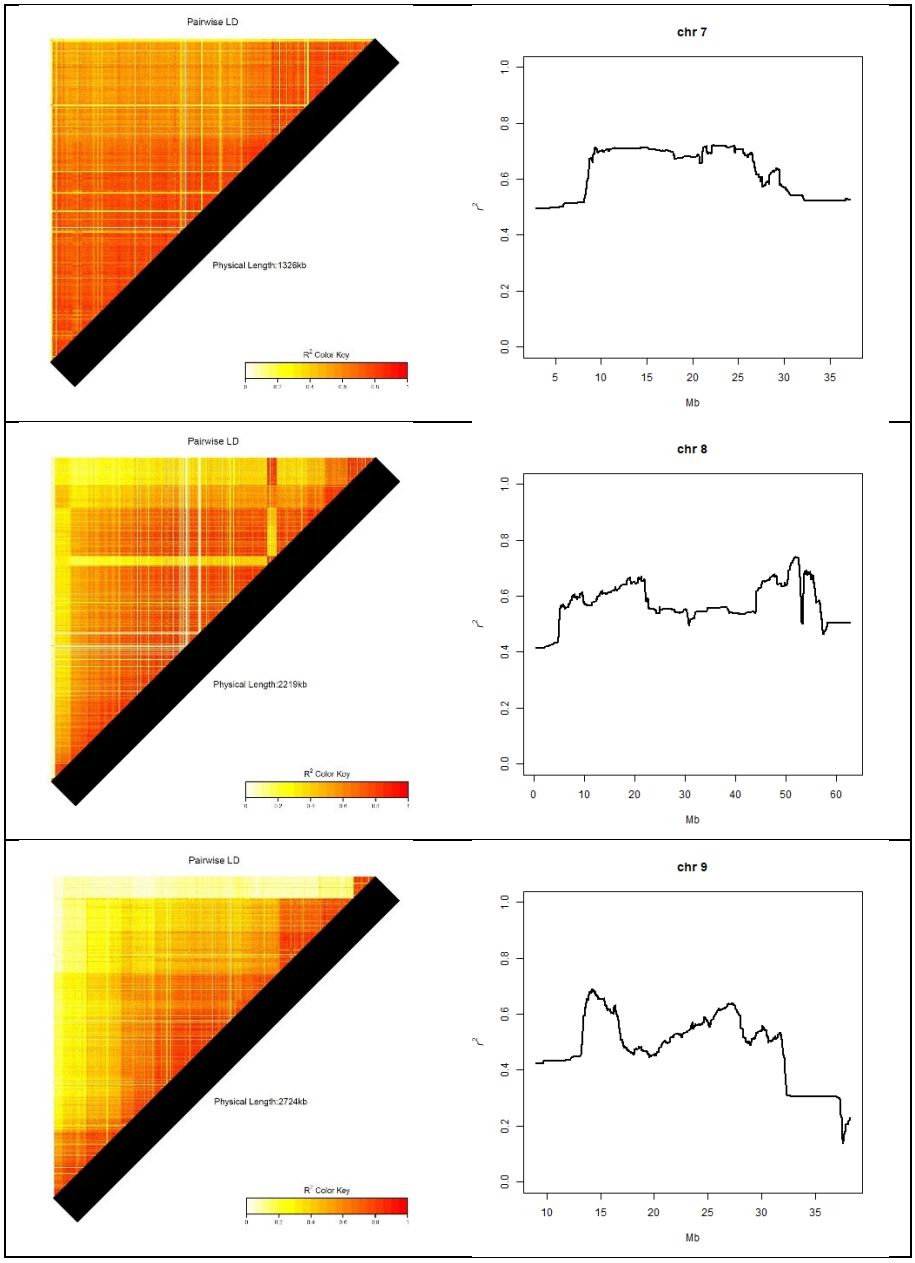
Appendix 6. LD heatmaps and LD evolution over chromosomes



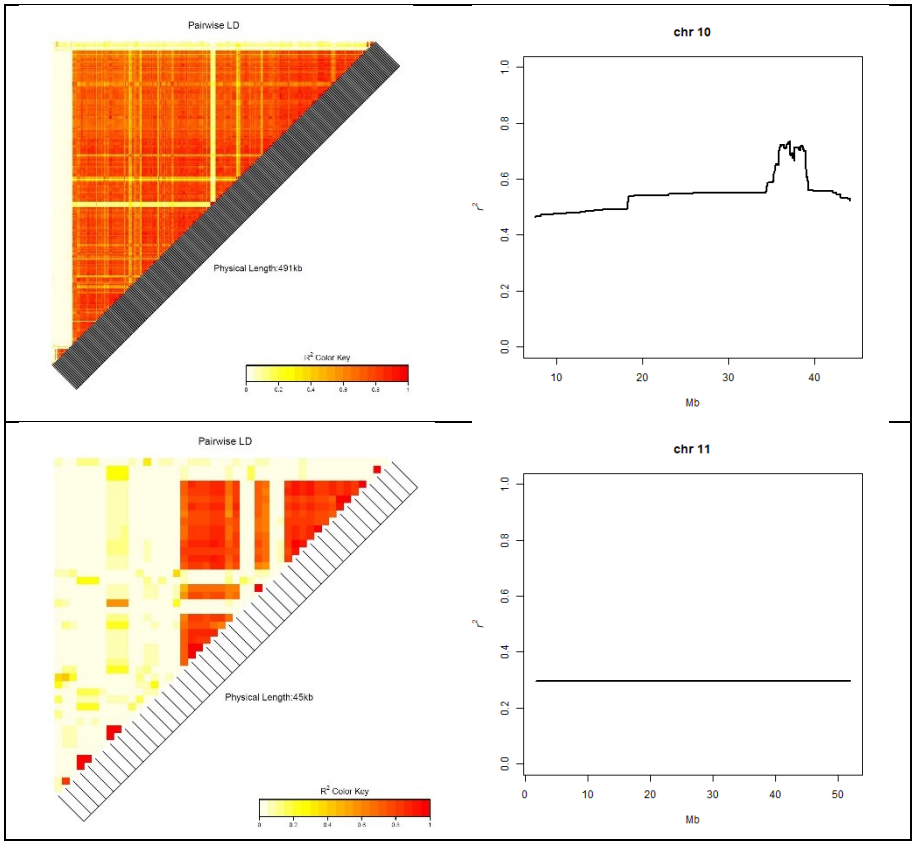
Appendix figure 6.1. LD heatmaps and LD plots for chromosomes 1 to 3.



Appendix figure 6.2. LD heatmaps and LD plots for chromosomes 4 to 6.



Appendix figure 6.3. LD heatmaps and LD plots for chromosomes 7 to 9.



Appendix figure 6.4. LD heatmaps and LD plots for chromosomes 10 and 11.