ABSTRACT

JIMENEZ MADRIGAL, JOSE PABLO. Next-Generation Sequencing Technologies in Tree Improvement and Conservation Genetics of *Dipteryx oleifera* Benth. (under the direction of Theodore H. Shear and Ross W. Whetten).

Dipteryx oleifera Benth. is a tropical tree species, endemic to the Caribbean lowlands, ranging from Nicaragua to Colombia. This keystone tree species provides food and shelter for many mammals and birds, including the endangered great green macaw. In addition, its high-quality wood has significant economic value. However, illegal logging and habitat fragmentation have diminished natural populations to the point that trade of the wood is now controlled by international treaty (CITES). To help this species conservation, GENFORES, a forestry industry – university co-operative program based in Costa Rica, started a *D. oleifera* breeding program. My research is part of that broader effort led by GENFORES. The goal of this dissertation was to explore innovative technologies and develop genomic resources to expedite the selection process and breeding program. Next-generation sequencing technologies provide an ideal platform to accomplish this goal, because it allows for whole-genome study at relatively low cost. Furthermore, it enables the identification of thousands of single nucleotide polymorphisms (SNPs) in multiple individuals simultaneously. SNPs are ubiquitous, codominant, and can be in functional parts of the genome, thus making them suitable markers for both tree improvement and conservation genetics.

The first objective of this dissertation was to determine *D. oleifera* genome size. Relative DNA nuclear content (2C) was estimated at 3.86 pg using flow cytometry and confirmed with sequencing data. Genome size variation is a common phenomenon among all organisms. In plants, many phenotypic traits show a correlation with genome size. One such trait is seed size, which in turn influences many aspects of plant ecology. In addition to genome size, I explored

the evolution of this trait along with seed size in the Dipterygeae clade. Although a small sample set, the results from this study show a moderate correlation between seed size and 2C-values, as well as a similar evolutionary history, *i.e.*, species with larger genomes also have bigger seeds. It is hypothesized that one or more polyploidization event may account for the variation seen in the traits.

The second objective of this dissertation was to generate the first draft genome sequence of *D. oleifera.* For the assembly, I used a combination of short Illumina reads and additional coverage in long PacBio reads. With a total of 1, 166,468,433 bp in 381,857 contigs, this assembly corresponds to 62% of the estimated 1C genome size. Although still fragmented, the resulting assembly contains 70.7% of complete single-copy and duplicated conserved orthologous genes (BUSCOs). In addition, I used short Illumina reads data and the resulting assembly to estimate ploidy level for this species. Based on biallelic markers frequency distribution, from the sequence data, *D. oleifera* is a tetraploid species, most likely an autopolyploid.

Finally, the third objective of this dissertation was to identify SNPs for marker-informed breeding. For DNA sequence variant discovery, I used a Genotyping-by-Sequencing approach in a *D. oleifera* progeny open-pollinated progeny trial. This resulted in 2,612 SNPs identified and 185 individuals genotyped. Marker data was used to estimate the realized genomic relationship among individuals in the progeny trial. Results between pedigree-based (A matrix) and pedigreed-based marker corrected (H matrix) models were compared for three traits: stem diameter, tree height, and total volume. Although a low-density panel, these markers were able to accurately estimate the genetic relationship among individuals in the progeny trial. The use of a marker-corrected relationship matrix improved model fit and parameter estimation accuracy.

More importantly, it highlighted a major constraint when working with open-pollinated progeny collected from natural populations. Under these conditions the assumptions of traditional pedigree-based models are most likely unrealistic, and marker data can capture better the true relationship among individuals.

Next-Generation Sequencing Technologies in Tree Improvement and Conservation Genetics of *Dipteryx oleifera* Benth.

by
Jose Pablo Jimenez Madrigal

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Forestry and Environmental Resources

Raleigh, North Carolina
2018

APPROVED BY:

_____	_____
Dr. Theodore H. Shear Co-chair of Advisory Committee	Dr. Ross W. Whetten Co-chair of Advisory Committee
_____	_____
Dr. Juan J. Acosta	Dr. Olman Murillo External Member
_____	
Dr. Qiuyun Xiang	

DEDICATION

To study genetics and disregard the key contribution of parents on their offspring would be the biggest mistake. Hence, I like to dedicate this work to my mom and dad, without them I would not be here, literally!

BIOGRAPHY

Jose P. Jimenez Madrigal was born in San Jose, Costa Rica. He attended Universidad de Costa Rica (UCR) and graduated with a Bachelor of Science degree in Biology. During his undergraduate studies he worked as a research assistant testing molecular markers for tropical trees. After graduation, he started working in crop breeding and genetic improvement at the Centro de Investigaciones Agronómicas (CIA).

Inspired by his experience at CIA, Jose looked for opportunities to further his education and in 2009 he was granted a Fulbright scholarship. He decided to go back to forest genetics and to pursue a master's degree in Forestry at NC State University (NCSU). He worked under the supervision of Dr. Gary Hodge and Dr. Ross Whetten developing molecular markers for tropical pine hybrid identification.

Upon getting his master's degree Jose went back to Costa Rica and started working for Instituto Tecnológico de Costa Rica (TEC) as a lecturer and researcher. In 2014, Jose was granted a full scholarship by his employer and returned to NCSU for his doctorate degree. He worked under the supervision of Dr. Ted Shear and Dr. Ross Whetten developing genomic resource for tree improvement and conservation of tropical tree *Dipteryx oleifera*. After finishing his doctorate degree Jose will go back to Costa Rica and start a tenure track position at TEC.

ACKNOWLEDGMENTS

I would like to express my profound appreciation to Dr. Ted Shear and Dr. Ross Whetten for their guidance and support; they are true mentors and amazing role models! I would also like to thank committee members Dr. Juan Acosta, Dr. Jenny Xiang, and Dr. Olman Murillo for their support and good suggestions.

I would like to acknowledge the guidance of Dr. Thomas Ranney and Research Technician Nathan Lynch at the Mountain Horticultural Crops Research and Extension Center; Dr. Steve McKeand and Dr. Fikret Isik, at the Tree Improvement Program, their classes were some of the best and informed much of what I did; the collaboration of Dr. Domingos Cardoso at the Herbarium of Universidade Estadual de Feira de Santana; and the help with lab work from Dr. Lilian Matallana at NC State and M.Sc. Fabiana Rojas at Instituto Tecnológico de Costa Rica.

Special thanks to GENFORES, Instituto Tecnológico de Costa Rica, and the Bruce and Barbara Zobel Endowment for International Forestry Studies, without their resources and financial support this study would not have been possible.

I would like to thank fellow graduate students Will Kohlway, Bruno Kanieski, Martin Pettersson, Matthew Jurjonas, and Patricia Maroto for helping me with the experiments, and sharing thoughts and experiences.

My deepest gratitude to Graduate Program Coordinator Sarah Slover for her patience and support throughout the entire process.

Last but not least, I would like to thank Vicky, who against her better judgment agreed to marry me. Her love and support are what got me through this ordeal.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2.

Chapter 3.

Chapter 4.

LIST OF FIGURES

Chapter 2.

Chapter 3.

Chapter 4.

CHAPTER 1: Next-Generation Sequencing Technologies and Forestry Applications

At the beginning of the decade a total of 55 plant genomes had been sequenced and new DNA sequencing technologies were expected to increase that number rapidly (Michael and Jackson 2013). Today, a quick search of the major databases (*e.g.*, NCBI, Phytozome, Ensembl, PlantGBD, etc.) shows the number of complete and partial plant genomes sequences in the hundreds. These genomic data have revolutionized the understanding of plant functions, both at the individual and population level, led to the discovery of new genes and metabolic pathways, and opened the doors to a new way of selection in plant breeding and genetic improvement programs (Desta and Ortiz 2014). However, of the plants that have had their genomes sequenced, only a few are timber species, such as eucalypt (*Eucalyptus grandis*) and loblolly pine (*Pinus taeda*) (Myburg et al. 2014; Neale et al. 2014).

In this chapter, I briefly describe the principal technologies involved in DNA sequencing, their applications in forestry, and present my case for their use in a *Dipteryx oleifera* Benth. breeding program. *D. oleifera* is a tropical timber species with ecological and commercial value. The main goal of this dissertation is to develop genomic resources that can help *D. oleifera* breeding and conservation.

A brief history of DNA sequencing

DNA sequencing refers to the process of determining the arrangement of nucleotides (nt) in a DNA molecule. The process consists of three basic steps: (i) sample preparation, during which the DNA strand is broken into multiple small fragments; (ii) the actual reading of the fragments by a combination of physical methods and chemical reactions to determine the precise

order of the nucleotides, and (iii) the reassembly of the sequence, currently involving the use of bioinformatics software to align the overlapping reads into a contiguous sequence (Schadt et al. 2010). The details of sample preparation, reading, and reassembly change considerably from technique to technique, but the overall process remains the same.

The so-called first generation of DNA sequencing started over 40 years ago with the seminal work of Sanger et al. (1977). Their publication describes a method for sequencing DNA by use of nucleotide analogs that acted as specific chain-terminating inhibitors of DNA polymerase. The product of these prematurely-terminated amplifications could be visualized by gel electrophoresis, compared, and recorded manually to generate a sequence. Although time-consuming and with limited throughput (up to 300 nt per run), it was a simple and accurate way to obtain sequence data. Nowadays, automated Sanger sequencing can generate sequence reads above 1000 nt and at a faster rate; however, the cost of this technique is still too high to be widely used for whole genome sequencing (Hert et al. 2008). Cost aside, Sanger sequencing dominated the field for almost 30 years. It continues to be considered the "gold standard" due to its accuracy, and it is still used for validation of many plant sequencing projects.

In response to the limitations of Sanger sequencing, and fueled by grants from the United States National Human Research Institute, new sequencing technologies were developed (Schloss 2008). These next-generation sequencing technologies (NGS), also known as second-generation technologies, vastly increased the output by sequencing many DNA fragments in parallel reactions. The principle behind these is commonly known as "wash-and-scan" since the process involves attaching several DNA fragments to a substrate, applying chemical agents such as labeled nucleotides and enzymes, and stopping the amplification reaction by washing away the excess reagent. This cycle is repeated until the reaction is no longer viable while scanning in

between cycles to read the newly incorporated nucleotides (Schadt et al. 2010). Several platforms used variants of this procedure; this review will focus on the most widely-used, *i.e.*, 454 Roche, Illumina/Solexa, Ion Torrent, SOLiD, and PacBio.

The first NGS to be commercially available was pyrosequencing, as implemented in the 454 Roche platform. This method generates close to 200,000 reads with up to 330 nt each. In pyrosequencing, the DNA fragments are captured on a bead. Each bead, along with amplification reagents (enzymes and primers), is dropped into a well of a fiberoptic slide and exposed to a flow of unlabeled nucleotides, each time a nucleotide is incorporated into the new DNA strand a pyrophosphate molecule is released, leading to a light emission that is monitored in real time, hence the name pyrosequencing. One advantage of the method is longer reads obtained in a short time. However, it has a high cost in term of reagents, and is prone to errors in homopolymer repeats (Metzker 2010; van Dijk et al. 2014).

Following shortly after the release of the 454 Roche platform was the Illumina/Solexa platform. Illumina works by the principle of sequencing by synthesis (SBS, as described by Ju et al. 2006). Like pyrosequencing, DNA fragments are bound to a solid surface coated with adapter oligonucleotides. These fragments are then treated with nucleotide analogs that possess both a fluorescent dye label and terminating/inhibiting group halting the reaction. At each cycle, the luminous signal is monitored, and then the terminating/inhibiting group is cleaved and washed away to allow the reaction to continue. Since each nucleotide is marked with a distinct color dye, it is possible to track the sequence of nucleotides. Illumina is currently the most used NGS technology due to its high throughput and low per-base cost. However, Illumina is technically challenging and prone to error in low complexity samples (Metzker 2010; van Dijk et al. 2014).

Another NGS technology that depends on binding the DNA fragments to a surface is Ion Torrent. This method differs from the 454 Roche platform by the use of an ion sensor instead of imaging technology to detect the release of protons during the nucleotide incorporation. This technology suffers from the same setbacks as pyrosequencing, mainly a high error rate associated with homopolymers. On the other hand, ion sensor detection reduces run time considerably (van Dijk et al. 2014).

The SOLiD platform, which stands for Sequencing by Oligo Ligation Detection, is comparable to Illumina's SBS. Though in this case sequencing is done by ligation (SBL), it uses DNA ligase and either one-base-encoded or two-base-encoded probes. The probes are labeled with a fluorophore and hybridized to their complementary DNA sample. The probes that do not ligate to a template are washed away, and the fluorescent signals from the ones that do are recorded. The process is repeated for multiple cycles, removing the probes by cleavage and adding new ones. The SOLiD platform has one of the highest accuracies, 99.94%. Nevertheless, run times are long and read lengths are among the smallest at approximately 75 nt (Metzker 2010; van Dijk et al. 2014). Despite its shortcomings, SOLiD has been used successfully for genome sequencing in model organisms such as the roundworm *Caenorhabditis elegans* (Valouev et al. 2008).

Finally, Pacific Bioscience has developed a new and exciting third-generation sequencing technology. The PacBio platform relies on single polymerase molecule sequencing (SMS) (Eid et al. 2009). What is revolutionary about this technology is the amplification reaction does not need to be halted between reading steps. Furthermore, nucleotide readings are done in real time. These allow for a reduction in run times. SMS can be accomplished in diverse ways: (i) with SBS technologies at a single molecule level, (ii) with nanopore-sequencing technologies, or (iii)

with advanced microscopy techniques that use direct imaging of individual DNA molecules. Regardless of the method, read lengths in this platform can reach averages of 10 to 20 Kilobases (Kb). Longer reads mean assembly and sequence alignment are more straightforward, which may be ideal for *de novo* genome sequencing projects. On the downside, this technology is more expensive than some of the second-generation technologies, and the reads have a higher error rate (Schadt et al. 2010).

NGS applications in tree breeding and forest management

Many authors have suggested that DNA sequencing technologies and genomic information can benefit the forest industry (Neale and Kremer 2011) by providing insight into growth traits (Grattapaglia et al. 2009), by illuminating the relationship between genotypic and phenotypic diversity (Neale and Ingvarsson 2008), and by improving breeding of domesticated trees (Neale 2007). So, how can NGS technologies be used in forest tree breeding? The answer is threefold: (i) gene discovery, (ii) next-generation Ecotilling in candidate genes, and (iii) high throughput genotyping.

Our study of genes and regulatory networks in trees has been limited, mainly because of their large genome size, long generation times, and limited molecular genetic knowledge base. NGS technology opens the door to a new set of tools for gene discovery. For example, in black cottonwood (*Populus trichocarpa*, the first tree species to have its genome sequenced) a high number of expressed sequence tags (EST) have been identified and are publicly available in gene databases worldwide. Since then, significant advances in poplar functional genomics have been accomplished, which range from transgenic trees used for biofuel production (*e.g.*, cellulose and lignin content modification) to understanding how adaptations to environmental factors and

productivity are controlled. Poplar is considered a model organism to study adaptive traits in woody plants, epigenetic regulation, and life history of trees (Brunner et al. 2004).

In the same way, high-throughput sequencing using the 454 Roche platform facilitated the discovery of genes and single nucleotide polymorphisms (SNP) in flooded gum (*Eucalyptus grandis*) (Novaes et al. 2008). The number of EST sequences available in the GenBank is in the tens of thousands; many are used in expression quantitative trait loci (eQTL) studies. Genes involved in xylem and wood-forming tissues, biotic stress resistance, and cold tolerance have been annotated and characterized. Identification of SNPs associated with complex traits, such as wood density and microfibril angle, has also been accomplished (Grattapaglia and Kirst 2008; Neale and Kremer 2011).

Norway spruce (*Picea abies*) was the first genome of a gymnosperm to be sequenced (Nystedt et al. 2013). The information provided great insight into the evolution of conifers, for instance, that the large genome size is not due to whole-genome duplication but rather to steady accumulation of transposable elements (mostly long terminal repeat-retrotransposons or LTR-RT's, though that may not be the case for other gymnosperms). This work also allowed the comparison of gene homologs between gymnosperms and angiosperms.

An alternative approach to gene discovery, where no reference genome exists, is transcriptome analysis. RNA sequencing, like NGS genomic approaches, allows *de novo* sequencing of an organism's transcriptome. While it will not represent the full set of genes encoded in the genome, it does provide insight into the metabolic capabilities of an organism in response to different environmental conditions. This approach was used successfully both in annual plants (*Arabidopsis thaliana*) and fungi (*Verticillium dahliae*) (Landesfeind and Meinicke 2014), but it is just beginning to be tested in tree species.

In cases where genes have already been identified, the interest may lie in determining the range of mutations present in each species. Sequencing candidate genes in a large natural population can be used to screen for possible variants of both common and rare alleles. Once all polymorphisms are identified, the information can be used in breeding programs. This technique is called next-generation Ecotilling and relies on the analysis of pooled samples. Four conditions are needed for Ecotilling: (i) validated genes for a trait of interest, (ii) a large population of trees that can be crossed, (iii) effective methods to evaluate the effect of different mutations on the individual's phenotype, and (iv) an appropriate crossing scheme to introduce the selected mutations in the breeding population. An example comes from a set of black poplar trees (*Populus nigra*), where a nonsense mutation was detected on a gene involved in the lignin biosynthesis pathway (HCT1). Homozygous individuals for this mutation were selected and evaluated for wood quality (Harfouche et al. 2012). Lower lignin content is a desirable trait in the pulp and paper industry as well as in biofuel production.

Finally, another application of NGS technologies in tree breeding is high-throughput genotyping. There are different procedures for genetic marker discovery and genotyping individuals from NGS data, for example, reduced-representation libraries (RRLs), complexity reduction of polymorphic sequences (CRoPS), and restriction-site-associated DNA sequencing (RAD-seq) (Davey et al. 2011). Furthermore, many bioinformatic and statistical tools have been developed for genotyping, SNP calling, and overall analysis of NGS data (Nielsen et al. 2011). In general, genotyping-by-sequencing (GBS) uses restriction endonucleases to target only a small portion of the genome, hence the reduction in complexity. The enzyme digestion is coupled with DNA barcoded adapters to produce multiplex libraries of samples ready for most NGS platforms. The approach is high-throughput and efficient. GBS results in thousand or even

millions of short DNA sequences which share a restriction site and can be compared within or among different individuals, allowing the identification of polymorphisms in the sequence that can be used as molecular markers. Moreover, it does not require previous knowledge of the genome sequence; *de novo* discovery is crucial in uncharacterized species, which is the case of most forest trees. If a well-defined reference genome exists, it can be combined with GBS data to create a genetic map, that makes population characterization easier (Poland and Rife 2012).

The best attribute of GBS is the vast amount of molecular markers that it generates and how those markers can be used for marker-assisted selection (MAS) or even genomic selection (GS). Genomic selection assumes that every trait locus, *i.e.*, the gene or location in the genome influencing the expression of a phenotypic trait, has the probability of being in linkage disequilibrium with at least one of the molecular markers in the entire target population. Therefore, the use of high-density markers is a fundamental feature, especially in tree species where low LD is commonly reported. Furthermore, GS could potentially accelerate breeding cycles, in some cases even reducing it by half, which represents considerable gain since most tree breeding cycles are measured in decades rather than months or years like most crops (Desta and Ortiz 2014). Another advantage is the accuracy of the predictions. For example, preliminary studies in loblolly pine found that the accuracy of estimations of breeding values for traits associated with wood properties (cellulose and lignin content) using GS were comparable with the accuracy of breeding values based on pedigree and phenotype information (Isik 2014). Overall, marker-assisted selection and genomic selection are promising applications in tree improvement, but we still need to be cautious of some of its caveats. For instance, it is not yet known how accurate the genomic predictions will be on trees several generations removed from the training or reference population (Grattapaglia and Resende 2011).

A lesser-explored but equally important application of NGS and GBS is in ecological restoration. Some of the practical applications of this discipline are: (i) the delineation of local genetic provenance seed sourcing zones, (ii) comparative assessment of genetic diversity in restored sites versus natural populations, and (iii) detection of genetic changes over generations (Williams et al. 2014). However, the most useful contribution could be in testing putatively adaptive markers associated with performance, not just for commercial traits but environmental resilience as well. With changes in global climate conditions, the ability to identify provenances best adapted to restoration or plantation sites will be essential.

The case for NGS in tree improvement and conservation of *Dipteryx oleifera*

Dipteryx oleifera Benth., formerly known as *D. panamensis*, is a large canopy-emergent tropical tree that can reach up to 50 m in height and 1.6 m in diameter. It is endemic to Nicaragua, Costa Rica, Panama, and Colombia. It can be found in humid and very humid tropical forests in the lowlands of the Atlantic plains, where annual precipitation ranges from 3500 to 5500 mm and temperatures fluctuate between 24 and 30 °C. It grows in a variety of soils, from sandy alluvial soils to acidic and clayey soils, at elevations ranging from 20 to 1300 m above mean sea level (masl), but it is most commonly found below 600 masl (Flores 1992; Vozzo 2010).

The trunk is straight with ample basal roots, the bark is yellowish and granular with vertical lenticels, and the crown is semispherical. It flowers annually, between late May and August, though blooming is highly dependent on weather conditions (temperature and precipitation). The tree is pollinated by up to 18 distinct species of bees but is also visited by hummingbirds and butterflies. Fruiting is annual with peak production between February and

March (González and Origgi 2003). Fruits have a single big seed with a thin layer of brown pulp surrounding it. The seed has an average fat content of 25%, making it a highly nutritious food (Murillo Gómez and Atehortúa 2013).

D. oleifera is considered a keystone species. It provides an ample food source during the dry season to 16 different species of mammals, including bats, rodents, and monkeys (Bonaccorso et al. 1980). Moreover, it is visited by over 100 different bird species, most notably the endangered great green macaw (*Ara ambiguus*). The great green macaw not only feeds on the fruits, but it nests almost exclusively in cavities in *D. oleifera* trunks. The relationship between *A. ambiguus* and *D. oleifera* has been thoroughly documented (Madriz Vargas 2004; Chun 2008; Gomez Figueroa 2009; Chassot and Arias 2012; Monge et al. 2012).

In addition to its ecological value, *D. oleifera* has very hard, dense wood with a specific gravity ranging from 0.83 to 1.09 (Vozzo 2010). The wood is durable and high in mechanical resistance; consequently, it is used for industrial floors, marine construction, machines, and sports equipment. The timber is harvested mostly from natural populations, although in Costa Rica this practice was restricted in 1996 and banned in 2008. When available in the local markets, *D. oleifera* wood is the most expensive, prized higher than native and introduced timber species such as acacia, eucalypt, and teak. Even the wood waste has potential economic value as fuel for energy generation (Gaitán-Álvarez 2015). Non-timber products are also valuable; in Colombia the seeds are roasted for food products, *e.g.*, candies and beverages (Murillo Gómez and Atehortúa 2013).

Despite the importance of *D. oleifera*, in both ecological and commercial terms, the amount planted is minimal, especially compared to non-native species like teak. Moreover, just a few studies have evaluated, directly or indirectly, its performance in a plantation setting

(Butterfield and Mariano 1995; Andrade Naveda 2002; Petit and Montagnini 2006; Schmidt 2009) or its potential for improvement (Martínez-Albán et al. 2016). NGS technologies can aid a *D. oleifera* tree improvement program by developing genomic resources. To develop such resources and direct applications, some fundamental information must be gathered first. The steps to follow are:

1. Determine the genome size and ploidy level of *D. oleifera*. For a breeding program to work, it is crucial to know ploidy because it influences fertility, crossability, and even gene expression (Adams and Wendel 2005). Correspondingly, genome size provides insight into the species genetic diversity, evolution, and taxonomic relationships (Balao et al. 2009; Shearer and Ranney 2013). Furthermore, knowledge of genome size is essential for genome sequencing since it determines the amount of effort and resources required to achieve the desired coverage or quality.
2. Generate a draft sequence for *D. oleifera* genome. Genome sequence data not only improves our understanding of tree genome structure and evolution, it also allows for the identification of new genes and metabolic pathways (Ellegren 2014). For breeding purposes, whole-genome sequence data can be used as a reference for discovery of variants and marker development.
3. Identify molecular markers for genetic characterization and marker-informed breeding. High-throughput sequencing techniques can be used to identify thousands of single nucleotide polymorphisms. SNPs are ubiquitous, codominant, and can be in functional parts of the genome, thus making them the ideal marker for tree improvement and conservation genetics (Poland and Rife 2012; Narum et al. 2013). Using molecular

markers could improve the accuracy of selection model predictions while reducing the breeding cycle time.

These steps represent the core objectives of this dissertation and will be addressed in the following chapters.

CHAPTER 2: Genome and Seed Size Evolution in the Dipterygeae Clade, Fabaceae-Papillionoideae

Abstract:

Genome size variation is a common phenomenon among all organisms. In plants, many phenotypic traits show a correlation with genome size. One such trait is seed size, which in turn influences many aspects of plant ecology. The objective of this chapter was to determine genome size in the Dipterygeae clade and to investigate the evolution of this trait along with seed size. Although a small sample set, the results from this study show a moderate correlation between seed size and 2C-values, as well as a similar evolutionary history. It is hypothesized that one or more polyploidization events may account for the variation in the traits. Other mechanisms involved are not clear.

Introduction

Genome size variation is a common phenomenon among all organisms. Genome size is usually reported in terms of C-values. The 2C-value refers to the total amount or DNA content of a cell nucleus expressed in picograms (pg). Alternatively, if ploidy level is known, genome size can be reported as 1C-values representing the unreplicated gametic chromosome set. In flowering plants, both genome size and chromosome number (ploidy) can vary greatly. For example, the difference between large bitter-cress, the smallest reported C-value plant (*Cardamine amara,* 1C = 0.05 pg), and fritillary, the largest reported C-value plant (*Fritillaria assyriaca,* 1C = 127.4 pg), is over 2500-fold (Soltis et al. 2003). The main known mechanisms

for genome size variation are polyploidization, transposable elements amplification, and different patterns of insertion and deletions (indels). However, the cause or driving forces behind these are less known. Theories that explain genome size variation can be framed in terms of maladaptive, neutral, or adaptive evolutionary models (Lynch and Conery 2003; Whitney et al. 2010).

Adaptive evolutionary models state that the accumulation of nuclear DNA may serve a purpose based on the amount of extra DNA, not only its informational content. Many different phenotypic traits, such as duration of mitosis and meiosis, minimum generation time, and response of annual plants to CO_2 are correlated with genome size (Petrov 2001). Another important trait correlated with relative genome size is seed size. Seed size, or mass, influences many aspects of plant ecology. In general, small seeds are produced at a lesser cost to the individual allowing for large numbers, while larger seeds are costly to produce but improve seedling establishment and survival rates in harsh or shifting environments. Seed mass correlates well with other traits such as dispersal syndrome, plant size and form, plant life-span, and the ability to form a persistent seed bank (Moles et al. 2005). Beaulieu *et al.* (2007) found that genome size could explain up to 6.2% of the variation in seed mass among 1,222 different plant taxa, making it the second most important factor for seed mass evolution. The relation is stronger with intraspecific variation. For example, the relative nuclear DNA content of soybean (*Glycine max* (L.) Merr.) is strongly correlated ($r = 0.97$) with seed size in twelve different cultivars (Chung et al. 1998).

The Dipterygeae clade is monophyletic and commonly placed as one of the earlier branching groups within the papilionoid legumes (Fabaceae - Papilionoideae). The group is comprised of four genera, following Cardoso *et al.* (2012) nomenclature: *Dipteryx* Schreb., *Monopteryx* Spruce ex Benth., *Pterodon* Vogel, and *Taralea* Aubl. This is an exclusively

Neotropical clade, with a distribution ranging from Nicaragua to Brazil (Fig. 2.1). The clade is comprised of 25 woody species, mostly trees but also shrubs. Characteristic features are a two-lipped calyx, monadelphous androecium, and the typical papilionate corolla differentiated into standard, keel, and wing petals (except in *Monopteryx*). *Dipteryx* is the most diverse genus within the clade, with twelve recognized species (Cardoso et al. 2012; Cardoso et al. 2013). Additionally, several species within this genus are important for their ecological and commercial value. For example, *D. oleifera* Benth. possesses high density wood, a desirable trait in the timber industry, and produces nutritious seeds which constitute a significant part of the diet of many animals in the forest (Bonaccorso et al. 1980). Other species within the genus also produce edible seeds with commercial value like the famously fragrant Tonka beans, mostly from *D. odorata* (Aubl.) Willd. *D. rosea* Spruce ex Benth. and *D. punctata* (S.F. Blake) Amshoff seeds possess similar traits (Ducke 1940).

Here I describe determination of *D. oleifera* DNA nuclear content and ploidy level, followed by comparison of genome size among species in the Dipterygeae clade and exploration of the relationship between genome size, seed size, and fruit size. It is expected that the ancestor of the Dipterygeae clade had a smaller genome than current taxa since gaining nuclear DNA content can be achieved more readily by the current known mechanisms (*e.g.*, polyploidization events) than reducing genome size. Correspondingly, if DNA nuclear content influences seed size it is expected that species with larger genomes also have larger seeds.

Materials and Methods

Flow cytometry

I determined relative DNA nuclear content for ten of the fifteen Dipterygeae species using flow cytometry. For a detailed review on genome size estimation through flow cytometry please see Doležel and Bartoš (2005). I collected silica-dried tissue samples from *D. oleifera* seedlings grown in a greenhouse in Costa Rica. Dr. Domingos Cardoso from the Herbarium of Universidade Estadual de Feira de Santana (HUEFS) in Bahia, Brazil, provided herbarium specimens for the other species. I followed the protocol for preparation of aqueous cell suspensions described by Shearer and Ranney (2013). The cell nuclei were stained using 4',6-diamidino-2-phenylindole (DAPI) staining buffer Sysmex CyStain® UV Precise. The samples were tested on different days, with two technical replicates each time, to avoid any experimental or equipment bias. I used fresh tissue from *Pisum sativum* var. Ctirad (2C = 8.75pg) as an internal control. The samples were processed and analyzed on a Partec PA II flow cytometer using the manufacturer's proprietary software. The resulting data represent the average 2C values (pg) per species, calculated from different day measurement and technical replicates for all accessions tested. The number of accessions tested per species ranged from one to six, with two accessions per species on average. I conducted the sample preparation and flow cytometry analysis at the Mountain Horticultural Crops Research & Extension Center at Mills River, NC. The 2C value for outgroup species *Bauhinia tomentosa* (2C = 1.26 pg) was retrieved from the Kew Royal Botanical Gardens - Plant DNA C-values database (release 6.0).

Chromosome counts

I collected fresh root tips from *D. oleifera* seedlings. The seedlings were grown in nursery beds and containers in Costa Rica. Root tips were cleaned of dirt and debris and then fixed, first using a 2mM 8-hydroxyquinoline and 0.24mM cycloheximide solution and then a three-parts 95% ethanol: one-part glacial acetic acid solution. Root tips were then washed and stored in 70% ethanol for later use (Shearer and Ranney 2013). I conducted the staining and visualization using a modified carbol fuchsin dye and a light microscope at the Mountain Horticultural Crops Research & Extension Center at Mills River, under Dr. Thomas Ranney's supervision.

Fruit and seed size data

The Fabaceae family is distinctive for its legume fruit. In Dipterygeae, the fruit is a single-seeded pod. I measured fruit length (FL) and fruit width (FW), from all available digital herbarium specimens (N = 183) at the Tropicos database (Missouri Botanical Garden), C.V. Starr Virtual Herbarium (New York Botanical Garden), and the REFLORA Virtual Herbarium (Rio de Janeiro Botanical Garden). Only mature fruit were considered. Similarly, I measured seed length (SL) and seed width (SW), from all digital specimens that presented an open fruit and exposed mature seeds (N = 75). I used image processing and analysis software ImageJ version 1.52a (Schneider et al. 2012) for measurements, calibrated with the scale included in each herbarium specimen. The fruit and seed size data were complemented with taxonomic descriptions available in the literature. Fruit and seed size values presented are the average of all measurements per species. Pearson's correlation coefficients were calculated among traits. I used a linear regression to predict missing seed size values. Linear regression and correlation

estimation were performed using statistical analysis software R version 3.5.0 (R Core Team 2018).

MatK gene sequence data

Previous studies have demonstrated that DNA barcode sequences from plastid matK protein-coding genes provide adequate resolution at many taxonomic levels in the legume family (Wojciechowski et al. 2004). Complete and partial matK sequences were retrieved from the GenBank, fifteen species within the Dipterygeae clade and one species from the Cercidoideae clade. The Cercidoideae clade is considered one of the earlier branching and basal groups in Fabaceae. I selected *Bauhinia tomentosa*, currently classified within the Cercidoideae clade, as the outgroup for phylogeny inference and reconstruction. Taxa and GenBank accession numbers are presented in Table 2.1. Dipterygeae sequence data, *i.e.* the matK accessions used in this analysis, correspond for the most part to those first presented by Cardoso et al. (2012; 2013; 2015).

Sequence alignment and phylogenetic analysis

I compiled the DNA barcode sequences into a single fasta format file and used the MUSCLE algorithm, with default parameters, as implemented in the freely available MEGA7 software package (Kumar et al. 2016) to perform multiple sequences alignment. Data manipulation, *i.e.*, sequence adjustments or edge trimming, was also done with MEGA7. The resulting alignment was used for phylogeny inference. Tree construction was performed in MEGA7 using two different methods:

(i) Maximum Parsimony (MP), using the Subtree-Pruning-Regrafting (SRG) algorithm. The consensus tree from the bootstrap analysis is presented.

(ii) Maximum Likelihood (ML) based on the Tamura-Nei model. The initial tree for the heuristic search was obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The consensus tree from the bootstrap analysis is presented.

In addition, I conducted phylogeny inference using the Randomized Accelerated Maximum Likelihood (RAxML) method, as implemented on CIPRES (RAxML-HPC2 on XSEDE), with default parameters. The best scoring ML tree with bootstrap support values is presented. For all methods, positions containing gaps and missing data were eliminated. There were a total 16 taxa and 1158 characters per taxa in the final dataset used for tree inference. *B. tomentosa* was set as the outgroup and used to root the trees.

Finally, I used Mesquite version 3.40 (Maddison and Maddison 2018) to trace seed and genome size evolutionary history. The phylogenetic tree was modified manually to resemble the ML phylogeny reconstructed. Genome and seed size were analyzed as continuous characters.

Results and Discussion

Genome size

Relative DNA nuclear content, expressed as 2C-values, ranged from 1.26 pg to 3.86 pg, with an average of 2.5 pg for the taxa in this study. The selected outgroup species, *B. tomentosa,* has the smallest genome size. Genome size variation within the Dipterygeae clade was small,

with *D. magnifica* (2C-value of 1.96 pg) and *D. oleifera* (2C-value of 3.86 pg) in opposite positions of the range. Base ploidy reported for the Dipterygeae tribe is n =8, that is 2n=16 chromosomes for diploid species and 4n=32 chromosomes for tetraploid species. While there are no specific reports on the ploidy level of *D. magnifica*, based on genome size it could be inferred to be a diploid. Chromosome counts from *D. oleifera* root tips were inconclusive (Fig. 2.2). However, based on molecular marker data, *D. oleifera* was reported as a potential tetraploid (Hanson et al. 2008a). The same is true for *D. odorata* (Vinson et al. 2009). A different ploidy level, in these cases tetraploid, could account for them having almost twice as much DNA content than *D. magnifica*. Taxa and 2C-values are summarized in Table 2.1.

I estimated relative DNA nuclear content from silica-dried tissues and herbarium samples. However, the age of some of the herbarium specimens and degradation level of the tissue made it impossible to isolate intact nuclei, properly stain them, or get a distinct signal in the flow cytometer. Best practices in flow cytometry dictate that fresh tissue should be used for genome size estimation; this represents a major constraint in the study of many taxa (*e.g.*, species that can only be found in tropical forest miles away from a research facility and are not easily grown in greenhouse conditions are difficult to study under laboratory conditions). Luckily, rapid desiccation of plant tissue using silica gel is an effective way to preserve tissue samples for flow cytometry analysis. Previous studies have demonstrated that rapid drying with silica gel introduces minor error (<10%), comparable to other sources of variation like instrument or staining protocol (Bainard et al. 2011). Traditionally, herbarium specimens are not subject to rapid desiccation but a rather slow pressing and drying process that may include a heat treatment, *i.e.*, warm air blowing from an electric fan. This results in degradation of the sample as evidenced by browning of the specimen. Tissue browning is commonly caused by polyphenols,

which contribute to the degradation of DNA as oxidizing agents. Compounds like flavonoids, terpenoids, and tannins occur widely in plants; such compounds can be released from the cell vacuole shortly after collection in older herbaria dried samples (Varma et al. 2007).

Fruit and seed size

I obtained fruit size data for all taxa (*N* = 16 species), but seed sizes for only a subset (*N* = 10 species). Pearson's correlation coefficients for pairwise comparisons among traits are summarized in Table 2.2. Both SL and SW showed moderate correlation with relative DNA nuclear content (2C) values (r = 0.83 and 0.70, respectively). These results support previous reports in the literature (Chung et al. 1998; Beaulieu et al. 2007). While most Dipterygeae species have one-seeded pods, *Monopteryx* spp is the exception with multiple seeds per pod. The same is true for the outgroup species *B. tomentosa*. This difference in fruit-seed ratio could account for the weaker correlations between fruit dimensions and 2C values. Since FW showed a stronger correlation with SL and SW I used a linear regression and FW values to extrapolate SL and SW missing data (Fig. 2.3). A summary of the average sizes per character is presented in Table 2.1.

Dipterygeae phylogeny

The phylogeny reconstruction using both Maximum Parsimony (MP) and Maximum Likelihood (ML) methods confirms the monophyly of the tribe, with species within each genus clustering together and supported by the bootstrap values (Fig. 2.4 to 2.6). However, relationships within the *Dipteryx* genus are not well resolved. For the MP tree (Fig. 2.4), the consistency index is 0.935, the retention index is 0.971, and the composite index is 0.908 for all

parsimony-informative sites. For the ML trees, the highest log likelihood are -2600.59 (Fig. 2.5) and -3395.81 (Fig. 2.6). There are no relevant differences in tree topology between the ML trees constructed using different software, *i.e.,* MEGA7 or CIPRES. Furthermore, tree topology strongly resembles that of previous published studies in the group that used similar Maximum Parsimony methods or sophisticated Bayesian inferences models (Cardoso et al. 2012; Cardoso et al. 2013; Cardoso et al. 2015). This should come as no surprise since the matK sequence data used in this study corresponds, for the most part, to the same accessions used by Cardoso *et al.*, only the analysis method differs. Since Maximum Likelihood represents a more robust method for phylogeny inference, only those results were used in subsequent analysis and discussion.

The genus *Dipteryx* and *Pterodon* were shown to be sister groups in the ML inferred phylogeny (Fig. 2.5, 26). The close relation between these two genera is also evident in their morphological features *Pterodon* only differs from *Dipteryx* in the more petaloid nature of the calycine lobes, flattened fruit, and in foliage (Hooker 1850). These minor differences led to the wrong identification of some *Pterodon* species. For example, *P. emarginatus* was previously placed in the genus *Dipteryx* or *Coumarouna* (*Dipteryx* basionym) based on vegetative features and geographical distribution. Molecular data from recent studies clearly separates the two genera. The genus *Taralea* was shown to be a sister to the *Dipteryx* and *Pterodon* clade. Some *Taralea* species are shrubs or have climbing habits which differ from the mostly large trees found in *Dipteryx* or *Pterodon*. Finally, *Monopteryx* represents the most basal and early diverging lineage within the tribe. In the past, this genus was placed within the Sophoreae clade (Pennington et al. 2001). *Monopteryx* differs from the previous genera in its non-papilionate corolla with wing petals reduced and keel petals connate and open out exposing the free stamens.

Despite the morphological differences, phylogenetic analysis of plastid matK and trnL intron sequences place *Monopteryx* as a sister group to the rest of Dipterygeae (Cardoso et al. 2015).

Genome and seed size evolution

Results from the evolutionary history trace analysis for genome and seed size within Dipterygeae support the initial hypothesis. The parsimony reconstruction indicates that the ancestor to Dipterygeae had a smaller genome size than current taxa (Fig. 2.7). This trend in DNA nuclear content increase was only reverted in *D. magnifica*. For seed size, the evolutionary trend was similar, that is, seed size tends to increase from the previous ancestor with few exceptions (Fig 2.8).

The increase in relative DNA nuclear content in this group may be due to polyploidization events. Polyploidization is the process of duplication of the whole or partial genome. Polyploid individuals possess more than two sets of chromosomes. This condition is heritable and can confer advantages in terms of evolutionary flexibility. For example, genes that are duplicated by a polyploidization events may retain their original function, undergo diversification in protein function/regulation, or get silenced through mutational or epigenetic mechanisms. Both duplicate gene expression and redundancy can influence fitness, thus creating a selective pressure that would favor polyploidization (Wendel 2000). Furthermore, polyploidy is an important force in angiosperm evolution. It is now believed that all angiosperm lineages have gone through at least one polyploidization episode; whole or partial genome duplication is both frequent and ubiquitous in angiosperms history. Polyploidy in angiosperms contributed greatly to the group diversification (Soltis et al. 2009). As stated before, cytological studies have estimated the base chromosome number for the Dipterygeae group as n=8, with diploids members having

2n=16, like *Pterodon* spp (Bandel 1974), while there are some tetraploid species in the *Dipteryx* genus.

Another source of nuclear DNA content variation, usually associated with genome size increase, is accumulation of repetitive DNA. There are diverse types of repetitive DNAs, but most notable are the transposable elements. In addition to increasing genome size following transposition, transposable elements can induce chromosomal rearrangements resulting in deletions, duplications, inversions, and reciprocal translocations. In plants, transposable elements activity can be controlled through epigenetic silencing by siRNA that initiates methylation of the transposable elements and limits transposition. Therefore, it is possible that genome size variation in plants can be the result of differential efficiency in transposable element silencing. For example, half the genome of *Arabis alpina* (genome size = 375 Mbp) is comprised of transposable elements. The difference in genome size with close relative *Arabidopsis thaliana* (genome size = 135 Mbp) was linked to a reduced capacity for silencing and removal of long terminal repeat retrotransposons (LTR-RTs) (Ågren and Wright 2015).

Conversely, there are mechanisms that could explain the reduction in DNA nuclear content. The indel bias refers to difference in patterns of insertion and deletions (indels). In plants, deletions are far more common than insertions in both protein-coding sequences and non-coding regions. However, DNA loss by indel bias is a slow process and it is questionable how relevant it can be to genome size variation (Gregory 2004). More plausible mechanisms that decrease DNA nuclear content are unequal homologous recombination and illegitimate recombination. Unequal homologous recombination between chromatids yields reciprocal duplication/deletions that do not provide any net change in DNA content. Nevertheless, when unequal homologous recombination occurs within a single chromatid it preferentially leads to

deletions (Bennetzen et al. 2005). In the case of illegitimate recombination, the mechanism of action is not fully understood; it could be by an error in DNA replication or by double-strand break repair. What has been established is the effect it can have on genome size variation. In *Arabidopsis*, for example, illegitimate recombination removes at least fivefold more DNA than unequal homologous recombination (Devos et al. 2002).

In the case of Dipterygeae, the most likely explanation for the variation in DNA nuclear content is one or more polyploidization events in the lineage, after those events, the different groups underwent a process of mutation, purging selection, and replication of transposable elements that would account for the small variation. Two notable exceptions are *D. magnifica* and *D. oleifera*. Based on data I present here, I hypothesize that *D. magnifica* is a diploid species, and may have undergone genome size reduction relative to other diploid species in the clade. However, the mechanisms involved are not clear. Conversely, *D. oleifera* showed the largest 2C-value and is likely to be a tetraploid. This may indicate that *D. oleifera* is prone to DNA accumulation or that it is the product of recent polyploidization, either whole genome duplication or a hybridization event.

Seed size was moderately correlated with 2C-values. The correlation is stronger than reported for distantly related taxa (Beaulieu et al. 2007), but weaker than at the intraspecific level (Chung et al. 1998). This result is relevant because it indicates that DNA nuclear content may indeed affect seed size in closely related taxa. In Dipterygeae, seed size increase seems to be the product of multiple independent events in the evolutionary history of the clade, but it does coincide with larger DNA nuclear content. However, it is not clear what drives this relation. If polyploidization is the main cause for increase in DNA nuclear content, with the current data it is impossible to distinguish whether increase seed size is a function of the additional DNA content

or a matter of gene dosage. According to the nucleoskeletal theory, the amount of DNA determines the nucleus and cell size. Theoretically, non-coding DNA has a structural role in the nucleus as a nucleoskeleton, in a way that the amount of DNA can determine the nucleus size which in turn influences the cell size. So, if genome size determines cell size, it could also influence seed size. Besides, a selective pressure on species that have large cells may explain the expansion in DNA content (Gregory 2001; Cavalier-Smith 2005). Future cytological studies should focus on whether the correlation I found between genome size and seed size is influenced by the size of the cells and cell's nucleus. This would help prove or disprove the nucleoskeletal theory. On the other hand, polyploid plants are known to be larger than their diploid counterparts, due to a gene dosage effect. The additional gene copies encoded in the homeologous chromosomes result in duplication of gene products, *i.e.*, double the genes may equal double the transcripts encoded, proteins, and byproducts. This could also translate into additional resources stored in the seed, making for bigger seeds.

Conclusion

DNA nuclear content increase is the evolutionary trend in the Dipterygeae clade, with just a few exceptions. One or more complete or partial polyploidization events are believed to drive this size increase. Seed size was moderately correlated with 2C-values, supported by the evolutionary history of both traits. However, this was just an exploratory study and is limited by a small dataset. Further study and a greater number of replicates are needed to fully resolve the evolution of these traits.

Table 2. 1. Average genome, fruit, and seed size for Dipterygeae taxa and outgroup.

Taxa	2C (pg)	Fruit Length (cm)	Fruit Width (cm)	Seed Length (cm)	Seed Width (cm)	GenBank Accession
Bauhinia tomentosa	1.26	10.50	1.75	0.78	0.63	AY386893
Dipteryx magnifica	1.96	4.03	3.25	2.62	1.69	JX295871
Monopteryx uaucu	2.10	10.94	3.12	2.50	1.61	KP177915
Taralea cordata	2.20	3.06	1.78	1.23	0.88	JX295872
Monopteryx inpae	2.40	9.19	2.38	1.57	0.92	JX295876
Dipteryx rosea	2.58	5.26	2.90	2.30	1.46	JF491268
Pterodon abruptus	2.60	4.51	2.59	2.11	1.22	JX295873
Taralea oppositifolia	2.82	4.07	3.07	2.11	1.62	JF491275
Dipteryx odorata	3.26	4.48	2.78	3.50	1.41	JF491266
Dipteryx oleifera	3.86	7.00	4.50	5.25	3.25	JX295933
Dipteryx alata	-	4.96	3.76	3.07	2.03	JF491265
Dipteryx polyphylla	-	4.73	2.85	2.26	1.43	JX295870
Dipteryx punctata	-	5.23	3.17	2.54	1.64	JF491267
Pterodon emarginatus	-	5.11	3.07	2.27	1.50	JF491272
Pterodon pubescens	-	5.64	3.21	3.47	2.31	JF491273
Taralea rigida	-	3.48	2.01	1.36	0.95	JX295934

Note: *2C* = average nuclear DNA content, *GenBank accession* = accession number for matK gene sequence Table is sorted by increasing genome size (2C). Values in red represent extrapolated data.

Table 2. 2. Pearson's correlation coefficient among genome, fruit, and seed size for Dipterygeae taxa and outgroup.

	2C	Fruit Length	Fruit Width	Seed Length	Seed Width
2C	1.00				
Fruit Length	-0.25*	1.00			
Fruit Width	0.27**	0.29***	1.00		
Seed Length	0.83***	0.03	0.72***	1.00	
Seed Width	0.70***	0.02	0.83***	0.83***	1.00

Note: *2C* = average nuclear DNA content. * = p-value < 0.05, ** = p-value < 0.01, *** = p-value < 0.001

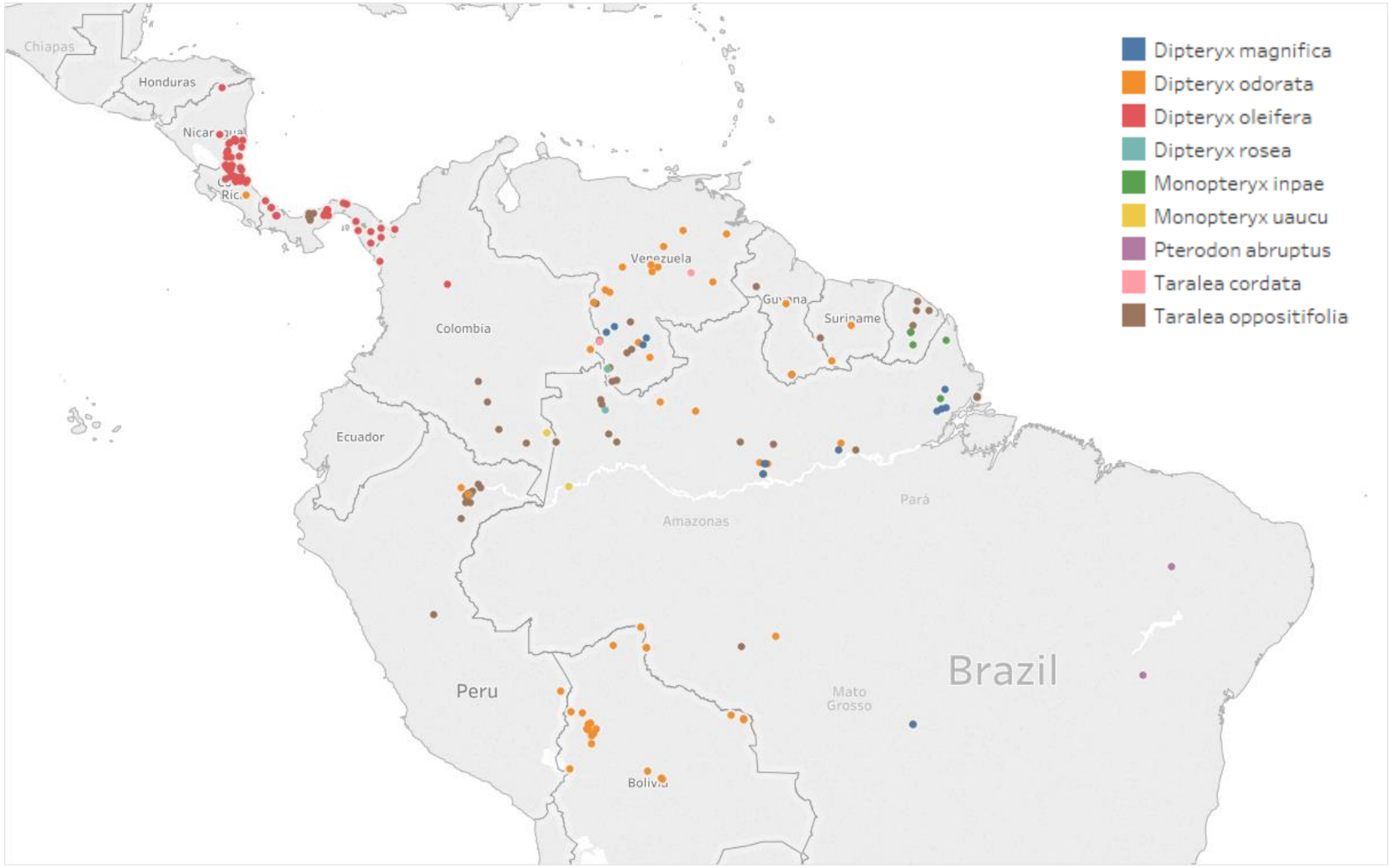

Figure 2. 1. Dipterygeae range distribution based on herbaria records.

Figure 2. 2. Chromosome count from *D. oleifera* root tip. Cells nuclei were stained with a solution of modified carbol fuchsin.

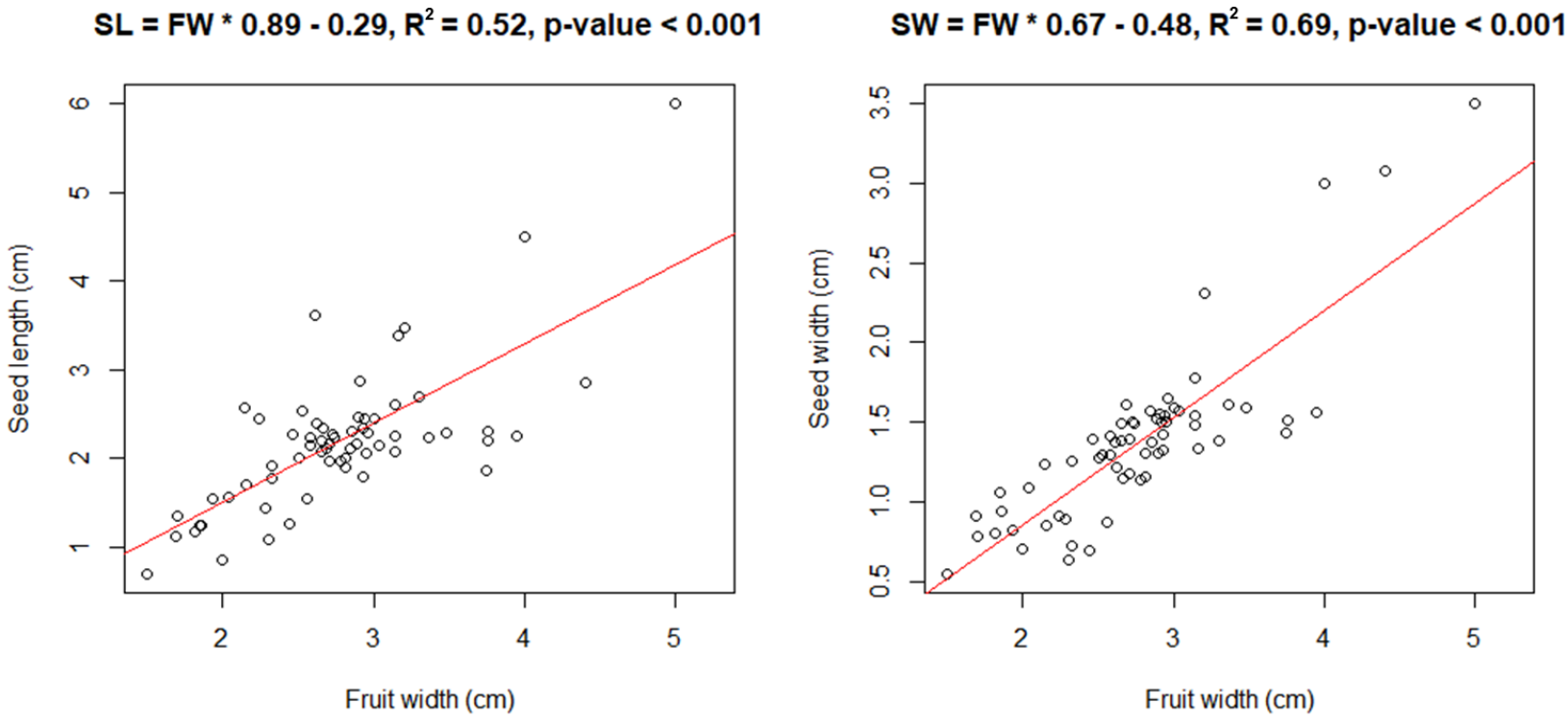

Figure 2. 3. Linear regression for fruit width (FW) and seed length (SL), and fruit width (FW) and seed width (SW). Red line represents linear fit.

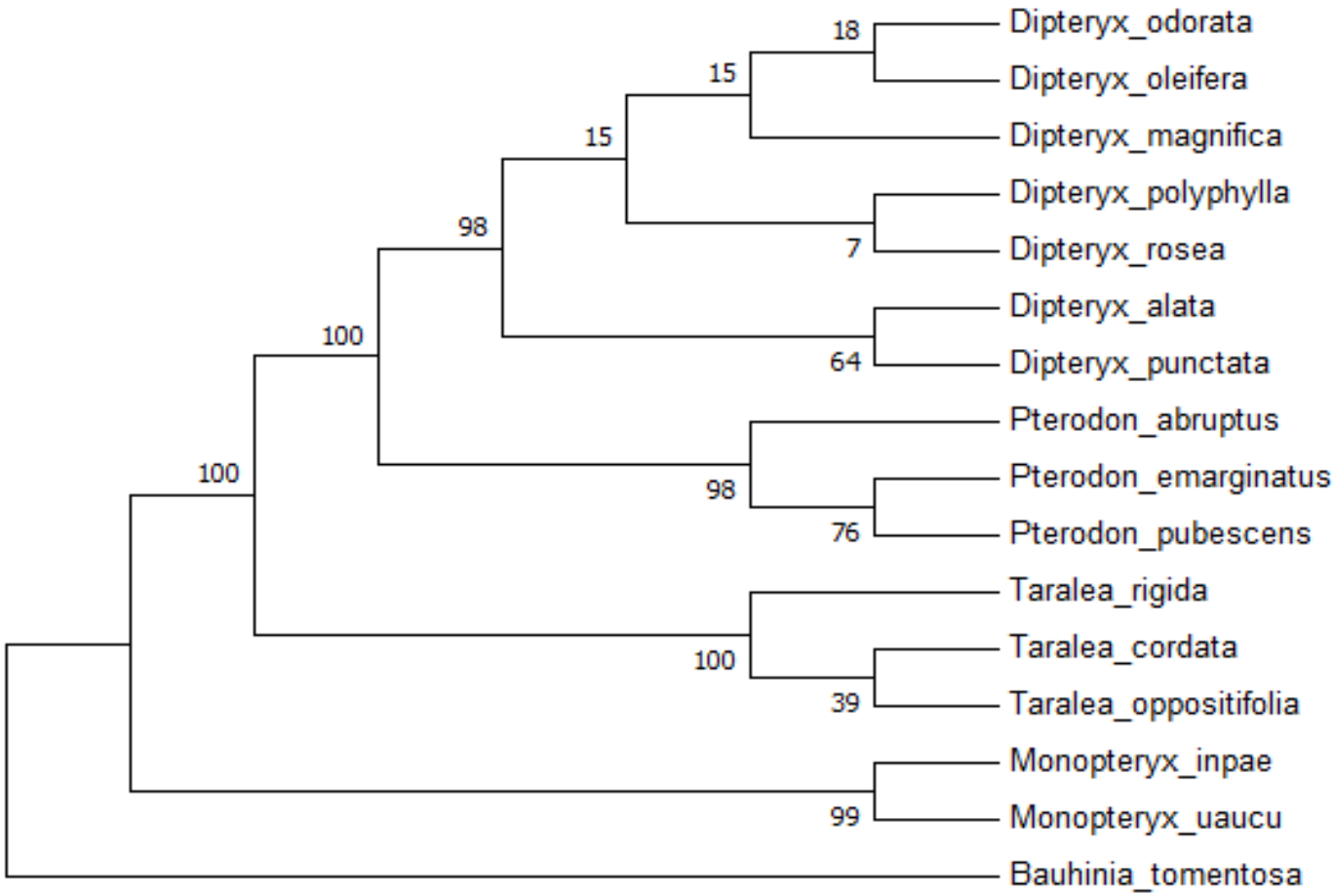

Figure 2. 4. Dipterygeae tribe phylogeny inferred using the Maximum Parsimony (MP) method, as implemented in Mega7 software package. The percentage of trees in which the associated taxa clustered together is shown next to the branches (bootstrap values). Tree rooted with *Bauhinia tomentosa* as an outgroup.

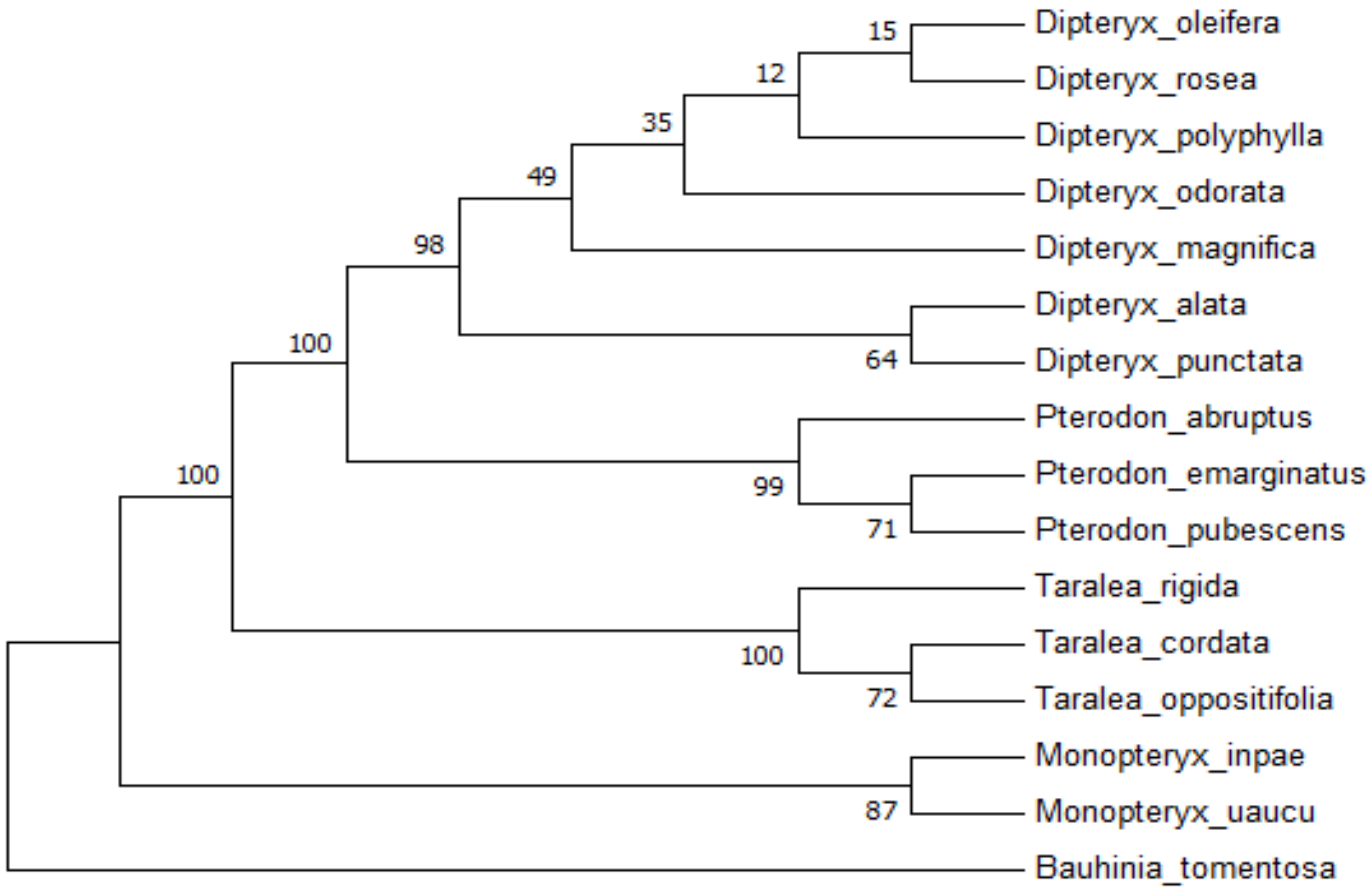

Figure 2. 5. Dipterygeae tribe phylogeny inferred using the Maximum Likelihood (ML) method based on the Tamura-Nei model, as implemented in Mega7 software package. The percentage of trees in which the associated taxa clustered together is shown next to the branches (bootstrap values). Tree rooted with *Bauhinia tomentosa* as an outgroup.

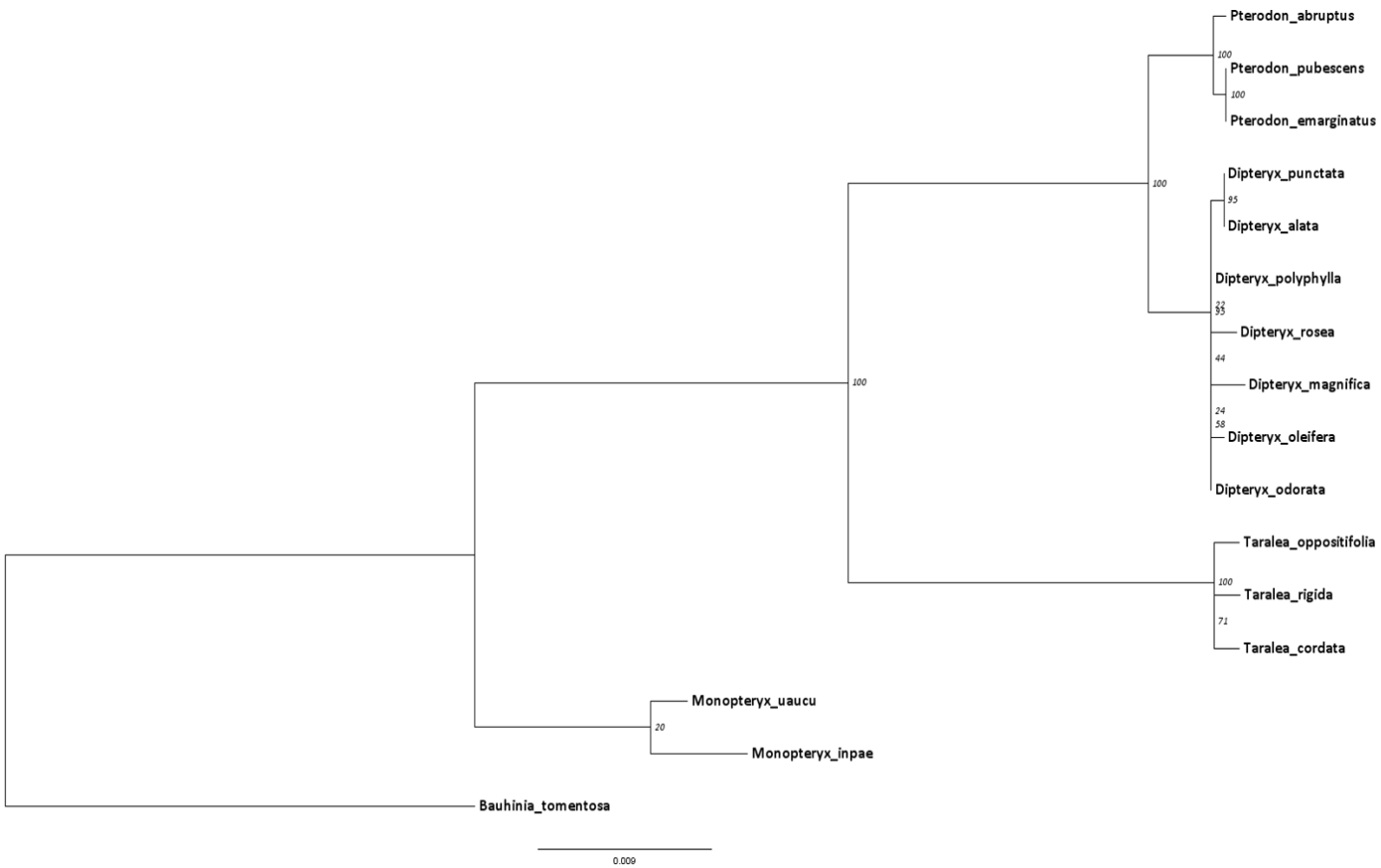

Figure 2. 6. Dipterygeae tribe phylogeny inferred using the Randomized Accelerated Maximum Likelihood (RAxML) method, as implemented on CIPRES (RAxML-HPC2 on XSEDE). Best scoring ML tree with bootstrap support values drawn as node labels. Tree rooted with *Bauhinia tomentosa* as an outgroup.

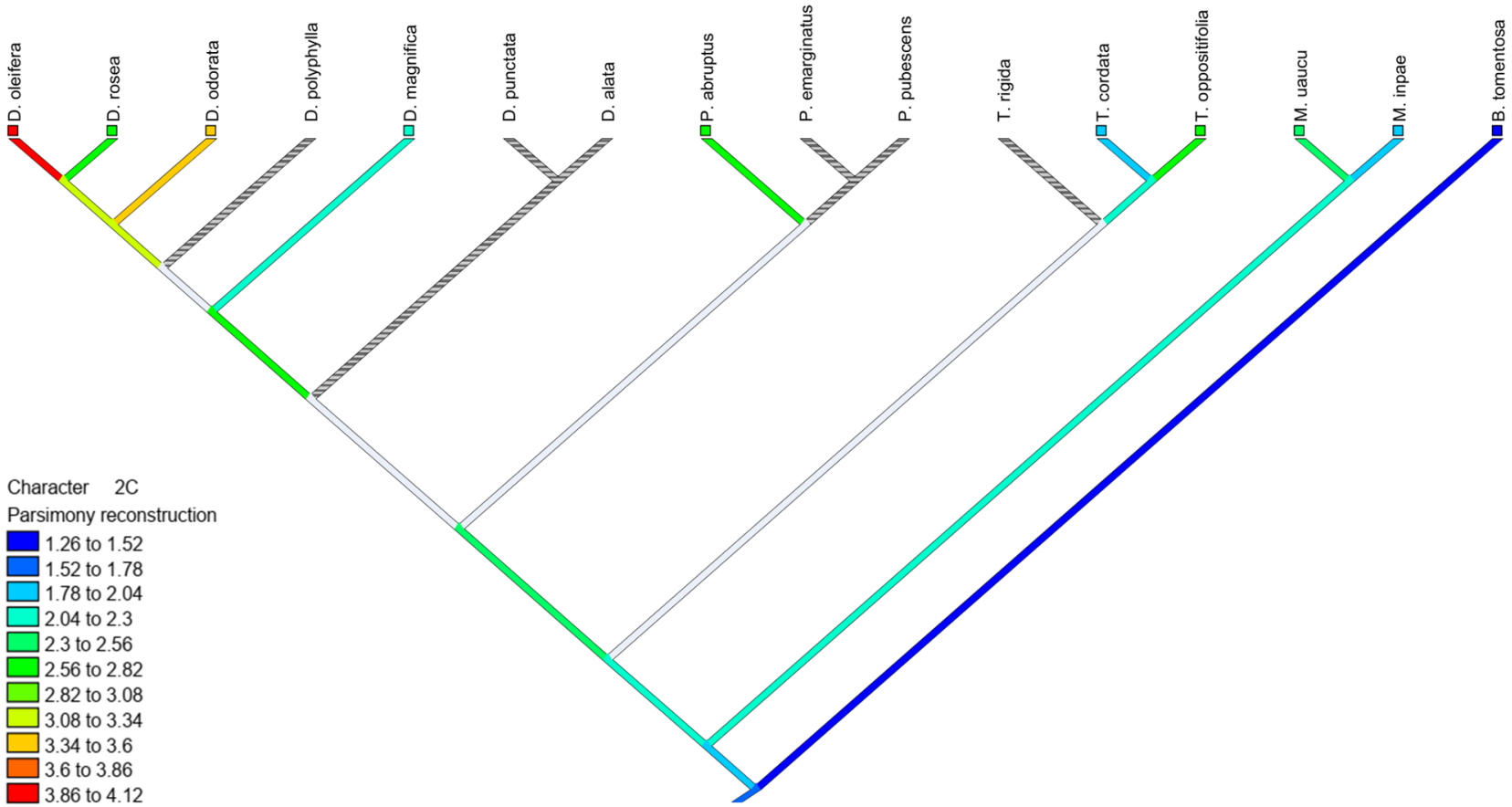

Figure 2. 7. Relative DNA nuclear content (2C-value) evolutionary history reconstruction using a Parsimony method. Character was treated as a continuous variable. Genome size ranged from 1.26 to 3.86 pg. Warmer colors (red) represent larger genome size, grey color represents missing data. Tree rooted with *Bauhinia tomentosa* as an outgroup.

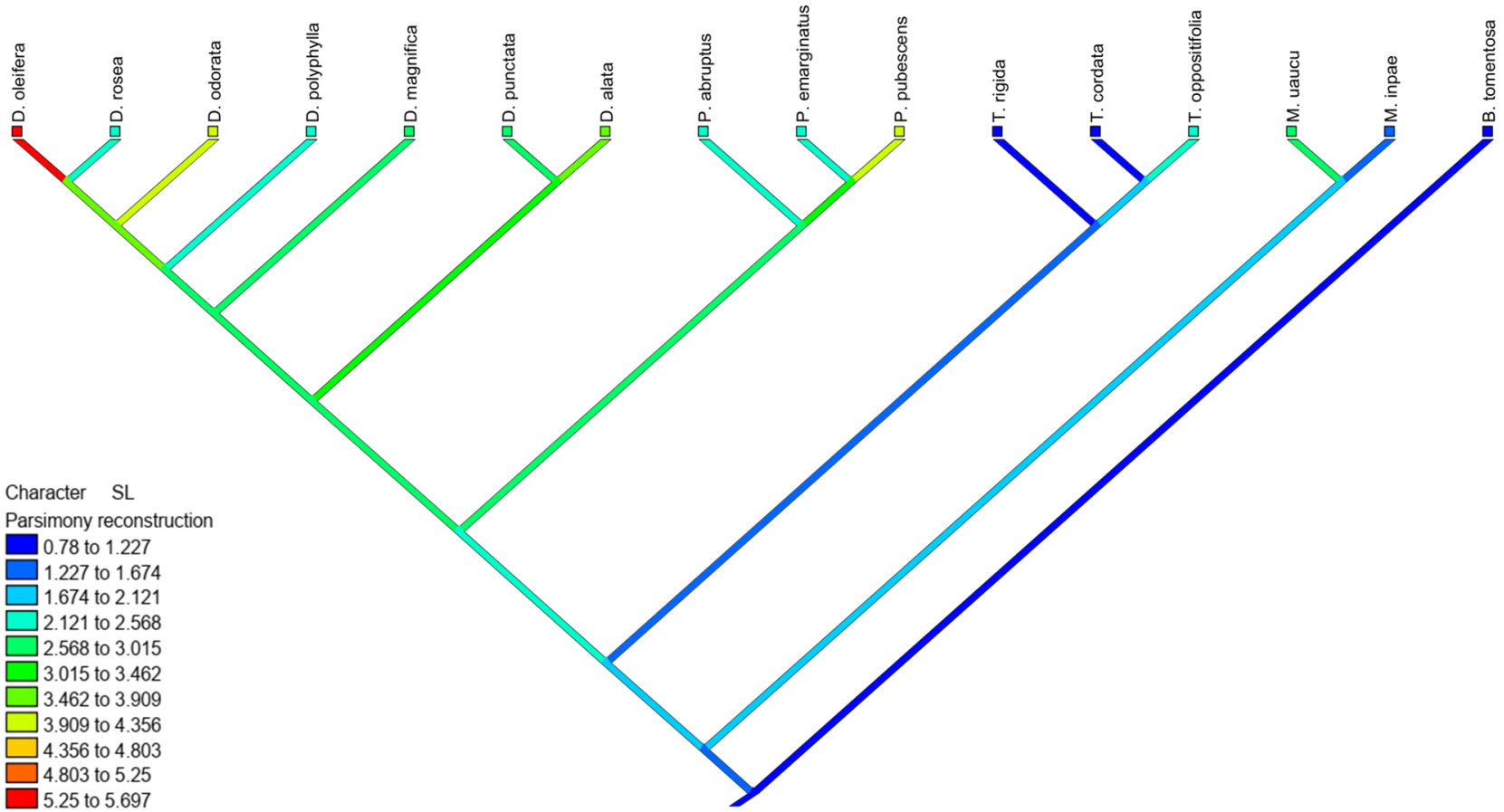

Figure 2. 8. Seed size evolutionary history reconstruction using a Parsimony method. Character was treated as a continuous variable. Seed length ranged from 0.78 to 5.25 cm. Warmer colors (red) represent larger seed length. Tree rooted with *Bauhinia tomentosa* as an outgroup.

CHAPTER 3: First Draft Genome Sequence of Tropical Tree *Dipteryx oleifera* Benth.

Abstract:

In this chapter I describe the sequencing and assembly of the first draft genome of *Dipteryx oleifera*, an ecologically and economically important tropical timber species. For the assembly, I used a combination of short Illumina reads and additional coverage in long PacBio reads. With a total of 1,166,468,433 bp in 381,857 contigs, this assembly corresponds to 62% of the estimated genome size. Although still fragmented, the resulting assembly contains 70.7% of expected component of complete single-copy and duplicated conserved orthologous genes. In addition, I used short Illumina reads data and the resulting assembly to corroborate genome size and ploidy level for this species.

Introduction

Dipteryx oleifera (Fabaceae) is a tropical tree species endemic to the Caribbean lowlands, ranging from Nicaragua to Colombia (Flores 1992). It is a keystone species, providing food and shelter to many mammals (Bonaccorso et al. 1980) and birds, including the endangered great green macaw (Chassot and Arias 2012; Monge et al. 2012). In addition to its ecological value, *D. oleifera* possesses high quality wood. However, illegal logging and habitat fragmentation have diminished natural populations to the point that trade of the wood is now controlled by international treaty (CITES Appendix III) (Hanson et al. 2008b). GENFORES, a forestry industry – university co-operative program based in Costa Rica, is working in the development of genomic resources to aid tree improvement and conservation of this species.

In general terms, a genome can be defined as all the information (*i.e.*, genes and regulatory regions) needed to build and maintain a cell or an organism, usually encompassed in the haploid chromosome set. Although this definition is an oversimplification (Goldman and Landweber 2016), it is still widely used. In the era of high-throughput DNA sequencing technologies, a genome has become synonymous with the actual DNA sequence coding all that information. Genome assemblies are composed of contigs and scaffolds. Contigs are contiguous consensus sequences derived from collections of overlapping reads. The maximum length of a contig is usually determined by the amount of repetitive sequences flanking it. If the length of the repetitive sequences exceeds that of the average read length, most assemblers would not be able to extend the contig sequence any further. Scaffolds are ordered and oriented sets of contigs that are linked to one another leaving gaps, filled with N's, where the sequences are unknown. To assemble scaffolds, most assemblers require libraries with large insert size, such as mate-pairs or long reads, to help bridge the gaps created by repetitive sequences. Hence, the growing popularity of hybrid assembly approaches, that combine accurate Illumina short-read with PacBio long-read.

In this chapter, I describe the sequencing and assembly strategy used to generate the first draft sequence for *D. oleifera*. In addition, I used the sequence data and the resulting assembly to corroborate *D. oleifera* genome size and ploidy level.

Materials and Methods

DNA isolation

I collected plant material in Costa Rica in July 2015. The leaf tissue was silica dried and then stored at -70°C. I isolated high molecular weight genomic DNA using a Wizard® Genomic

DNA purification kit (Promega, US) following the manufacturer instructions and assessed DNA quality and quantity using a NanoDrop™ (Thermo Scientific ™, US). The DNA sample used comes from a single individual, a 5-year-old tree planted as part of a provenance-progeny trial managed by GENFORES (Martínez-Albán et al. 2016).

Library preparation and sequencing

The short-read library was prepared using Illumina's DNA TruSeq Nano Library preparation kit and following the manufacturer instructions. Final average library size, *i.e.*, after DNA fragmentation, end repair, size selection, and adapter ligation, was 550 bp. The library was normalized to 2 nM prior to loading onto the flow cell, then sequenced in the Illumina HiSeq2500 platform using SBS sequencing kit version 4. The 150 bp paired-ended (PE) single index library was sequenced for 308 cycles. In addition, a long-read library was prepared using PacBio's template preparation kit and following the manufacturer instructions. The library was size-selected for fragments ranging between 10 Kb to 50 Kb. The library was normalized to two different concentrations, 10 pM and 15 pM, and loaded onto two separate SMRT cells respectively. The PacBio run length was set to 600 minutes for both cells. Next-generation sequencing was performed by the NC State University Genomic Sciences Laboratory (Raleigh, NC, USA).

Draft sequence assembly

I performed quality assessment of the raw Illumina reads with FastQC version 0.10.1 (Andrews 2010). Quality control identified a small percentage (< 1%) of TruSeq Adapter, Index 2 contamination. I used BBMap/BBtools suite with BBduk version 36.27 package (Bushnell

2015) for quality trimming and adapter removal. The raw PacBio subreads were filtered for reads with sequences length ≥ 1000 bp and then converted from bam format to fastq format using BamTools (Barnett et al. 2011). I used both the clean Illumina reads and the filtered PacBio reads as input for sequence assembly.

I tested three different software packages for genome sequence assembly, two open source and freely available: SOAPdenovo2 version 2.04 (Luo et al. 2012) and ABySS 2.0 (Jackman et al. 2017); and one licensed and proprietary: CLC Genomics Workbench version 10.1.1 with the Finish Module version 1.7 plug-in (https://www.qiagenbioinformatics.com). Both ABySS 2.0 and CLC Genomics workbench incorporate short and long reads into the assembly, however, SOAPdenovo2 only works with short read data. Despite its limitations, SOAPdenovo2 was tested since it is generally regarded as a good benchmark assembler. For each software, I conducted a series of experiments to optimize parameter setting, for example testing different k-mer sizes. As a final step, I used Sealer (Paulino et al. 2015) to close gaps in the assemblies. Optimal parameter setting for each assembler are detailed in Appendix 1, Assembly scripts. The best resulting assemblies from each software are presented in the results.

Draft sequence evaluation

I used QUAST version 4.6.0 (Gurevich et al. 2013) to estimate summary statistics for each assembly. Quality statistics include: total length, number of contigs, largest contig, N50, N75, L50, L75, GC (%), and number of N's per 100 Kbp. The *Nx* statistics refers to the largest contig or scaffold length, *y*, such that using contigs/scaffolds of length ≥ *y* accounts for at least *x*% of the bases of the assembly. The *Lx* count refers to the smallest number of contigs or scaffolds whose length sum produces *Nx*. These are commonly regarded as measurements of the

assembly contiguity. Further properties of the assembly composition and quality were obtained using KAT (Mapleson et al. 2017). Finally, to assess the completeness of the assembly I ran a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis using the Embryophyta orthologs database as a reference (Simão et al. 2015). The embryophyta_odb9 is comprised of 1440 conserved orthologous genes commonly found in species across the plant kingdom.

Genome size and ploidy validation

For genome size estimation, I used the clean Illumina reads as input and Jellyfish version 2.2.7 (Marçais and Kingsford 2011) for k-mer counting with k-mer size ranging from 18 to 22. For each k-mer size, the depth distribution was counted, and the peak values identified. Since the Illumina short reads are randomly generated, the depth of coverage should follow a Poisson distribution. The genome size can then be calculated as:

$$Genome\ size = \frac{Total\ number\ of\ K-mer}{Peak\ value\ of\ k-mer\ frequency\ distribution}$$

The genome size presented in the results is the average from all k-mer sizes tested.

To determine ploidy, I aligned the clean short reads to the CLC Genomic Workbench assembly using BWA version 0.7.17-r1194-dirty (Li and Durbin 2009), marked and removed duplicates with SAMBLASTER version 0.1.24 (Faust and Hall 2014), and removed the unaligned reads using SAMtools version 1.3.1 (Li et al. 2009). The resulting bam file was used as input for nQuire (Weiß et al. 2018). nQuire estimates ploidy by assessing the distribution of allele frequencies at biallelic single nucleotide polymorphism (SNPs). It assumes that allele frequency distributions occur at different ratios for each ploidy level, *i.e.* 0.5/0.5 for diploids, 0.33/0.67 in triploids, and a mixture of 0.5/0.5 and 0.25/0.75 for tetraploids. I used nQuire to clean the data and estimate ploidy. The nQuire lrdmodel subcommand uses a Gaussian Mixture

Model (GMM) and maximum likelihood to assess empirical data under the assumptions of diploidy, triploidy, and tetraploidy. In addition, I used a linear regression test against the three fixed models (nQuire histotest subcommand).

Results and Discussion

Draft sequence assembly

Using flow cytometry (see chapter 2), I determined *Dipteryx oleifera* relative DNA nuclear content (2C) as 3.86 picograms (pg), or 1.93 pg for the 1C = 2n genome. Genome size can also be expressed in terms of base pairs (bp), where 1 pg DNA equals 0.978×10^9 bp, hence *D. oleifera* 1C =2n genome size can be expressed as 1,887,540,000 bp or 1.89 Gb. The Illumina HiSeq run generated 275 million 150 bp paired-ended reads; based on the previously estimated genome size, this represents a coverage of approximately 44x. The PacBio Sequel run generated 5.82 Gb and 3.71 Gb respectively for each SMRT cell; combined this represents a coverage of approximately 5x. After quality trimming and adapter removal, 98% of the total reads were retained. These were used as input for the assembly. All the software produced an assembly, although of varying quality. Table 3.1 presents QUAST summary statistics for each assembly.

Assembly quality is evaluated in terms of contiguity, completeness, and correctness. Contiguity refers to the length of the sequences assembled. *Nx* statistics are commonly used to evaluate contiguity, in particular N50 which represents the smallest scaffold or contig length above which 50% of the assembly would be represented. Completeness can be interpreted in two senses: genome coverage and gene coverage. Genome coverage is the percentage of the genome that is contained in the assembly and can be estimated by comparing the assembly total length to the estimated genome size calculated from a different source, like flow cytometry. Gene

coverage is the percentage of genes in the genome that are contained in the assembly. This can be assessed using a BUSCO analysis (Yandell and Ence 2012; Simão et al. 2015). Finally, correctness is the hardest to estimate since it refers to misassemblies and errors in the sequence. A widespread practice is to compare the assembly to a reference sequence. Unfortunately, that is impossible for *de novo* assemblies of non-model organisms. One alternative is to use the read spectrum and assembly copy number to validate the assembly. These provides insight into the composition and quality of the assembly (Mapleson et al. 2017).

Despite its reputation, SOAPdenovo2 produced the assembly with the least favorable statistics. With 354,749,605 bp, it yielded the smallest and most incomplete assembly. Moreover, it is highly fragmented, as evidenced by the low N50 value and large number of contigs. ABySS 2.0 resulted in a similar assembly, although less fragmented as evidenced by the larger N50 values and smaller number of scaffolds. At 414,066,313 bp, it is slightly larger but still significantly incomplete. Since neither of these assemblies represents more than 20% of the estimated genome size, I conducted no further quality assessment.

The CLC Genomic Workbench assembly is the most complete and least fragmented with 1,166,468,433 bp in length, 62% of the estimated 1C genome size. Accordingly, universal orthologs analysis identified 1,019 complete conserved orthologous genes out of 1,440 total groups searched. Of the identified conserved orthologous genes, 889 are complete and single copy (S, 61.7%) and 130 are complete and duplicated (D, 9.0%). Fragmented (F, 10.6%) and missing orthologous genes (M, 18.7%) account for roughly 30%, which may correspond to the missing part of the assembly based on sequence length. Conserved orthologous genes in the OrthoDB come from sampling hundreds of genomes and selecting orthologous groups with single-copy orthologs in more than 90% of the species. A small number of duplicates reflects the

quality of the assembly since conserved orthologous genes are evolving under single-copy control. In the case of *D. oleifera,* the duplicates can be a result of its ploidy level. In terms of contiguity, the CLC Genomics assembly has the largest N50 values, the largest scaffold length, and less than 2% gaps in the whole sequence. Since there is no reference genome sequence available for *D. oleifera,* it is difficult to assess the correctness of the draft sequence. However, I validated the assembly by comparing the short reads k-mer spectrum and the assembly copy number. Fig. 3.1 represents how many elements of each frequency in the reads spectrum were included in the assembly once (red), twice (purple), three (green) or four (blue) times, or none (black). The quality of the assembly can be inferred from this distribution. A small amount of non-incorporated reads is desirable, since it indicates that most of the information was included in the assembly. For a haploid genome, when the assembler is performing well, sequencing errors should be absent and the genuine content present once, with duplications or repeated content centered around multiples of the sampling frequency for unique content. In the case of *D. oleifera*, the distribution of elements repeated up to four times is consistent with the ploidy level, *i.e.*, tetraploid, and indicates that the assembly is correctly representing the multiple haplotypes present.

Genome size and ploidy validation

The average genome size estimated from the k-mer frequency distribution is 3.78 Gb. This value is almost identical to the genome size (2C=3.86 pg or 3.78 Gb) estimated from flow cytometry data (see chapter 2). Ploidy estimates, from both the lrdmodel (lowest delta-log likelihood) and histotest (highest R^2 value), support *D. oleifera* tetraploidy. nQuire uses a Gaussian Mixture Model (GMM) approach in its lrdmodel subcommand. The GMM models read

frequency as a mixture of Gaussian distributions that are scaled by a mixture proportion. The GMM along with an Expectation-Maximization (EM) algorithm can be used for parameter estimation and model comparison when specific expectations about the data are known. For *D. oleifera*, the expectation is a mixture of three Gaussian distributions with means of 0.25, 0.5, and 0.75 for a tetraploid. The delta log-likelihoods estimated by the lrdmodel represent the distances between each fixed model and the best fit under the assumption of the GMM. The histotest subcommand uses a simple linear regression-based test against the three fixed models, *i.e.* diploidy, triploidy, and tetraploidy. Results from lrdmodel and histotest are summarized in Table 3.2 and Table 3.3 respectively.

Conclusion

I produced an assembly that could be considered a high-quality draft, *sensu* Chain et al. (2009), and prove useful for downstream application, although it may still lack full genome coverage. Short-read sequencing data corroborate *D. oleifera* estimated genome size and ploidy level. Future work should focus on improving the assembly quality by increasing coverage with both short and long read libraries. In addition, Hi-C libraries should be implemented to assemble scaffolds into pseudo-chromosome scale sequences (Burton et al. 2013; Kaplan and Dekker 2013). Finally, RNA sequencing experiments and *de novo* transcriptome assembly have not been conducted for this species but would greatly aid the genome annotation process and should be a priority.

Table 3. 1. *Dipteryx oleifera* draft assembly summary statistics per software used.

	SOAPdenovo2[*]	**ABySS 2.0**[**]	**CLC Genomics**[**]
No. of contigs/scaffolds	386,703	213,641	381,857
Largest contig/scaffold (bp)	43,094	167,819	397,587
Total length (bp)	354,749,605	414,066,313	1,166,468,433
GC (%)	32	32	33
N50 (bp)	878	8,084	8,194
N75 (bp)	629	902	2,385
L50	105,744	10,205	31,993
L75	227,249	53,597	100,363
No. of N's per 100 Kb	0	1,362.23	1,460.97

Note: *No. of contigs/scaffolds* = the total number of contigs or scaffold in the assembly. *Largest contig/scaffold* = the length, in base pair (bp), of the largest contig or scaffold in the assembly. *Total length* = the total number of bases in the assembly. *GC (%)* = the total number of G and C nucleotides divided by the total length of the assembly. *Nx* (where $0 \leq x \leq 100$) = the largest contig or scaffold length, *y*, such that using contigs/scaffolds of length $\geq y$ accounts for at least *x*% of the bases of the assembly. *Lx* (where $0 \leq x \leq$ *No. of contigs/scaffolds*) = the smallest number of contigs or scaffolds whose length sum produces *Nx*. No. of N's per 100 Kbp = is the average number of uncalled bases (N's) per 100,000 assembly bases, usually related to average gap size in the scaffolds of the assembly. [*] SOAPdenovo2 assembly is comprised of contigs only. [**] ABySS 2.0 and CLC Genomics assemblies are comprised of both contigs and scaffolds.

Table 3. 2. Log likelihood estimates from the Gaussian Mixture Model test for ploidy level of *Dipteryx oleifera*, as implemented by the nQuire software package (lrdmodel).

Free model maximized log-likelihood	7,211,920
Diploid fixed model maximized log-likelihood	407,096
Triploid fixed model maximized log-likelihood	2,786,286
Tetraploid fixed model maximized log-likelihood	6,689,204
Diploid delta log-likelihood	6,804,823
Triploid delta log-likelihood	4,425,634
Tetraploid delta log-likelihood	522,716

Note: Lowest delta log-likelihood indicates most likely ploidy level based on biallelic frequency distribution observed.

Table 3. 3. Simple linear regression test for ploidy level of *Dipteryx oleifera*, as implemented by the nQuire software package (histotest).

	Diploid	**Triploid**	**Tetraploid**
Norm SSR	0.0706	0.0386	0.0051
Slope	-0.2197	-0.1781	1.0072
Slope Std. Error	0.0684	0.1054	0.1051
R^2	0.1487	0.04628	0.6089

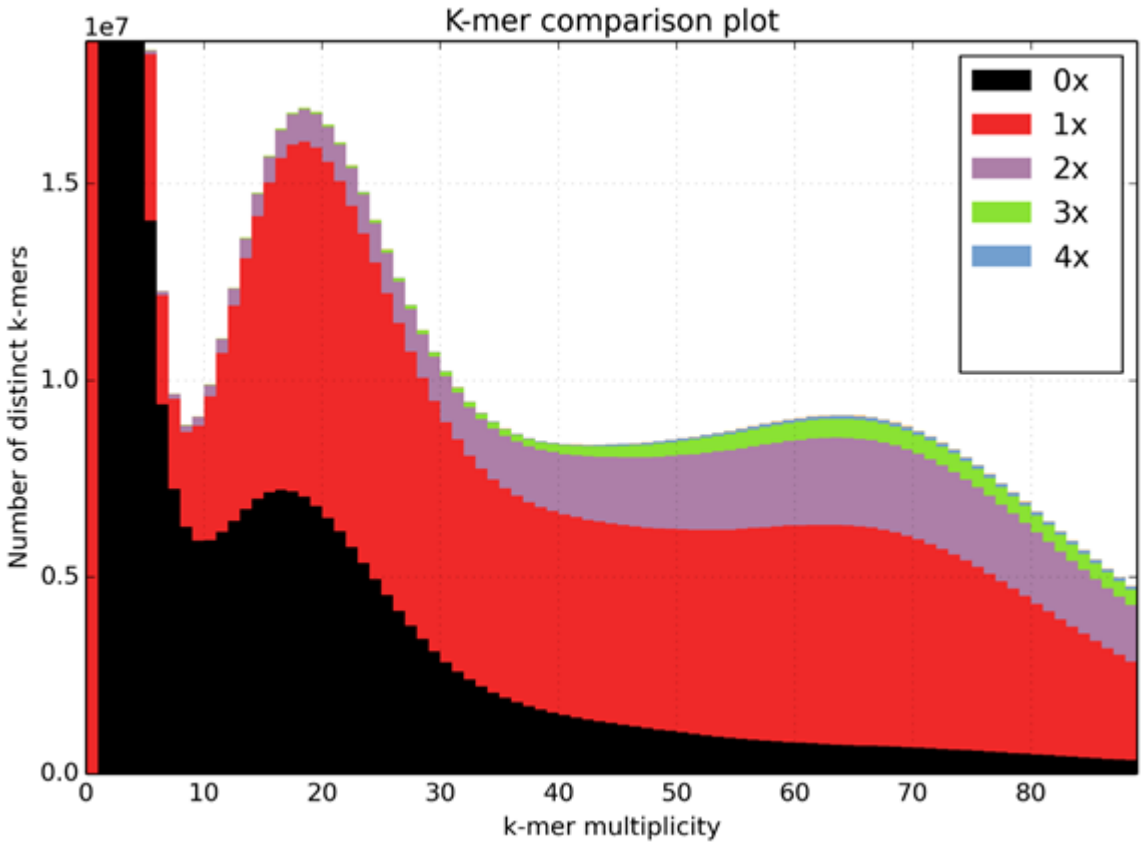

Figure 3. 1 Read k-mer frequency versus assembly copy number stacked histograms for the CLC Genomics Workbench assembly of *Dipteryx oleifera*. Read content in black is absent from the assembly, red occurs once, purple twice, green three times and blue four times. K-mer spectra show an error distribution under 10x, heterozygous content around 20x and homozygous content around 70x. Distribution is consistent with a tetraploid organism.

CHAPTER 4: Single Nucleotide Polymorphism (SNP) Discovery and Marker-Informed Breeding in a *Dipteryx oleifera* Benth. Open-Pollinated Progeny Trial

Abstract:

This chapter describes the discovery of DNA sequence variants, using a Genotyping-by-Sequencing approach, in a *Dipteryx oleifera* progeny trial. This resulted in 2,612 SNPs identified and 185 individuals genotyped. Marker data were used to estimate the realized genomic relationship among individuals in the trial. Results between pedigree-based (A-matrix) and pedigreed-based marker-corrected (H-matrix) models were compared. Estimated genetic parameters and predicted breeding values for tree diameter, total height, and volume are similar between models; however, marker-corrected relationships resulted in increased accuracy in the predictions. Ranking of the individuals based on volume breeding values and selection of the top 30 in the ranking results in an expected genetic gain of 7.7% in volume.

Introduction

Dipteryx oleifera Benth. is a large canopy-emergent tree naturally occurring in the Caribbean lowlands and very humid tropical forest of Costa Rica. The tree possesses hard, dense wood with a specific gravity of 0.83 to 1.09 (Vozzo 2010). The wood is durable and rates high in mechanical resistance, which makes it highly sought after in local markets. Based on its energy properties, *i.e.*, combustibility index and calorific value, even *D. oleifera* sawdust has potential economic value as fuel for energy generation (Gaitán-Álvarez 2015). Non-timber products are

also valuable; in Colombia the seeds are roasted for food products, such as candies and beverages (Murillo Gómez and Atehortúa 2013).

D. oleifera timber was harvested mostly from natural populations, although in Costa Rica this practice was restricted in 1996 and banned in 2008. There are exploratory studies evaluating *D. oleifera* performance in a plantation setting (Butterfield and Mariano 1995; Petit and Montagnini 2006; Gamboa-Badilla and Arias Le Claire 2008), however, the amount planted is still small compare to non-native species like teak. To reverse this trend, GENFORES, a forestry industry – university co-operative program based in Costa Rica, has started a breeding program for *D. oleifera*. The research I describe here is part of GENFORES efforts to develop molecular markers that can aid *D. oleifera* selection process.

In recent years, next-generation sequencing (NGS) technologies have shown a trend of increasing availability and throughput together with decreasing overall cost. This has allowed breeding programs to incorporate such technologies in their marker discovery and genotyping pipelines. For example, reduced representation techniques like genotyping-by-sequencing (GBS) (Elshire et al. 2011), and double digest restriction-site-associated DNA sequencing (RAD-seq) (Peterson et al. 2012) allow for easy discovery of single nucleotide polymorphism (SNP). SNPs are ubiquitous, codominant, and can be in functional parts of the genome, thus making them ideal markers for tree improvement and conservation genetics (Poland and Rife 2012; Narum et al. 2013). In general, these techniques use restriction endonucleases to target only a small portion of the genome, hence the reduction in complexity. The enzyme digestion is coupled with DNA barcoded adapter ligation to produce multiplex libraries of samples ready for most NGS platforms. The approach is rapid, robust, and high-throughput, producing thousands of molecular markers. Furthermore, these markers can be used to estimate the genetic relationships among

individuals in a breeding program and contribute to the accuracy of the selection model predictions (Habier et al. 2007).

A previous study demonstrated the potential for tree improvement in *D. oleifera* (Martínez-Albán et al. 2016), however, its conclusions were based on a traditional model that relied only on pedigree information. Here, I compare the use of pedigree and molecular markers in modeling the relationship among individuals in a *D. oleifera* progeny trial and discuss the advantages of using marker data in breeding value predictions.

Materials and Methods

Progeny trial data

An open pollinated (OP) progeny trial comprised of two plantings of *Dipteryx oleifera* Benth. was established in Costa Rica in 2010 (Fig.4.1). The first planting (NA) is located within the species natural range in the North Atlantic lowlands (10.370372° N, 84.515622° W). It is comprised of 29 families from 3 provenances. The provenances are Coopesanjuan (La Gloria de Aguas Zarcas, San Carlos, Costa Rica), Crucitas (Pocosol de San Carlos, Costa Rica), and Puerto Viejo (Sarapiquí, Costa Rica). These provenances represent natural populations, with a mixture of trees found in old growth forest patches, secondary forest, and remnant trees in cattle grazing pastures. The second planting (SP) is in the South Pacific lowlands (8.661508° N, 83.471102° W). Due to seed availability and mortality, this planting includes 19 of the 29 families present at the NA planting, with representatives of all provenances.

Both plantings have a randomized complete block design (RCB), with 6 blocks and 3 pairs of trees per family per block. More information on the experimental design, planting conditions, and provenances can be found in Martínez-Albán *et al.* (2016). The NA planting was

thinned at year four, and only the best tree of each pair was left. The SP did not undergo thinning.

GENFORES provided measurements for diameter at breast height (1.3. m above the ground level) over bark (DBH, cm) and total height (H, m) at year six for all surviving trees. I estimated total volume (V, m^3) using the following taper equation (León et al. 2017):

$$V = (\frac{DBH}{100})^2 \times \frac{\pi}{4} \times H \times FF \qquad \text{Eq. 4. 1}$$

where FF (form factor) is a constant (here 0.6) that accounts for the conical shape of the trunk.

Molecular markers

I collected plant material from a subset of the progeny trial at both plantings between June-July 2015 and June-July 2016. The leaf tissue was dried with silica gel and then stored at -70°C. I isolated high molecular weight genomic DNA using two commercial kits, Wizard® Genomic DNA purification kit (Promega) and DNeasy® Plant Mini kit (Qiagen), and a CTAB (Cetyl trimethylammonium bromide) method (Lodhi et al. 1994). DNA quality and quantity were assessed using a NanoDrop™ spectrophotometer (Thermo Scientific).

I prepared a reduced representation library according to the two-enzyme protocol described in Poland et al. (2012), with minor modifications. In brief, restriction was carried out followed by adapter ligation, then by PCR amplification and size selection. First, 10 µl of DNA (30-60 ng/ µl, 300-600 ng total) per sample were pipetted into a well on a 96-well plate that contained the restriction master mix. The restriction master mix consisted of 2 µl 10x Promega 4-CORE® Buffer System (Buffer B, final concentration 1x), 0.2 µl acetylated BSA (10 µg/µl), 0.2 µl (2 U) of each Promega Restriction Enzyme (PstI and MspI), and 6.8 µl of sterile deionized water per reaction. The samples were placed in a thermal cycler at 37°C for two hours. After

restriction, 20 µl of ligation master mix was added to each well and incubated overnight at room temperature. The ligation master mix consisted of 4 µl Promega T4 DNA Ligase 10x Buffer (final concentration 1x), 0.4 µl 100 mM ATP (final concentration 1 mM), 1 µl Promega T4 DNA ligase (3 Weiss Units), 4.6 µl of sterile deionized water, and 5 µl of each adapter. The barcoded adapters were designed to anneal to the PstI cut site. The common adapter annealed to the MspI cut site. A complete list of the sequences of the adapters used is in Appendix 2, GBS Adapters. Next, 5-10 µl of each restriction-ligation sample were pooled, cleaned, and concentrated using a DNA Clean & Concentrator™ kit (Zymo Research). I prepared pools of 32 samples or three pools per 96-well plate. PCR amplification was performed using 10 µl of pooled sample, 25 µl NEB Q5® High-Fidelity 2X Master Mix (final concentration 1x), 10 µl of nuclease free water, and 5 µl 10 µM forward and reverse primers (2.5 µl each, final concentration 0.5 µM). I prepared a separate reaction for each pooled sample and used different NEBNext Illumina primers for each reaction, so samples can be identified by their specific barcode-index combination. PCR settings for amplification were 98 °C for 30s, 16 cycles alternating temperatures (98°C for 10s, 62°C for 30s, 72°C for 30s), 72°C for 2 m, followed by 10°C until sample recovery. PCR products were pooled, cleaned, and concentrated using a DNA Clean & Concentrator™ kit (Zymo Research). Finally, the pooled PCR products were selected for fragments ranging from 400 to 600 bp, using the BluePippin System (Sage Science). The recovered fragments were used as the input library. The library was normalized to 2 nM prior to loading onto the flow cell, then sequenced in the Illumina HiSeq2500 platform using SBS sequencing kit version 4. The multiple index library ran for 132 cycles to produce single-end 124-nt reads and 8-nt index reads. Next-generation sequencing was performed by the NC State University Genomic Sciences Laboratory

(Raleigh, NC, USA). Only the samples with the highest coverage were retained. This resulted in sequence reads from a total of 185 trees from 19 different families in the progeny trial.

I used Stacks version 2.0 for demultiplexing, data cleaning, and variant calling (Catchen et al. 2011; Catchen et al. 2013). Stacks software was developed for the analysis of reduced representation Illumina sequence data such as GBS or RAD-seq. It uses short read sequence data to identify and genotype loci in a set of individuals. Stacks works by assembling loci from each sample, either by comparison to a reference genome or *de novo*, grouping together loci across samples to build a catalog, and then comparing each sample to the catalog for variant calling and genotype inference. I used the step by step *de novo* pipeline with these parameters: -m, minimum stack depth or minimum depth of coverage, set to 3; -M, distance allowed between stacks, set to 4; and -n, distance allowed between catalog loci, set to 4. For more details on the commands and parameters used for demultiplexing, cleaning, and variant calling see Appendix 3, Demultiplexing and variant calling script.

Finally, I used statistical analysis software R (R Core Team 2018) and R package Updog (Gerard et al. 2018) to validate genotypes, impute missing data, and get allele dosage values. Updog uses a variational Bayes approach to estimate the allele dosage and associated posterior probabilities for each genotype per individual. Furthermore, Updog accounts for locus-specific allele bias, locus-specific sequencing error, locus-specific overdispersion, and correlation between samples while jointly estimating each genotype. The resulting genotypes were filtered, retaining only loci with mean posterior probabilities ≥ 0.9.

Statistical analysis

For the computation of the realized relationship matrix (G-matrix) I used the allele dosage data from the 185 genotyped trees. For genetic parameters estimation I used a subset of the total phenotyped trees; this subset includes only trees in families with genotyped individuals (N = 625).

I estimated variance components with restricted maximum likelihood (REML) and breeding values (BVs) with best linear unbiased prediction (BLUP), as implemented in statistical analysis software ASReml® version 4.0 (VSNi), for stem diameter, total height, and total volume. The univariate linear mixed model used for the analysis corresponds to the Individual ("Animal") Model:

$$y_{ijkl} = \mu + B_i + P_j + S_k + T_l + \varepsilon_{ijkl} \qquad \text{Eq. 4. 2}$$

where, y_{ijkl} is the vector of observations of the trait (*i.e.* DBH, height, volume),

μ is the overall mean,

B_i is the random *i*th block effect ~N (0, σ_b^2),

P_j is the fixed provenance effect (j = 1, 2, 3),

S_k is the fixed site effect (k = 1, 2),

T_l is the random *l*th tree effect ~ N (0, σ_T^2),

and ε_{ijkl} is the residual term ~ N (0, σ_ε^2).

Both P_j and S_k were included in the model to account for provenance effect, and site effect due to the difference between environment and silvicultural treatment in the two plantings.

The model can be written in matrix form as:

$$y = Xb + Zu + e \qquad \text{Eq. 4. 3}$$

where, y is the vector of observations,

b is a vector of fixed effects,

u is a vector of random effects,

e is a vector of random residuals,

and *X* and *Z* are incidence matrices that assign each element of *b* and *u* to their corresponding element in *y*.

The *Z* matrix specifies the covariance structure use to model the tree effects in relation to the observations. This relationship can be based on the pedigree, molecular markers, or a combination of both. I used R package AGHmatrix (Amadeu et al. 2016) to compute the relationship matrices used in the linear mixed models. AGHmatrix was design to work with polyploids, thus it was adapted to account for tetrasomic inheritance patterns in tetraploid species, such as *D. oleifera*. I tested the model using three different relationship matrices: (i) a pedigree-based relationship matrix, A-matrix; (ii) a pedigree-based, molecular-marker-corrected relationship matrix, H-matrix, considering only additive effects; and (iii) a pedigree-based, molecular-marker-corrected relationship matrix, H-matrix, considering both additive and non-additive effects. The models were compared based on the converged log likelihood (LogL), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). The R script used to generate the relationship matrices, as well as the ASReml script to run the model can be found in Appendix 4, Relationship matrix and linear mixed model.

The heritability for each trait was calculated from the model's variance components as follows:

$$h_i^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\varepsilon^2} \qquad \text{Eq. 4. 4}$$

where, σ_T^2 is the variance of the tree effects, and σ_ε^2 is the residual variance.

The accuracy of the breeding value predictions for individuals in the model was calculated as:

$$r = \sqrt{\frac{S^2}{(1+F)\times\sigma_T^2}} \qquad \text{Eq. 4. 5}$$

where, *S* is the standard error of the prediction, *F* is the inbreeding coefficient of the individual being predicted, and σ_T^2 is the tree effect variance. Equations 4.2 to 4.5 follow Isik *et al.* (2017) notation.

I used volume BVs and accuracies from the best fitting model to construct a ranking of the individuals in the progeny trial. Genetic gain was estimated using the breeder's equation:

$$G = h_i^2 \times selection\ differential \quad \text{Eq. 4. 6}$$

where, *G* is the expected genetic gain, h_i^2 is the individual tree heritability estimated from the model, and the *selection differential* represents the difference between the mean value of the selected individuals and the mean value of the population. Genetic gain was estimated based on selection of the top 30 individuals and top 100 individuals in the ranking.

Results and Discussion

Molecular marker development

The reduced representation library contained 722,398,562 short read sequences from the sampled individuals. After demultiplexing and cleaning, 90.9 % of the reads were retained, 4.9 % were dropped due to ambiguous barcodes, 0.9 % were dropped due to low quality, and 3.2 % were dropped to ambiguous RAD-tags. Only the individuals with the highest coverage were retained, resulting in 185 samples with a mean coverage per sample of 21.7x. These samples were used as input in the Stacks pipeline. Despite the low coverage, Stacks identified a total of 4,676 SNPs with missing data ≤ 10 %. The Stacks filter parameters were set to look for SNPs

present in individuals from at least two of the three provenances in this study. This reduces, although it does not eliminate, the occurrence of private alleles or alleles with low frequencies in the progeny trial.

There are low to moderate levels of allelic bias, sequencing error, and overdispersion (Fig. 4.2.). These factors are expected to modify the allele proportion at each genotype but are accounted for and corrected in the Updog model. Moreover, Gerard and collaborators (2018) found Updog accurately estimates the allele frequency even for large levels of overdispersion and bias. For example, Fig. 4.3 depicts an average SNP from the *D. oleifera* dataset. Colored dots indicate individual genotypes based on the allele dosage (number of copies of the alternate allele). Color intensity indicates the posterior probability, *i.e.* the probability of correctly assigning the allele dosage or genotype. This example shows low levels of allelic bias (0.84), sequencing error (0.02), and overdispersion (0.01). As a result, most of the individuals are genotyped with high confidence (max posterior probability ≥ 0.75). Although Updog was able to estimate allele dosage values for all individuals and all loci, in some cases the posterior probabilities of the imputed genotypes were low. Therefore, the dataset was filtered for genotypes with max posterior probabilities ≥ 0.9. This resulted in a subset of 2,612 SNPs with high confidence genotype calls.

Relationship matrix effects on genetic parameters and breeding values

The resulting SNP markers were used to construct the genomic relationship matrix, or G-matrix. The main difference between a relationship matrix based on pedigree records and one based on SNP marker genotypes is that the former represents the expected relationship, while the latter represents the realized relationships. Put differently, pedigree-based estimators of the

relationship provide theoretical expectations of genetic similarity between individuals, for example among full-sib individuals the expected proportion of the genome that is identical-by-descent is 0.5. Conversely, SNP marker-based estimators provide a better representation of the real proportion of the genome shared between individuals. Moreover, using dense SNP marker data it is possible to estimate identical-by-descent probabilities even without knowledge of the pedigree (Powell et al. 2010).

For open-pollinated (OP) families, as in the *D. oleifera* progeny trial, a relationship matrix based on SNP marker genotypes provides an accurate and precise representation of the relationship among individuals. A pedigree-based approach would assume the family to be composed exclusively of half-sibs and each pair-wise kinship within family assigned a fixed expected value for allele-sharing based on that similarity (0.25 for half-sibs), however, this may not be the case. An OP family may include a mixture of offspring from crosses with unrelated male parents (true half-sibs), offspring from crosses with related male parents (half-cousins, full-cousins, etc.); multiple offspring from crosses to single individual males (full-sibs), and even some offspring from self-pollination. In this scenario, a SNP marker relationship matrix would be able to better capture the true relationship among individuals and should therefore increase the accuracy with which genetic parameters are estimated and breeding values are predicted.

This is exemplified in Fig. 4.4, which compares the results of the A-matrix (pedigree-based) and G-matrix (marker-based) from the *D. oleifera* dataset. The A-matrix in the figure depicts the idealized scenario with 19 distinct families. The G-matrix depicts quite a different scenario, where the relationships among individuals are much more complex. The G-matrix-based analysis detects individuals from different families that show various levels of relationships, and even individuals within families that differ in their degree of relationship. The

family boundaries and structure imposed by the pedigree information are no longer as clear, thus the assumptions of unrelated individuals between families and only half-sibs within families are not realistic. The G-matrix-based analysis also points to potential errors in the pedigree, *e.g.* mislabeled individuals assigned to families to which they are not truly related.

I tested the utility of the relationship matrix based on pedigree and corrected with marker information on three phenotypic traits: diameter, height, and volume. Table 4.1 presents summary statistics for the traits. I used a univariate linear mixed model to estimate the variance components for each trait. I compared three different relationship matrices in the model, a pedigree-based (A-matrix) and two pedigree-based, marker-corrected (H-matrix). The two H-matrices differ in the elements included; H_a considers only additive effects while H_f consider both additive and non-additive effects. Table 4.2 presents estimators of relative quality for the statistical models, as well as variance components and individual heritability estimated for each trait and each model. The models do not differ in the number of parameter or sample size. The only change between models was the covariance structure used to estimate the tree effects. Therefore, the difference in AIC or BIC can be attributed to the model's goodness of fit to the data. In general, using the H matrix results in a significant improvement in model fit, except for diameter where the differences are minor. H_f provides the best fit for height, while H_a provides the best fit for volume (Table 4.2). Individual heritability estimates are lower with the H-matrices than the A-matrix. This is likely to be the result of individuals having increased kinship that accounts for similarity of phenotypes in the H-matrix-based analyses. For individuals more closely related, as indicated by the H-matrix, a larger proportion of the variance could be ascribed to environmental factors rather than genetic contributions. More important, while the

tree effect variance and heritability values may decrease, the accuracy of those estimations increased; that is, standard errors are reduced by the used of either H-matrix (Table 4.2).

Provenance did not have a significant effect in any of the models or traits. Again, this is likely to be the result of individuals having increased kinship, and evidence of gene flow between the provenances, despite the distance. Planting site did have a significant effect on diameter (A: $F = 114.44$, p-value < 0.001; H_a: $F = 107.82$, p-value < 0.001; H_f: $F = 111.89$, p-value < 0.001) and volume (A: $F = 70.03$, p-value < 0.001; H_a: $F = 65.47$, p-value < 0.001; H_f: $F = 68.11$, p-value < 0.001), but not in tree height. The trees in the SP planting (site 2) have, on average, smaller diameter and volume estimates. However, it is impossible to determine whether the effect is due to environmental factors (*e.g.* soils, precipitation, etc.) or silvicultural management (thinning). Both factors are confounded.

Tables 4.3 to 4.5 present the mean breeding value and accuracy per family for diameter, height, and volume, respectively. In general, mean breeding value predictions do not change much between A and H models. However, H_a does result in higher accuracy values, which indicates a better prediction of the true value.

Finally, since volume is a function of the tree diameter and height, as well as a commercially relevant trait, I created a ranking based on volume breeding values. Expected genetic gain by selecting the top 30 individuals in the ranking was 0.008 m^3, that is a 7.7% increase over the population mean. However, the top 30 individuals come from 7 families with an overrepresentation of individuals from family PV3 (20/30 individuals). If selection is expanded to the top 100 individuals, the expected genetic gain drops to 0.006 m^3 or 5.3% over the population mean, but 14 of the original 19 families are included. There is still an overrepresentation of individuals from only 3 families (PV3 = 33 individuals, 10 = 17

individuals, and CSJ2 = 13 individuals) among the top 100 individuals. A complete list of the top 100 trees from the progeny trial ranked by their H_f volume breeding values is presented in Appendix 5, Ranking. Moving forward in the *D. oleifera* breeding program, careful consideration should be put in selecting individuals for controlled crosses to maintain a genetically diverse population (to avoid inbreeding depression) while still making genetic gain on the traits of interest. The breeding program may consider linear deployment (Lindgren 1993), stratified sub-lining (Ruotsalainen and Lindgren 2000), or two-stage selection strategies (Danusevicius and Lindgren 2002) to manage genetic gain and diversity in the breeding population. Further study is necessary to determine which strategy will constitute the best approach.

Conclusion

Discovery of variants and marker development from sequencing data is a challenging endeavor in polyploid species. Distinct inheritance patterns, like tetrasomic inheritance and double reduction in autotetraploids, increase the complexity of the genotyping process. In this research I have proved that Stacks, a bioinformatic tool originally designed for variant detecting and genotyping in diploid species, can be used in combination with R software package Updog to produce accurate genotypes in tetraploid *D. oleifera*. Although a low-density panel (N = 2,612 SNPs), the molecular marker data has the potential to improve model fit and accuracy of the genetic parameters, even when only a fraction of the population was genotyped. This is consistent with previous reports in animal (Badke et al. 2014) and plant (Beaulieu et al. 2014) breeding programs. Moreover, marker data proves particularly useful when working with natural population for which little to no information is available on their genetic background and mating

patterns. Relationship matrices corrected with marker data provide a better representation of the covariance structure of tree effects in those population. Future work should focus on increasing the number of markers and individuals genotyped. As the genetic information on this progeny trial increases, the accuracy and predictive power of the models are expected to improve as well.

Table 4. 1. *Dipteryx oleifera* open-pollinated progeny trial summary statistics for diameter (DBH), total height (H) and volume (V) at year six for genotyped and non-genotyped individuals in the dataset.

	N	Mean (SD)		
		DBH (cm)	H (m)	V (m^3)
All trees with phenotype	1032	11.59 (3.15)	13.41 (2.18)	0.0957 (0.05)
Trees in families with genotyped individuals	625	11.70 (3.12)	13.33 (2.16)	0.0965 (0.05)
Genotyped individuals	185	11.60 (3.25)	13.26 (2.41)	0.0957 (0.05)
Total number of families	29	-	-	-
Total number of families with genotyped individuals	19	-	-	-

Note: *N* = number of records, SD = standard deviation

Table 4. 2. Linear mixed model comparison, variance components, and heritability estimation for diameter, height, and volume in a *Dipteryx oleifera* progeny trial. Models differ in the relationship matrix used to estimate the individual tree effects.

Trait:	Diameter (cm)			Height (m)			Volume (m^3)		
Model:	A	H_a	H_f	A	H_a	H_f	A	H_a	H_f
LogL	-966.43	-966.42	-966.47	-791.385	-790.90	-788.92	1554.93	1553.23	1553.69
AIC	1938.85	1938.83	1938.958	1588.77	1587.79	1583.83	-3103.85	-3100.46	-3101.37
BIC	1952.15	1952.13	1952.24	1602.06	1601.09	1597.13	-3090.56	-3087.17	-3088.08
σ_T^2 (SE)	2.64 (1.27)	1.57 (0.56)	1.84 (0.67)	0.85 (0.49)	0.53 (0.26)	0.92 (0.37)	$8.6x10^{-4}$ ($4.0x10^{-4}$)	4.4x10-4 ($1.6x10^{-4}$)	$5.3x10^{-4}$ ($1.9x10^{-3}$)
σ_ε^2 (SE)	5.75 (1.07)	6.66 (0.58)	6.08 (0.74)	3.76 (0.46)	4.03 (0.31)	3.51 (0.41)	$1.6x10^{-3}$ ($3.4x10^{-4}$)	$2.0x10^{-3}$ ($1.7x10^{-4}$)	$1.8x10^{-3}$ ($2.1x10^{-4}$)
h_i^2 (SE)	0.31 (0.14)	0.19 (0.06)	0.23 (0.08)	0.18 (0.10)	0.12 (0.05)	0.21 (0.08)	0.34 (0.15)	0.18 (0.06)	0.23 (0.08)

Note: A = model using A matrix, H_a = model using H matrix, considering only additive effects; H_f = model using H matrix, considering both additive and non-additive effects, LogL = log likelihood at which the model converged, AIC = Akaike information criterion, BIC = Bayesian information criterion, σ_T^2 = variance of the tree effects, σ_ε^2 = residual variance, h_i^2 = individual tree heritability, and SE = standard error.

Table 4. 3. *Dipteryx oleifera* open-pollinated progeny trial mean breeding values and accuracy per family for diameter (cm).

	Mean BV (r)		
Family	A	H_a	H_f
4	12.70 (0.80)	12.73 (0.85)	12.72 (0.84)
6	13.22 (0.80)	13.17 (0.85)	13.19 (0.84)
8	11.54 (0.80)	11.65 (0.82)	11.64 (0.86)
9	11.94 (0.80)	12.10 (0.84)	12.03 (0.84)
10	13.23 (0.80)	13.17 (0.85)	13.20 (0.83)
CSJ1	12.56 (0.79)	12.64 (0.83)	12.58 (0.84)
CSJ2	13.27 (0.79)	13.28 (0.83)	13.24 (0.83)
CSJ3	12.43 (0.79)	12.52 (0.83)	12.45 (0.83)
CSJ4	12.18 (0.79)	12.31 (0.84)	12.24 (0.83)
CSJ6	12.68 (0.79)	12.74 (0.83)	12.70 (0.83)
CSJ7	12.78 (0.79)	12.85 (0.84)	12.79 (0.82)
CSJ8	12.29 (0.79)	12.34 (0.83)	12.29 (0.82)
SM9	11.99 (0.79)	12.23 (0.84)	12.09 (0.83)
PV2	11.37 (0.79)	11.55 (0.83)	11.49 (0.85)
PV3	14.39 (0.79)	14.05 (0.84)	14.18 (0.83)
PV4	11.54 (0.80)	11.69 (0.83)	11.63 (0.86)
PV8	12.70 (0.79)	12.75 (0.85)	12.73 (0.83)
PV9	12.44 (0.79)	12.47 (0.84)	12.45 (0.84)
PV10	12.76 (0.79)	12.65 (0.83)	12.73 (0.84)

Note: BV = breeding value, r = accuracy of breeding value (larger values indicate a better estimation), A = model using A matrix, H_a = model using H matrix, considering only additive effects; H_f = model using H matrix, considering both additive and non-additive effects.

Table 4. 4. *Dipteryx oleifera* open-pollinated progeny trial mean breeding values and accuracy per family for height (m).

	Mean BV (r)		
Family	A	H_a	H_f
4	13.28 (0.86)	13.28 (0.90)	13.30 (0.85)
6	13.45 (0.86)	13.44 (0.90)	13.49 (0.85)
8	12.75 (0.87)	12.76 (0.88)	12.70 (0.87)
9	12.7 (0.87)	12.79 (0.89)	12.68 (0.86)
10	13.56 (0.86)	13.50 (0.90)	13.58 (0.85)
CSJ1	13.02 (0.86)	13.10 (0.88)	13.03 (0.85)
CSJ2	13.34 (0.86)	13.37 (0.88)	13.37 (0.85)
CSJ3	12.96 (0.86)	13.04 (0.88)	12.96 (0.85)
CSJ4	13.19 (0.86)	13.23 (0.89)	13.21 (0.84)
CSJ6	13.1 (0.86)	13.17 (0.88)	13.11 (0.84)
CSJ7	13.38 (0.86)	13.42 (0.89)	13.43 (0.84)
CSJ8	13.18 (0.86)	13.22 (0.89)	13.20 (0.84)
SM9	13.01 (0.86)	13.11 (0.89)	13.03 (0.85)
PV2	12.67 (0.86)	12.76 (0.88)	12.65 (0.87)
PV3	14.03 (0.86)	13.84 (0.89)	14.04 (0.85)
PV4	12.39 (0.87)	12.48 (0.88)	12.32 (0.88)
PV8	13.07 (0.86)	13.16 (0.89)	13.11 (0.84)
PV9	13.34 (0.86)	13.32 (0.89)	13.35 (0.86)
PV10	13.40 (0.86)	13.28 (0.88)	13.37 (0.85)

Note: BV = breeding value, r = accuracy of breeding value (larger values indicate a better estimation), A = model using A matrix, H_a = model using H matrix, considering only additive effects; H_f = model using H matrix, considering both additive and non-additive effects.

Table 4. 5. *Dipteryx oleifera* open-pollinated progeny trial mean breeding values and accuracy per family for volume (m^3).

	Mean BV (r)		
Family	A	H_a	H_f
4	0.11 (0.78)	0.11 (0.86)	0.11 (0.84)
6	0.12 (0.78)	0.12 (0.86)	0.12 (0.84)
8	0.09 (0.78)	0.09 (0.83)	0.09 (0.86)
9	0.10 (0.78)	0.10 (0.85)	0.10 (0.84)
10	0.12 (0.78)	0.12 (0.86)	0.12 (0.84)
CSJ1	0.11 (0.77)	0.11 (0.84)	0.11 (0.84)
CSJ2	0.12 (0.77)	0.12 (0.83)	0.12 (0.84)
CSJ3	0.11 (0.77)	0.11 (0.84)	0.11 (0.84)
CSJ4	0.10 (0.77)	0.1 (0.85)	0.10 (0.83)
CSJ6	0.11 (0.77)	0.11 (0.84)	0.11 (0.83)
CSJ7	0.11 (0.77)	0.11 (0.84)	0.11 (0.83)
CSJ8	0.10 (0.77)	0.10 (0.84)	0.10 (0.83)
SM9	0.10 (0.77)	0.10 (0.85)	0.10 (0.84)
PV2	0.09 (0.78)	0.09 (0.83)	0.09 (0.85)
PV3	0.14 (0.78)	0.14 (0.85)	0.14 (0.84)
PV4	0.09 (0.78)	0.09 (0.83)	0.09 (0.86)
PV8	0.11 (0.78)	0.11 (0.85)	0.11 (0.83)
PV9	0.10 (0.78)	0.11 (0.85)	0.11 (0.85)
PV10	0.11 (0.78)	0.11 (0.84)	0.11 (0.84)

Note: BV = breeding value, r = accuracy of breeding value (larger values indicate a better estimation), A = model using A matrix, H_a = model using H matrix, considering only additive effects; H_f = model using H matrix, considering both additive and non-additive effects.

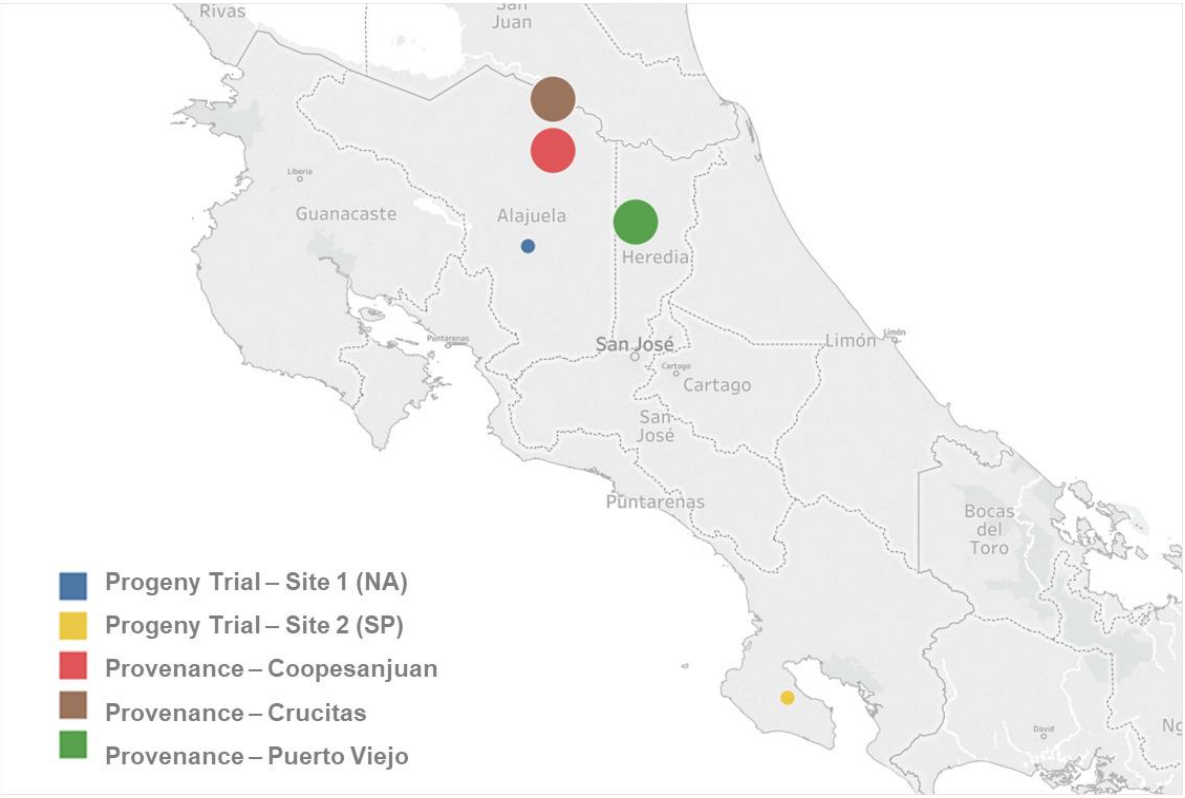

Figure 4. 1. *Dipteryx oleifera* progeny trial planting sites and provenances location.

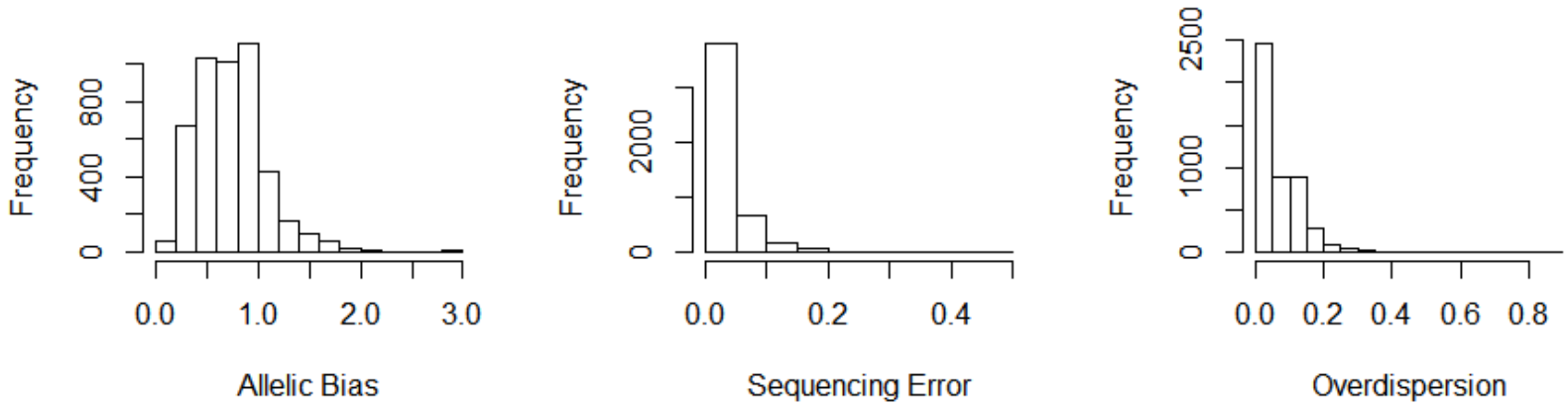

Figure 4. 2. Histograms of allelic bias, sequencing error, and overdispersion for the *Dipteryx oleifera* genotyped samples. Values estimated from 4676 loci and 190 individuals using R package Updog. Allelic bias value center around 1, where 0.5 means that a reference allele read is twice as probable to be correctly observed than the alternate allele read, while 2 means the opposite scenario. Sequencing error rates are considered low for values between 0.5 and 1%. Overdispersion ratio, values closer to 0 indicate less overdispersion and values closer to 1 indicate greater overdispersion.

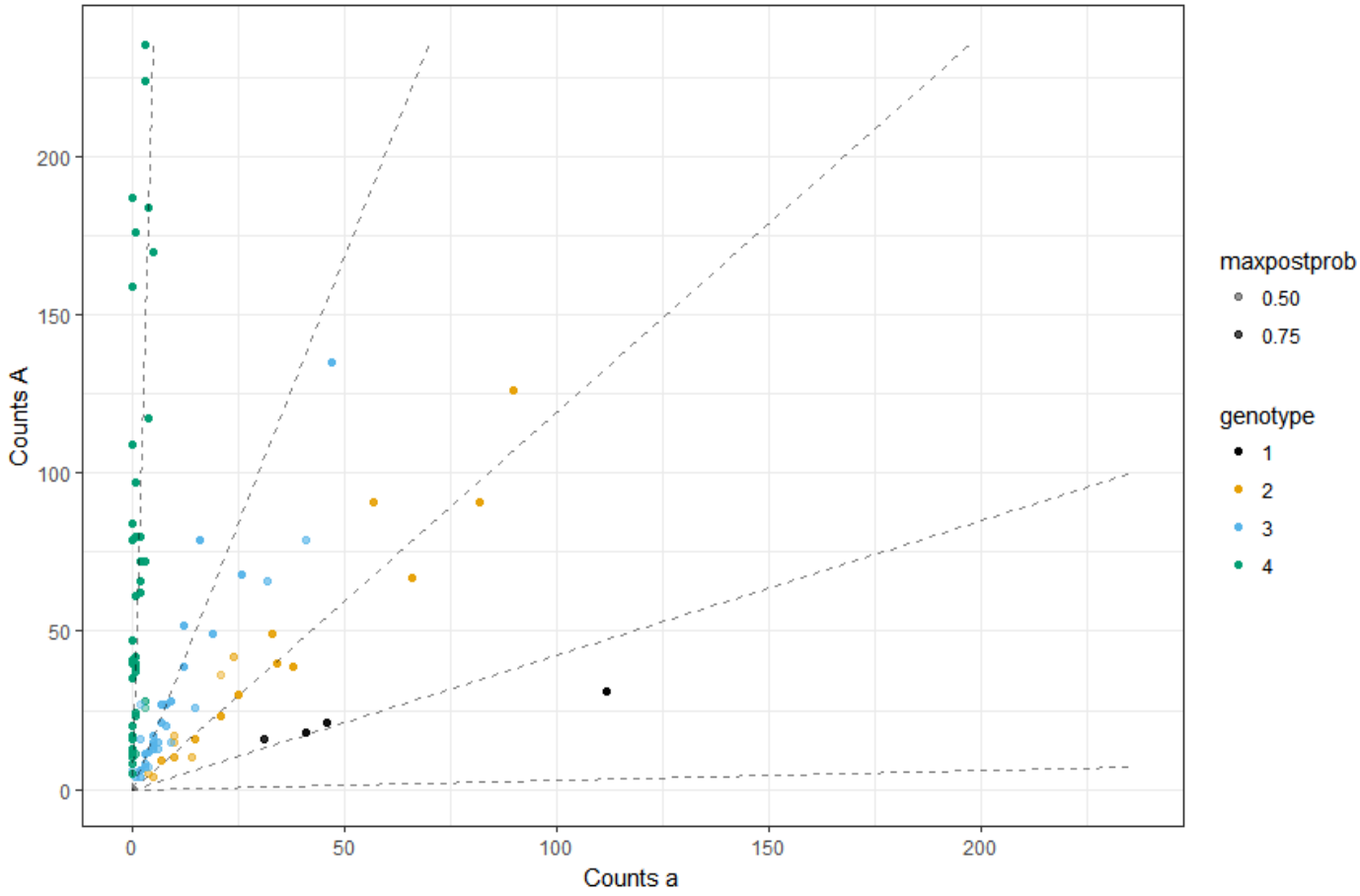

Figure 4. 3. Genotype plot of one SNP in a tetraploid *Dipteryx oleifera*. Each point is an individual with the number of alternative reads along the x-axis and the number of reference reads along the y-axis. The dashed lines represent the expected proportions for each genotype (aaaa, Aaaa, …, AAAA).

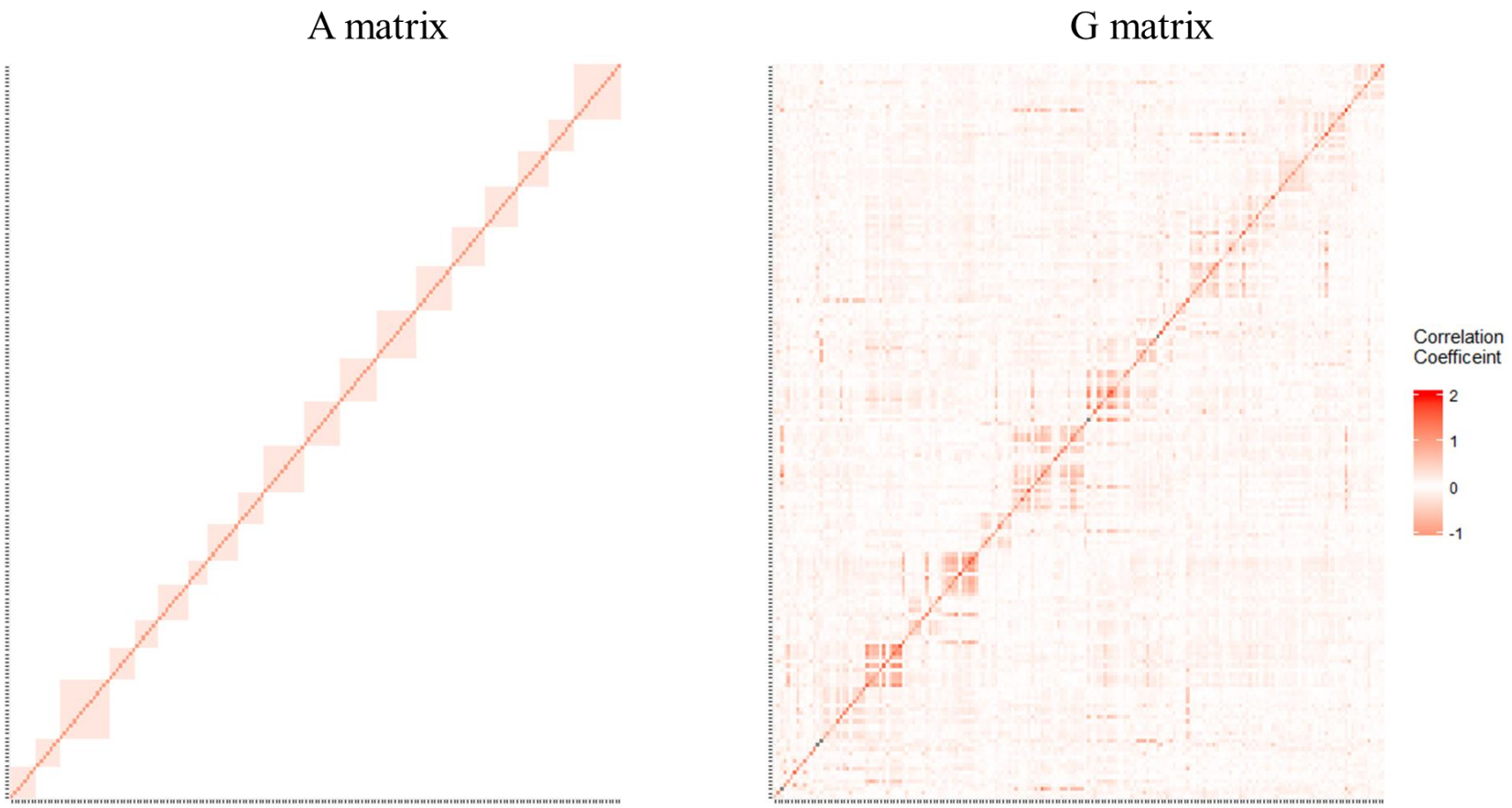

Figure 4. 4. Heat map of additive relationship matrix based on pedigree (A-matrix) and realized genomic relationship matrix (G-matrix) based on marker information (2,612 SNPs) from 185 individuals in the *Dipteryx oleifera* progeny trial.

CHAPTER 5: Concluding Remarks

Dipteryx oleifera Benth. is an important timber species that has been, paradoxically, overexploited and overlooked at the same time. Illegal logging and habitat fragmentation have diminished *D. oleifera* natural populations to the point that trade of the wood is now controlled by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), Appendix III (Estrada-Chavarria et al. 2005). Yet, the amount planted by the forest industry in Costa Rica is minimal, despite its positive performance in plantation settings (Butterfield and Mariano 1995; Andrade Naveda 2002; Petit and Montagnini 2006; Gamboa-Badilla and Arias Le Claire 2008; Schmidt 2009). According to Costa Rica's Oficina Nacional Forestal (National Forestry Office), the market share of *D. oleifera* wood has steadily decreased since 2011. Today, the volume sold is so low that it is not included in the country's wood price statistics.

One reason why *D. oleifera* was overlooked by plantation forestry was the absence of a breeding program to provide improved plant material and reduce the impact on the natural populations. This changed when GENFORES, a forestry industry – university co-operative program that works in tree improvement and germplasm conservation of native timber species in Costa Rica (Murillo et al. 2001; Murillo et al. 2010), planted its first *D. oleifera* provenance-progeny trial in 2010. My research is part of that broader effort led by GENFORES. The goal of this dissertation was to explore innovative technologies and develop genomic resources to expedite the selection process and breeding program.

In doing so, I have determined *D. oleifera* genome size with flow cytometry and corroborated it with sequencing data. Ploidy level was also determined with sequencing data, which indicates *D. oleifera* is tetraploid and most likely an autopolyploid. Furthermore, 62% of

the genome has been sequenced and assembled into scaffolds. This first draft of the genome sequence, although incomplete, has proved useful for downstream application, such as marker discovery. In the future this draft could be used, in combination with gene expression and transcriptome analysis, to improve our understanding of the genetic mechanisms involved in high density wood formation, one of the most appealing traits of *D. oleifera*.

Information generated in this research on genome size, ploidy level, and seed size, may also be relevant if considering hybrid crosses among *D. oleifera* congeners. For example, *D. odorata's* fragrant fruits are harvested for their coumarin content, which is used in perfumes and as food additive. Hybrid crosses with *D. odorata,* which has similar genome size and ploidy, could add value to *D. oleifera* non-timber products. Breeding for higher coumarin content in *D. oleifera* fruits could provide an additional source of income and serve as an incentive to plant this species.

However, the genomic resource with the most direct application to *D. oleifera* breeding program is the panel of 2,612 SNPs markers identified using a Genotyping-by-Sequencing (GBS) approach. These markers were able to accurately estimate the genetic relationship among individuals in the progeny trial. The use of a marker-corrected relationship matrix improved model fit and parameter estimation accuracy. More importantly, it highlighted a major constraint in the *D. oleifera* progeny trial. The traditional pedigree-based models can provide accurate estimations of genetic parameters and predicted breeding values; however, they rely on deep pedigree information. When dealing with an open-pollinated progeny trial, where the seeds were collected from natural population and at best only the female tree is known, these models must rely on unrealistic assumptions. In the case of this *D. oleifera* open-pollinated progeny trial the assumption is that every individual in a family (*i.e.* seeds collected under the same alleged

mother tree) are half-sibs. Now, *D. oleifera* is primarily pollinated by bees (Perry and Starrett 1980), and pollen dispersal can reach distances of up to 2.3 Km for trees growing in isolated pastures. While previously considered an obligate outcrosser species, a study by Hanson et al. (2008b) found increased inbreeding rates due to selfing for those trees growing in isolated pastures, and moderate structure in the overall population. Their study area overlaps with one of the provenances in the *D. oleifera* progeny trial. Hence, the assumption of half-sibs within family and completely unrelated families are likely to be inaccurate. The SNPs markers developed provided a better representation of the true relationships (Fig. 4.4).

As stated before, marker data can be used to infer coancestry among individuals, even in the absence of pedigree information (Powell et al. 2010). Using the same genotyping strategy described here, we could still estimate breeding values, make selections, and predict genetic gains from new collections or older *D. oleifera* trials for which pedigree information is not available or was lost. This would allow expansion of the genetic base of the *D. oleifera* breeding program.

In terms of discovery of variants and marker development for *D. oleifera*, much work is still needed. As the price of DNA sequencing continues to decrease, I hope to expand the SNP marker panel and number of genotyped individuals to get us closer to a true genomic selection model. The ideal scenario would be a breeding program as technologically advanced as *Eucalyptus* or pine species. In addition, new traits should be included, in particular traits related to stress responses and climate resilience.

The GENFORES *D. oleifera* breeding program is just starting, and more work is needed before the first generation of improved material can be deployed for commercial planting. Nevertheless, their efforts combined with this research could, in a near future, alleviate logging

pressure on natural populations by providing better-quality seeds for plantation forestry. An additional opportunity is to breed for climate resilient genotypes, a much needed trait since extreme weather conditions - higher temperatures, drier summers, and wetter winters - are expected to become more frequent in the tropical and subtropical areas (IPCC 2014; Fu 2015).

Costa Rica made the global news by announcing their commitment to carbon neutrality by 2021. Tropical forest has an important role in climate change mitigation, as it accounts for a 25% of the terrestrial carbon pool (Corlett 2016). Out of seven native tree species evaluated, Redondo-Brenes (2007) found *D. oleifera* to be the best option for long term carbon sequestration. By advancing *D. oleifera* breeding program GENFORES is not only aiding this species conservation but also the country's climate mitigation goals.

REFERENCES

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Curr Opin Plant Biol. 8:135–141. doi:10.1016/j.pbi.2005.01.001.

Ågren JA, Wright SI. 2015. Selfish genetic elements and plant genome size evolution. Trends Plant Sci. 20:195–196. doi:10.1016/j.tplants.2015.03.007.

Amadeu RR, Cellon C, Olmstead JW, Garcia AAF, Resende MFR, Muñoz PR. 2016. AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. Plant Genome. 9:1–10. doi:10.3835/plantgenome2016.01.0009.

Andrade Naveda MC. 2002. Evaluacion del crecimiento de ocho especies forestales nativas en la Universidad EARTH, La Mercedes de Guacimo, Costa Rica. EARTH.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.

Badke YM, Bates RO, Ernst CW, Fix J, Steibel JP. 2014. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 Genes|Genomes|Genetics. 4:623–631. doi:10.1534/g3.114.010504.

Bainard JD, Husband BC, Baldwin SJ, Fazekas AJ, Gregory TR, Newmaster SG, Kron P. 2011. The effects of rapid desiccation on estimates of plant genome size. Chromosom Res. 19:825–842. doi:10.1007/s10577-011-9232-5.

Balao F, Casimiro-Soriguer R, Talavera M, Herrera J, Talavera S. 2009. Distribution and diversity of cytotypes in Dianthus broteri as evidenced by genome size variations. Ann Bot. 104:965–973. doi:10.1093/aob/mcp182.

Bandel G. 1974. Chromosome numbers and evolution in the Leguminosae. Caryologia. 27:17–32. doi:10.1080/00087114.1974.10796558.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. Bamtools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 27:1691–1692. doi:10.1093/bioinformatics/btr174.

Beaulieu J, Doerksen T, Clément S, Mackay J, Bousquet J. 2014. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity (Edinb). 113:343–352. doi:10.1038/hdy.2014.36.

Beaulieu JMJ, Moles ATA, Leitch IIJ, Bennett MDM, Dickie JJB, Knight CAC. 2007. Correlated evolution of genome size and seed mass. New Phytol. 173:422–437. doi:10.1111/j.1469-8137.2006.01919.x.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. Ann Bot. 95:127–132. doi:10.1093/aob/mci008.

Bonaccorso FJ, Glanz WE, Sandford CM. 1980. Feeding assemblages of mammals at fruiting Dipteryx panamensis (Papilionaceae) trees in Panama: seed predation, dispersal, and parasitim. Rev Biol Trop. 28:61–72.

Brunner AM, Busov VB, Strauss SH. 2004. Poplar genome sequence: functional genomics in an ecologically dominant plant species. Trends Plant Sci. 9:49–56. doi:10.1016/j.tplants.2003.11.006.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 31:1119–1125. doi:10.1038/nbt.2727.

Bushnell B. 2015. BBMap short read aligner, and other bioinformatic tools.

Butterfield RP, Mariano EC. 1995. Screening trial of 14 tropical hardwoods with an emphasis on species native to Costa Rica: fourth year results. New For. 9:135–145.

doi:10.1007/BF00028686.

Cardoso D, Pennington RT, Queiroz LP De, Boatwright JS, Wyk B Van, Wojciechowski MF, Lavin M. 2013. Reconstructing the deep-branching relationships of the papilionoid legumes. South African J Bot. 89:58–75. doi:10.1016/j.sajb.2013.05.001.

Cardoso D, De Queiroz LP, Pennington RT, De Lima HC, Fonty E, Wojciechowski MF, Lavin M. 2012. Revisiting the phylogeny of papilionoid legumes: new insights from comprenhensively sampled early-branching lineages. Am J Bot. 99:1991–2013. doi:10.3732/ajb.1200380.

Cardoso D, São-Mateus WMB, da Cruz DT, Zartman CE, Komura DL, Kite G, Prenner G, Wieringa JJ, Clark A, Lewis G, et al. 2015. Filling in the gaps of the papilionoid legume phylogeny: the enigmatic Amazonian genus Petaladenium is a new branch of the early-diverging Amburaneae clade. Mol Phylogenet Evol. 84:112–124. doi:10.1016/j.ympev.2014.12.015.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Mol Ecol. 22:3124–3140. doi:10.1111/mec.12354.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks : building and genotyping loci de novo from short-read sequences. G3 Genes|Genomes|Genetics. 1:171–182. doi:10.1534/g3.111.000240.

Cavalier-Smith T. 2005. Economy, speed and size matter: Evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 95:147–175. doi:10.1093/aob/mci010.

Chain PSG, Grafham D V., Fulton RS, FitzGerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. 2009. Genome project standards in a new era of sequencing. Science (80-). 326:236–237. doi:10.1126/science.1180614.

Chassot O, Arias GM. 2012. Connectivity conservation of the great green macaw's landscape in Costa Rica and Nicaragua (1994- 2012). Parks. 18:61–70.

Chun SLM. 2008. The utility of digital aerial surveys in censuing Dipteryx panamensis, the key food and nesting tree of the endangered great green macaw (Ara ambigua) in Costa Rica. Duke University.

Chung J, Lee JH, Arumuganathan K, Graef GL, Specht JE. 1998. Relationships between nuclear DNA content and seed and leaf size in soybean. Theor Appl Genet. 96:1064–1068. doi:10.1007/s001220050840.

Corlett RT. 2016. The impacts of droughts in tropical forests. Trends Plant Sci. 21:584–593. doi:10.1016/j.tplants.2016.02.003.

Danusevicius D, Lindgren D. 2002. Two-stage selection strategies in tree breeding considering gain, diversity, time and cost. For Genet. 9:145–157.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 12:499–510. doi:10.1038/nrg3012.

Desta ZA, Ortiz R. 2014. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19:592–601. doi:10.1016/j.tplants.2014.05.006.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 12:1075–1079. doi:10.1101/gr.132102.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. Trends Genet. 30:418–426. doi:10.1016/j.tig.2014.07.001.

Doležel J, Bartoš J. 2005. Plant DNA flow cytometry and estimation of nuclear genome size.

Ann Bot. 95:99–110. doi:10.1093/aob/mci005.

Ducke A. 1940. Revision of the species of the genus Coumarouna Aubl. or Dipteryx Schreb. Trop Woods. 61:1–10.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-Time DNA sequencing from single polymerase molecules. Science (80-). 323:133–138. doi:10.1126/science.1162986.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 29:51–63. doi:10.1016/j.tree.2013.09.008.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 6:1–10. doi:10.1371/journal.pone.0019379.

Estrada-Chavarria A, Rodriguez-Gonzales A, Sanchez-Gonzalez J. 2005. Evaluacion y categorizacion del estado de conservacion de plantas en Costa Rica. San Jose, Costa Rica.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 30:2503–2505. doi:10.1093/bioinformatics/btu314.

Flores E. 1992. Arboles y Semillas del Neotropico. vol. 1. San Jose: Museo Nacional de Costa Rica.

Fu R. 2015. Global warming-accelerated drying in the tropics. Proc Natl Acad Sci. 112:3593–3594. doi:10.1073/pnas.1503231112.

Gaitán-Álvarez J. 2015. Propiedades energéticas de biomasa torrefaccionada de Dipteryx panamensis Pittier y Gmelina arborea Roxb. ex Sm. Rev For Mesoam Kurú. 13:57–62. doi:10.18845/rfmk.v13i30.2461.

Gamboa-Badilla N, Arias Le Claire H. 2008. Regeneración de Dipteryx panamensis en bosques

bajo manejo forestal en el paisaje fragmentado del noreste de Costa Rica. In: V Simposio Internacional sobre Manejo Sostenible de los Recursos Forestales. Universidad de Pinar del Rio: SIMFOR. p. 1–12.

Gerard D, Ferrao LFV, Garcia AAF, Stephens M. 2018. Harnessing empirical bayes and mendelian segregation for genotyping autopolyploids from messy sequencing data. bioRxiv.:1–29. doi:10.1101/281550.

Goldman AD, Landweber LF. 2016. What is a genome? PLoS Genet. 12:1–7. doi:10.1371/journal.pgen.1006181.

Gomez Figueroa P. 2009. Ecología y conservación de la lapa verde (Ara ambigua) en Costa Rica. Rev Posgrado y Soc. 9:58–80. doi:ISSN 1659-178X.

González IM, Origgi LAF. 2003. Comportamiento fenológico del almendro en la zona norte de Costa Rica. Tecnol en Marcha. 16:52–60.

Grattapaglia D, Kirst M. 2008. Eucalyptus applied genomics: from gene sequences to breeding tools. New Phytol. 179:911–929. doi:10.1111/j.1469-8137.2008.02503.x.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR. 2009. Genomics of growth traits in forest trees. Curr Opin Plant Biol. 12:148–156. doi:10.1016/j.pbi.2008.12.008.

Grattapaglia D, Resende MD V. 2011. Genomic selection in forest tree breeding. Tree Genet Genomes. 7:241–255. doi:10.1007/s11295-010-0328-4.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev. 76:65–101. doi:10.1111/j.1469-185X.2000.tb00059.x.

Gregory TR. 2004. Insertion-deletion biases and the evolution of genome size. Gene. 324:15–34. doi:10.1016/j.gene.2003.09.030.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome

assemblies. Bioinformatics. 29:1072–1075. doi:10.1093/bioinformatics/btt086.

Habier D, Fernando RL, Dekkers JCM. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177:2389–2397. doi:10.1534/genetics.107.081190.

Hanson TR, Brunsfeld SJ, Finegan B, Waits LP. 2008a. Characterization of microsatellite markers for the almendro (Dipteryx panamensis), a tetraploid rainforest tree. Mol Ecol Resour. 8:425–427. doi:10.1111/j.1471-8286.2007.01980.x.

Hanson TR, Brunsfeld SJ, Finegan B, Waits LP. 2008b. Pollen dispersal and genetic structure of the tropical tree Dipteryx panamensis in a fragmented Costa Rican landscape. Mol Ecol. 17:2060–2073. doi:10.1111/j.1365-294X.2008.03726.x.

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 Plant Sci. 17:64–72. doi:10.1016/j.tplants.2011.11.005.

Hert DG, Fredlake CP, Barron AE. 2008. Advantages and limitations of next-generation sequencing technologies: A Comparison of electrophoresis and non-electrophoresis methods. Electrophoresis. 29:4618–4626. doi:10.1002/elps.200800456.

Hooker WJ. 1850. Hooker's journal of botany and Kew Garden miscellany. London: Reeve, Benham, and Reeve.

IPCC. 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II, and III to the Fifth Assessement Report of the Intergovermental Panel on Climate Change [Core Writting Team; R.K. Pachauri and L.A. Meyer (eds)]. Geneva, Switzerland.

Isik F. 2014. Genomic selection in forest tree breeding: the concept and an outlook to the future. New For. 45:379–401. doi:10.1007/s11056-014-9422-z.

Isik F, Holland J, Maltecca C. 2017. Genetic Data Analysis for Plant and Animal Breeding. Springer.

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. 2017. ABySS 2 .0 : resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 27:768–777. doi:10.1101/gr.214346.116.Freely.

Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, et al. 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci. 103:19635–19640. doi:10.1073/pnas.0609513103.

Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat Biotechnol. 31:1143–1147. doi:10.1038/nbt.2768.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 33:1870–1874. doi:10.1093/molbev/msw054.

Landesfeind M, Meinicke P. 2014. Predicting the functional repertoire of an organism from unassembled RNA–seq data. BMC Genomics. 15:1003. doi:10.1186/1471-2164-15-1003.

León N, Murillo O, Badilla Y, Ávila C, Murillo R. 2017. Expected genetic gain and genotype by environment interaction in almond (Dipteryx panamensis (Pittier) Rec. and Mell) in Costa Rica. Silvae Genet. 66:9–13. doi:10.1515/sg-2017-0002.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–

2079. doi:10.1093/bioinformatics/btp352.

Lindgren D. 1993. Quantitative comparison between truncation selection and a beter procedure. Hereditas. 118:289–292.

Lodhi M a, Ye G-N, Weeden NF, Reisch BI. 1994. A simple and efficient method for DNA extraction from grapevine cultivars and Vitis species. Plant Mol Biol Report. 12:6–13. doi:10.1007/BF02668658.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 1:1–6.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science (80-). 302:1401–1404. doi:10.1073/pnas.1017572108.

Maddison WP, Maddison DR. 2018. Mesquite: a modular system for evolutionary analysis.

Madriz Vargas B. 2004. Relacion de dependencia directa para la alimentacion y anidacion de la lapa verde (ara Ambigua) y el almendro (Dipteryx panamensis) en la zona norte de Costa Rica. San Jose.

Mapleson D, Accinelli GG, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 33:574–576. doi:10.1093/bioinformatics/btw663.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27:764–770. doi:10.1093/bioinformatics/btr011.

Martínez-Albán V, Fallas-Valverde L, Murillo-Gamboa O, Badilla-Valverde Y. 2016. Potencial de mejoramiento genético en Dipteryx panamensis a los 33 meses de edad en San Carlos, Costa Rica. Rev For Mesoam Kurú. 13:03-12. doi:10.18845/rfmk.v13i30.2455.

Metzker ML. 2010. Sequencing technologies the next generation. Nat Rev Genet. 11:31–46. doi:10.1038/nrg2626.

Michael TP, Jackson S. 2013. The first 50 plant genomes. Plant Genome. 6:1–7. doi:10.3835/plantgenome2013.03.0001in.

Moles AT, Ackerly DD, Webb CO, Tweddle JC, Dickie JB, Westoby M. 2005. A brief history of seed size. Science (80-). 307:576–580. doi:10.1126/science.1104863.

Monge G, Chassot O, Ramírez O, Alemán I. 2012. Temporada de nidificación 2009 de Ara ambiguus y Ara macao en el sureste de Nicaragua y Norte de Costa Rica. Zeledonia. 16:3–14.

Murillo Gómez PA, Atehortúa L. 2013. Cultivos celulares de Choibá Dipteryx oleifera Benth. Rev Colomb Biotecnol. 15:124–131.

Murillo O, Badilla Y, Rojas F. 2010. GENFORES, desde el TEC hacia el desarrollo empresarial internacional. Investig TEC. Setiembre:10–11.

Murillo O, Obando G, Badilla Y, Araya E. 2001. Estrategia de mejoramiento genetico para el programa de conservacion y mejoramiento genetico de especies forestales del ITCR/FUNDECOR, Costa Rica. Rev For Latinoam. 16:273–285.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al. 2014. The genome of Eucalyptus grandis. Nature. 510:356–362. doi:10.1038/nature13308.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol. 22:2841–2847. doi:10.1111/mec.12350.

Neale DB. 2007. Genomics to tree breeding and forest health. Curr Opin Genet Dev. 17:539–

544. doi:10.1016/j.gde.2007.10.002.

Neale DB, Ingvarsson PK. 2008. Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol. 11:149–155. doi:10.1016/j.pbi.2007.12.004.

Neale DB, Kremer A. 2011. Forest tree genomics: Growing resources and applications. Nat Rev Genet. 12:111–122. doi:10.1038/nrg2931.

Neale DB, Wegrzyn JL, Stevens KA, Zimin A V., Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15:1–13. doi:10.1186/gb-2014-15-3-r59.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 12:443–451. doi:10.1038/nrg2986.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics. 9:312. doi:10.1186/1471-2164-9-312.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature. 497:579–584. doi:10.1038/nature12211.

Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, Birol I. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics. 16:1–8. doi:10.1186/s12859-015-0663-4.

Pennington R, Lavin M, Ireland H, Klitgaard B, Preston J, Hu J-M. 2001. Phylogenetic relationships of basal Papilionoid legumes based upon sequences of the chloroplast trnL

intron. Syst Bot. 26:537–556. doi:10.1043/0363-6445-26.3.537.

Perry DR, Starrett A. 1980. The pollination ecology and blooming strategy of a neotropical emergent tree, Dipteryx panamensis. Biotropica. 12:307–313. doi:10.2307/2387702.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 7:e37135. doi:10.1371/journal.pone.0037135.

Petit B, Montagnini F. 2006. Growth in pure and mixed plantations of tree species used in reforesting rural areas of the humid region of Costa Rica, Central America. For Ecol Manage. 233:338–343. doi:10.1016/j.foreco.2006.05.030.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. Trends Genet. 17:23–28. doi:10.1016/S0168-9525(00)02157-0.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One. 7:e32253. doi:10.1371/journal.pone.0032253.

Poland JA, Rife TW. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome. 5:92–102. doi:10.3835/plantgenome2012.05.0005.

Powell JE, Visscher PM, Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet. 11:800–805. doi:10.1038/nrg2865.

R Core Team. 2018. R: A language and environment for statistical computing.

Redondo-Brenes A. 2007. Growth, carbon sequestration, and management of native tree plantations in humid regions of Costa Rica. New For. 34:253–268. doi:10.1007/s11056-007-9052-9.

Ruotsalainen S, Lindgren D. 2000. Stratified sublining: a new option for structuring breeding

populations. Can J For Res. 30:596–604. doi:10.1139/cjfr-30-4-596.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 74:5463–5467. doi:10.1073/pnas.74.12.5463.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. Hum Mol Genet. 19:R227–R240. doi:10.1093/hmg/ddq416.

Schloss JA. 2008. How to get genomes at one ten-thousandth the cost. Nat Biotechnol. 26:1113–1115. doi:10.1038/nbt1008-1113.

Schmidt F. 2009. The effect of site selection on the growth of Dipteryx panamensis in timber plantations in Costa Rica and Panama. University of Technology, Dresden, Germany.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ : 25 years of image analysis HISTORICAL commentary NIH Image to ImageJ : 25 years of image analysis. Nat Methods. 9:671–675. doi:10.1038/nmeth.2089.

Shearer K, Ranney TG. 2013. Ploidy levels and relative genome sizes of species, hybrids, and cultivars of Dogwood (Cornus spp .). Hort Sci. 48:825–830.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi:10.1093/bioinformatics/btv351.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, DePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. Am J Bot. 96:336–348. doi:10.3732/ajb.0800079.

Soltis DE, Soltis PS, Bennett MD, Leitch IJ. 2003. Evolution of genome size in the angiosperms. Am J Bot. 90:1596–1603. doi:10.3732/ajb.90.11.1596.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G,

McKernan K, et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 18:1051–1063. doi:10.1101/gr.076463.108.

Varma A, Padh H, Shrivastava N. 2007. Plant genomic DNA isolation: an art or a science. Biotechnol J. 2:386–392. doi:10.1002/biot.200600195.

Vinson CC, Ribeiro DO, Harris SA, Sampaio I, Ciampi AY. 2009. Isolation of polymorphic microsatellite markers for the tetraploid Dipteryx odorata, an intensely exploited Amazonian tree species. Mol Ecol Resour. 9:1542–1544. doi:10.1111/j.1755-0998.2009.02694.x.

Vozzo J. 2010. Manual de Semillas de Arboles Tropicales. Washington, D.C.: USDA - Forest Service.

Weiß CL, Pais M, Cano LM, Kamoun S, Burbano HA. 2018. nQuire: a statistical framework for ploidy estimation using next generation sequencing. BMC Bioinformatics. 19:22. doi:10.1186/s12859-018-2128-z.

Wendel JF. 2000. Genome evolution in polyploids. Plant Mol Evol. 42:225–249.

Whitney KD, Baack EJ, Hamrick JL, Godt MJW, Barringer BC, Bennett MD, Eckert CG, Goodwillie C, Kalisz S, Leitch IJ, et al. 2010. A role for nonadaptive processes in plant genome size evolution? Evolution (N Y). 64:2097–2109. doi:10.1111/j.1558-5646.2010.00967.x.

Williams A V., Nevill PG, Krauss SL. 2014. Next generation restoration genetics: Applications and opportunities. Trends Plant Sci. 19:529–537. doi:10.1016/j.tplants.2014.03.011.

Wojciechowski MF, Lavin M, Sanderson MJ. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades

within the family. Am J Bot. 91:1846–1862. doi:10.3732/ajb.91.11.1846.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329–342. doi:10.1038/nrg3174.

APPENDICES

Appendix 1: Assembly scripts.

This section contains the scripts used for assembly with the different software and the last step with Sealer. Pieces of bash code are enclosed by {} brackets and in monospaced font, comments are preceded by the # symbols, while commands are preceded by the $ symbol. Unless otherwise specified the parameters presented here represent the optimal after a series of trials. All computing was performed on a Linux-based system with 64 processors and 345 Gb of available RAM.

SOAPdenovo2 v2.04:

```
{
$ SOAPdenovo-63mer all -s <path to config file> -K 63 -p 32 -R -o <output dir
name>
#Configuration file parameters:
                        max_rd_len=150
                        [LIB]
                        avg_ins=350
                        reverse_seq=0
                        asm_flags=3
                        rd_len_cutoff=150
                        rank=1
                        pair_num_cutoff=3
                        map_len=32
                        q1=/path/to/short/reads/R1.fq.gz
                        q2=/path/to/short/reads/R2.fq.gz
}
```

SOAPdenovo2 took approximately 1440 processor hours (45 hours, 32 threads) to finish and reached peak memory consumption at 345Gb (RAM).

ABySS v2.0.2:

```
{
$ abyss-pe j=12 name=<output dir name> k=64 in='/path/to/short/reads/R1.fq.gz
/path/to/short/reads/R2.fq.gz' long=pacbio
pacbio='/path/to/long/reads/long.fq' B=60G H=3 kc=3 v=-v
}
```

ABySS 2.0 took approximately 672 processor hours (56 hours, 12 threads) to finish and reached peak memory consumption at 60Gb (RAM).

CLC Genomics Workbench v10.1.1 with Finish Module v1.7 plug-in (trial version):

CLC Genomics is implemented through a graphical user interface (GUI), not the command line. The following represent the menus, submenus, and options selected at each step. For every process if a parameter or option value is not specified then assume defaults.

File > Import > Illumina

input = short reads, with remove failed reads = enable and PE insert = 1-1000, output = PE_reads (R1.fq.gz and R2.fq.gz paired in a single file).

De Novo Sequencing > De Novo Assembly 1.3

input = PE_reads, word size = auto (26), bubble size = auto (50), contig length = 200, mapping mode = create simple contig sequence (fast), Perform scaffolding = Yes, Auto-detect paired distances = Yes, Guidance only reads = No, Min distance = 1, Max distance = 1000, output = PE_contigs

Finishing Module > Error correction

input = pacbio subreads, coverage = 40% (corrects 40% of the longest reads using the remaining ones), output = pacbio_corrected

Finishing Module > Join Contigs

input = PE_contigs + pacbio_corrected + pacbio subreads, based contig joining on long reads support, Output = clc_assembly.fa

CLC Genomics took approximately 119 processor hours (119 hours, single thread) and reached peak memory consumption at 48Gb.

Sealer (as implemented in ABySS 2.0.2):

```
{
#First run: test broad range of k-mer sizes.
#-k needs to be: k-mer used in assembly < -k < 100.
#For example, optimal k-mer size +11, +21, +31, etc.
$ /abyss-2.0.2/Sealer/abyss-sealer -b90G -k<+11> -k<+21> -k<+31> -k<+41> -o
<output dir name> -S <path/to/assembly> -j 32 -P 10
<path/to/short/reads/R?.fq.gz>

#Second run: test narrow range of -k, based on best results from 1st run.
#For example, k-mer size which close the most gaps -4, -2, +2, +4, etc.
$ /abyss-2.0.2/Sealer/abyss-sealer -b90G -k<-4> -k<-2> -k<+2> -k<+4> -o
<output dir name> -S <path/to/assembly/from/first/run> -j 32 -P 10
<path/to/short/reads/R?.fq.gz>
}
```

The k-mer size selected at each run and for each assembly varied depending on the optimal parameters used to generate the assembly. For example, in the case of the CLC Genomics assembly the first run used k-mer sizes values (-k) of 31, 41, 51, 61, 71, 81 and 91. This resulted in 299,629 gaps closed (44.22% of total gaps detected), with -k = 51 closing the most gaps. For the second run -k values were set to 47, 49, 53 and 55. This resulted in 1,786 additional gaps closed (0.47% of total gaps detected). The process can be continued for a third or fourth run, but the gains are minimal.

Appendix 2: GBS Adapters.

Barcoded adapter sequences:

PstI_topA01	5' - CAC GAC GCT CTT CCG ATC TAT GTC CTG CA - 3'
PstI_botA01	5' - GGA CAT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topA02	5' - CAC GAC GCT CTT CCG ATC TAG ATG CAT GCA - 3'
PstI_botA02	5' - TGC ATC TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topA03	5' - CAC GAC GCT CTT CCG ATC TTT CTG AGG TGC A - 3'
PstI_botA03	5' - CCT CAG AAA GAT CGG AAG AGC GTC GTG - 3'
PstI_topA04	5' - CAC GAC GCT CTT CCG ATC TAG GTG TAC GTG CA - 3'
PstI_botA04	5' - CGT ACA CCT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topA05	5' - CAC GAC GCT CTT CCG ATC TTC CTA AGC ACT GCA - 3'
PstI_botA05	5' - GTG CTT AGG AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topA06	5' - CAC GAC GCT CTT CCG ATC TCG CCA GAC TTA TGC A - 3'
PstI_botA06	5' - TAA GTC TGG CGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topA07	5' - CAC GAC GCT CTT CCG ATC TGT TGG ATG CA - 3'
PstI_botA07	5' - TCC AAC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topA08	5' - CAC GAC GCT CTT CCG ATC TCG CTG ATT GCA - 3'
PstI_botA08	5' - ATC AGC GAG ATC GGA AGA GCG TCG TG - 3'
PstI_topA09	5' - CAC GAC GCT CTT CCG ATC TCA CAG ACT TGC A - 3'
PstI_botA09	5' - AGT CTG TGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topA10	5' - CAC GAC GCT CTT CCG ATC TAC CAG TCC ATG CA - 3'
PstI_botA10	5' - TGG ACT GGT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topA11	5' - CAC GAC GCT CTT CCG ATC TAA GTG TGA ACT GCA - 3'
PstI_botA11	5' - GTT CAC ACT TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topA12	5' - CAC GAC GCT CTT CCG ATC TGT CGC AGA GAA TGC A - 3'
PstI_botA12	5' - TTC TCT GCG ACA GAT CGG AAG AGC GTC GTG - 3'
PstI_topB01	5' - CAC GAC GCT CTT CCG ATC TAA TCG CTG CA - 3'
PstI_botB01	5' - GCG ATT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topB02	5' - CAC GAC GCT CTT CCG ATC TCG CAA TTT GCA - 3'
PstI_botB02	5' - AAT TGC GAG ATC GGA AGA GCG TCG TG - 3'
PstI_topB03	5' - CAC GAC GCT CTT CCG ATC TCT CAG AAG TGC A - 3'
PstI_botB03	5' - CTT CTG AGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topB04	5' - CAC GAC GCT CTT CCG ATC TCT CAT GCA GTG CA - 3'
PstI_botB04	5' - CTG CAT GAG AGA TCG GAA GAG CGT CGT G - 3'
PstI_topB05	5' - CAC GAC GCT CTT CCG ATC TCA TGG CGA ATT GCA - 3'
PstI_botB05	5' - ATT CGC CAT GAG ATC GGA AGA GCG TCG TG - 3'
PstI_topB06	5' - CAC GAC GCT CTT CCG ATC TCA ATC TCA GGA TGC A - 3'
PstI_botB06	5' - TCC TGA GAT TGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topB07	5' - CAC GAC GCT CTT CCG ATC TTA TGC GTG CA - 3'
PstI_botB07	5' - CGC ATA AGA TCG GAA GAG CGT CGT G - 3'
PstI_topB08	5' - CAC GAC GCT CTT CCG ATC TAA GCT GCT GCA - 3'
PstI_botB08	5' - GCA GCT TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topB09	5' - CAC GAC GCT CTT CCG ATC TTA AGC GCA TGC A - 3'
PstI_botB09	5' - TGC GCT TAA GAT CGG AAG AGC GTC GTG - 3'

PstI_topB10 5' - CAC GAC GCT CTT CCG ATC TTC GGA CAA CTG CA - 3'
PstI_botB10 5' - GTT GTC CGA AGA TCG GAA GAG CGT CGT G - 3'
PstI_topB11 5' - CAC GAC GCT CTT CCG ATC TAA CCT CGC ACT GCA - 3'
PstI_botB11 5' - GTG CGA GGT TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topB12 5' - CAC GAC GCT CTT CCG ATC TAA TCC ACC AGT TGC A - 3'
PstI_botB12 5' - ACT GGT GGA TTA GAT CGG AAG AGC GTC GTG - 3'
PstI_topC01 5' - CAC GAC GCT CTT CCG ATC TAA GCG ATG CA - 3'
PstI_botC01 5' - TCG CTT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topC02 5' - CAC GAC GCT CTT CCG ATC TAA TCA GGT GCA - 3'
PstI_botC02 5' - CCT GAT TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topC03 5' - CAC GAC GCT CTT CCG ATC TCA ACC GTA TGC A - 3'
PstI_botC03 5' - TAC GGT TGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topC04 5' - CAC GAC GCT CTT CCG ATC TAT GAG GAA CTG CA - 3'
PstI_botC04 5' - GTT CCT CAT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topC05 5' - CAC GAC GCT CTT CCG ATC TTC ATC GGA ATT GCA - 3'
PstI_botC05 5' - ATT CCG ATG AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topC06 5' - CAC GAC GCT CTT CCG ATC TAC AAC TCC AAC TGC A - 3'
PstI_botC06 5' - GTT GGA GTT GTA GAT CGG AAG AGC GTC GTG - 3'
PstI_topC07 5' - CAC GAC GCT CTT CCG ATC TTC GAC TTG CA - 3'
PstI_botC07 5' - AGT CGA AGA TCG GAA GAG CGT CGT G - 3'
PstI_topC08 5' - CAC GAC GCT CTT CCG ATC TAC TGA GCT GCA - 3'
PstI_botC08 5' - GCT CAG TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topC09 5' - CAC GAC GCT CTT CCG ATC TCA TTC GTC TGC A - 3'
PstI_botC09 5' - GAC GAA TGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topC10 5' - CAC GAC GCT CTT CCG ATC TAG TGT GCC ATG CA - 3'
PstI_botC10 5' - TGG CAC ACT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topC11 5' - CAC GAC GCT CTT CCG ATC TTA AGC GGC ATT GCA - 3'
PstI_botC11 5' - ATG CCG CTT AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topC12 5' - CAC GAC GCT CTT CCG ATC TCA TCA GGA CAC TGC A - 3'
PstI_botC12 5' - GTG TCC TGA TGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topD01 5' - CAC GAC GCT CTT CCG ATC TGA ACG TTG CA - 3'
PstI_botD01 5' - ACG TTC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topD02 5' - CAC GAC GCT CTT CCG ATC TGG ACA AGT GCA - 3'
PstI_botD02 5' - CTT GTC CAG ATC GGA AGA GCG TCG TG - 3'
PstI_topD03 5' - CAC GAC GCT CTT CCG ATC TTC GTG CAT TGC A - 3'
PstI_botD03 5' - ATG CAC GAA GAT CGG AAG AGC GTC GTG - 3'
PstI_topD04 5' - CAC GAC GCT CTT CCG ATC TTT CTA TCC GTG CA - 3'
PstI_botD04 5' - CGG ATA GAA AGA TCG GAA GAG CGT CGT G - 3'
PstI_topD05 5' - CAC GAC GCT CTT CCG ATC TTC GAC TAC ATT GCA - 3'
PstI_botD05 5' - ATG TAG TCG AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topD06 5' - CAC GAC GCT CTT CCG ATC TAC AAG GCA CGT TGC A - 3'
PstI_botD06 5' - ACG TGC CTT GTA GAT CGG AAG AGC GTC GTG - 3'
PstI_topD07 5' - CAC GAC GCT CTT CCG ATC TCG GAA TTG CA - 3'
PstI_botD07 5' - ATT CCG AGA TCG GAA GAG CGT CGT G - 3'
PstI_topD08 5' - CAC GAC GCT CTT CCG ATC TGT TAC GTT GCA - 3'
PstI_botD08 5' - ACG TAA CAG ATC GGA AGA GCG TCG TG - 3'

PstI_topD09	5' - CAC GAC GCT CTT CCG ATC TGG TTG AAC TGC A - 3'
PstI_botD09	5' - GTT CAA CCA GAT CGG AAG AGC GTC GTG - 3'
PstI_topD10	5' - CAC GAC GCT CTT CCG ATC TCC TGA CAC ATG CA - 3'
PstI_botD10	5' - TGT GTC AGG AGA TCG GAA GAG CGT CGT G - 3'
PstI_topD11	5' - CAC GAC GCT CTT CCG ATC TTC TCA AGA ACT GCA - 3'
PstI_botD11	5' - GTT CTT GAG AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topD12	5' - CAC GAC GCT CTT CCG ATC TAA GGT AAC CAC TGC A - 3'
PstI_botD12	5' - GTG GTT ACC TTA GAT CGG AAG AGC GTC GTG - 3'
PstI_topE01	5' - CAC GAC GCT CTT CCG ATC TTG ACC ATG CA - 3'
PstI_botE01	5' - TGG TCA AGA TCG GAA GAG CGT CGT G - 3'
PstI_topE02	5' - CAC GAC GCT CTT CCG ATC TTG GAG TAT GCA - 3'
PstI_botE02	5' - TAC TCC AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topE03	5' - CAC GAC GCT CTT CCG ATC TGT GTC CTT TGC A - 3'
PstI_botE03	5' - AAG GAC ACA GAT CGG AAG AGC GTC GTG - 3'
PstI_topE04	5' - CAC GAC GCT CTT CCG ATC TCC GTT AAG GTG CA - 3'
PstI_botE04	5' - CCT TAA CGG AGA TCG GAA GAG CGT CGT G - 3'
PstI_topE05	5' - CAC GAC GCT CTT CCG ATC TTG GTT GGA ATT GCA - 3'
PstI_botE05	5' - ATT CCA ACC AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topE06	5' - CAC GAC GCT CTT CCG ATC TGG ACA TGA TGT TGC A - 3'
PstI_botE06	5' - ACA TCA TGT CCA GAT CGG AAG AGC GTC GTG - 3'
PstI_topE07	5' - CAC GAC GCT CTT CCG ATC TCC ACA ATG CA - 3'
PstI_botE07	5' - TTG TGG AGA TCG GAA GAG CGT CGT G - 3'
PstI_topE08	5' - CAC GAC GCT CTT CCG ATC TAG AGT GGT GCA - 3'
PstI_botE08	5' - CCA CTC TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topE09	5' - CAC GAC GCT CTT CCG ATC TCG GTA ATC TGC A - 3'
PstI_botE09	5' - GAT TAC CGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topE10	5' - CAC GAC GCT CTT CCG ATC TAG TGT CAA CTG CA - 3'
PstI_botE10	5' - GTT GAC ACT AGA TCG GAA GAG CGT CGT G - 3'
PstI_topE11	5' - CAC GAC GCT CTT CCG ATC TTT CAG AAC AGT GCA - 3'
PstI_botE11	5' - CTG TTC TGA AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topE12	5' - CAC GAC GCT CTT CCG ATC TAA CAC CGC TTC TGC A - 3'
PstI_botE12	5' - GAA GCG GTG TTA GAT CGG AAG AGC GTC GTG - 3'
PstI_topF01	5' - CAC GAC GCT CTT CCG ATC TCT CAC ATG CA - 3'
PstI_botF01	5' - TGT GAG AGA TCG GAA GAG CGT CGT G - 3'
PstI_topF02	5' - CAC GAC GCT CTT CCG ATC TCA AGA TGT GCA - 3'
PstI_botF02	5' - CAT CTT GAG ATC GGA AGA GCG TCG TG - 3'
PstI_topF03	5' - CAC GAC GCT CTT CCG ATC TAC ACC TCA TGC A - 3'
PstI_botF03	5' - TGA GGT GTA GAT CGG AAG AGC GTC GTG - 3'
PstI_topF04	5' - CAC GAC GCT CTT CCG ATC TGA CGT GAT GTG CA - 3'
PstI_botF04	5' - CAT CAC GTC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topF05	5' - CAC GAC GCT CTT CCG ATC TGT CAA TGC ACT GCA - 3'
PstI_botF05	5' - GTG CAT TGA CAG ATC GGA AGA GCG TCG TG - 3'
PstI_topF06	5' - CAC GAC GCT CTT CCG ATC TGT GTT ACC TGA TGC A - 3'
PstI_botF06	5' - TCA GGT AAC ACA GAT CGG AAG AGC GTC GTG - 3'
PstI_topF07	5' - CAC GAC GCT CTT CCG ATC TGT GAT GTG CA - 3'
PstI_botF07	5' - CAT CAC AGA TCG GAA GAG CGT CGT G - 3'

PstI_topF08	5' - CAC GAC GCT CTT CCG ATC TTG AGT CCT GCA - 3'
PstI_botF08	5' - GGA CTC AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topF09	5' - CAC GAC GCT CTT CCG ATC TGT CGT TAC TGC A - 3'
PstI_botF09	5' - GTA ACG ACA GAT CGG AAG AGC GTC GTG - 3'
PstI_topF10	5' - CAC GAC GCT CTT CCG ATC TGA TCA GGT GTG CA - 3'
PstI_botF10	5' - CAC CTG ATC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topF11	5' - CAC GAC GCT CTT CCG ATC TGT AGG CGA ACT GCA - 3'
PstI_botF11	5' - GTT CGC CTA CAG ATC GGA AGA GCG TCG TG - 3'
PstI_topF12	5' - CAC GAC GCT CTT CCG ATC TCA ACC TAC TCT TGC A - 3'
PstI_botF12	5' - AGA GTA GGT TGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topG01	5' - CAC GAC GCT CTT CCG ATC TGT TAG GTG CA - 3'
PstI_botG01	5' - CCT AAC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topG02	5' - CAC GAC GCT CTT CCG ATC TGC ACG AAT GCA - 3'
PstI_botG02	5' - TTC GTG CAG ATC GGA AGA GCG TCG TG - 3'
PstI_topG03	5' - CAC GAC GCT CTT CCG ATC TGT CCT TGA TGC A - 3'
PstI_botG03	5' - TCA AGG ACA GAT CGG AAG AGC GTC GTG - 3'
PstI_topG04	5' - CAC GAC GCT CTT CCG ATC TGA CAT ACA CTG CA - 3'
PstI_botG04	5' - GTG TAT GTC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topG05	5' - CAC GAC GCT CTT CCG ATC TGG TCT CCA AGT GCA - 3'
PstI_botG05	5' - CTT GGA GAC CAG ATC GGA AGA GCG TCG TG - 3'
PstI_topG06	5' - CAC GAC GCT CTT CCG ATC TGC TGG AAG TGT TGC A - 3'
PstI_botG06	5' - ACA CTT CCA GCA GAT CGG AAG AGC GTC GTG - 3'
PstI_topG07	5' - CAC GAC GCT CTT CCG ATC TGA CTT CTG CA - 3'
PstI_botG07	5' - GAA GTC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topG08	5' - CAC GAC GCT CTT CCG ATC TAC GAT CCT GCA - 3'
PstI_botG08	5' - GGA TCG TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topG09	5' - CAC GAC GCT CTT CCG ATC TGC GCA TAT TGC A - 3'
PstI_botG09	5' - ATA TGC GCA GAT CGG AAG AGC GTC GTG - 3'
PstI_topG10	5' - CAC GAC GCT CTT CCG ATC TGG TAT GGA ATG CA - 3'
PstI_botG10	5' - TTC CAT ACC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topG11	5' - CAC GAC GCT CTT CCG ATC TTT GCG CCA AGT GCA - 3'
PstI_botG11	5' - CTT GGC GCA AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topG12	5' - CAC GAC GCT CTT CCG ATC TCC TCT TCT TCC TGC A - 3'
PstI_botG12	5' - GGA AGA AGA GGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topH01	5' - CAC GAC GCT CTT CCG ATC TCC ATA CTG CA - 3'
PstI_botH01	5' - GTA TGG AGA TCG GAA GAG CGT CGT G - 3'
PstI_topH02	5' - CAC GAC GCT CTT CCG ATC TTC CAC CTT GCA - 3'
PstI_botH02	5' - AGG TGG AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topH03	5' - CAC GAC GCT CTT CCG ATC TCC TAA CAG TGC A - 3'
PstI_botH03	5' - CTG TTA GGA GAT CGG AAG AGC GTC GTG - 3'
PstI_topH04	5' - CAC GAC GCT CTT CCG ATC TTG GTA CGT CTG CA - 3'
PstI_botH04	5' - GAC GTA CCA AGA TCG GAA GAG CGT CGT G - 3'
PstI_topH05	5' - CAC GAC GCT CTT CCG ATC TAG AAC CAC ATT GCA - 3'
PstI_botH05	5' - ATG TGG TTC TAG ATC GGA AGA GCG TCG TG - 3'
PstI_topH06	5' - CAC GAC GCT CTT CCG ATC TCG AAC GCG TAT TGC A - 3'
PstI_botH06	5' - ATA CGC GTT CGA GAT CGG AAG AGC GTC GTG - 3'

PstI_topH07	5' - CAC GAC GCT CTT CCG ATC TGA TGA CTG CA - 3'
PstI_botH07	5' - GTC ATC AGA TCG GAA GAG CGT CGT G - 3'
PstI_topH08	5' - CAC GAC GCT CTT CCG ATC TGG TAT TGT GCA - 3'
PstI_botH08	5' - CAA TAC CAG ATC GGA AGA GCG TCG TG - 3'
PstI_topH09	5' - CAC GAC GCT CTT CCG ATC TAC AGC ACT TGC A - 3'
PstI_botH09	5' - AGT GCT GTA GAT CGG AAG AGC GTC GTG - 3'
PstI_topH10	5' - CAC GAC GCT CTT CCG ATC TTT GTA CCG GTG CA - 3'
PstI_botH10	5' - CCG GTA CAA AGA TCG GAA GAG CGT CGT G - 3'
PstI_topH11	5' - CAC GAC GCT CTT CCG ATC TTT ACA CCA ACT GCA - 3'
PstI_botH11	5' - GTT GGT GTA AAG ATC GGA AGA GCG TCG TG - 3'
PstI_topH12	5' - CAC GAC GCT CTT CCG ATC TGG AGT TAG TCC TGC A - 3'
PstI_botH12	5' - GGA CTA ACT CCA GAT CGG AAG AGC GTC GTG - 3'

Common adapter sequence:

MspI_top2.01	5' - GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T - 3'
MspI_Newbot	5' - CGA GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT C - 3'

Appendix 3: Demultiplexing and variant calling script.

This section contains the scripts used for demultiplexing, cleaning, and variant calling with Stacks version 2.0. Pieces of bash code are enclosed by {} brackets and in monospaced font, comments are preceded by the # symbols, while commands are preceded by the $ symbol. Unless otherwise specified the parameters presented here represent the optimal after a series of trials. All computing was performed on a Linux-based system with 64 processors and 345 Gb of available RAM.

Cleaning and demultiplexing:

GBS files were split by index by the GSL. The process of demultiplexing and cleaning is the same for each library index, just the barcode file changes. Barcode files are composed of two columns, tab separated; the first column contains the barcode sequence while the second column has the corresponding sample name.

```
{

$ mkdir clean_reads
$ process_radtags -f path/to/raw/reads/*.fq.gz -i gzfastq -o clean_reads/ -b
path/to/barcode_file -e pstI -E phred33 -c -q -r -t 80
}
```

After demultiplexing and cleaning, 656,660,293 or 90.9 % of the reads were retained, 4.9 % were dropped due to ambiguous barcodes, 0.9 % were dropped due to low quality, and 3.2 % were dropped to ambiguous RAD-tags.

Variant calling:

```
{
#Call variants using Stacks v2.0, without a reference, step by step
#Create popmap with 1=PV, 2=CSJ+SM, and 3=Crucitas
#Create unique stacks (ustacks)
$ mkdir stacks_results
$ ID=1
$ for file in clean_reads/*.fq.gz; do ustacks -f ${file} -o stacks_results/ -
i $ID -t gzfastq -p12 -d -m 3 -M 4 --max_locus_stacks 4 --gapped; ID=$(expr
$ID + 1); done
#Build catalog (cstacks) using only a fraction of the samples (n=85).
$ cstacks -b 1 -P stacks_results/ -M path/to/popmap/ -n 4 -p 12
#Create population stacks (sstacks)
$ sstacks -b1 -P stacks_results/ -M path/to/popmap/ -p 12
#Transpose the data so it is stored by locus (tsv2bam)
$ tsv2bam -P stacks_results/ -M path/to/popmap/ -t 12
#Build a contig from the metapopulation data, align stacks per sample, #call
variant (gstacks)
$ gstacks -P stacks_results/ -M path/to/popmap/ -t 12
#Run populations module Calculate F statistics and output a vcf file.
$ populations -P stacks_results/ -M path/to/popmap/ -t 12 -r 0.9 -p 2 --
fstats --hwe --vcf
}
```

The mean coverage per sample was 21.7x. Despite the low coverage, this produced a total of 4,691 SNPs calls with missing data ≤ 10 %.

Appendix 4: Relationship matrix and linear mixed model

This section contains the scripts used to estimate the allele dosage information with R package Updog, generate the relationship matrices with R package AGHmatrix, and to run the linear mixed model with ASReml. Pieces of code are enclosed by {} brackets and in monospaced font, comments are preceded by the # symbols. Unless otherwise specified the parameters presented here represent the optimal after a series of trials. All computing was performed on a Windows10 based system with 4 processors and 16 Gb of available RAM.

Allele dosage

```
{
#Set working directory
setwd("Path/to/working/directory/”)
#Load required library
library(ggplot2)
library(reshape2)
library(dplyr)
library(tibble)
library(updog)
library(AGHmatrix)

#Load data into R, a table with total counts and reference allele counts
#per individual per locus extracted from the Stacks vcf file. Also, header
#information per variable.
snps <- read.table("stacks.snp.data", as.is=TRUE)
header <- scan("stacks.header.line", what="character", sep="\t")
names(snps) <- header
```

```
#Format data for mupdog (Multi-SNP updog). Requires a list of 3 matrix:
#sizemat= a matrix of total read depth (DP), refmat= a matrix of reference
#allele counts (AD), and ploidy= value or vector. Individuals in rows and
#SNPs in cols.
loci.ID <- paste(snps$POS, snps$ID, sep="_")    #A vector of snp ID
sample.ID <- header[6:195]                      #A vector of samples ID
ref <- snps[,196:385]
ref <- as.matrix(as.data.frame(lapply(ref, as.numeric)))
colnames(ref) <- sample.ID
rownames(ref) <- loci.ID
ref <- t(ref)                                   #=refmat
size <- snps[,6:195]
size <- as.matrix(as.data.frame(lapply(size, as.numeric)))
colnames(size) <- sample.ID
rownames(size) <- loci.ID
size <- t(size)                                 #=sizemat

#First run, use default parameters, includes missing data.
test.all <- mupdog(refmat=ref, sizemat=size, ploidy=4, verbose=TRUE,
control=list(obj_tol=10^-4))
#Finished after 29 iterations, takes about 6 hours.
#iteration: 29, objective: -1755856, err: 9.762214e-05

#Summaries of output
plot(test.all, 10)
hist(test.all$bias, xlab="Allelic Bias", main=NULL)
hist(test.all$seq, xlab="Sequencing Error", main=NULL)
hist(test.all$od, xlab="Overdispersion", main=NULL)
hist(test.all$inbreeding, xlab="Inbreeding", main=NULL)
```

```
hist(test.all$allele_freq)

#Filter data by max post prob. First, transpose matrix so loci are in rows
#and ind.id in columns. Add rownames as a column and transform to data
#frame for easier manipulation.
mpp <- test.all$maxpostprob
colnames(mpp) <- loci.ID
mpp <- t(mpp)
mpp <- as.data.frame(mpp)
mpp <- rownames_to_column(mpp)
colnames(mpp) <- c("loci", sample.ID)
#Remove 5 ind (trees were cut in 2015)
mpp$do_322 <- NULL
mpp$do_368 <- NULL
mpp$do_505 <- NULL
mpp$do_545 <- NULL
mpp$do_690 <- NULL

#Estimate average max post prob per locus across all ind.
means.mpp <- data.frame(loci=mpp[,1], mean.mpp=rowMeans(mpp[,-1]))

#Filter the map dosage matrix by max post prob.
md <- test.all$map_dosage
colnames(md) <- loci.ID
md <- t(md)
md <- as.data.frame(md)
md <- rownames_to_column(md)
colnames(md) <- c("loci", sample.ID)
#Remove 5 ind (trees were cut in 2015)
```

```
md$do_322 <- NULL

md$do_368 <- NULL

md$do_505 <- NULL

md$do_545 <- NULL

md$do_690 <- NULL

#then add new column with mean mpp

md$mean.mpp <- means.mpp$mean.mpp


#Filter by mpp (max.prob > 0.9), map dosages, transpose so ind are in row

#and markers in columns


md9 <- filter(md, mean.mpp>=0.9)

md9 <- column_to_rownames(md9, var="loci")

md9 <- md9[,1:185]

md9 <- t(md9)                                    #Results in 2612 SNPs

}
```

Relationship matrix

```
{
###G-Matrix
#Additive relationship matrix
Gv9 <- Gmatrix(SNPmatrix=md9, method="VanRaden", maf=0, ploidy=4)
#Full-autopolyploid matrix
Gf9 <- Gmatrix(SNPmatrix=md9, method="Slater", maf=0, ploidy=4)

###A-Matrix
#Load data, pedigree file from 19 parents + 625 progeny
ped.dipteryx19 <- read.csv("ped_dipteryx19.csv", header = TRUE)
#Compute A-matrix, as in Kerr et al (2012) - additive
Ak19 <- Amatrix(ped.dipteryx19, ploidy=4)              #
#Compute A-matrix, as in Slater (2014) - full
As19 <- Amatrix(ped.dipteryx19, ploidy=4, slater=TRUE) #

###H-Matrix
#Compute H-matrix (additive = Ha)
H.ak19.gv9.maf0 <- Hmatrix(A=Ak19, G=Gv9, markers=md9, maf=0, ploidy=4)
#Compute H-matrix (full = Hf)
H.as19.gf9.maf0 <- Hmatrix(A=As19, G=Gf9, markers=md9, maf=0, ploidy=4)

#Format the matrix to import into ASReml
formatmatrix(H.ak19.gv9.maf0, round.by=12, exclude.0=TRUE, name="Ha.grm")
formatmatrix(H.as19.gf9.maf0, round.by=12, exclude.0=TRUE, name="Hf.grm")
}
```

Linear mixed model

```
{
!WORKSPACE 100 !RENAME !DOPART 1
Title: Almendro_2016.
#treeid,fam,prov,site,block,diameter,height,volume
 Treeid     !P     #Tree ID=do_xxx, !P=link to pedigree file
 Fam        !A 19  #Fam=Female/Family ID, !A=alphanumeric
 prov       !I 3   #Provenance (1=#, 2=CSJ+SM, 3=PV), !I=integer
 site       !I 2   #Planting sites (1=NA, 2=SP), !I=integer
 block      !I 6   #Blocks, 6 per site, !I=integer
 diameter          #DBH (in cm)
 height            #Total height (in m)
 volume            #Total volume (in m3)

ped_dipteryx19.csv   !ALPHA !SKIP 1 #Alphanumeric, skip header
Ha.grm               !NSD           #!NSD=Allow negative numbers in matrix
#Hf.grm              !NSD           #!NSD=Allow negative numbers in matrix
Almendro_2016.csv    !SKIP 1        #Data file, skip header

#Individual model, A-matrix
#use nrm=numerator matrix, estimated from the pedigree file

!PART 1
!CYCLE diameter height volume
$A ~ mu prov site !r block nrm(treeid)
     residual units
#Calculates heritability from variance components, output in .pvc file
VPREDICT !DEFINE
```

```
F Additive nrm(treeid)
F Pheno nrm(treeid) + Residual
H h2i Additive Pheno

#Individual model, H-matrix
#use grm=genomic relationship matrix (Ha=additive, Hf=full)
#for Hf uncomment line and comment out Ha.grm, redo part 2
!PART 2
!CYCLE diameter height volume
$A ~ mu prov site !r block grm(treeid)
     residual units
#Calculates heritability from variance components, output in .pvc file
VPREDICT !DEFINE
F Additive grm(treeid)
F Pheno grm(treeid) + Residual
H h2i Additive Pheno
}
```

Appendix 5: Ranking.

Top 100 individuals ranked based on H_f model volume breeding values:

Rank	Tree ID	Family	BV	Accuracy
1	do_536	PV3	0.1593	0.81
2	do_310	PV3	0.1584	0.84
3	do_802	PV3	0.1548	0.84
4	do_381	PV3	0.1500	0.84
5	do_050	6	0.1496	0.83
6	do_828	PV3	0.1493	0.84
7	do_005	PV3	0.1485	0.85
8	do_521	PV3	0.1480	0.84
9	do_569	PV3	0.1456	0.84
10	do_658	CSJ2	0.1454	0.83
11	do_155	PV3	0.1452	0.83
12	do_326	PV10	0.1452	0.75
13	do_057	10	0.1446	0.85
14	do_818	PV3	0.1431	0.84
15	do_592	PV3	0.1427	0.84
16	do_475	PV3	0.1421	0.84
17	do_705	PV3	0.1421	0.84
18	do_509	PV3	0.1414	0.84
19	do_835	PV3	0.1407	0.84
20	do_308	PV3	0.1407	0.84
21	do_510	10	0.1404	0.83
22	do_234	PV3	0.1397	0.84
23	do_024	CSJ2	0.1395	0.86
24	do_441	CSJ6	0.1391	0.83
25	do_819	PV3	0.1383	0.84
26	do_574	10	0.1377	0.84
27	do_613	PV3	0.1365	0.83
28	do_085	PV3	0.1349	0.83
29	do_034	CSJ7	0.1346	0.85
30	do_826	CSJ2	0.1344	0.83
31	do_809	PV3	0.1340	0.84
32	do_309	PV3	0.1339	0.84
33	do_524	CSJ2	0.1334	0.83
34	do_540	9	0.1333	0.82
35	do_225	10	0.1332	0.84
36	do_655	PV3	0.1331	0.84
37	do_678	PV3	0.1324	0.84

38	do_191	CSJ7	0.1324	0.86
39	do_068	CSJ3	0.1323	0.84
40	do_372	10	0.1323	0.84
41	do_422	CSJ1	0.1321	0.88
42	do_079	10	0.1316	0.78
43	do_734	10	0.1316	0.84
44	do_233	PV3	0.1313	0.84
45	do_227	10	0.1311	0.84
46	do_548	4	0.1309	0.87
47	do_074	4	0.1307	0.85
48	do_638	PV3	0.1303	0.84
49	do_253	CSJ1	0.1303	0.83
50	do_721	10	0.1301	0.84
51	do_064	PV8	0.1298	0.86
52	do_745	PV3	0.1297	0.87
53	do_215	6	0.1290	0.84
54	do_620	PV9	0.1288	0.82
55	do_829	4	0.1287	0.84
56	do_363	6	0.1283	0.84
57	do_256	CSJ2	0.1281	0.83
58	do_462	PV3	0.1280	0.84
59	do_834	CSJ2	0.1279	0.83
60	do_391	PV8	0.1278	0.84
61	do_427	CSJ2	0.1277	0.83
62	do_485	PV8	0.1276	0.85
63	do_616	CSJ1	0.1275	0.83
64	do_841	CSJ7	0.1268	0.83
65	do_840	CSJ2	0.1268	0.83
66	do_031	CSJ6	0.1268	0.84
67	do_727	PV3	0.1266	0.84
68	do_006	PV3	0.1266	0.84
69	do_822	4	0.1265	0.84
70	do_146	10	0.1265	0.84
71	do_343	CSJ7	0.1265	0.82
72	do_838	CSJ7	0.1263	0.83
73	do_420	10	0.1262	0.84
74	do_424	CSJ1	0.1262	0.83
75	do_328	CSJ2	0.1259	0.83
76	do_711	10	0.1259	0.84
77	do_562	CSJ3	0.1256	0.88
78	do_636	PV8	0.1256	0.74
79	do_380	PV3	0.1256	0.84
80	do_561	10	0.1255	0.86
81	do_702	6	0.1253	0.84
82	do_319	PV8	0.1252	0.84

83	do_396	PV10	0.1251	0.76
84	do_448	SM9	0.1250	0.90
85	do_696	10	0.1248	0.84
86	do_692	PV3	0.1247	0.84
87	do_717	CSJ2	0.1247	0.83
88	do_378	PV3	0.1247	0.84
89	do_418	10	0.1241	0.84
90	do_549	CSJ2	0.1239	0.83
91	do_676	6	0.1237	0.84
92	do_515	CSJ7	0.1235	0.88
93	do_598	10	0.1235	0.84
94	do_597	CSJ3	0.1235	0.86
95	do_459	CSJ2	0.1232	0.83
96	do_147	10	0.1230	0.84
97	do_176	CSJ2	0.1228	0.87
98	do_527	6	0.1228	0.85
99	do_641	6	0.1227	0.84
100	do_106	CSJ1	0.1226	0.83