ABSTRACT

JACKSON, COLIN ALAN. Single Nucleotide Polymorphism Discovery and Probe Design Using RNA-Seq and Targeted Capture in Tropical and Subtropical Pine Species (Under the direction of Dr. Juan Jose Acosta and Dr. Fikret Isik).

Mexico and Central America are two primary pockets of diversity for tropical and subtropical pine species. Of the species found there, many are considered threatened or endangered, which in turn has spurred conservation efforts and the establishment of landrace populations for commercial use across South America and Southern Africa. The development of genomic resources in *Pinus taeda* has created a platform to extend these technologies into tropical pines. The information gained from these resources will prove to be valuable for the continued improvement of pure and hybrid breeding populations, as well as gene conservation efforts. In this study, we performed targeted genome and gene-based SNP discovery for the development of a high-throughput genome-wide genotyping array for tropical pines. Pooled RNA-Seq data was generated from shoot tissues of seedlings from five species: *Pinus patula*, *Pinus tecunumanii*, *Pinus oocarpa*, *Pinus greggii* and *Pinus maximinoi*. Pooled RNA was isolated in triplicate from 8 to 16 seedlings from two or more families per species. Paired-end RNA sequencing generated between 29.4 and 67.7 million raw reads per pool. Reads were trimmed and mapped to each species’ respective transcriptome assembly. SNP detection was performed under standard and strict criteria using a min-alternative-fraction of ≥ 0.05. Using standard criteria, we identified between 687K and 1.3M transcriptome-based SNPs per species, while strict calls produced between 426K and 824K SNPs per species. SNP probe design (using Affymetrix design criteria) resulted in 113K to 186K candidate probes per species for a total of 803K probes designed. Probes were further assessed for unique vs. repetitive mapping against
the v2.01 *P. taeda* genome assembly. Of the 803K probes that mapped to the genome, 60% to 67% mapped uniquely in the reference genome assembly.

In addition, targeted sequence data for the five species mentioned above, plus *Pinus caribaea*, was generated using a custom set of 40K capture probes (RAPiD Genomics Gainesville, FL). About 30K were designed from single copy regions of the v2.01 *P. taeda* genome and 10K were designed from *P. patula* and *P. tecunumanii* transcriptome assemblies. Across the six species, a total of 81 pooled samples were sequenced. Each pooled sample contained DNA from 4 to 8 trees from a single provenance. The 81 provenances cover the natural range of the six species in Mexico and Central America. Capture-seq analysis generated between 3.1 and 7.7 million reads per sample with 20-30X coverage of each capture region. We identified a total of 1.3 million SNPs suitable for probe design, from which, 562K probes were successfully designed. Between the two datasets, there are ~1.6 million unique locations represented by ~1 million probes. This study shows that RNA-Seq and targeted capture sequencing are valuable technologies in identifying variation within non-model species and produce a high quality dataset for implementation of probe design for use on a commercial genotyping array.
Single Nucleotide Polymorphism Discovery and Probe Design Using RNA-Seq and Targeted Capture in Tropical and Subtropical Pine Species

by
Colin Alan Jackson

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Master of Science

Forestry and Environmental Resources

Raleigh, North Carolina 2019

APPROVED BY:

Juan Jose Acosta
Committee Co-Chair

Fikret Isik
Committee Co-Chair

Gary Hodge

Ross Whetten
BIOGRAPHY

Colin Jackson, son of Jeffrey and Carol Jackson, was born on May 12, 1993 in Texarkana, TX. Colin was raised in Broken Bow, OK where he spent much of his time helping raise livestock and being active outdoors: hunting, fishing, and hiking. He attended Broken Bow High School and graduated in the class of 2011. After graduating, he attended Oklahoma State University where he was active in undergraduate research from his first through final year researching the field of microbial ecology. In 2016, Colin graduated with his B.S. in microbiology with an emphasis in cell and molecular biology along with a minor in forestry. After graduation, Colin worked a year in the forestry industry in the field of tree improvement. This led him to his path at NCSU working jointly with Camcore and the Cooperative Tree Improvement Program for his graduate degree.
ACKNOWLEDGMENTS

First, I would like to thank my family and particularly my parents for their support throughout all of my endeavors in life. Without them, there is no way I would be in the position I am in today. Thank you for shaping me into who I am today and know that my successes are equally yours.

Next, I would like to thank my co-advisors Drs. Juan Acosta and Fikret Isik, and committee members Gary Hodge and Ross Whetten, for their guidance and constructive criticisms as I worked through this project. I am also tremendously grateful for my collaborators at the University of Pretoria, Drs. Alexander Myburg and Nanette Christie and Dr. Jill Wegrzyn at the University of Connecticut. This project has been hugely collaborative and would not have been possible without the direction and resources given by them and others at their institutions.

Lastly, I would like to acknowledge my fellow TIP, Camcore, and FER GSA student counterparts for not only providing advise on the project but also providing a sense of community that helps keep the ever present challenge of work life balance in check.
| LIST OF TABLES | .................................................................................................................. v |
| LIST OF FIGURES | ................................................................................................................. vi |

Chapter 1. Pines, Tree Improvement, and Genetics ................................................................. 1  
1.1. Introduction ...................................................................................................................... 1  
1.2. Genetics and Breeding of Pines ....................................................................................... 4  
1.3. Molecular Markers .......................................................................................................... 9  
1.4. Summary .......................................................................................................................... 18  
1.5. Literature Cited ............................................................................................................... 20  

Chapter 2. Single Nucleotide Polymorphism Discovery in Tropical and Subtropical Pines  
Using RNA-Seq .................................................................................................................. 35  
2.1. Introduction ..................................................................................................................... 35  
2.2. Materials and Methods .................................................................................................. 37  
2.2.1. Plant Material ............................................................................................................. 37  
2.2.2. RNA Isolation and Sequencing .................................................................................. 38  
2.2.3. SNP Discovery ........................................................................................................... 39  
2.2.4. Probe Design ............................................................................................................. 40  
2.2.5. Probe Description ...................................................................................................... 40  
2.3. Results ............................................................................................................................ 43  
2.3.1. RNA-Sequencing and Alignment .............................................................................. 43  
2.3.2. Variant calling and Probe Design ............................................................................. 43  
2.3.3. Probe Description ...................................................................................................... 45  
2.4. Discussion ....................................................................................................................... 46  
2.5. Literature Cited ............................................................................................................... 51  

Chapter 3. Single Nucleotide Polymorphism Discovery and Probe Design using Targeted  
Sequencing Data of Tropical and Subtropical Pine Species .................................................. 69  
3.1. Introduction ..................................................................................................................... 69  
3.2. Methods and Materials ................................................................................................ 71  
3.2.1. Plant Material and DNA Isolation ........................................................................... 71  
3.2.2. Capture Probe Design and Target Enrichment ......................................................... 72  
3.2.3. Quality Control and Alignment ................................................................................ 73  
3.2.4. Variant Calling and SNP Probe Design ................................................................... 74  
3.2.5. SNP and Probe Description ...................................................................................... 75  
3.3. Results ............................................................................................................................ 77  
3.3.1. Target Capture Sequencing and Alignment ............................................................. 77  
3.3.2. Variant Calling and Probe Design ........................................................................... 77  
3.3.3. SNP and Probe Characterization ............................................................................. 78  
3.4. Discussion ....................................................................................................................... 79  
3.5. Literature Cited ............................................................................................................... 84  

Chapter 4 Conclusions .......................................................................................................... 99  
4.1. Literature Cited ............................................................................................................... 101  

APPENDIX ............................................................................................................................ 102
Chapter 2

Table 2.1  Sequencing and alignment metrics generated from RNA sequencing. Sequencing generated millions (M) of reads and billions of bases on average per species. Raw reads underwent quality control (QC) requiring a base quality score of 30 and minimum read length of 50bp. Average read length decreased slightly from the 150bp raw reads generated from sequencing. Average alignment percentage across species was consistent. .............................................................57

Table 2.2  SNP numbers and probe design metrics for SNPs generated under standard parameters. The number of SNPs identified from variant calling generally numbered in the millions with the exception of *P. greggii* which had approximately 680K identified. Of the SNPs generated, less than one third were successfully converted to a probe. Probes tended to have a high number of unique left and right flanking regions, showing little redundancy within species...58

Table 2.3  Catalogue of the number of transitions (Ts) and transversions (Tv) per species. Ts:Tv ratios were consistent across species and within expected ranges for SNPs generated from transcriptome sequences. .............................................................59

Chapter 3

Table 3.1  Provenance and tree sampling structure for each of the nine subspecies selected for targeted capture. Each provenance represents a pooled sample submitted for sequencing and contained between 4 to 8 trees in the sample. The 81 total provenances represented the natural range of the species.................................89

Table 3.2  Sequencing metrics generated from targeted capture sequencing. Sequencing generated millions of reads and bases on average per species. Raw reads underwent quality control trimming, requiring a base quality score of 30 and minimum read length of 50bp. Average read length decreased slightly from the 150bp raw reads generated from sequencing to ~143bp. ........................................90
LIST OF FIGURES

Chapter 2

Figure 2.1 Work flow for SNP discovery using RNA-seq data. Work flow includes two phases, bioinformatics (blue) and probe design (purple). Software used and file outputs are shown in the ovals underneath the colored boxes. ........................................... 60

Figure 2.2 Sharing pattern for each species’ left and right flanking regions when allowing for zero, one, or two differences within a flanking region to be considered shared. For each species, a large portion of flanking regions are unique to that species regardless of differences allowed. Allowing for sequence differences increases the number of flanking regions shared by two or three species at a larger proportion than sharing between four or all. .................................................. 61

Figure 2.3 Sharing of left flanking regions allowing for zero differences across species. P. maximinoi has the highest number of unique flanking regions while P. tecunumanii and P. oocarpa have a high number of shared flanks between the two as expected from phylogeny................................................................. 62

Figure 2.4 Principal component analysis of left flanking (a) and right flanking (b) regions when allowing for zero differences in flanking sequence. Correlation between species that are more closely related phylogenetically is strong as indicated by loading angles. The more closely related P. tecunumanii / P. oocarpa and P. patula / P. greggii pairs have angles that are close to one another. The more distantly related P. maximinoi segregates itself. .............................................................. 63

Figure 2.5 Repetitive mapping of left and right flanking regions for each species when mapped to the v2.01 P. taeda genome assembly. Over 50% of flanking regions in all species map to a unique location within the genome reference. ...................... 64

Figure 2.6 Sharing pattern for each species’ left and right flanking regions when mapped to v2.01 of the P. taeda genome assembly. Flanking regions are considered shared if they mapped within two bases of another. A large portion of flanking regions are unique to a given species while very few map to the same location in all species. ................................................................................... 65

Figure 2.7 Distribution of left flanking regions shared mapping locations. Each species contains a large number of regions unique to the individual. Pinus patula / P. greggii and P. tecunumanii / P. oocarpa species combinations sharing mapping locations at a higher proportion between themselves than with other species........ 66

Figure 2.8 Distribution of right flanking regions shared mapping locations. A relatively small number of regions are shared between four or five species. Pinus maximinoi has the largest number of uniquely mapping regions and species groupings are consistent with that found in the left regions.............................................. 67
Chapter 3

Figure 3.1 Work flow for SNP discovery using targeted sequencing data. Work flow includes two phases, bioinformatics (blue) and probe design (purple). Software used and file routines are shown in ovals underneath the colored boxes......................91

Figure 3.2 The distribution of SNPs shared between the Top-down and Bottom-up filtering datasets. Between the two sets of SNPs generated from filtering approximately 403K SNPs were shared between the two. This resulted in approximately 1.36 million unique SNPs selected for probe design..............................................92

Figure 3.3 Sharing pattern for SNPs which had probes successfully designed for them by species. SNP sharing patterns among the species are relatively similar with the exception of *P. maximinoi* which contains a large number of private SNPs. There are a total of just over 25K SNPs that are shared across all species .........................93

Figure 3.4 Distribution of SNP sharing across species for SNPs with successful probes. SNP sharing patterns among the species are relatively similar with the exception of *P. maximinoi* which contains a large number of private SNPs when compared to the others ........................................................................................................94

Figure 3.5 Sharing pattern for SNPs with successful probes by subspecies. Subspecies exhibited a relatively similar number of private SNPs (~20K) with the exception of *P. maximinoi* (~120K). All species shared just under 25K SNPs between each other.................................................................95

Figure 3.6 Sharing pattern for SNPs with successful probes by provenance. Provenances exhibited a relatively similar number of private within them. *Pinus maximinoi* provenances had a much larger number of SNPs shared between 2 to 14 provenances than other species. .................................................................96

Figure 3.7 Principal component analysis of SNP sharing for species (a) and subspecies (b). Correlation of SNP sharing at the species level appears to be high as indicated by the narrow loading angles. *Pinus maximinoi* segregates itself from the other species showing little correlation. Subspecies follow a pattern as expected from phylogeny with subspecies of the same species sharing a similar loading angle.....97

Figure 3.8 Repetitive mapping of left and right flanking regions for SNP probes when mapped to the "tropicalized" v2.01 *P. taeda* genome assembly. Over 75% of
flanking regions in all species map to a unique location within the genome reference.
Chapter 1. Pines, Tree Improvement, and Genetics

1.1 Introduction

The genus *Pinus* consists of over 100 species and is the largest genus of conifers (Price et al. 1998). Pine species are widely distributed exclusively in the northern hemisphere, with the one exception being a population of *Pinus merkusii*, which is found in the southern hemisphere (Mirov 1967). All pines are evergreen trees or shrubs that can be found in climatic regions varying from tropical to boreal and across elevation ranges from sea level to alpine in North and Central America along with most of Europe and Asia. Across this widespread range, there are two primary pockets of diversity, one in Mexico and Central America and another in China (Plomion et al. 2007). Individual species have varying natural ranges, with some such as *P. resinosa*, *P. taeda*, and *P. contorta* having very large geographical ranges, while *P. radiata*, *P. greggii*, and *P. patula* have much smaller natural distributions.

Many of these species have found value in commercial timber markets. Globally, forest plantations are increasing in area. In 2015, the United States and China had 47 percent and 88 percent, respectively, more planted forests than they did in 1990 (FAO 2015). Within these plantations, pines play a key role and provide a significant portion of wood harvested on planted lands. In 2000, planted forests globally consisted of 20 percent *Pinus* species, with those making up significant percentages of planted lands in Asia (30%), North and Central America (88%), and Africa (28%) (FAO 2015). Utilization of pines in plantation forestry is common due to their adaptability, diversity, and ability to be planted on a wide variety of sites. *Pinus taeda*, for instance, has shown the ability to undergo morphological and physiological changes in response to unfavorable conditions such as moderate and long-term
drought (Samuelson et al. 2014; Maggard et al. 2016). Additionally, pines can perform well on marginal sites and are, good candidates for afforestation and reforestation efforts (Pinto et al. 2011; López-Tirado and Hidalgo 2016).

Even though the natural ranges of almost all pines are restricted to the northern hemisphere, their high productivity and success in plantations has resulted in the establishment of land race populations of many species south of the equator. A land race can be described as a population of individuals that has become adapted to an environment in which it has been planted (Zobel and Talbert 1984). In many cases, these land races are more productive than naturally occurring species. One of the most prevalent is *P. radiata*, which has a small natural range off the coast of California and is now one of the most widely planted species of pines across the globe, with populations established in New Zealand, Australia, and portions of South America (Sutton 1999). In South Africa, numerous species such as *P. patula*, *P. oocarpa*, *P. tecunumanii*, *P. caribaea*, and others have land race populations established and now play an important role in the country's commercial timber markets (Griess et al. 2016).

Species of tropical and subtropical pines native to Mexico and Central America are of particular interests when it comes to land race populations. The large amount of diversity within and between species has made them candidates for small scale deployments in areas where they may be most adapted. In the case of *P. caribaea*, there are approximately 1 million hectares of plantations globally, with the majority of that being in Venezuela (Dieters and Nikles 1997). Valued for its rapid growth, *P. caribaea* provides a cautionary tale when establishing land race populations. When first established in South Africa, it proved to have much less desirable wood quality characteristics than expected (Zobel and Talbert 1984).
Pinus greggii has shown widespread adaptability across testing sites at a variety of elevations and latitudes. It has also shown potential to survive on colder dryer sites than P. patula. At this point, P. greggii is almost exclusively planted in South Africa for commercial uses (Dvorak et al. 2000). Pinus patula has been planted on over 1 million hectares worldwide with much of that in eastern and southern Africa. Despite not being as cold tolerant as P. greggii, it performs well on sites that are less tropical than other species and has good wood quality that is useful for a number of different purposes. Pinus tecunumanii has been established on approximately 10,000 hectares worldwide with most plantations occurring in Brazil and Colombia (Dvorak et al. 2000). Pinus tecunumanii has shown fast growth and wood quality characteristics when compared to other tropical species and has been the focus of hybridization studies to potentially incorporate growth traits and disease resistance with other species (Kanzler et al. 2012; Hongwane et al. 2018). Two other species P. maximinoi and P. oocarpa have been deployed on a much smaller scale. Although, these species show utility in niche deployments potentially offering advantages through future study (Dvorak et al. 2000).

Ultimately, the demand for high value timber products such as lumber, poles, pulp and paper has driven the increase in plantation area and establishment of such land race populations. In addition to these standard timber products, there has been a growing interest in using pine slash and residuals from harvest for the manufacture of various phenolic compounds, essential oils, pellet for bioenergy and other secondary products (Kelkar et al. 2006; Rodrigues-Corrêa et al. 2012). Even more recently, there has been extensive research into the use of pine as a source of feedstock for lignocellulosic biofuel production (Lu et al.
2015; Vaid et al. 2018; Dong et al. 2018). The value of these products combined with the versatility of pines makes these species an important focus of study.

1.2 Genetics and Breeding of Pines

Tree breeding and genetic improvement in pines is in its relative infancy when compared to commercial agricultural crops. The first generation breeding population of loblolly pine (\textit{P. taeda}) was established from selections of phenotypically superior individuals in naturally occurring stands, in the 1950's (Zobel and Talbert 1984). These selections were used to create full-sib families from breeding orchards and families were subsequently tested in progeny tests. Similar steps have been taken in other species of pine as well in establishing breeding programs (Ruotsalainen 2014). In forest trees, estimation of trait heritability and breeding values are done almost exclusively using phenotypic data from progeny testing and pedigree information through the use of best linear unbiased predictions (BLUP).

Field testing of progeny is both time consuming and expensive, due to test establishment, maintenance, and measurement. Typically, testing requires relatively large areas of land and a large number of progenies to be tested across a number of sites to make accurate estimations about genetic parameters. It has been shown that accurate estimations for selection can be made on uniform sites at age six for loblolly pine (McKeand 1988). New breeding strategies coupled with efficient testing designs can reduce some labor and cost, but the process remains costly primarily due to the maintenance and duration of the tests. These challenges create an opportunity to leverage molecular data as a way to reduce costs associated with testing.
All *Pinus* species are diploid (2n=24), many with very large genome sizes, ranging from 20-30Gb (De La Torre et al. 2014). This is much larger than other tree genome references such as a poplar, (~500Mbp) (Tuskan et al. 2006), eucalyptus (~640Mbp) (Myburg et al. 2014), and have genomes between 133-296 times bigger than *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000). At the whole genome level, it has been shown that approximately 14% of the pine genome is made up of low- or single-copy sequences, while 86% are highly repeated sequences (Wakamiya et al. 1993). Within the genus *Pinus*, one of the most characterized genomes is that of *P. taeda* which had its draft genome assembled in 2014 (Zimin et al. 2014) and then an improved assembly released in 2017 (Zimin et al. 2017). Although still an incomplete reference assembly, the *P. taeda* genome provides an attractive platform for the implementation evolutionary, population structure, and diversity studies in other species of pines.

Pine genome mapping efforts began as early as the 1970's and 1980's using isozyme loci that were quite small and limited in number (Rudin 1978; Conkle 1979). With the advent of restriction enzymes, PCR technologies and greater sequencing capabilities, DNA markers were used to expand upon the earlier maps. The first maps to use DNA marker technology were assembled using restriction fragment length polymorphisms (RFLPs) in *P. taeda* and *P. radiata*, finding 20 and 22 linkage groups respectively (Devey et al. 1994, 1996). Since then, the use of more advanced sequencing technology has allowed scientists to use single nucleotide polymorphisms (SNPs) and other molecular markers to enhance coverage and detail of these maps (Eckert et al. 2009; Echt et al. 2011; Martínez-García et al. 2013).

Quantitative trait loci (QTLs) involved in many traits important to the improvement of pines have been identified. Many traits important to forestry, such as volume growth,
disease resistance, and wood quality are complex quantitative traits with complex inheritance. Quantitative genetics assumes that the genetic control of a complex trait can be attributed to the additive effects of the many loci involved and is then affected by the environment (White et al. 2007). Quantitative trait loci are locations within a genome that are associated with a given quantitative trait. Many different quantitative traits have been mapped, including wood quality traits, such as wood specific gravity and density, pulp yield, cell wall content, and average microfibril angle (Groover et al. 1994; Kumar et al. 2000; Ball 2001; Sewell et al. 2002; Brown et al. 2003). Growth traits have also been of interest. QTL studies assessing many growth traits including average numbers of branch per whorl, average branch diameter, and average whorl spacing (Shepherd et al. 2002a) have been performed. Studies examining forking (Xiong et al. 2016) and tree height and diameter (Kaya et al. 1999; Weng et al. 2002) have been done as well. Links between qualitative traits and markers have been shown and have led to the discovery of disease resistance markers such as in the case of fusiform rust (Cronartium quercuum f. sp. fusiforme) in P. taeda (Wilcox et al. 1996) and pitch canker (Fusarium circinatum) in P. radiata (Moraga-Suazo et al. 2014).

In contrast to crop and animal species, there seems to be a more limited ability to detect QTLs in pines and other forest trees. Detection of QTLs largely depends on the quality of phenotypic assessment. In many cases, these assessments are challenging because of the high environmental and developmental variation in tree plantations or the small population sizes used in the analyses (Plomion et al. 2007). In addition, the results from Brown et al. (2003), when compared to other studies in tree species, suggests that in many cases, the amount of variance described by QTLs in trees are overestimated. Even with potentially inflated estimations, the amount of total variation explained by these QTL experiments have
been quite low. This suggests that there is a more complex gene structure at work, having fewer major effect loci and many more of low effect. This is not surprising, given the highly complex inheritance patterns suggested for growth and yield traits (Hill et al. 2008).

Additionally, QTL stability across tree age can vary (Lerceteau et al. 2001). This indicates that the loci responsible for characteristics may change depending upon stage of growth or environmental effects being encountered.

Leveraging QTL information for the use of marker assisted selection (MAS) was an attractive proposition during the time of QTL mapping. MAS was of particular interest in forest trees due to their long breeding cycles and was seen as a way to reduce the time associated with them. Evaluation of the technology showed that within-family selection using MAS was not useful unless used in weakly heritable traits and used with very large numbers of samples along with a high selection intensity (Strauss et al. 1992). They also concluded that there may be value in MAS in interspecies hybrids and hard to measure phenotypic data such as wood quality. It was suggested that MAS could be utilized in commercial breeding programs if families with evidence of major gene effects could be identified. However, the testing capacity of a program would have to increase in order to meet the large number of samples needed to utilize MAS (O’Malley and Mckeand 1994). Ultimately, despite a significant amount of time and resources being put into QTL research, there has been limited to no success in utilizing MAS for breeding in forest trees (Isik 2014).

Despite the issues in using MAS in forest trees, genomic selection, a new frontier in utilizing molecular genetic data is being explored. Using simulated data it was shown that selections can be made with very accurate genomic information (Meuwissen et al. 2001). Genomic selection is defined as the selection of individuals based on genomic estimated
breeding values, which are estimated by solving prediction equations based on single nucleotide polymorphism (SNP) markers (Hayes et al. 2009). Genomic selection is of great interest to tree breeders due to the large number of potentially informative SNPs and the potential to reduce progeny testing efforts and the time taken to complete a breeding cycle.

Genomic selection is being deployed primarily in populations that exhibit high linkage disequilibrium and that are highly structured (Thavamanikumar et al. 2013). Successful genomic selection programs have been deployed in cattle (Hayes et al. 2009), rice (Oryza sativa) (Spindel et al. 2015), and others (Oh, Na, & Park, 2017; Zhao et al., 2012). There has been little work done in pines and other forest trees, however. In forest trees, under simulated data, it has been shown that genomic selection is possible and advantageous compared to traditional BLUP based-selection, given a genotyping density above 10 markers/cM for effective population sizes of 60 and up to 20 markers/cM for populations with an effective size of 100 (Grattapaglia and Resende 2011). These results were then supported in study examining two unrelated elite populations of eucalyptus by matching or exceeding the traditional phenotypic selection method accuracies in all traits (Resende et al. 2012b). A study performed on P. pinaster produced a lower estimates of the correlation of true and estimated breeding values obtained from progeny tests using genomic selection (Isik et al. 2015). In P. taeda, it was shown that selection efficiency per unit time was 53-112% higher using genomic selection and could significantly reduce the time required to complete a breeding cycle (Resende et al. 2012a). These results, along with reduced genotyping costs and readily available molecular markers through next generation sequencing, are making genomic selection the next area of focus in pine and tree breeding.
1.3 Molecular Markers

Polymorphisms that are found in populations and reveal neutral sites of variation in DNA are called molecular markers (Kumar et al. 2009). Molecular markers provide scientists and plant breeders a valuable tool to help understand the architecture of these complex quantitative traits. According to Gupta et al. (2001), there are three classifications of molecular markers: first generation, second generation, and third generation markers. He defines first generation markers as restriction fragment length polymorphisms (RFLPs) and random amplified polymorphic DNAs (RAPDs). He categorizes second generation markers include amplified fragment length polymorphisms (AFLPs) and simple sequence repeats (SSRs) and the third generation of molecular markers as single nucleotide polymorphisms (SNPs).

Restriction fragment length polymorphisms (RFLPs) were first used in the construction of the genetic map of humans (Botstein et al. 1980). This technique differentiates organisms by analysis patterns of their cleaved DNA. Utilization of a specific restriction endonuclease allows for comparison of individuals based on the distance between cut sites in their DNA (Kumar et al. 2009). The resulting fragments are then separated via electrophoresis. During electrophoresis, high voltage is added so that the DNA fragments are separated by size and then detected using a camera system. Using software, the similarity of the size of cut DNA to the sample of interest is determined and indicates similarity between the samples.

Despite being, co-dominantly inherited and very reproducible, the use of RFLPs is costly, time consuming, and requires a large quantity of highly purified DNA for analysis (Agarwal et al. 2008). RFLPs have been extensively used in crop plants for mapping.
purposes and to assess diversity of breeding populations (Schön et al. 1994; Kim and Ward 2000). In pines, despite finding a number of markers that could be used for high density mapping, only two such studies were done. One was performed to create a linkage map for loblolly pine based on a third generation outbred population (Devey et al. 1994). The other was to create a linkage map for *P. radiata* (Devey et al. 1996). Ultimately, these studies were the basis of a comparative mapping project which showed high levels of synteny (similarity of genetic maps possibly due to shared ancestry) between the two species (Devey et al. 1999).

Random amplified polymorphic DNA (RAPD) markers utilize PCR and random primers to detect variation in DNA. The procedure is used to detect nucleotide sequence polymorphisms through the use of a given primer (Williams et al. 1990). Through the use of a short primer of random sequence, genomic DNA is bound and amplified using PCR technology. The resulting amplified strands are stained, run through gel and then viewed under ultraviolet light. The presence or absence of a band indicates differences between the DNA template strands due to sequence differences in the priming sites. The use of RAPD markers is fairly easy due to the low amount of purified DNA needed for the reaction, easily accessible primers, and no prior knowledge of the genome being studied needed. However, concerns about reproducibility exist between and within labs because of the sensitivity to reaction conditions and the risk of contamination due to the nonspecific nature of primer binding (Rabouam et al. 1999).

Despite the issues with reproducibility, RAPD analysis was optimized in pines using *P. radiata* as a template (Devey et al. 1996; Ostrowska et al. 1998). However genomic mapping efforts using RAPDs had been made prior to their optimization. Genetic maps were
RAPDs have been used to highlight genetic diversity in highly conserved genomes such as the case with *P. pinaster* (Plomion et al. 1995), *P. elliottii* (Nelson et al. 1993), and others (Nelson et al. 1994; Yazdani et al. 1995). RAPD markers have been shown to be useful in highlighting genetic diversity in highly conserved genomes such as the case with *P. resinosa* (DeVerno and Mosseler 2008) and have been used to assess the genetic variation within endangered and threatened species (Newton et al. 2002; Zhang et al. 2005).

Amplified fragment length polymorphisms (AFLPs) are categorized as a second generation marker and selectively target sequences of DNA through the use of specific restriction enzymes for amplification. Polymorphisms are then detected based on differences in length of the amplified segments on gel. Through the use of two restriction enzymes, genomic DNA is fragmented. Then double stranded adaptors are ligated to the end of the fragments generating a template for PCR reactions downstream. Ligated fragments are amplified using PCR reactions. The amplified DNA is then separated via gel electrophoresis. AFLPs present an useful combination of RFLP and PCR technologies (Agarwal et al. 2008).

AFLPs are highly reproducible and have the ability to generate large amounts of fragment markers without prior knowledge of genomic sequence. This fact makes them useful in the investigation of non-model organisms which may have limited genomic information available. Two challenges when using AFLPs are the need for highly purified DNA and the ambiguity of analyzing the bands present in the gel (Kumar et al. 2009). In forest trees AFLPs have shown utility in population structure and genome mapping (Cervera et al. 2000). In pines, linkage maps for *P. radiata* (Moraga-Suazo et al. 2014), *P. taeda* (Remington et al. 1999), and *P. edulis* (Travis et al. 1998) have been created using AFLPs. Numerous population structure, diversity, and hybridization studies have been performed
using the technique as well (Diaz et al. 2001; Ribeiro et al. 2002; Di and Wang 2013; Vasilyeva and Semerikov 2014).

Another type of second generation makers are simple sequence repeats (SSRs), also known as microsatellites. SSRs are short segments of repetitive DNA that generally consist of 1-5 nucleotides and are dispersed throughout eukaryotic genomes (Powell et al. 1996). Microsatellite makers can be found in both non-transcribed and transcribed regions of the genome. In transcribed regions of the genome expressed sequence tags can be mined for SSR markers (Varshney et al. 2005). SSRs tend to be much more informative than AFLPs or RAPDs since they are co-dominant markers. The ability to segregate based on zygosity is one of the primary draws of SSR technology. In a given population, there are generally several alleles for the same locus. For example, using hypothetical numbers, a given individual in a population may contain 6, 7, 8, or 10 repeats. This type of information can be useful in identifying different individuals within the population (Oliveira et al. 2006), while there are differences in the number of repeats, the upstream and downstream sequences flanking them are generally highly conserved (Jarne and Lagoda 1996).

SSR markers are classified into three categories: perfect, imperfect and complex/compound. These classifications are based on the fragments being repeated. A perfect SSR is where the repeat sequence is unbroken by any additional nucleotides, i.e. the sequence AT repeated 12 time (AT$_{12}$). A SSR is imperfect when the repeat sequence is broken by other nucleotides, i.e. the sequence AT repeated six times followed by GC and then AT another six times (AT$_6$GC$_1$AT$_6$). A complex/compound SSR is where two or more repetitive sequences are in line with each other, i.e. the sequence AT eight times followed by GC repeated 10 times (AT$_8$GC$_{10}$). Additionally, markers are generally classified as
SSR/microsatellites if they are between 10 and 30 nucleotides long, while SSRs are classified as minisatellites between 30 and 100 bases long (Weber 1990).

SSRs have been shown to be highly transferable between species (Shepherd et al. 2002b). One such study found 251 and 168 SSRs in unigene sets (transcripts that appear to stem from the same locus) of *P. taeda* and *P. pinaster* respectively with the most common repeat type being trinucleotides. For these markers they designed 72 primers and found single band amplification rates in non-source species ranging from 64.6% to 94.2% (Chagné et al. 2004). Transferability has been found in organelle DNA within pines as well. Variation in length of mononucleotide repeats in chloroplast DNA were shown in 11 species of pine (Powell et al. 1996). Within pines, AT repeats seem to be the most common motif across species (Smith and Devey 1994; Bai et al. 2014).

SSRs are highly polymorphic and can easily be used for population structure experiments (Smith and Devey 1994). Since microsatellites exhibit Mendelian inheritance, it is possible to use them for fingerprinting (Liewlaksaneeyanawin et al. 2004), and evolutionary history of species (Boys et al. 2005). A linkage map for *P. taeda* has been created using microsatellites as well (Echt et al. 2011). In addition, SSRs have been widely used in other areas of genetics: population genetics, pedigree construction, marker assisted breeding, and genetic conservation.

The use of SSR analysis does have some difficulties however. Microsatellites can have a high mutation rate which can cause incorrect conclusions to be drawn about the genetic history of an individual (Queller et al. 1993). Additionally, homoplasy (the occurrence of similar gene structures without being present in a common ancestor) occur at SSR loci due to mutation, which can cause a misrepresentation of genetic divergence.
(Wheeler et al. 2014). Several mutation mechanisms can affect the results received from DNA SSRs. There are two primary models used to characterize the mutations in SSRs, infinite allele model and stepwise model, each having their own shortcomings (Jarne and Lagoda 1996).

The third generation of molecular makers are single nucleotide polymorphisms (SNPs), pronounced "snips". When compared with other molecular markers, SNPs appear with much higher frequency and are far more prevalent (Gupta et al. 2001). SNPs are defined as a single nucleotide change when compared to a reference sequence. There are two types of SNPs, transitions and transversions. Transitions occur when a purine base is replaced with another purine base. A transversion occurs when a purine base is replaced with a pyrimidine or vice versa (Slonczewski and Foster 2009). Given their abundance, they have become the marker of choice for genetic studies. These markers are also easy to produce and cost efficient. It has been shown that in elite populations of maize, polymorphisms can be as common as 1 per 31 bases in non-coding regions and 1 per 124 bases in coding regions (Ching et al. 2002). A SNP frequency of 1 SNP per 103 bases using expressed sequence tag (ETS) data was found in P. pinaster (Dantec et al. 2004). Despite being very common throughout the genome, SNP association with traits tend to occur at a low frequency or account for very little of the genetic variance of a trait. A recent study found approximately 2.8 million SNPs using exome capture which included 34 SNPS and 11 SNP/SNP interactions that were associated with important growth traits in P. taeda (Lu et al. 2017). In another exome capture study using around 67K and 55K SNPs in P. tadea and P. elliottii respectively, extensive similarity between species was shown when looking at linkage disequilibrium and levels of genetic diversity (Acosta et al. 2019).
Before the implementation of next generation sequencing (NGS) technologies, SNPs were discovered though a variety of techniques to help cope with the repetitive nature of plant genomes. The development of SNP markers through EST mining followed by PCR validation and the resequencing of unigene amplicons using Sanger's sequencing method were two of the more popular methods for SNP discovery within genic regions (Mammadov et al. 2012). DNA sequencing is the process of identifying the order DNA bases through synthesis of a complementary strand of DNA and determining which nucleotides have been added.

In short, Sanger sequencing involves using a primer to bind next to the region of interest to facilitate the extension of the complementary strand of DNA. The use of a dideoxynucleotide causes the polymerase to terminate replication at various lengths along the amplicon. Each type of dideoxynucleotide was radioactively marked and were detected by autoradiography (Sanger et al. 1977).

SNP mining through the use of ESTs avoids the problems of de novo discovery by using available cDNA libraries from sources such as NCBI. ESTs are mRNA sequences derived from single sequencing reactions performed on randomly selected clones from cDNA libraries (Parkinson and Blaxter 2009). EST sequences are compiled and SNPs identified and filtered using bioinformatics software. The SNPs are then validated through software and re-sequencing. PCR is used to amplify sequences of interest and those that yield well defined bands are the re-sequenced for validation (Dong et al. 2010). EST mining techniques have been used for SNP discovery in both *P. pinaster* (Dantec et al. 2004; Chancerel et al. 2011) and *P. taeda* (Eckert et al. 2009).
Next generation sequencing (NGS) technologies, for the most part, utilize sequencing by synthesis (SBS) methods. SBS is where the sequence of DNA is read while the complementary strand is being synthesized. The sequence can be determined by the use of labeled nucleotides which can be differentiated from one another (Goodwin et al. 2016). Methods used by Illumina for their NGS technology have similarities to Sanger sequencing through the use of a terminator nucleotide. As labeled bases are incorporated, the sequence is then determined. Pyrosequencing uses synthesis reactions to emit light which can then be used to identify the base being added to the complementary DNA strand (Ronaghi et al. 1998). These technologies generally prove shorter read lengths than older technologies but provide a higher number of reads, allowing for greater sequencing depth. This in turn allows for higher confidence in the base call and reduces error (Sims et al. 2014).

In SNP discovery there are two main components, read mapping and SNP calling. Read mapping is a process that aligns reads generated from sequencing to that of a reference set of DNA. This reference is generally a complete genome reference of a well-characterized organism. However, with NGS technology, references can now be de novo assemblies of more closely related organisms or even the target organism itself (Kumar et al. 2012). The use of read assembly algorithms such as, Bowtie (Langmead et al. 2009) and Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) are good tools for handling short-read, high-throughput data. High throughput methods use the DNA sequences from multiple individuals to compare their sequences to a reference. When the sequenced reads have been aligned to the reference, they can then be scanned for potential SNPs.

SNP calling, or variant calling, is the process of locating where the physical positions of the SNPs are located in the genome, based on a reference sequence. Potential SNPs are
identified by SNP calling software such as SAMTools (Li 2011), FreeBayes (Garrison & Marth 2012), or VarScan (Koboldt et al. 2009). These programs use algorithms to scan reads for mismatches that could be potential SNPs. When assessing these mismatches, sequencing depth and read map accuracies are assessed to determine the quality of the SNP and identify it as a possible error (You et al. 2012). These variant callers each are biased towards different types of SNP calling errors, meaning careful consideration must be made when selecting a program based on the data being used (Hwang et al. 2016). This SNP calling process requires a high level of confidence in the quality and accuracy of the reference sequence to prevent false positives. Despite the utility of these variant callers, often errors in read mapping and variant calling are compounded, which requires additional custom screening to be done on the compiled set of called variants.

The ability to identify SNPs is an important tool in human, animal, and plant genetics. SNPs can be utilized in a number of different ways. In humans SNP identification has been used to identify disease susceptibility and other complex traits (Zheng et al. 2009; Frazer et al. 2009). In plants, SNPs have the ability to produce dense genetic maps (Rafalski 2002) which can be used in marker assisted selection, help understand genome organization, or to identify genotypes. SNPs can also be used to assess linkage disequilibrium in populations, for genome wide association studies, and have shown great potential in genomic selection.

SNP genotyping arrays can be a useful tool in utilizing SNP data for the above mentioned work. SNP genotyping arrays are chips containing a set of selected probe sequences. A probe is a DNA sequence that is complementary to the region of interest. Although array processes vary between platforms, they all involve the production of a signal in the presence of a targeted marker. Currently, ThermoFisher Scientific’s Axiom technology
is being used for the construction of a commercial array for pine species. Under this platform, four steps are used: amplification and fragmentation of DNA samples, hybridization, ligation, and signal amplification (ThermoFisher). This technology hybridizes amplified fragmented DNA to complementary capture probes. Following hybridization, bound target is washed to remove noise and ligated with a multi-color solution probe. Following ligation, arrays are stained and imaged to quantify marker presence.

In pine, the implementation of SNP array technology is in its infancy when compared to crop and animal species. A high density 9K SNP chip was created for *P. pinaster* (Plomion et al. 2016), however, it gives a much narrower view than the 57K array created for maize (Ganal et al. 2011), 60K chip for eucalyptus (Silva-Junior et al. 2015a), or 50K chip in cattle (Matukumalli et al. 2009). Utilization of very large SNP genotyping arrays have been used for genome-wide association studies in humans and other organisms (McCarthy et al. 2008). In animals, the use of these large number of markers have been used in genomic selection and have accurately predicted the breeding values in non-phenotyped populations (Erbe et al. 2012). This type of utility could have huge implications in tree improvement and breeding by reducing the labor and time associated with the process.

**1.4 Summary**

Given their commercial value and the projected increase in demand for wood products, genetic improvement of pine species is an important area of study. Traditional tree improvement methods have been effective in creating faster growing, straighter, and more disease-resistant trees, however, the process is still time consuming and costly. With the advent of next generation sequencing platforms, decreased genotyping costs and more efficient software for analysis, utilization of SNP marker data obtained from sequence data
for tree breeding is more accessible than ever before. These technologies have the potential to revolutionize tree breeding and improvement through reducing breeding cycles and increasing the efficiency in estimating genetic parameters.

This study aims to contribute to this effort by performing targeted genome and gene-based SNP discovery for the development of a high-throughput, genome wide, genotyping array for six tropical pine species. Two custom bioinformatics pipelines were developed for SNP discovery using datasets generated from RNA-sequencing and targeted capture sequencing, generating a high quality set of SNP probes for assessment.
1.7 Literature Cited


FAO (2015) Global Forest Resources Assessment


Kim HS, Ward RW (2000) Patterns of RFLP-based genetic diversity in germplasm pools of common wheat with different geographical or breeding program origins. Euphytica 115:197–208


O’Malley DM, Mckeand SE (1994) MARKER ASSISTED SELECTION FOR BREEDING VALUE IN FOREST TREES. For Genet 1:207–218


using genomic selection: accuracy of prediction models across ages and environments.


10.1111/j.1469-8137.2011.04038.x


Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors.


Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. PLoS Genet 11:1004982. doi: 10.1371/journal


Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features
Vasilyeva G, Semerikov V (2014) Application of amplified fragment length polymorphisms
markers to study the hybridization between Pinus sibirica and P. pumila. Ann For Res
Weber JL (1990) Informativeness of human (dC-dA)n · (dG-dT)n polymorphisms. Genomics
7:524–530. doi: 10.1016/0888-7543(90)90195-Z
early growth in a (longleaf pine × slash pine) × slash pine BC1 family. Theor Appl
Caveats of Using Chloroplast Simple Sequence Repeats for Studies of Plant Biology.
resistance to fusiform rust disease in loblolly pine by genomic mapping. Proc Natl Acad
Sci U S A 93:3859–64
arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535


Chapter 2. Single Nucleotide Polymorphism Discovery in Tropical and Subtropical Pines Using RNA-Seq

2.1 Introduction

The genus *Pinus* is the largest genus of conifers in the world consisting of over 100 individual species (Price et al. 1998). Their ability to adapt, grow and reproduce under a large number of ecological conditions is evidenced by their large natural distribution in the northern hemisphere. Due to this adaptability and variation, pines have come to play a crucial role in sustaining the global forestry market. In 2000, approximately 20% of all trees planted in commercial settings were pines (FAO 2015).

Within the genus, the subsection *Australes* contains many species that are important to pulp and sawtimber markets. In particular, some species of tropical and subtropical pines that are native to Mexico and Central America, have had landrace populations established in Africa and South America for commercial applications (Dvorak et al. 2000a; Griess et al. 2016). Of these species, *P. patula* and *P. greggii* are successfully planted on colder sites in higher elevations due to their frost tolerance, while species such as *P. caribaea*, *P. maximinoi*, are valued for their rapid growth, and *P. oocarpa* and *P. tecunumanii* for their wood properties (Dvorak et al. 2000a). The large amount of genetic diversity present within and between species, along with their ability to mate and produce viable offspring, has generated interest in breeders to potentially exploit hybrid crosses for niche commercial deployments, adding to their importance (Gwaze 1999; Hongwane et al. 2018).

Despite traditional breeding and testing methods being effective in consistently producing better trees, the cost and labor associated with the methods are spurring breeding programs to explore the implementation of genomic technologies. Much of the ability to
understand the relationship between genotype and phenotype revolves around the ability to effectively identify genomic variation (Piskol et al. 2013). Before next generation sequencing (NGS) technologies were available, efforts to understand this variation started in the 1980’s through the utilization of isozymes (Conkle 1979; Eckert et al. 1981) and continued through the use of expressed sequence tags (EST), microsatellites and simple sequence repeats (SSR) (Elsik et al. 2000; Temesgen et al. 2001; Bérubé et al. 2007). ESTs have also been a valuable resource for locating markers within related species (Liewlaksaneeyanawin et al. 2004). Ultimately, EST mining gave way to other methods such as genotyping by sequencing and other advancements as they were developed.

Currently, with the advent of NGS technologies, the ability to identify and exploit genomic variation has been revolutionized. Whole genome and whole exome sequencing studies have become attainable for large-scale projects looking to identify variation in both human and plant species (Michael and Jackson 2013; Gudbjartsson et al. 2015). Whole genome sequencing would be ideal in most cases looking for genetic variability, however, the implementation of such studies remains costly and computationally prohibitive for most pine species. Given the large size of pine genomes, which are generally more than 20Gb (De La Torre et al. 2014), and their highly repetitive nature (Wegrzyn et al. 2014), reduced representation sequencing methods are an attractive alternative to whole genome sequencing. Of these reduced sequencing methods, RNA sequencing (RNA-Seq) is among the most popular.

RNA-Seq is a popular alternative to whole genome studies because of its reduced cost when compared to whole genome studies and its versatility to address many different questions involving variation, gene expression, and alternative splicing (Oshlack et al. 2010;
Ozsolak and Milos 2011). More recently, it has been shown to be an effective means of identifying SNPs in transcribed regions of many different species (Thumma et al. 2012; Liu et al. 2014; Telfer et al. 2018). SNPs located in expressed regions are found at a lower frequency and are more likely to have an effect on protein structure and function, making them useful for explaining phenotypic variation. Utilizing RNA-Seq data for variant discovery does pose some unique challenges, primarily due to the complex nature of the transcripts themselves. Intron-exon boundaries present a computational challenge, especially in the absence of a same species or complete reference (De Wit et al. 2015). Therefore, special care must be taken to account for this when selecting SNPs for validation downstream.

In this manuscript, the design and characterization of more than 1.3 million SNP probes designed for five species of tropical or subtropical pines is described. Next generation RNA sequencing generated between 29 and 67 million raw reads per sequencing pool across the species. Reads were used to perform transcriptome-wide identification of more than 5.3 million SNP markers. The SNP probes designed in this study are to be assessed for their potential use on a commercial genotyping array in a future study with the ultimate goal of the array being used to facilitate genomic selection, population structure, hybrid studies, and more in the field of tree improvement and conservation.

2.2 Materials and Methods

2.2.1 Plant Material

RNA sequence data was generated for five species of pine through a pathogen challenge experiment for pitch canker *Fusarium circinatum*. The pathogen challenge, RNA extraction and sequencing steps outlined in this project were performed by Erik Visser at the University of
Pretoria as part of a previous unrelated study. The species tested were: *P. greggii*, *P. maximinoi*, *P. oocarpa*, *P. patula*, and *P. tecunumanii*. *Pinus patula* seedlings consisting of a single open pollinated family were sourced from Mondi Forests (Trahar Technology Center, South Africa). *Pinus tecunumanii* seedlings consisting of four open pollinated families were sourced from SAPPI (Shaw Research Center, South Africa). *Pinus maximinoi* and *P. oocarpa* seedlings were sourced from York Timbers (South Africa) and *P. greggii* seedlings were sourced from Komatiland Forests (South Africa). Seedlings were grown and challenged at the Forestry and Agriculture Biotechnology Institute (FABI) disease screening facility (University of Pretoria, South Africa). Pathogen inoculation and testing was performed as described in Visser et al. (2018, 2015). In short, the top 1cm of shoot tissue from 16 seedlings were pooled and used to create three biological replicates for *P. patula* and *P. tecunumanii*, while tissue from 8 seedlings were used per pool for the remaining species. Tissue was harvested from both inoculated and mock-inoculated seedlings at 3 and 7 days post inoculation to create a more robust expression dataset.

### 2.2.2 RNA Isolation and Sequencing

Isolation of RNA was done utilizing the Norgen Biotek Corp.’s Plant/Fungi RNA Purification Kit using ground homogenized samples. Extraction was performed using the manufacturer’s recommendation, adding acid washed beads to improve cell lysis. The integrity and purity of extracted samples were assessed using electrophoresis.

Samples from all species from both time points for the inoculated and mock-inoculated samples were sent to Novogene (Novogene Corporation Inc., Chula Vista, CA) for strand specific RNA sequencing using Illumina HiSeq2500. *Pinus patula* and *P. tecunumanii* samples were sequenced as described in Visser et al. (2018). *Pinus greggii*, *P.
maximinoi, and P. oocarpa samples were sequenced using 300bp insert, paired end 150bp (PE150) sequencing. Similar to that of sample set three in Visser et al. (2018).

2.2.3 SNP Discovery

The paired end RNA-seq libraries received from Novogene were put through a custom bioinformatics pipeline for SNP discovery (Figure 2.1). Reads were quality checked using FastQC (Andrews 2010). Reads from each library were trimmed and filtered using Sickle v1.33 (Joshi & Fass 2011), requiring a base quality score of 30 and a minimum read length of 50. The trimmed reads from each library for each species were aligned using Burrows-Wheeler Aligner’s (BWA) v0.7.15 (Li and Durbin 2009), mem routine. Using standard parameters, reads were aligned back to their respective species’ transcriptome assembly (Visser et al. 2018; Visser et al. 2015; unpublished data). The SAM alignment files generated were converted to BAM files and subsequently sorted using Samtools’ v1.7 (Li et al. 2009) view and sort routines.

Variant detection was done under two sets of parameters, strict and standard, and performed using Freebayes v1.1.0 (Garrison & Marth, 2012). SNP calls were done jointly, using all alignment files from an individual species as an input and utilized the pooled continuous parameter. Under the pooled continuous parameter, all alleles that pass input filters are output, regardless of the genotyping outcome or model. The standard parameters used are as follows: min-mapping-quality = 0, min-base-quality = 15, min-alternate-fraction = 0.05, min-alternate-count = 3, and min-coverage = 10. The strict parameters were: min-mapping-quality = 20, min-base-quality = 25, min-alternate-fraction = 0.05, min-alternate-count = 5, and min-coverage = 12. The resulting outputs were filtered to include only SNP variants using vcflib v1.0.0-rc1 (Garrison 2016).
2.2.4 Probe Design

Successfully called SNPs from each species were then subject to probe design. A probe is defined as the identified SNP plus a number of base pairs of sequence on either side of it. In this study, a target of 35 bases on either or both sides of the SNP was used. Probes were designed for use on ThermoFisher Scientific’s Axiom genotyping array platform.

Custom Perl scripts were used on both strict and standard datasets to extract the 35 bases upstream and downstream of each SNP also referred to as left and right flanking regions respectively. International Union of Pure and Applied Chemistry (IUPAC) codes were inserted into the probe flanking regions to identify proximity to another SNP and the flanks were trimmed to the innermost IUPAC code on both sides. Probes were then filtered to include only flanking regions for SNPs that had either the left, right, or both flanking regions clear of an IUPAC code for a full 35 bases in order to meet technical specifications.

To address the potential issue of probes being located in or around intron/exon boundaries, probes were further filtered. The transcriptome assemblies for each species were translated to amino acid sequences using EMBOSS v6.6.0 (Rice et al. 2000) and aligned to v2.01 of the *P. taeda* genome using GenomeThreader v1.7.1 (Gremme et al. 2005). The resulting alignment file was used to inform the probes. Probes that were within a region that mapped to the genome and that fell within 5 bases of a splice site were removed. Probes from a region of the transcriptome assembly that did not map to the genome and those that did not fall within 5 bases of a splice site were selected for downstream analysis.

2.2.5 Probe Description

Probes that survived filtering based on the previously mentioned technical specifications were further characterized. This was done to determine their similarity
between species and their potential behavior once submitted for scoring for the screening array. Probes for each species underwent three separate characterization steps: 1) the number of times each flanking sequence mapped to the *P. taeda* genome assembly, 2) sequence similarity of flanking regions between each species, 3) number of flanking sites that map to the same location in the *P. taeda* genome.

1. Comparison of the flanking sequence similarity between species was done under three conditions: allowing for zero, one, or two bases different to be considered shared. A flanking sequence was considered shared between two species if it did not exceed the number of bases different allowed. For condition one, allowing no differences (i.e. sequences must be identical to be considered shared), custom R and SAS scripts were used to create an incidence matrix (with flanking regions as the rows and species as columns). For conditions two and three, a custom Python script was used to calculate similar incidence matrices.

2. The characterization of a probe’s flanking site’s ability to map to more than one location to the *P. taeda* genome was done utilizing BWA mem (Li and Durbin 2009). A browser extensible data (bed) file was created for both the right and left flanking regions for each probe using a custom R script and then modified using Bedtools v2.27 (Quinlan and Hall 2010) slop routine to select the 35 bases either up and downstream from the SNP, specifying the positions of either the left or right flanking region. The resulting bed files were converted to fasta format using Bedtools v2.27.1 getfasta routine and then converted to fastq format via BBmap v.37.41 (Bushnell, sourceforge.net/projects/bbmap). An arbitrary base quality score of 35 was used in the fastq generation step. Once formatted into fastq
format, probe sequences were aligned back to the genome using BWA mem with
the –a (all alignment) option, which gives all possible alignments for a read
regardless of its mapping score or redundancy.

3. The left and right flanking regions of the probes were assessed for shared
mapping locations using BWA aln/samse (Li and Durbin 2009). For each species
specific probe set, a bed file was generated using a custom R script. Flank specific
bed files were created in the same manner as above, using Bedtools' slop routine
and targeting the 35 bases up or downstream of the SNP location. Flanking
specific bed files were then converted to fasta file format using Bedtools getfasta
and subsequent fasta files were converted to fastq files using BBMap v.37.41
(Bushnell, sourceforge.net/projects/bbmap) and a base quality score of 35 was
assigned to all bases in a read. The resulting fastq files were aligned to the P.
taeda genome using the optimized short read aligner BWA aln/samse routine. The
output alignment files were converted to bed files using Bedtools bamtobed
routine. The species specific bed files were then concatenated to create a “master”
alignment bed file. Using Bedtools’ intersect routine, the species specific bed files
were compared to the “master” file. If mapping locations were within two bases
of each other, the site was considered shared by those flanking regions. An
incidence matrix was constructed for downstream analysis.

Principal component analysis (PCA) was performed to assess probe correlation
between species. PCA was performed on the incidence matrices generated for sequence
similarity and shared mapping assessments. PCA for sequence similarity was performed
using the zero difference allowed matrix for both the left and right flanking regions. A
random sample of 50,000 observations from the matrices were selected and used for the analysis.

2.3 Results

2.3.1 RNA-Sequencing and Alignment

RNA-sequencing of the five tropical pine species generated between 6.3 billion and 5.0 billion raw bases from between 33.7 million and 45.4 million paired end reads on average for each species. After quality control, trimming reads for base quality and read length, an average of between 4.1 billion and 5.6 billion bases from between 29.1 million to 40.7 million reads were retained for further analysis. Average read length across the species ranged from 133 to 143 bases. Alignment of each species' quality controlled reads to their respective transcriptome assemblies showed between 73.6% to 78.3% of reads successfully aligned to their reference (Table 2.1). Of the reads that successfully aligned to their transcriptome, more than 90% of reads mapped in their proper pair.

2.3.2 Variant Calling and Probe Design

Initial variant calling for strict and standard parameters generated a list of variants including, complex events, multi-nucleotide polymorphisms, SNPs, and indels. Variant calling was not restricted to SNPs alone per recommended best practices for variant calling using Freebayes (Garrison & Marth, 2012) in order to increase detection power. The number of variants called ranged from 845K to 1.6 million under standard parameters and between 531K to 974K variants under strict per species. The raw variant vcf files were filtered to include only SNP markers, removing complex events, multi-nucleotide polymorphisms, and indels. This filtering produced reduced the dataset of between 686K to 1.1 million and between 426K to 824K SNPs per species for standard and strict calls respectively. Pinus
"patula" had the highest number of SNPs discovered within the species while "P. greggii" had the lowest number of SNPs generated from sequencing data (Table 2.2).

Strict and standard parameters were used to create a more robust catalogue of SNPs for probe design than using either one set of parameters individually. SNPs generated from strict parameters created a higher confidence catalogue of SNPs of which we expect a larger proportion of true SNPs rather than artifacts from sequencing or alignment errors than standard parameters. However, lower confidence SNPs generated from standard parameter were included for probe design as validation of SNP probes will be performed using a screening genotyping array, allowing for some leniency in the design phase. From the two datasets, between 175K to 301K probes were successfully designed per species (Table 2.2) that were then further characterized. In total 1,319,369 probes were successfully created from the design phase. These probes fell into three categories based on presence of a full 35mer flanking site: left flank only (442K), right flank only (425K), or both (451K).

Of the ~1.3 million SNPs across all species that had probes successfully designed for them, approximately 61% were transitions (C/T or A/G) and 39% were transversions (A/C, A/T, C/G, or G/T). Across all species, the C/T transition was the most prevalent making up 31% of the total number of observations while A/T transversions were the least observed making up 9% of observations. At the species levels, transition/transversion ratios ranged from 1.46 in "P. greggii" to 1.59 in "P. patula" (Table 2.3). Architecture of transition and transversion numbers were consistent across species.
2.3.3 Probe Description

The left and right flanking regions of the probes generated were assessed separately for sequence similarity, repetitive mapping, and shared mapping locations. In order to be included in the dataset a flank must have reached the full 35bp length for that flank side.

Sequence similarity assessment of the probe's left and right flanks showed a high proportion of unique flanking regions within each species. As the number of allowed base differences to be considered shared was increased, the proportion of flanking regions shared between two and three species increased while those shared between four or more species remained fairly constant. The sharing had a similar pattern for both right and left flank datasets (Figure 2.2). Looking at all possible combinations of sharing between species, patterns become apparent between the species. For the left flanking regions allowing for zero base differences *P. maximinoi* had the largest proportion of unique flanks, while sharing between species such as *P. tecunumanii* and *P. oocarpa* occurred at a higher proportion than with other species (Figure 2.3). This pattern holds constant across left and right flanks for each number of differences allowed. Diagrams for the other combinations of flanks and differences allowed are found in the Appendix (Appendix 1-5). Principal component analysis (PCA) was performed on the left and right flanking regions when allowing for zero differences datasets. PCA showed a strong correlation between species that are more closely phylogenetically related (Figure 2.4). Right and left regions produced similar results with ~29% and ~23% of variance explained by component 1 and 2 respectively.

Repetitive mapping analysis of the left and right flanking 35mer regions to the *P. taeda* genome revealed that over 50% of flanking regions mapped to a unique location for each species (Figure 2.5). Approximately 17% of flanking regions mapped to more than five
locations in the genome reference with a maximum number of 511 locations for a single
flank found. Alignment percentage for all species' left and right flanking regions was greater
than 90%.

Analysis of the shared mapping locations of left and right flanking regions revealed a
high proportion of flanks mapping to a unique location in the v2.01 *P. taeda* genome
assembly (Figure 2.6). *Pinus maximinoi* once again exhibited the largest number of unique
locations distinguishing it as an outlier in behavior when compared to the other species. More
than 83% of flanks successfully mapped to the reference for all species. When examining all
possible combinations of sharing, a relatively small number of mapping locations were
shared for all species. *Pinus tecunumanii* and *P. oocarpa* tended to have a larger number of
shared locations between the species when compared with the others as did *P. greggii* and *P.
patula* for both left and right flanking datasets (Figures 2.7 & 2.8).

Principal component analysis revealed 29% and 22% of variance associated with
shared mapping location were explained by component one and two respectively for both the
left and right flanking region. Correlation between pairs of species were shown with *P.
patula* / *P. greggii* and *P. tecunumanii* / *P. oocarpa* pairs having loading angles which are
close to one another (Figure 2.9). Left and right flanks paired in a similar manner indicating
biological consistency between datasets.

2.4 Discussion

Pines, as important long-term rotation crops, have been researched extensively to
develop useful statistical models to estimate genetic parameters using phenotypic data. More
recently, implementation of molecular technologies has aimed to improve the rate and
efficiency associated with tree improvement. However, when compared with annual rotation
crop species, molecular markers and genetic maps are much less developed (Isik 2014). The use of NGS technology has the potential to facilitate whole genome or whole transcriptome level molecular markers to be used in breeding programs. While there have been genome assemblies of several pine species developed (Nystedt et al. 2013; Stevens et al. 2016; Zimin et al. 2017), the complexity and size of pine genomes make whole genome level sequencing efforts difficult when compared to other plant species. Due to the cost and labor associated with whole genome assembly, work to develop a reference has not been undertaken for the species in this study. Therefore, this study focused on using de-novo assembled transcriptome references and NGS RNA-Seq data to identify SNP markers using a custom bioinformatics pipeline.

Challenges associated with using RNA-Seq for create a number of potential avenues for the introduction of errors into an experiment. Despite NGS sequencing technologies have an error rate as low as 0.1% (Glenn, 2011), incorrectly sequenced bases are still a source of false SNPs. Additionally, in plants and especially pine species where repetitive sequences can make up more than 80% of a genome (Wegrzyn et al. 2014), de novo assembly of transcriptomes and subsequent variant calling efforts can be effected by paralogous gene families and multi-mapping reads that may be present (Pepke et al. 2009). Of the strategies available to handle multi-mapping reads, this study utilized the best match method, reporting the alignment with the fewest mismatches and randomly assigning a location between those with equivalent mapping scores.

Assessment of the transition (Ts) vs transversion (Tv) ratio of the study species yielded ratios of between 1.46 to 1.59. Even though Ts:Tv ratios have not been reported in the species present in this study, the numbers are consistent with the ratios found in
transcriptomic studies of other tree species. SNPs generated from the analysis apex shoot tissue in eucalyptus has shown a Ts:Tv ratio of 1.62 and foliar analysis of Norway spruce produced a ratio of 1.49 (Singh et al. 2011; Chen et al. 2012). The divergence from the 1:2 ratio expected from random mutations indicates the SNPs called in this study are probably true SNPs and not errors from sequencing or mapping. Bias due to higher rates of C/T mutations because of CpG effects and the deamination of methylcytosines in CpG regions produces a much higher Ts rate and skews the ratio (Vignal et al. 2002; Morton et al. 2006). Results from this study follow this trend, with C/T transitions being found in the highest proportion, making up 31% of SNPs with successfully designed probes.

A unique challenge associated with this study was to identify SNPs and probes that are potentially shared between species when generated from different reference sequences. To address this issue, probe flanking regions were used as a surrogate of the SNP they surrounded. Comparative transcriptomic studies have shown reasonable alignment of transcripts of relatively diverged pine species to the *P. taeda* genome (Baker et al. 2018). Therefore, probe flanking regions which originated from transcriptome assemblies were mapped to v2.01 *P. taeda* genome to be used as a common reference and assessment of sequence similarity between probes was performed to gain insight to the potential performance of a probe on the commercial array. Successful alignment of flanking regions to the genome was higher than shown in Baker et al. (2018). However, this is not surprising given the read length being mapped is lower in this study. Very little expansion of pine transcriptomic regions has been shown when compared to other plant species (Kirst et al. 2003) so many of the non-aligning flanking regions could be attributed to the fragmented and incomplete nature of the *P. taeda* genome rather than absence of sequence.
Sequence similarity between probes and multiple locations within a genome has been shown to cause nonspecific binding and subsequent cross reactivity. This cross reactivity can generate false positive signals which will skew downstream analysis. It has been shown that sequences not targeted by probes which have >70-80% similarity with target sequences will affect signals generated for 50mer probes (Kane et al. 2000). The presence of repeat regions in the genome either through paralogous gene families or pseudogenes also has an effect on probe specificity (Chen et al. 2011). This is evident in this study by the >40% of flanking regions mapping to multiple sites within the P. taeda genome.

Applying the logic above to the probe sequences themselves allows for identification of probes that may behave in a similar manner and thus may be redundant. Both the sequence similarity and shared mapping locations assessments performed to assess species sharing and redundancy required greater than 94% similarity in flanking regions to be considered shared, much higher than the 80% mark mentioned above. However, direct comparisons cannot be made given the different probe lengths between studies which would lead to different dynamics in hybridization. While sequence similarity is one factor that affects nonspecific hybridization, others such as location of base differences within the flanking region and if the difference is A/T or C/G can also influence binding (Chen et al. 2011) and were not taken into account in this study.

SNPs generated from this study are expected to follow the phylogenetic pattern found in biology through flanking region similarity as well as shared mapping locations. Lacking a common reference and due to the unknown family sampling in the study, a true population structure analysis was not achievable. Therefore, PCA was performed on the incidence matrices for both sequence similarity and shared mapping to assess if SNPs are being shared.
between species in a similar pattern as expected from biology. A phylogenetic study of Mexican and Central American pines using RAPD makers showed relatedness between *P. greggii* and *P. patula* species is higher than with other species of tropical pines. The same was shown for *P. oocarpa* and *P. tecunumanii* (Dvorak et al. 2000b). Another study supports this representation of phylogeny, and places *P. maximinoi* as the most distantly related to the study species (Vargas-Mendoza et al. 2011). PCA on both shared mapping and sequence similarity between flanking regions followed this pattern shown in the phylogenetic studies, further indicating the SNPs found in this study are likely true and will have utility on the screening array for population genetics and hybrid studies.
2.5 Literature Cited


FAO (2015) Global Forest Resouces Assessment


Gwaze DP (1999) Performance of some F1 interspecific Pine hybrids in Zimbabwe. For


Liu JJ, Sniezko RA, Sturrock RN, Chen H (2014) Western white pine SNP discovery and
high-throughput genotyping for breeding and conservation applications. BMC Plant Biol. doi: 10.1186/s12870-014-0380-6

Michael TP, Jackson S (2013) The First 50 Plant Genomes. Plant Genome. doi: 10.3835/plantgenome2013.03.0001 in


nucleotide polymorphisms and insertions/deletions in expressed sequence tag data of
Eucalyptus. J Genet 90:e34–e38

Genetics 204:1613–1626. doi: 10.1534/genetics.116.193227

transcriptome sequencing. PLoS One. doi: 10.1371/journal.pone.0205835

polymorphism (ESTP) markers in loblolly pine (Pinus taeda L.). Theor Appl Genet
102:664–675

camaldulensis seedlings subjected to water stress reveals functional single nucleotide
polymorphisms and genes under selection. BMC Genomics. doi: 10.1186/1471-2164-
13-364

of Mexican Pine Species Based on Three Loci from Different Genomes (Nuclear,
Mitochondrial, and Chloroplast). In: Agboola J (ed) Relevant Perspectives in Global
Environmental Change. InTech, pp 139–154

Vignal A, Milan D, Sancristobal M, Eggen A (2002) A review on SNP and other types of
molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305. doi:
10.1051/gse:2002009


Table 2.1. Sequencing and alignment metrics generated from RNA sequencing. Sequencing generated millions (M) of reads and billions of bases on average per species. Raw reads underwent quality control (QC) requiring a base quality score of 30 and minimum read length of 50bp. Average read length decreased slightly from the 150bp raw reads generated from sequencing. Average alignment percentage across species was consistent.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>P. greggii</th>
<th>P. maximinoi</th>
<th>P. oocarpa</th>
<th>P. patula</th>
<th>P. tecunumanii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Reads</td>
<td>33.7 M</td>
<td>36.0 M</td>
<td>35.8 M</td>
<td>45.9 M</td>
<td>42.9 M</td>
</tr>
<tr>
<td>Raw Bases</td>
<td>5,100 M</td>
<td>5,400 M</td>
<td>5,400 M</td>
<td>6,400 M</td>
<td>6,000 M</td>
</tr>
<tr>
<td>QC’d Reads</td>
<td>29.1 M</td>
<td>32.3 M</td>
<td>31.7 M</td>
<td>40.7 M</td>
<td>37.4 M</td>
</tr>
<tr>
<td>QC’d Bases</td>
<td>4,200 M</td>
<td>4,700 M</td>
<td>4,500 M</td>
<td>5,600 M</td>
<td>5,130 M</td>
</tr>
<tr>
<td>QC’d Read Length</td>
<td>143bp</td>
<td>145bp</td>
<td>142bp</td>
<td>133bp</td>
<td>133bp</td>
</tr>
<tr>
<td>Alignment Percent</td>
<td>76.3</td>
<td>78.3</td>
<td>77.5</td>
<td>73.6</td>
<td>77.1</td>
</tr>
</tbody>
</table>
Table 2.2. SNP numbers and probe design metrics for SNPs generated under standard parameters. The number of SNPs identified from variant calling generally numbered in the millions with the exception of *P. greggii* which had approximately 680K identified. Of the SNPs generated, less than one third were successfully converted to a probe. Probes tended to have a high number of unique left and right flanking regions, showing little redundancy within species.

<table>
<thead>
<tr>
<th>Metrics</th>
<th><em>P. greggii</em></th>
<th><em>P. maximinoi</em></th>
<th><em>P. oocarpa</em></th>
<th><em>P. patula</em></th>
<th><em>P. tecunumanii</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs identified</td>
<td>680 K</td>
<td>1,110 K</td>
<td>1,030 K</td>
<td>1,390 K</td>
<td>1,110 K</td>
</tr>
<tr>
<td>Probes designed</td>
<td>175 K</td>
<td>293 K</td>
<td>301 K</td>
<td>290 K</td>
<td>259 K</td>
</tr>
<tr>
<td>Unique Left Flanks</td>
<td>87 K</td>
<td>185 K</td>
<td>135 K</td>
<td>138 K</td>
<td>109 K</td>
</tr>
<tr>
<td>Unique Right Flanks</td>
<td>86 K</td>
<td>182 K</td>
<td>132 K</td>
<td>135 K</td>
<td>107 K</td>
</tr>
</tbody>
</table>
Table 2.3. Catalogue of the number of transitions (Ts) and transversions (Tv) per species.

Ts:Tv ratios were consistent across species and within expected ranges for SNPs generated from transcriptome sequences.

<table>
<thead>
<tr>
<th>Model</th>
<th>P. greggii</th>
<th>P. maximinoi</th>
<th>P. oocarpa</th>
<th>P. patula</th>
<th>P. tecunumanii</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/C</td>
<td>17 K</td>
<td>27 K</td>
<td>28 K</td>
<td>28 K</td>
<td>25 K</td>
</tr>
<tr>
<td>A/G</td>
<td>48 K</td>
<td>82 K</td>
<td>85 K</td>
<td>79 K</td>
<td>72 K</td>
</tr>
<tr>
<td>A/T</td>
<td>16 K</td>
<td>25 K</td>
<td>25 K</td>
<td>24 K</td>
<td>21 K</td>
</tr>
<tr>
<td>G/C</td>
<td>18 K</td>
<td>30 K</td>
<td>31 K</td>
<td>29 K</td>
<td>26 K</td>
</tr>
<tr>
<td>C/T</td>
<td>52 K</td>
<td>89 K</td>
<td>91 K</td>
<td>93 K</td>
<td>81 K</td>
</tr>
<tr>
<td>G/T</td>
<td>18 K</td>
<td>29 K</td>
<td>30 K</td>
<td>27 K</td>
<td>25 K</td>
</tr>
<tr>
<td>Ts:Tv</td>
<td>1.46</td>
<td>1.54</td>
<td>1.55</td>
<td>1.59</td>
<td>1.57</td>
</tr>
</tbody>
</table>
Figure 2.1. Work flow for SNP discovery using RNA-seq data. Work flow includes two phases, bioinformatics (blue) and probe design (purple). Software used and file outputs are shown in the ovals underneath the colored boxes.
Figure 2.2. Sharing pattern for each species' left and right flanking regions when allowing for zero, one, or two differences within a flanking region to be considered shared. For each species, a large portion of flanking regions are unique to that species regardless of differences allowed. Allowing for sequence differences increases the number of flanking regions shared by two or three species at a larger proportion than sharing between four or all.
Figure 2.3. Sharing of left flanking regions allowing for zero differences across species. *P. maximinoi* has the highest number of unique flanking regions while *P. tecunumanii* and *P. oocarpa* have a high number of shared flanks between the two as expected from phylogeny.
Figure 2.4. Principal component analysis of left flanking (a) and right flanking (b) regions when allowing for zero differences in flanking sequence. Correlation between species that are more closely related phylogenetically is strong as indicated by loading angles. The more closely related *P. tecunumanii* / *P. oocarpa* and *P. patula* / *P. greggii* pairs have angles that are close to one another. The more distantly related *P. maximinoi* segregates itself.
Figure 2.5. Repetitive mapping of left and right flanking regions for each species when mapped to the v2.01 *P. taeda* genome assembly. Over 50% of flanking regions in all species map to a unique location within the genome reference.
Figure 2.6. Sharing pattern for each species’ left and right flanking regions when mapped to v2.01 of the *P. taeda* genome assembly. Flanking regions are considered shared if they mapped within two bases of another. A large portion of flanking regions are unique to a given species while very few map to the same location in all species.
**Figure 2.7.** Distribution of left flanking regions shared mapping locations. Each species contains a large number of regions unique to the individual. *Pinus patula* / *P. greggii* and *P. tecunumanii* / *P. oocarpa* species combinations sharing mapping locations at a higher proportion between themselves than with other species.
Figure 2.8. Distribution of right flanking regions shared mapping locations. A relatively small number of regions are shared between four or five species. *Pinus maximinoi* has the largest number of uniquely mapping regions and species groupings are consistent with that found in the left regions.
Figure 2.9. Principal component analysis of left flanking (a) and right flanking (b) regions shared mapping locations. Correlation between species that are more closely related phylogenetically is strong as indicated by loading angles. The more closely related *P. tecunumanii* / *P. oocarpa* and *P. patula* / *P. greggii* pairs have angles that are close to one another, while the more distantly related *P. maximinoi* distant from any other species.
Chapter 3. Single Nucleotide Polymorphism Discovery and Probe Design using Targeted Sequencing Data of Tropical and Subtropical Pine Species

3.1 Introduction

The genus *Pinus* consists of over 100 species across the globe which are found naturally almost exclusively in the northern hemisphere (Price et al. 1998). Across their global range, there are two primary hotspots of biodiversity; one in China and another in Mexico and Central America (Plomion et al. 2007). Amongst the tropical and subtropical pine species found in Mexico and Central America, many are considered to be threatened or endangered in their natural range. The threat to their natural populations has in turn spurred conservation efforts and led to the establishment of commercial land race populations for species in South America and in southern Africa (Dvorak et al. 2000a; Griess et al. 2016). Of these tropical and subtropical pine species, six species were selected for this study due to their importance to commercial timber markets and ecological significance: *Pinus greggii*, *Pinus maximinoi*, *Pinus oocarpa*, *Pinus patula*, *Pinus tecunumanii*, and *Pinus caribaea*. Being able to better understand the amount of genetic diversity present within these species and leveraging that information is crucial to effectively manage them for commercial use and gene conservation efforts. Utilizing molecular genetic data generated by the use of next generation sequencing technologies shows great promise in helping to expand our knowledge of such species and allow for more efficient management of genetic resources.

Next generation sequencing technologies have revolutionized the field of genetics by facilitating the use of high throughput genotyping to make inferences about population
structure and diversity of many different species. Whole genome resequencing studies have been performed on a number of important plant species to date (Tuskan et al. 2006; Myburg et al. 2014). However, these sorts of studies remain cost prohibitive and computationally inefficient for many species with large repetitive genomes such as those found in conifers. Within conifers, only a handful of whole genome sequencing projects have been undertaken. Reference genomes of *Pinus taeda* (Zimin et al. 2017), *Picea abies* (Nystedt et al. 2013), and *Pinus lambertiana* (Stevens et al. 2016) are a few that have been assembled to date. The prohibitive cost and complexity associated with whole genome sequencing has generated methods that can be used to overcome these issues. The utilization of transcriptome and reduced representation sequencing alleviates much of the complexity associated with whole genome sequencing by focusing on coding and regulatory sequences in the genome.

Reduced representation sequencing such as targeted sequence capture have three advantages over sequencing of an entire genome: reduction of sequencing space allows for multiplexing of samples, and targets only areas of the genome that are informative. It also aides in identifying regions of interest across populations using intraspecific genomic assays (Grover et al. 2012). Targeted sequence capture uses the hybridization of synthetic probes to sheared genomic DNA to selectively target regions of interest in the genome, which are then enriched and sequenced for further study (Vidalis et al. 2018). Hybridization based methods of sequence capture using synthetic designed probes have been successfully exhibited in a number of plant species with complex genome structures: *P. taeda* (Neves et al. 2013; Lu et al. 2016), *P. abies* (Vidalis et al. 2018), *Pinus albicaulis* (Syring et al. 2016), and *Picea mariana* (Pavy et al. 2016).
Additionally, target capture probes that have been developed for *P. taeda* have shown high transferability across pine species (Neves et al. 2013). This makes the genomic resources developed for *P. taeda* an attractive platform to extend these technologies into pine species that lack genome reference assemblies of their own, such as tropical pines mentioned before. Information gained from such experiments will prove to be valuable for the continued improvement of pure and hybrid breeding populations, as well as gene conservation efforts, for tropical and subtropical pine species.

The objectives of this study were to 1) utilize targeted sequence capture methods for single nucleotide polymorphic (SNP) loci discovery in six tropical or subtropical pine species, 2) assess SNPs and their probes for their potential use on a commercial high throughput genome wide genotyping array.

3.2 Methods and Materials

3.2.1 Plant Material and DNA Isolation

The population used in this study originated from the Pine Genomic Atlas Study that characterizes the genetic diversity and structure of commercially important pine species in southern Africa (Tii-kuzu, MSc in preparation). A total of six species were selected for this study: *P. greggii*, *P. maximinoi*, *P. oocarpa*, *P. patula*, *P. tecunumanii*, and *P. caribaea*. Within these species, sub-species were represented for *P. greggii* (north and south), *P. patula* (var. *patula* and var. *longipedunculata*), and *P. tecunumanii* (high and low elevation). Foliage samples were taken from first generation genetic material established in Camcore field trials across South Africa for the extraction of genomic DNA. For the six species, a total of 81 pooled samples were created. Each pooled sample contained DNA between 4 to 8 trees from different families and represented a single provenance of a given species. A total of 567
trees were sampled (Table 3.1). The 81 provenances (Appendix 6) were selected, covering the natural range of the six species in Mexico and Central America in order to create a more robust and diverse dataset.

Genomic DNA extraction was processed by the Forest Molecular Genetics Program’s DNA Fingerprinting Platform (University of Pretoria, South Africa). DNA was isolated from 50mg of fresh needle tissue using the NucleoSpin® Plant II DNA extraction kit (Machery-Nagel, Germany) according to the manufacturer’s specifications. DNA was quantified using a MultiSkan™ Go Microplate Spectrophotometer (Thermo Scientific, MA, USA). Approximately 100ng of genomic DNA per sample was shipped to RAPiD Genomics (Gainesville, FL, USA) where equimolar pooled samples for each provenance were created for targeted capture sequencing.

### 3.2.2 Capture Probe Design and Target Enrichment

DNA was mechanically sheared to an average size of 400bp using a Covaris E210 focused ultrasonicator (Covaris, Woburn, MA). Libraries were constructed by repairing the ends of the sheared fragments using the End-It DNA End-Repair kit (Epicentre Biotechnologies, Madison, WI), producing blunt end fragments. Ligation of a single adenine residue to the 3’ end of the blunt end fragment was performed using 15-U Klenow fragment (New England Biolabs Inc., Ipswich, MA) and deoxyadenosine triphosphate (dATP) (Promega, Madison, WI). Barcoded adapters that are suited for Illumina Sequencing were ligated to the libraries and the ligated fragments were PCR-amplified using standard cycling protocols (Mamanova et al. 2010). A custom set of 40K target capture probes were developed by RAPiD Genomics to facilitate target capture sequencing. Of these 40K probes, 30K were designed from single copy regions of the v2.01 *Pinus taeda* genome assembly.
(Zimin et al. 2017) and 10K were designed from the *P. patula* and *P. tecunumanii* transcriptome assemblies (Visser et al. 2015, 2018). A total of 16 barcoded libraries were pooled with equimolar amounts to a total of 500ng of DNA for hybridization to the target capture probes.

For target capture enrichment, the Select XT Target Enrichment System for solution-based target enrichment of Illumina paired-end data by Agilent (Palo Alto, CA, USA) was used. After enrichment, samples were re-amplified for an additional 6-14 cycles. All samples were sequenced using Illumina HiSeq 3000 with paired end 150bp reads. Samples were then de-multiplexed using Illumina’s BCLtofastq.

### 3.2.3 Quality Control and Alignment

The paired end reads generated from the 40K capture probes underwent processing through a custom SNP discovery pipeline (Figure 3.1). Quality control was performed on the reads using FastQC (Andrews 2010) for visualization. Raw reads that did not meet base quality or length standards were trimmed and filtered using Sickle v1.33 (Joshi & Fass 2011). Standard procedures used required a base quality score of 30 and a minimum read length of 50bp. Trimmed reads from the 81 libraries were aligned to a modified version of the v2.01 *P. taeda* genome assembly.

Using SNP calls generated from a pilot study of the current study (unpublished data), v2.01 of the *P. taeda* genome assembly was modified to create the modified reference used in this study. The modified reference was created by switching the alternative and reference alleles at any location where the alternative allele appeared in more than 60% of the total number of observations. The alternative allele coverage was calculated per SNP across all provenances by \( \frac{AO}{(AO+RO)} \), where AO is the number of alternative observations and RO
is the number of reference observations. Approximately 25K bases were changed and the resulting modified *P. taeda* genome will be referred to as tropical pine reference genome going forward. Using Burrows-Wheeler Aligner (BWA) v0.7.15 (Li and Durbin 2009), the tropical pine reference genome was indexed using the index routine and trimmed reads aligned using the mem routine under standard parameters. The SAM files generated from BWA were converted to BAM files and sorted using Samtools v1.7 (Li et al. 2009) view and sort routines respectively.

3.2.4 Variant Calling and SNP Probe Design

Reads that did not align within 500bp on either side of a capture probe location were removed from the datasets prior to variant calling. Reads were filtered using Samtools v1.7 view routine, to include only those reads which had the sam flag 2. Including only reads mapped in their proper pair and within the window of a capture probe were kept for downstream analysis. Variant calling was performed using Freebayes v1.2.0-2-g29c4002 (Garrison & Marth 2012). Variant calls were done jointly; using each of the 81 filtered sorted BAM files as input and utilized the pooled-discrete and cnv_map parameters to specify ploidy in each of the 81 pooled samples. Additional parameters used include: min-alternative-fraction = 0.01, min-alternative-count = 2, min-coverage = 8, min-mapping-quality = 1, min-base-quality = 20, and report-genotype-likelihood-max.

The raw variants that were called were further filtered. Using vcflib v1.0.0-rc1 (Garrison 2016), indels, multi-nucleotide polymorphisms, and complex events were removed, leaving only SNPs for analysis. This SNP only dataset was filtered using two processes, each filtering step was done independently and did not influence the results of the other. The top down process filtered SNPs across all samples based on the following criteria:
an overall alternative allele fraction (AAF) > 0.1 and overall coverage > 100X. The bottom up process used the following criteria: maximum within sample AAF > 0.2 and maximum within sample alternative observation > 4. If a SNP was in either post filter dataset, it was selected for probe design. The bottom up process filtered SNPs within sample, where if any sample for a given SNP met criteria the SNP was kept.

SNPs that passed filtering underwent probe design. A probe in this case is defined as a SNP plus the sequence of a set number of DNA bases on its upstream and downstream flank. Probes were designed using technical specifications outlined for use on ThermoFisher Scientific’s Axiom genotyping array platform. Technical specifications required at least one region, from a SNP, either upstream or downstream, to be free of another SNP or indel for 35 bases. These upstream and downstream regions around the SNP are also referred to as left and right flanking regions.

A set of custom Perl scripts was used to extract the 35 bases upstream and downstream of each SNP. International Union of Pure and Applied Chemistry (IUPAC) codes were then inserted into the probe flanking regions in order to denote the location of a SNP within the flanking region. The flanking regions for each SNP probe were then trimmed to the innermost IUPAC code and those probes without at least a right or left flanking region of 35 bases were removed.

3.2.5 SNP and Probe Description

SNPs that had probes successfully designed were further characterized, examining their relatedness between and across species to gain insight on how they may behave when used on the screening array. SNPs were characterized at the species, subspecies and provenance levels through creation of incidence matrices and principal component analysis.
(PCA). Furthermore, left and right SNP probe flanking regions that reached the full 35mer length were assessed for repetitive mapping to the tropical pine reference genome.

Species, subspecies, and provenance level incidence matrices were created for all SNPs generated by variant calling. The species and subspecies matrices required a within sample (species or subspecies) AAF > 0.05 and AO > 4 for a SNP to be considered present in the specific species or subspecies. The provenance matrix required a within sample (provenance) AAF > 0.1 and AO > 2 for a SNP to be considered present within a given provenance. The generated matrices were then subset to include only those SNPs that had a probe successfully designed for it and were used to illustrate how SNPs were shared among the species. Principal component analysis was performed at the species and subspecies levels on a random sample of 50K observations within the incidence matrices to assess concurrence of SNP between species and subspecies.

The repetitive mapping of flanking regions to the tropical pine genome assembly was done utilizing a series of software. Left and right flanking regions were assessed independently. A bed file denoting the SNP location was created using a custom R script for each dataset. Using Bedtools’ v2.27.1 (Quinlan and Hall 2010) slop routine, the bed files were adjusted to target the regions of interest for each dataset, either the 35 bases left or right of the SNP. The bed files were converted to fasta and then fastq file formats using Bedtools v2.27.1 getfasta routine and BBMap v37.41 (Bushnell, sourceforge.net/projects/bbmap) respectively. Arbitrary base scores of 35 were used in the generation of the fastq files. The fastq files for each species were aligned to the tropical pine genome assembly using BWA’s mem routine and the -a parameter, outputting all alignments for a read regardless of score or redundancy.
3.3 Results

3.3.1 Target Capture Sequencing and Alignment

Targeted sequencing capture of the six species generated between 4.6 million to 5.3 million raw paired end reads made up of between 690 million to 790 million bases on average per species. After trimming reads to for a minimum base quality score and minimum read length, between 4.0 to 4.6 million trimmed reads consisting of between 590 to 650 million bases were used for downstream analysis (Table 3.2). Average read length ranged from between 141 to 146 bases for each species after trimming. Alignment of the trimmed reads to the tropical pine genome yielded very high alignment percentages of >98% for each species. Filtering of the reads to include only those that mapped in proper pair and within 500 bases on either side of an area targeted for sequencing reduced the number of reads to be used in variant calling by ~57% across all species.

3.3.2 Variant Calling and Probe Design

Variant calling within properly aligned and paired reads using the tropical pine genome as reference generated a list of 4.3 million variants including complex events, indels, multi-nucleotide polymorphisms, and SNPs. Filtering of this set to include only SNP markers yielded a list of 3.4 million raw SNPs. Results from the top down filtering of the raw SNPs created a list of about 418K SNPs while bottom up generated a list of approximately 1.3 million SNPs. A total of 403K SNPs were common across both filtered datasets (Figure 3.2). Sets were merged to create a list of 1,356,522 SNPs for probe design.

Probe design processes performed on the approximately 1.3 million SNPs selected yielded a total of 562K probes. Of these probes, 228K had only a left flanking region clear of another SNP while 227K were generated with only a right flanking region meeting technical
specifications. 107K probes were generated with both a full left and right 35mer flanking sequence clear of another SNP.

### 3.3.3 SNP and Probe Characterization

SNPs with successful probes were further characterized at the species, subspecies, and provenance level. About 26K SNPs are shared across all study species (Figure 3.3). *Pinus caribaea* and *P. maximinoi* acted as outliers with a much lower proportion of SNPs shared between two, three, or four species, making up 29% and 13% of shared SNPs respectively. *Pinus maximinoi* contained a much higher proportion of private SNPs (63%) when compared to all other species. Looking at the distribution of all possible sharing combinations across species, *P. oocarpa* and *P. tecunumanii* shared SNPs at a slightly greater proportion (13%) with each other than with other species, none of which exceeded 10% sharing. *Pinus maximinoi* once again exhibits a low proportion of sharing across all other species and has a high number of private SNPs (Figure 3.4). SNP sharing at the subspecies level exhibited the same overall pattern as at the species level. Across subspecies, about 24K SNPs were shared while about 26K SNPs were shared between eight subspecies. Subspecies of the same species had very similar proportion of SNP sharing (Figure 3.5). *Pinus maximinoi* exhibited a much higher proportion (63%) of private SNPs than other subspecies. At the provenance level, a small number of SNPs were present across all 81 provenances or were unique to a single provenance. *Pinus maximinoi* once again acted as an outlier. When looking at the pattern of sharing across provenances it had a much higher proportion being shared between 2-14 than the other categories (Figure 3.6). All other provenances exhibited a similar pattern between them.
Principal component analysis performed on the species and subspecies SNP sharing matrices both showed *P. maximinoi* behaving very differently than the other species (Figure 3.7). At the species level, 46% and 17% of the variance was explained by component one and two respectively. Loading angles of the species, except for *P. maximinoi*, were fairly similar indicating correlation between them. Looking at the subspecies level, 48% and 11% of variance was explained by component one and two respectively. Loadings segregated by subspecies within species, grouping *P. greggii*, *P. patula*, and *P. tecunumanii* subspecies together, indicating stronger correlation between closer related species.

Successfully designed probes had both their left and right flanking regions assessed for repetitive mapping to the tropical pine genome reference. The range of times mapped for a flanking region ranged from 1 to >500. However, the majority of left and right flanking regions mapped to a unique location to the genome (Figure 3.8).

3.4 Discussion

Reduced representation sequencing models such as targeted capture sequencing are attractive alternatives to whole genome resequencing studies for non-model organisms that have large, repetitive, and incomplete genome assemblies such as conifers. Despite their massive size, many over 20Gb (De La Torre et al. 2014), coding regions in pine species seem to have not expanded much beyond that of other vascular plants (Kirst et al. 2003). Instead, much of the size seems to stem from the presence of repetitive elements. An estimated 82% of the *P. taeda* genome is assumed to be repetitive with 62% of the repeats stemming from retrotransposons (Neale et al. 2014) with the presence of pseudogenes contributing although to a lesser known degree (Kovach et al. 2010). Therefore, being able to selectively target single copy coding and regulatory sequences greatly reduces the computation and challenges
associated with these repetitive elements. This, coupled with the transferability of enrichment probes between related species (Neves et al. 2013), makes targeted sequencing an ideal method for this study.

This study exhibited one of the challenges associated with implementation of targeted sequencing in incomplete and poorly characterized genomes, that being off target capture. Off-target capture occurs either through the inadvertent targeting of multi-copy regions in the genome due to mistakes in the assembly the probes were designed from or due to evolutionary divergence between the target and reference species (Grover et al. 2012; Syring et al. 2016). In this study, an off-target capture of greater than 50% was shown across all species when enrichment libraries were mapped to the full tropical pine reference genome. This is a higher percentage than seen in other studies (Fu et al. 2010; Vidalis et al. 2018); although not unreasonable given most of the target capture probes were designed from a somewhat diverged species' genome assembly.

Investigation into this number by restricting mapping locations to only regions targeted by probes, showed large coverage differences in approximately 2,000 of the 40,000 targeted regions when compared to mapping to the full genome. Even though the target probes were designed to target single copy locations within the genome, it is possible that due to an error in annotation or the fragmented nature of the assembly, the 2K probes targeted one of the many repetitive elements in the genome. If the 2K probes were generated from the transcriptome assemblies of _P. patula_ or _P. tecunumanii_, it is possible that the probe locations appeared single copy but due to either the divergence of gene structure from _P. taeda_ or the incomplete assemblies for both the transcriptome and genome, they were not recognized as such. These reads that mapped off target were subsequently filtered in an effort
to reduce potential sources of false SNPs during the variant calling probe design stages (Vidalis et al. 2018).

In an effort to increase efficacy of the SNP probes when used on the screening array, the original 3.3 million SNPs were filtered based on allele frequency and depth statistics. The mean pre-filter depth per SNP was 22X when averaged across all individuals and 1700X when summed across individuals. Filtering metrics were based on eight trees per pooled sample. The bottom down filtering assumes good coverage for a SNP in each of the 81 provenance pools is 20X. Using the selected number of alternative observations AO of >4 and alternative fraction AF > 0.2, this equates to at least two of eight trees carrying the alternate allele. This effectively reduces SNPs that may be present in a single tree or that were generated due to sequencing and mapping errors in the downstream analysis. Top down filtering uses global (across species) metrics. Again, assuming a well-covered SNP is represented by 20X coverage in a provenance pool, an estimated 1600X total coverage would be expected if the SNP is present in all provenances. Using the selected AF > 0.1, this would correspond to an AO of 160 and indicate a SNP is present in at least 8 provenance pools. To reduce the stringency, a depth of 100X was selected indicating an AO of 10. Thus, filtering for SNPs that may be present in at least five provenance pools. These filtering criteria should reduce the number of provenance specific SNPs present in the dataset and produce SNPs shared by multiple species.

One of the challenges associated with multispecies and single species fixed genotyping arrays is the presence of ascertainment bias. Ascertainment bias stems from either nonrandom sampling or by SNP discovery protocols that favors those SNPs with higher minor allele frequencies (Lachance and Tishkoff 2013). This bias can become
troublesome when assessing population structure and diversity, which depends on allele frequencies and become problematic when genotyping across species (Albrechtsen et al. 2010; Garvin et al. 2010). SNPs called from this study are intrinsically biased towards shared regions within *P. taeda* genome due to the design of the target capture probes.

Although true ascertainment bias metrics will be assessed with the construction of the SNP screening array, in an effort to gain insight into this potential problem, SNP sharing at the species, subspecies and provenance level was assessed. This showed a large portion of SNPs being shared across many species, indicating the potential for ascertainment bias downstream. Additionally, PCA analysis on SNP sharing at the species and subspecies levels showed departure from what is expected from phylogeny. Phylogenetic analysis of many Mexican and Central American pines using RAPD markers has shown a strong relationship between *P. greggii* and *P. patula* species along with *P. tecunumanii* and *P. oocarpa* being more closely related. *Pinus caribaea* seems to be genetically equidistant from the other species in the study (Dvorak et al. 2000b). An additional study places *P. maximinoi* as a distinct genetic group in relation to the others (Vargas-Mendoza et al. 2011). At the species level, correlation of SNP sharing projected by PCA does not show a high level differentiation between species that are not closely related. Loading angles for all species other than *P. maximinoi* were fairly similar. At the subspecies level, subspecies of the same species shared almost identical angles. However, there was very little variation in the angles overall. This indicates that SNPs generated from this dataset have the potential to skew the screening array toward SNPs that are commonly shared and may not be useful as a single species identifier. However, with careful consideration on SNP selection, multispecies SNP chips have exhibited little bias and utility across species (Silva-Junior et al. 2015).
Eventually, after scoring of the probes and validation through a screening array, the final outcome of this project will be a multispecies commercial genotyping SNP array. The primary purpose for this array will be to accelerate breeding efforts through genomic selection, its utility in population genetics and hybrid species studies will also provide valuable resources to tree improvement and forest genetics professionals. It is estimated that genomic selection has the potential to increase selection efficiency per year by between 50-87% when compared to traditional pedigree based methods in scots pine (Calleja-Rodriguez et al. 2019). Recently, genomic selection has shown utility in hybrid populations of *eucalyptus* (Resende et al. 2017) underscoring the need for both markers that are transferrable across and unique to individual species. Additionally, species and breed specific markers have been shown to be useful in marker-trait association studies (Meuwissen et al. 2013; Pareek et al. 2017). Such SNP markers could prove valuable for the species in this study that have hybrid populations deployed commercially.
3.5 Literature Cited


Table 3.1. Provenance and tree sampling structure for each of the nine subspecies selected for targeted capture. Each provenance represents a pooled sample submitted for sequencing and contained between 4 to 8 trees in the sample. The 81 total provenances represented the natural range of the species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Provenances (pooled samples)</th>
<th>Trees sampled</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. caribaea</em></td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td><em>P. oocarpa</em></td>
<td>11</td>
<td>75</td>
</tr>
<tr>
<td><em>P. greggii N</em></td>
<td>10</td>
<td>67</td>
</tr>
<tr>
<td><em>P. greggii S</em></td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td><em>P. maximinoi</em></td>
<td>13</td>
<td>87</td>
</tr>
<tr>
<td><em>P. patula var. patula</em></td>
<td>7</td>
<td>56</td>
</tr>
<tr>
<td><em>P. patula var. longipedunculata</em></td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td><em>P. tecunumnaii HE</em></td>
<td>16</td>
<td>109</td>
</tr>
<tr>
<td><em>P. tecunumanii LE</em></td>
<td>9</td>
<td>65</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>81</strong></td>
<td><strong>567</strong></td>
</tr>
</tbody>
</table>
Table 3.2. Sequencing metrics generated from targeted capture sequencing. Sequencing generated millions of reads and bases on average per species. Raw reads underwent quality control trimming, requiring a base quality score of 30 and minimum read length of 50bp. Average read length decreased slightly from the 150bp raw reads generated from sequencing to ~143bp.

<table>
<thead>
<tr>
<th>Species</th>
<th>Raw Reads (millions)</th>
<th>Raw Bases (millions)</th>
<th>Trimmed Reads (millions)</th>
<th>Trimmed Bases (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. caribaea</td>
<td>4.60</td>
<td>693</td>
<td>4.00</td>
<td>588</td>
</tr>
<tr>
<td>P. greggii</td>
<td>5.30</td>
<td>795</td>
<td>4.58</td>
<td>652</td>
</tr>
<tr>
<td>P. maximinoi</td>
<td>4.92</td>
<td>737</td>
<td>4.20</td>
<td>598</td>
</tr>
<tr>
<td>P. oocarpa</td>
<td>4.83</td>
<td>724</td>
<td>4.12</td>
<td>586</td>
</tr>
<tr>
<td>P. patula</td>
<td>5.10</td>
<td>776</td>
<td>4.38</td>
<td>622</td>
</tr>
<tr>
<td>P. tecunumanii</td>
<td>5.27</td>
<td>791</td>
<td>4.52</td>
<td>641</td>
</tr>
</tbody>
</table>
Figure 3.1. Work flow for SNP discovery using targeted sequencing data. Work flow includes two phases, bioinformatics (blue) and probe design (purple). Software used and file routines are shown in ovals underneath the colored boxes.
Figure 3.2. The distribution of SNPs shared between the Top-down and Bottom-up filtering datasets. Between the two sets of SNPs generated from filtering approximately 403K SNPs were shared between the two. This resulted in approximately 1.36 million unique SNPs selected for probe design.
Figure 3.3. Sharing pattern for SNPs which had probes successfully designed for them by species. SNP sharing patterns among the species are relatively similar with the exception of *P. maximinoi* which contains a large number of private SNPs. There are a total of just over 25K SNPs that are shared across all species.
Figure 3.4. Distribution of SNP sharing across species for SNPs with successful probes. SNP sharing patterns among the species are relatively similar with the exception of *P. maximinoi* which contains a large number of private SNPs when compared to the others.
Subspecies exhibited a relatively similar number of private SNPs (~20K) with the exception of *P. maximinoi* (~120K). All species shared just under 25K SNPs between each other.

**Figure 3.5.** Sharing pattern for SNPs with successful probes by subspecies. Subspecies exhibited a relatively similar number of private SNPs (~20K) with the exception of *P. maximinoi* (~120K). All species shared just under 25K SNPs between each other.
Figure 3.6. Sharing pattern for SNPs with successful probes by provenance. Provenances exhibited a relatively similar number of private within them. *Pinus maximinoi* provenances had a much larger number of SNPs shared between 2 to 14 provenances than other species.
Figure 3.7. Principal component analysis of SNP sharing for species (a) and subspecies (b).

Correlation of SNP sharing at the species level appears to be high as indicated by the narrow
loading angles. *Pinus maximinoi* segregates itself from the other species showing little
correlation. Subspecies follow a pattern as expected from phylogeny with subspecies of the
same species sharing a similar loading angle.
Figure 3.8. Repetitive mapping of left and right flanking regions for SNP probes when mapped to the "tropicalized" v2.01 *P. taeda* genome assembly. Over 75% of flanking regions in all species map to a unique location within the genome reference.
Chapter 4 Conclusions

Between the two datasets utilized in this study a large number of SNPs were discovered and selected for probe design. A total of 5.3 million SNPs were identified between the five pine species assessed in the RNA-Seq study, from which 1.3 million probes were successfully designed for assessment. For targeted capture sequencing, a total of 3.4 million SNPs were identified within six pine species, from which 562K probes were designed. This generated a total of 1.8 million probes for assessment from the two datasets. Mapping flanking sites from the RNA-Seq generated probes to the tropical pine reference indicates a total of 1.6 million unique locations targeted between the two datasets by approximately 1 million probes. This shows some overlap between the two SNP calling sets, which is expected due to the selective targeting of single copy coding or regulatory regions in the targeted capture design and the use of RNA-Seq.

Assessment of SNP sharing between species shows that there are a large proportion of SNPs that are unique to a given species for the RNA-Seq dataset and to a lesser extent in the targeted sequencing. Additionally, correlation of SNP sharing between more closely related species indicates that the SNPs identified and probes designed are behaving as expected from biology. This correlation between species could indicate that the SNPs generated from this study are true and not arbitrary calls due to sequencing or alignment errors. This is further supported by the Ts:Tv ratios that are consistent with previously published studies in tree species. The deviation from the 1:2 ratio expected from random events indicates that the SNPs are more likely to be true SNPs or at least following a mutation pattern influenced by biological mechanisms.
Ultimately, the assessment of the probes generated in this study through scoring and use on a screening array will provide definitive answers to their utility on a commercial genotyping array. However, results from this study will be valuable in helping to decide on SNPs that give the appropriate species sharing distribution as well as identifying species specific SNPs to be included on the final version of the array. These sorts of species distribution considerations will be an important part of the design phase moving forward. Particularly, in the future of this study, there is a need for a set of SNP probes that are shared across species, making the array valuable for populations of each species assessed. However, the need for transferability will ideally be balanced with species specific SNPs that will be valuable in species and hybrid cross identification. Also, as shown in cattle, the need for breed specific markers, or in the case of this study species specific, are important in determining marker trait associations which differentiate populations (Pareek et al. 2017).

Despite the challenges, a number of successful multi-species genotyping arrays have been designed and shown to be effective (Hulse-Kemp et al. 2015; Silva-Junior et al. 2015b).

The completion of the commercial array for these species will provide tree improvement managers, breeders, and scientists a powerful tool to aide in the management and conservation of tropical pine species. The array will provide users with a highly transferrable, reproducible, and computationally friendly technique to assess their populations. This makes the array useful to a wide range of individuals and applications by removing the need for special bioinformatics knowledge or infrastructure, helping usher in the use of molecular breeding technology in commercial pine breeding programs.
4.1 Literature Cited


APPENDIX
Appendix 1. Sharing of left flanking regions allowing for one difference across species. *P. maximinoi* has the highest number of unique flanking regions while *P. greggii* and *P. patula* have a high number of shared flanks between the two as expected from phylogeny.
Appendix 2. Sharing of left flanking regions allowing for two differences across species.

The number of flanking regions only marginally increases when compared with the one difference allowed scenario.
Appendix 3. Sharing of right flanking regions allowing for zero differences across species.

Right flanking regions appear to follow the same distribution as left regions, with many regions unique to each species and few shared across all.
Appendix 4. Sharing of right flanking regions allowing for one difference across species.

There is an increase in the number of flanking regions shared between species in all categories with species sharing regions as you would expect from biology.
Appendix 5. Sharing of right flanking regions allowing for two differences across species.

Flanking regions are more commonly shared between *P. patula* / *P. greggii* and *P.oocarpa* / *P.tecunumanii* pairs than with other species.
Appendix 6. Table of geographic and climatic information for each of the 81 provenances sampled across all species used in targeted capture sequencing.

<table>
<thead>
<tr>
<th>Species</th>
<th>Provenance</th>
<th>Country</th>
<th>Lat (N)</th>
<th>Long (W)</th>
<th>Elevation (m)</th>
<th>Rainfall (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. caribaea</em></td>
<td>Gualjoco</td>
<td>Honduras</td>
<td>14° 55'</td>
<td>88° 14'</td>
<td>240 - 355</td>
<td>1200</td>
</tr>
<tr>
<td></td>
<td>Santa Cruz de Yojoa</td>
<td>Honduras</td>
<td>14° 53'</td>
<td>86° 56'</td>
<td>530 - 720</td>
<td>2758</td>
</tr>
<tr>
<td></td>
<td>Limón</td>
<td>Honduras</td>
<td>15° 51'</td>
<td>85° 23'</td>
<td>20 - 85</td>
<td>2452</td>
</tr>
<tr>
<td></td>
<td>Trincheras</td>
<td>Guatemala</td>
<td>15° 27'</td>
<td>89° 03'</td>
<td>150 - 500</td>
<td>2000</td>
</tr>
<tr>
<td><em>P. greggii N</em></td>
<td>El Madroño</td>
<td>Mexico</td>
<td>21° 16'</td>
<td>99° 10'</td>
<td>1500 - 1660</td>
<td>1100</td>
</tr>
<tr>
<td></td>
<td>Jamé</td>
<td>Mexico</td>
<td>25° 21'</td>
<td>100° 37'</td>
<td>2500 - 2590</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>Laguna Seca</td>
<td>Mexico</td>
<td>21° 02'</td>
<td>99° 10'</td>
<td>1750 - 1900</td>
<td>820</td>
</tr>
<tr>
<td></td>
<td>Las Placetas</td>
<td>Mexico</td>
<td>24° 55'</td>
<td>100° 11'</td>
<td>2370 - 2520</td>
<td>750</td>
</tr>
<tr>
<td></td>
<td>Los Lirios</td>
<td>Mexico</td>
<td>25° 22'</td>
<td>100° 29'</td>
<td>2300 - 2400</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>Jamé</td>
<td>Mexico</td>
<td>25° 21'</td>
<td>100° 37'</td>
<td>2500 - 2590</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>Cerro El Potosí</td>
<td>Mexico</td>
<td>24° 54'</td>
<td>100° 12'</td>
<td>2430 - 2500</td>
<td>750</td>
</tr>
<tr>
<td></td>
<td>Ojo de Agua</td>
<td>Mexico</td>
<td>24° 53'</td>
<td>100° 13'</td>
<td>2115 - 2400</td>
<td>750</td>
</tr>
<tr>
<td></td>
<td>La Tapona</td>
<td>Mexico</td>
<td>24° 37'</td>
<td>100° 10'</td>
<td>2090 - 2350</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>Country</td>
<td>Coordinates</td>
<td>Elevation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P. greggii S</strong></td>
<td>Magueyes</td>
<td>Mexico</td>
<td>--</td>
<td>--</td>
<td>2250 - 2350</td>
<td>1200</td>
</tr>
<tr>
<td>El Madroño</td>
<td>Mexico</td>
<td>21° 16'</td>
<td>99° 10'</td>
<td>1500 - 1660</td>
<td>1100</td>
<td></td>
</tr>
<tr>
<td>Laguna Atezca</td>
<td>Mexico</td>
<td>20° 49'</td>
<td>98° 46'</td>
<td>1250 - 1420</td>
<td>1642</td>
<td></td>
</tr>
<tr>
<td>Laguna Seca</td>
<td>Mexico</td>
<td>21° 02'</td>
<td>99° 10'</td>
<td>1750 - 1900</td>
<td>820</td>
<td></td>
</tr>
<tr>
<td>Valle Verde</td>
<td>Mexico</td>
<td>21° 29'</td>
<td>99° 10'</td>
<td>1150 - 1250</td>
<td>1400</td>
<td></td>
</tr>
<tr>
<td>San Joaquín</td>
<td>Mexico</td>
<td>20° 56'</td>
<td>99° 34'</td>
<td>2130 - 2350</td>
<td>1109</td>
<td></td>
</tr>
<tr>
<td>Carrizal Chico</td>
<td>Mexico</td>
<td>20° 26'</td>
<td>98° 20'</td>
<td>1360 - 1770</td>
<td>1855</td>
<td></td>
</tr>
<tr>
<td><strong>P. maximinoi</strong></td>
<td>Cobán</td>
<td>Guatemala</td>
<td>15° 28'</td>
<td>90° 24'</td>
<td>1300 - 1440</td>
<td>2075</td>
</tr>
<tr>
<td>San Jerónimo</td>
<td>Guatemala</td>
<td>15° 03'</td>
<td>90° 15'</td>
<td>1280 - 1860</td>
<td>970</td>
<td></td>
</tr>
<tr>
<td>San Juan Sacatepéquez</td>
<td>Guatemala</td>
<td>14° 41'</td>
<td>90° 38'</td>
<td>1580 - 2000</td>
<td>1138</td>
<td></td>
</tr>
<tr>
<td>Dulce Nombre de Copán</td>
<td>Honduras</td>
<td>14° 50'</td>
<td>88° 51'</td>
<td>1100 - 1300</td>
<td>1386</td>
<td></td>
</tr>
<tr>
<td>Marcala</td>
<td>Honduras</td>
<td>14° 10'</td>
<td>88° 01'</td>
<td>1600 - 1800</td>
<td>1670</td>
<td></td>
</tr>
<tr>
<td>Tatumbla</td>
<td>Honduras</td>
<td>14° 02'</td>
<td>87° 07'</td>
<td>1400 - 1600</td>
<td>1153</td>
<td></td>
</tr>
<tr>
<td>San Jerónimo CH</td>
<td>Mexico</td>
<td>17° 03'</td>
<td>92° 08'</td>
<td>940 - 1020</td>
<td>1417</td>
<td></td>
</tr>
<tr>
<td>San Jerónimo OA</td>
<td>Mexico</td>
<td>16° 10'</td>
<td>97° 00'</td>
<td>1220 - 1480</td>
<td>1350</td>
<td></td>
</tr>
<tr>
<td>Candelaria</td>
<td>Mexico</td>
<td>16° 00'</td>
<td>96° 31'</td>
<td>1370 - 1480</td>
<td>1117</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>Country</td>
<td>Latitude</td>
<td>Longitude</td>
<td>Altitude Range</td>
<td>Elevation</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------</td>
<td>------------</td>
<td>------------</td>
<td>----------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Las Compuertas</td>
<td>Mexico</td>
<td>17° 10'</td>
<td>99° 59'</td>
<td>1050 - 1200</td>
<td>1400</td>
<td></td>
</tr>
<tr>
<td>El Portillo</td>
<td>Honduras</td>
<td>14° 28'</td>
<td>89° 01'</td>
<td>1400 - 1600</td>
<td>1325</td>
<td></td>
</tr>
<tr>
<td>Yuscarán</td>
<td>Honduras</td>
<td>13° 50'</td>
<td>86° 55'</td>
<td>1500 - 1700</td>
<td>1300</td>
<td></td>
</tr>
<tr>
<td>La Lagunilla</td>
<td>Guatemala</td>
<td>14° 42'</td>
<td>89° 57'</td>
<td>1540 - 1860</td>
<td>1017</td>
<td></td>
</tr>
<tr>
<td>El Portillo</td>
<td>Mexico</td>
<td>20° 24'</td>
<td>98° 25'</td>
<td>2320 - 2480</td>
<td>1350</td>
<td></td>
</tr>
<tr>
<td>Corralitla</td>
<td>Mexico</td>
<td>18° 38'</td>
<td>97° 06'</td>
<td>2000 - 2230</td>
<td>2500</td>
<td></td>
</tr>
<tr>
<td>Conrado Castillo</td>
<td>Mexico</td>
<td>23° 56'</td>
<td>99° 28'</td>
<td>1500 - 2060</td>
<td>1012</td>
<td></td>
</tr>
<tr>
<td>Tlacotla</td>
<td>Mexico</td>
<td>19° 40'</td>
<td>98° 05'</td>
<td>2750 - 2915</td>
<td>1097</td>
<td></td>
</tr>
<tr>
<td>Pinal de Amoles</td>
<td>Mexico</td>
<td>21° 07'</td>
<td>99° 41'</td>
<td>2380 - 2550</td>
<td>1350</td>
<td></td>
</tr>
<tr>
<td>Zacualtipán</td>
<td>Mexico</td>
<td>20° 39'</td>
<td>98° 40'</td>
<td>1980 - 2200</td>
<td>2047</td>
<td></td>
</tr>
<tr>
<td>Llano de las Carmonas</td>
<td>Mexico</td>
<td>19° 48'</td>
<td>97° 54'</td>
<td>2530 - 2880</td>
<td>1097</td>
<td></td>
</tr>
<tr>
<td>El Manzanal</td>
<td>Mexico</td>
<td>16° 06'</td>
<td>96° 33'</td>
<td>2350 - 2660</td>
<td>1348</td>
<td></td>
</tr>
<tr>
<td>El Tlacuache</td>
<td>Mexico</td>
<td>16° 44'</td>
<td>97° 09'</td>
<td>2300 - 2620</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Ixtlán</td>
<td>Mexico</td>
<td>17° 24'</td>
<td>96° 27'</td>
<td>2600 - 2870</td>
<td>1750</td>
<td></td>
</tr>
<tr>
<td>Santa María Papalo</td>
<td>Mexico</td>
<td>17° 49'</td>
<td>96° 48'</td>
<td>2270 - 2720</td>
<td>1100</td>
<td></td>
</tr>
<tr>
<td>Camotán</td>
<td>Guatemala</td>
<td>14° 49'</td>
<td>89° 22'</td>
<td>740 - 960</td>
<td>926</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>Country</td>
<td>Latitude</td>
<td>Longitude</td>
<td>Minimum Elevation</td>
<td>Maximum Elevation</td>
<td>Altitude</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>-----------</td>
<td>------------------</td>
<td>-------------------</td>
<td>---------</td>
</tr>
<tr>
<td>El Castaño (Bucaral)</td>
<td>Guatemala</td>
<td>15° 01'</td>
<td>90° 09'</td>
<td>930</td>
<td>1330</td>
<td>900</td>
</tr>
<tr>
<td>San Luis Jilotepeque</td>
<td>Guatemala</td>
<td>14° 37'</td>
<td>89° 46'</td>
<td>950</td>
<td>1010</td>
<td>895</td>
</tr>
<tr>
<td>San Sebastian Coatlán</td>
<td>Mexico</td>
<td>16° 11'</td>
<td>96° 50'</td>
<td>1750</td>
<td>1750</td>
<td>598</td>
</tr>
<tr>
<td>El Jicaro</td>
<td>Mexico</td>
<td>16° 32'</td>
<td>94° 12'</td>
<td>1000</td>
<td>1000</td>
<td>1684</td>
</tr>
<tr>
<td>Taretan</td>
<td>Mexico</td>
<td>19° 25'</td>
<td>102° 04'</td>
<td>1610</td>
<td>1610</td>
<td>1622</td>
</tr>
<tr>
<td>Huayacocotla</td>
<td>Mexico</td>
<td>20° 30'</td>
<td>98° 25'</td>
<td>1190</td>
<td>1410</td>
<td>1711</td>
</tr>
<tr>
<td>San Pedro Solteapán</td>
<td>Mexico</td>
<td>18° 15'</td>
<td>94° 51'</td>
<td>602</td>
<td>602</td>
<td>1812</td>
</tr>
<tr>
<td>La Petaca</td>
<td>Mexico</td>
<td>23° 25'</td>
<td>105° 48'</td>
<td>1560</td>
<td>1710</td>
<td>1155</td>
</tr>
<tr>
<td>Chinipas</td>
<td>Mexico</td>
<td>27° 18'</td>
<td>108° 35'</td>
<td>1140</td>
<td>1780</td>
<td>822</td>
</tr>
<tr>
<td>Dipilto</td>
<td>Nicaragua</td>
<td>13° 43'</td>
<td>86° 32'</td>
<td>1075</td>
<td>1320</td>
<td>--</td>
</tr>
<tr>
<td>Pachoc</td>
<td>Guatemala</td>
<td>14° 51'</td>
<td>91° 16'</td>
<td>2000</td>
<td>2500</td>
<td>1350</td>
</tr>
<tr>
<td>San Jerónimo</td>
<td>Guatemala</td>
<td>15° 03'</td>
<td>90° 18'</td>
<td>1620</td>
<td>1850</td>
<td>1200</td>
</tr>
<tr>
<td>San Lorenzo</td>
<td>Guatemala</td>
<td>15° 05'</td>
<td>89° 40'</td>
<td>1900</td>
<td>2100</td>
<td>1700</td>
</tr>
<tr>
<td>San Vicente</td>
<td>Guatemala</td>
<td>15° 05'</td>
<td>90° 07'</td>
<td>1690</td>
<td>2200</td>
<td>1700</td>
</tr>
<tr>
<td>Celaque</td>
<td>Honduras</td>
<td>14° 33'</td>
<td>88° 40'</td>
<td>1540</td>
<td>2030</td>
<td>1273</td>
</tr>
<tr>
<td>Las Trancas</td>
<td>Honduras</td>
<td>14° 07'</td>
<td>87° 49'</td>
<td>2075</td>
<td>2185</td>
<td>1579</td>
</tr>
<tr>
<td>Location</td>
<td>Country</td>
<td>Latitude</td>
<td>Longitude</td>
<td>Elevation</td>
<td>Year</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Chanal</td>
<td>Mexico</td>
<td>16° 42'</td>
<td>92° 23'</td>
<td>2010 - 2350</td>
<td>1238</td>
<td></td>
</tr>
<tr>
<td>Chempil</td>
<td>Mexico</td>
<td>16° 45'</td>
<td>92° 25'</td>
<td>2020 - 2220</td>
<td>1146</td>
<td></td>
</tr>
<tr>
<td>El Carrizal</td>
<td>Mexico</td>
<td>15° 24'</td>
<td>92° 18'</td>
<td>2130 - 2280</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Jitotol</td>
<td>Mexico</td>
<td>17° 02'</td>
<td>92° 51'</td>
<td>1660 - 1750</td>
<td>1701</td>
<td></td>
</tr>
<tr>
<td>Montebello</td>
<td>Mexico</td>
<td>16° 06'</td>
<td>91° 45'</td>
<td>1660 - 1750</td>
<td>1909</td>
<td></td>
</tr>
<tr>
<td>Napite</td>
<td>Mexico</td>
<td>16° 34'</td>
<td>92° 19'</td>
<td>2070 - 2350</td>
<td>1350</td>
<td></td>
</tr>
<tr>
<td>Rancho Nuevo</td>
<td>Mexico</td>
<td>16° 41'</td>
<td>92° 35'</td>
<td>2280 - 2340</td>
<td>1238</td>
<td></td>
</tr>
<tr>
<td>Chiul</td>
<td>Guatemala</td>
<td>15° 20'</td>
<td>91° 04'</td>
<td>2440 - 2680</td>
<td>1996</td>
<td></td>
</tr>
<tr>
<td>San José</td>
<td>Mexico</td>
<td>16° 42'</td>
<td>92° 41'</td>
<td>2245 - 2400</td>
<td>1252</td>
<td></td>
</tr>
<tr>
<td>Juquila</td>
<td>Mexico</td>
<td>16° 15'</td>
<td>97° 13'</td>
<td>2090-2260</td>
<td>1325</td>
<td></td>
</tr>
<tr>
<td>San Rafael del Norte</td>
<td>Nicaragua</td>
<td>13° 14'</td>
<td>86° 08'</td>
<td>910 - 1170</td>
<td>1366</td>
<td></td>
</tr>
<tr>
<td>Yucul</td>
<td>Nicaragua</td>
<td>12° 55'</td>
<td>85° 47'</td>
<td>1080 - 1330</td>
<td>1394</td>
<td></td>
</tr>
<tr>
<td>San Francisco</td>
<td>Honduras</td>
<td>14° 57'</td>
<td>86° 07'</td>
<td>900 - 1590</td>
<td>1491</td>
<td></td>
</tr>
<tr>
<td>Villa Santa</td>
<td>Honduras</td>
<td>14° 12'</td>
<td>86° 17'</td>
<td>800 - 1000</td>
<td>1302</td>
<td></td>
</tr>
<tr>
<td>Culmí</td>
<td>Honduras</td>
<td>15° 08'</td>
<td>85° 36'</td>
<td>400 - 950</td>
<td>1491</td>
<td></td>
</tr>
<tr>
<td>Los Planes</td>
<td>Honduras</td>
<td>14° 48'</td>
<td>87° 53'</td>
<td>1100 - 1650</td>
<td>2287</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>Country</td>
<td>Latitude</td>
<td>Longitude</td>
<td>Elevation</td>
<td>Temperature</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Cerro Cusuco</td>
<td>Honduras</td>
<td>15° 29'</td>
<td>88° 12'</td>
<td>1350 - 1630</td>
<td>2287</td>
<td></td>
</tr>
<tr>
<td>Gualaco</td>
<td>Honduras</td>
<td>15° 03'</td>
<td>86° 08'</td>
<td>600 - 800</td>
<td>1491</td>
<td></td>
</tr>
<tr>
<td>Cerro la Joya</td>
<td>Nicaragua</td>
<td>12° 25'</td>
<td>85° 59'</td>
<td>940 - 1160</td>
<td>1394</td>
<td></td>
</tr>
</tbody>
</table>