ABSTRACT

TOWNSEND, LAURA ASHLEY. Identifying Genetic Variability in Loblolly Pine (Under the direction of Ross Whetten).

Loblolly pine is the leading southern pine species in the U.S., accounting for approximately 80 percent of US commercial timber production. The NC State University Cooperative Tree Improvement Program has been breeding loblolly pine for almost 60 years, using its diverse gene pool and its ability to adapt to various environments to select for increased productivity, disease resistance, and wood quality. One way to employ genomic methods in breeding is to identify associations of individual genetic marker loci with growth/quality characteristics. To enrich for functional genetic variation in the 22-gigabase pine genome, two approaches to genotyping-by-sequencing (GBS) were attempted. The first employed methylation-sensitive restriction enzymes to enrich gene regions, and the second used isolated nuclei to identify nuclease hypersensitive sites. The methylation-sensitive enzyme approach was used to genotype a set of progeny from a study consisting of 140 pollen mix families representing the entire loblolly pine range. DNA was extracted from 1471 individual tree samples using an optimized high-throughput and cost effective protocol adapted from the Canadian Center for DNA Barcoding. All samples were sequenced using Illumina-HiSeq and aligned to the newly published reference genome. Analysis of the sequence alignments shows 1887 of the 50172 annotated genes in the pine genome contain SNP variants detected by methylation-sensitive GBS. When tested with height data at age eight years, a key predictor of pine productivity, 180 SNPs were shown to be significantly associated (after Bonferroni correction). Due to high amounts of missing data across multiple samples, it was not possible to generate haplotype based kinship matrices to determine

unknown pollen parents. For the isolated nuclei, a test was performed to determine concordance of nuclease-sensitive sites in four distinct tissues and organs across three different trees. All four tissue types together mapped to approximately 1/3 of the reference genome and each tissue type generated a large number of uniquely mapped sites. However, the data generated are not currently informative enough to be used for future designs of SNP chip arrays or capture baits.

Discovering Genetic Variability in Loblolly Pine

by
Laura Ashley Townsend

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

2015

APPROVED BY:

_____	_____
Ross Whetten Committee Chair	Kevin Potter
_____	
Brian Reich	

DEDICATION

For my parents who always believed in and supported my dreams.

BIOGRAPHY

Laura Townsend was born in Wyandotte, Michigan on November 20, 1989 to John and Margaret Townsend. Laura received a B.S in Environmental Studies from Florida International University in 2010 with a concentration in plant conservation. She moved to Texas to intern with U.S Fish and Wildlife doing reforestation surveys and realized that she wanted to pursue a career in Forestry and Genomics. She began graduate school at North Carolina State University in 2012 to pursue a degree in Forestry focusing on forest genetics under the direction of Dr. Ross Whetten. Laura hopes to travel some more in the world and continue discovering the vast mysteries of the universe.

ACKNOWLEDGMENTS

I would like to thank everyone at the Tree Improvement Program with a special thanks to Will Kohlway, who taught me all the lab skills I know, Lillian, and a special thanks to my head advisor Ross Whetten for his patience and guidance.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1 1

1.1 Importance of Loblolly Pine 1

1.2 Climate Change and Forestry 2

1.3 Current Breeding Strategies 2

1.4 Genetic Markers in Pine Breeding 4

1.4.1 QTL and MAS 4

1.4.2 Association Genetics 4

1.4.3 Genomic Selection 5

1.4.4 Next Generation Sequencing 6

1.5 Literature Cited 8

CHAPTER 2 13

2.1 Introduction 13

2.2 Materials and Methods 16

2.2.1 Sample Material 16

2.2.2 Tissue Sampling and Storage 17

2.2.3 DNA Extraction 17

2.2.4 GBS Library Preparations 18

2.2.5 Sequencing 19

2.2.5 Processing of Illumina data 19

2.2.6 SNP discovery 19

2.2.7 Populations-F_{ST} Values 19

2.2.8 Alignment to reference 20

2.2.9 Association Testing 20

2.3 Results 20

2.3.1 Quality Time and Cost 20

2.3.2 Sequencing Reads 21

2.3.3 Populations 22

2.3.4 Alignment and Association Testing 23

2.4 Discussion/Conclusions 24
2.4.1 DNA Quality and Cost Optimization 24
2.4.2 Sequencing Reads and Genetic Variability 26
2.4.3 Usage of the Reference Genome 27
2.4.4 Further Steps 28
2.5 Tables 30
2.6 Figures 33
2.7 Literature Cited 41
CHAPTER 3 45
3.1 Introduction 45
3.2 Materials and Methods 47
3.2.1 Plant Material 47
3.2.2 Nuclei Extraction 47
3.2.3 DNA Extraction 48
3.2.4 GBS Library Preparation and Sequencing 48
3.2.5 Processing of Illumina Data 49
3.3 Results and Discussion 50
3.3.1 Nuclei Extraction and Library Preparation 50
3.3.2 Mapping Reads and Comparisons 50
3.3.3 Steps for the Future 51
3.4 Tables 53
3.5 Figures 56
3.6 Literature Cited 62
APPENDIX 65
4.1 Chapter 2 Detailed Sequencing Analysis Code 66
4.1.1-Filtering and Trimming Reads for Quality 66
4.1.2- Split Data from Four Lanes of Pooled Plates into Separate Directories and Files 66
4.1.3-Create pool.key file in GBS_barcode.pl Format 67
4.1.4-A script to run GBS_barcode.pl 68
4.1.5 Command line for read count information 69

4.1.6-Example of UStacks for One Plate 70
4.1.7-cstacks Command 70
4.1.8-sstacks Command 70
4.1.9-cstacks Done for Population Calculations 70
4.1.10-Populations Command 71
4.1.11-Finding SNP Counts for Each match.tsv File *71*
4.1.13-Alignment to Reference Using BWA and SAMtools *71*
4.1.14-Filtering Half-Present Samples Using Freebayes in .VCF Format 71
4.1.15-Association Testing Using RVTest 73
4.2 Chapter 3 Detailed Sequencing Analysis Code 73
4.2.1-Trim Adapters, Filter on Quality, and Split Reads into Barcode Sets Using Flexbar 73
4.2.2-Example of BWA Alignment for One Tissue Type 74
4.2.3- Use BEDtools to Convert BAM Files to BED Format for Multi-Intersect Function 74
4.2.4-Making Multiple Comparisons 74

LIST OF TABLES

Table 2.1: Comparison of four concentration methods. 30
Table 2.2: Pairwise FST values of the seven PSSSS seed source regions. 30
Table 2.3: List of P-value associations. 30
Table 3.1: List of volumes/weights used per tissue type during nuclei isolations 53
Table 3.2: List of volumes used per tissue type during digestion of the isolated nuclei . 53
Table 3.3: DNA ligation parameters 53
Table 3.4: Ligation master-mix parameters 53
Table 3.5: Concentration of ligated DNA before and after size selection 54
Table 3.6: PCR volume parameter 54
Table 3.7: PCR master-mix volume parameters 54
Table 3.8: Comparison of unique mapped regions per tissue type. 54
Table 3.9: Comparison of number of mapped sites in each tissue type. 55
Table 3.10: Table comparison of mapped sites sums. 55

LIST OF FIGURES

Figure 2.1: GBS Library Preparation. 33
Figure 2.2: GBS Library Pooling. 34
Figure 2.3: Storage Solution Degradation Over Time. 35
Figure 2.4: Read Counts by Concentration Strategies. 36
Figure 2.5: Read Count by Barcodes. 36
Figure 2.6: SNPs versus Reads. 37
Figure 2.7: Dendrogram /Heatmap. 37
Figure 2.8: Seed Sources. 38
Figure 2.9: F_{ST} Distribution. 38
Figure 2.10: Distribution of Size Classes per Plate. 39
Figure 2.11: Distribution of Reads of All Samples in Plate 4. 40
Figure 2.12: T7 Vs. T4 Ligase. 40
Figure 3.1: 6-Order Folding Structure of DNA. 56
Figure 3.2: Digestion of all Tissue Types. 57
Figure 3.3: PCR Amplification of Four Tissue Types. 58
Figure 3.4: Bioanalyzer Results. 59
Figure 3.5: Mapped Sites Tissue Comparison. 60
Figure 3.6: Comparison of Tree Replicates. 61

CHAPTER 1

1.1 Importance of Loblolly Pine

Pine trees are large, long lived, and genetically variable. Their lifespans can range from decades to centuries, making adaption and breeding programs a slow process. Pine tree plantations are an important economic commodity and thus need to be protected. Planted pine forests in the southeast U.S alone account for 16% of the supplied global industrial wood and 7.5% of the industrial economic activity of the region ("PINEMAP Year 2 Annual Report," 2013). Of the over 30 million acres of pines planted, loblolly pine (*Pinus taeda* L.) in particular accounts for approximately 80% of produced wood in the U.S (Dorman, 1976; Wear, Carter, & Prestemon, 2007). The pine has a very diverse gene pool which is applicable for adaptation to various environments (Shiver, Rheney, & Hitch, 2000). Starting in the mid-1950s, breeding programs were initiated to improve overall gain of selected desired phenotypic characteristics of loblolly pine (Easley, 1963). Gains in the past have generally been achieved through recurrent selection and breeding of superior trees (Zobel & Talbert, 1984). As of 2003, it has been reported that approximately 95% of all planting stock of southern U.S pine industrial forest is the product of genetically improved seed (McKeand, Mullin, Byram, & White, 2003). These breeding stocks have been improved for economically important traits such as straightness, wood density, and rust resistance. Even with these achievements, there are problems to overcome in breeding programs. A large majority of the desired breeding traits are complex and have low heritability, which makes selection difficult, and with the rising threat of climate change more emphasis should be put on breeding for traits that will increase the adaptability of trees to varying climates (Zobel & Talbert, 1984).

1.2 Climate Change and Forestry

The United States is experiencing climate change believed to be the result of increased greenhouse gases in the atmosphere (Glick, Stein, Edelson, & National Wildlife Federation, 2011). Since the 1960s, temperatures have been increasing, with pronounced significance occurring in the summer season along the Gulf and Atlantic coasts (Kunkel et al., 2013). Increases in wildfires, insects, disease, drought, and extreme weather events as a result of these changes are expected to occur. The combination of drought and higher temperatures can cause high mortality rates in trees (Joyce et al., 2013). A study conducted on eastern temperate forests of the US (including the planted range of loblolly pine), found rates of loblolly pine mortality were most sensitive to summer maximum temperature (Dietze & Moorcroft, 2011). Given these developing conditions, trees breeders have even more reason to improve their breeding strategies in terms of both adaptive trait selection and length of breeding cycles. Many of the desired adaptive traits for the timber industry for resistance to climate change are the same. These traits include frost hardiness, growth, water usage, and disease/pest resistance. Unfortunately, most of these adaptive traits are under polygenetic control and have low heritability, which means that making genetic gains using traditional breeding techniques is slow and difficult (Gailing, Vornam, Leinemann, & Finkeldey, 2009; Namkoong, 1979).

1.3 Current Breeding Strategies

Traditional pine recurrent selection programs initiated by tree improvement programs are costly and very time consuming (Namkoong, 1988). The three main components of the pine improvement cycle are breeding, testing, and propagation. Breeding is where multiple

selected crosses are made among selected parents based on estimates of their breeding values from phenotypic measurements or from previous test results. Testing is where the general and specific combining abilities of those parents are quantified more accurately. Propagation is where the actual seeds and/or seedlings used in deployment are produced. Methods such as top-grafting, controlled pollination, and flowering enhancing techniques have been shown to reduce the length of the breeding phase down to 4 years or less (Resende et al., 2012; Zobel & Talbert, 1984). Techniques such as somatic embryogenesis and rooted cuttings can also help to decrease propagation times. However, it is important to mention that these propagation techniques have not yet been perfected in pines, in terms of time and expense (Grattapaglia & Resende, 2010). This leaves the testing phase as the most time consuming portion of the pine improvement cycle. The testing phase can range from 6-10 years in loblolly pine (Resende et al., 2012). One reason testing is so time consuming is that pine trees need to mature to some degree before final measurements can be made. Mature trees and juvenile trees differ in traits such as wood properties, needle growth, and reproductive abilities (Zobel & Van Buijtenen, 1989). Pines do not typically reach maturity until age 10, and optimal selection ages for economically important traits have been shown to range from between 5-10 years depending on the species and the characteristics being measured (Gwaze, Harding, Purnell, & Bridgwater, 2002; Gwaze, 2008). With the aforementioned breeding advancements, the breeding cycle of Loblolly Pine can be reduced from an average of 26 years or more to approximately 13 years. The further acceleration of this process could help in the fight against climate change by improving adaption to changing or unpredictable environments, and increasing resistance to pests and disease (Harfouche et al., 2012).

1.4 Genetic Markers in Pine Breeding

1.4.1 QTL and MAS

Further acceleration of the tree improvement cycle has become possible through genomic based techniques. Experiments in these techniques, such as genetic mapping of Quantitative Trait Loci (QTL) and marker assisted selection (MAS), were first described for pine tree breeding by Neale & Williams (1991). MAS is when a molecular marker, known to be genetically linked to the QTL in question, is used indirectly for selection of a desired phenotypic trait. Markers used are generally DNA based and include restriction fragment length polymorphisms (RFLPs), randomly amplified polymorphic DNA (RAPD), microsatellites and more (O'Malley, 1996). One marker type that has gained popularity and widespread use is the single nucleotide polymorphism (SNP). SNPs are the smallest unit of polymorphism, and are available in great abundance, making them an ideal candidate to use in MAS (Ganal & Roder, 2007). Regrettably, there are some inherent problems involved with using MAS with QTL mapping of pine trees. A common genomic attribute of pine species is their rapid decay of linkage disequilibrium (LD) due to their large genomes and highly outcrossing nature (Eckert et al., 2009). The use of MAS with QTL requires the marker in use to remain linked to the QTL. This rapid decay of LD limits the use of MAS to only within family selections, which results in only being able to use small population sizes (Grattapaglia & Resende, 2010).

1.4.2 Association Genetics

Association genetics has been suggested as an alternative to QTL-based MAS in forest trees because it can be applied at the population level. This is because there is no breakdown of LD on the large scale; it has already occurred throughout the centuries of random mating in the loblolly pine populations. This results in LD only being found in

markers that are less than a few Kb away from each other (Lepoittevin, Harvengt, Plomion, & Garnier-Géré, 2011; Neale, 2007). Association genetics uses candidate genes in the identification of QTLs. Pines are ideal for the strategy, given their large and unstructured populations (Neale & Savolainen, 2004). Association genetics has been used in pine trees to successfully map SNPs in candidate genes that are associated with climate adaption traits. Some studies include the association of cold hardiness traits in Douglas-fir, and the association of wood and drought related genes in loblolly pine. However the small effect of each individual association is a problem in association genetic studies of pine. Each association rarely exceeds 5% of the phenotypic variance (Eckert et al., 2009; Gonzalez-Martinez, Wheeler, Ersoz, Nelson, & Neale, 2006; Gonzalez-Martinez, Krutovsky, & Neale, 2006). The quantitative traits being looked at in trees are generally affected by many loci, requiring dense marker maps that cannot be feasibly made with association genetics (Meuwissen, Hayes, & Goddard, 2001).

1.4.3 Genomic Selection

Genomic selection (GS) using genome wide association studies (GWAS) to combat problems with association genetics in trees was first proposed by Meuwissen et al. (2001). This method involves first phenotyping for desired traits, and then genotyping to create dense molecular marker maps from a large number of individuals, which are known as the "training" population. The combined effect of all the molecular markers on phenotypes is estimated simultaneously, thus allowing for the disregard of QTL locations (Meuwissen et al., 2001). A study using genomic selection and the random regression best linear unbiased prediction (BLUP) model found the selection efficiency for loblolly pine increased by more than half when compared to using phenotypic selection alone. The model's levels of

accuracy to correctly predict phenotypic traits remained high even across different environments (Harfouche et al., 2012).

Current studies suggest that GS can be used to predict optimal mating pairs of a breeding population while bypassing traditional testing techniques (Grattapaglia & Resende, 2010; Resende et al., 2012). The newly developed crosses can either be re-used in a new breeding cycle to create better genetically improved seeds, or put out into deployment populations. Either way, these techniques show great potential to create well-adapted pine trees rapidly while reducing the cost and complexity of lengthy field trials (Harfouche et al., 2012). Efforts using GS are currently being performed in species such as *Populus*, but whole genome wide selections have not previously been considered possible due to the large sequencing efforts required as conifers have an incredibly large genome (Neale & Savolainen, 2004; Tuskan et al., 2006). Therefore, the feasibility of using GS or molecular markers in general relies on the technological advances and declining cost of genotyping large, complex genomes.

1.4.4 Next Generation Sequencing

Previously, high-throughput genotyping methods for pine were too expensive to allow routine application in breeding populations. However, a new form of sequencing technology termed next-generation sequencing (NGS) has been developed. NGS encompasses a large number of sequencing platforms from Illumina MiSeqs to Life Technology's Ion Torrents. A review of NGS can be found in the published paper by Grada & Weinbrecht (2013). Essentially, it has been stated to "produce an enormous volume of data cheaply" and has been used in large volume SNP discoveries and GWAS (Metzker, 2009). DNA sequencing costs are dropping so rapidly that it has been predicted that the cost to sequence a base of DNA will be less than the cost to store the resulting data (Stein, 2010). A

method known as genotyping-by-sequencing (GBS) can be used with NGS technology to produce a cheaper way to sequence high density, large genome species such as pines. GBS works by generating large quantities of SNP data to be used in GWAS. When sequencing large, complex genomes such as pine, one of two methods needs to be used to reduce the sections of the genome sequenced to only genetically important areas. One method is through target enrichment such as hybridization based methods using microarrays or soluble oligonucleotides. The other method is through the use of restriction enzymes (REs). REs fragment long strands of DNA at specified sites by digestion. Some REs use common sites and thus cut frequently, while others cut infrequently. The REs leave what are termed "sticky ends" where there are small overhangs of two to four bases of single-stranded DNA, making ligation of barcode adapters possible. Barcode adapters are known, short DNA sequences that are ligated onto the ends of digested DNA to tag each individual sample, allowing samples to be pooled together and sequenced as one sample. The REs, along with barcode adaptors, increase the sample numbers capable of being sequenced at one time, while reducing genome complexity, thereby making GBS a high throughput method (Elshire et al., 2011). Application of GBS to create high density maps has already been proven successful in the large and complex genomes of barley and wheat, thus its application to large and complex pine genomes looks promising (Poland, Brown, Sorrells, & Jannink, 2012).

1.5 Literature Cited

Dietze, M. C., & Moorcroft, P. R. (2011). Tree mortality in the eastern and central United States: patterns and drivers. *Global Change Biology*, *17*(11), 3312–3326. http://doi.org/10.1111/j.1365-2486.2011.02477.x

Dorman, K. W. (1976). *The Genetics and breeding of southern pines*. Washington: U.S Dept. of Agriculture, Forest Service: for sale by Supt. of Docs., U.S. Govt.

Easley, L. . (1963). Growth of loblolly pine from seed production in a seed production area vs. nursery run stock. *Journal of Forestry*, *61*(5), 338–389.

Eckert, A. J., Wegrzyn, J. L., Pande, B., Jermstad, K. D., Lee, J. M., Liechty, J. D., … Neale, D. B. (2009). Multilocus Patterns of Nucleotide Diversity and Divergence Reveal Positive Selection at Candidate Genes Related to Cold Hardiness in Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii). *Genetics*, *183*(1), 289–298. http://doi.org/10.1534/genetics.109.103895

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, *6*(5), e19379. http://doi.org/10.1371/journal.pone.0019379

Gailing, O., Vornam, B., Leinemann, L., & Finkeldey, R. (2009). Genetic and genomic approaches to assess adaptive genetic variation in plants: forest trees as a model. *Physiologia Plantarum*, *137*(4), 509–519. http://doi.org/10.1111/j.1399-3054.2009.01263.x

Ganal, M. W., & Roder. (2007). Genomics-assisted crop improvement. Volume 2, Genomics applications in crops- Microsatellite and SNP Markers in wheat Breeding. Retrieved April 20, 2013, from http://site.ebrary.com/id/10217506

Glick, P., Stein, B. A., Edelson, N. A., & National Wildlife Federation. (2011). *Scanning the conservation horizon : a guide to climate change vulnerability assessment*. Washington, D.C.: National Wildlife Federation.

Gonzalez-Martinez, S. C., Wheeler, N. C., Ersoz, E., Nelson, C. D., & Neale, D. B. (2006). Association Genetics in Pinus taeda L. I. Wood Property Traits. *Genetics*, *175*(1), 399–409. http://doi.org/10.1534/genetics.106.061127

Gonzalez-Martinez, Santiago C., Krutovsky, K. V., & Neale, D. B. (2006). Forest-tree population genomics and adaptive evolution. *New Phytologist*, *170*(2), 227–238. http://doi.org/10.1111/j.1469-8137.2006.01686.x

Grada, A., & Weinbrecht, K. (2013). Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology*, *133*(8), e11. http://doi.org/10.1038/jid.2013.248

Grattapaglia, D., & Resende, M. D. V. (2010). Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, *7*(2), 241–255. http://doi.org/10.1007/s11295-010-0328-4

Gwaze, D. (2008). Optimum selection age for height in shortleaf pine. *New Forests*, *37*(1), 9–16. http://doi.org/10.1007/s11056-008-9104-9

Gwaze, D. P., Harding, K. J., Purnell, R. C., & Bridgwater, F. E. (2002). Optimum selection age for wood density in loblolly pine. *Canadian Journal of Forest Research*, *32*(8), 1393–1399. http://doi.org/10.1139/x02-064

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

Joyce, L. A., Running, S. W., Breshears, D. D., Dale, V. H., Malmsheimer, R. W., Sampson, R. N., … Woodall, C. W. (2013, January). Federal Advisory Committee Draft Climate Assessment Reprot Released for Public Review. National CLimate Assessment and Development Advisory Committee. Retrieved from http://www.ncadac.globalchange.gov/

Kunkel, K. E., Stevens, L. E., Stevens, S. E., Sun, L., Janssen, E., Fuhrman, C. M., … Dobson, J. G. (2013, January). NOAA Technical Report NESDIS 142-2 Regional Climate Trends and Scenarios for the U.S. National Climate Assessment Part 2. Climate of teh Southeast U.S. U.S Department of Commerce National Oceanic and Atmospheric Administration.

Lepoittevin, C., Harvengt, L., Plomion, C., & Garnier-Géré, P. (2011). Association mapping for growth, straightness and wood chemistry traits in the Pinus pinaster Aquitaine breeding population. *Tree Genetics & Genomes*, *8*(1), 113–126. http://doi.org/10.1007/s11295-011-0426-y

McKeand, S., Mullin, T., Byram, T., & White, T. (2003). Deployment of Genetically Improved Loblolly and Slash Pines in the South. *Journal of Forestry*, *101*(3), 32.

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. http://doi.org/10.1038/nrg2626

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, *157*, 1819–1829.

Namkoong, G. (1979). *Introduction to Quantitative Genetics in Forestry*. Washington, D.C. USA: Forest Service United States Department of Agriculture.

Namkoong, G. (1988). *Tree breeding: principles and strategies*. New York: Springer-Verlag.

Neale, D. B. (2007). Genomics to tree breeding and forest health. *Current Opinion in Genetics & Development*, *17*(6), 539–544. http://doi.org/10.1016/j.gde.2007.10.002

Neale, D. B., & Savolainen, O. (2004). Association genetics of complex traits in conifers. *Trends in Plant Science*, *9*(7), 325–330. http://doi.org/10.1016/j.tplants.2004.05.006

Neale, D. B., & Williams, C. G. (1991). Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. *Canadian Journal of Forest Research*, *21*(5), 545–554. http://doi.org/10.1139/x91-076

O'Malley, D. M. (1996). *The impact of plant molecular genetics*. Cambridge, MA, U.S.A: Birkhaüser.

PINEMAP Year 2 Annual Report. (2013, February). Retrieved from http://www.pinemap.org/reports/annual-reports/

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE*, *7*(2), e32253. http://doi.org/10.1371/journal.pone.0032253

Resende, M. F. R., Munoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., … Kirst, M. (2012). Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). *Genetics*, *190*(4), 1503–1510. http://doi.org/10.1534/genetics.111.137026

Shiver, B. D., Rheney, J. W., & Hitch, K. L. (2000). Loblolly Pine Outperforms Slash Pine in Southeastern Georgia and Northern Florida. *Southern Journal of Applied Forestry*, *24*(1), 31–36.

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biology*, *11*(5), 207. http://doi.org/10.1186/gb-2010-11-5-207

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., … Salamov, A. (2006). The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, *313*(5793), 1596–1604. http://doi.org/10.1126/science.1128691

Wear, D. N., Carter, D. R. C., & Prestemon, J. (2007). *The U.S. South's timber sector in 2005: a prospective analysis of recent change*. Asheville, NC, USA: U.S. Dept. of Agriculture, Forest Service, Southern Research Station.

Zobel, B., & Talbert, J. (1984). *Applied Forest Tree Improvement*. United States of America: John Wiley & Sons.

Zobel, B., & Van Buijtenen, J. (1989). Wood Variation: Its Causes and Control. *Springer-Verlag*.

CHAPTER 2

2.1 Introduction

Loblolly pine is an important source of income and is a resource used around the world. Of about 1.1 billion tree seedlings planted each year, loblolly pine comprises approximately 84% of the pines planted in the southern United States (McKeand et al., 2003). With its long generation time and low heritability, classical breeding strategies can only make so much gain per generation (Dorman, 1976). An alternative approach to classical breeding strategies is breeding using molecular genetic information. Experiments in techniques such as genetic mapping of Quantitative Trail Loci (QTL) and marker assisted selection (MAS) were first described for pine tree breeding by Neale & Williams (1991). Detection of single nucleotide polymorphisms, or SNPs, is a common form of genetic variation useful both for understanding diversity in native populations, and for applied breeding programs aimed at developing better planting stock.

Loblolly pine is a member of the gymnosperms, and similar to many other conifer species, it has an incredibly large and complex genome. The genome size of loblolly pine was determined to be approximately 21.6 Gb, making it almost 7 times larger than the human genome (O'Brien, Smith, Gardner, & Murray, 1996). Even with this complexity, the economic and environmental importance of loblolly pine as a species has pushed researchers to publish the first reference genome of any pine species in 2014, making it the largest reference genome in existence to date (Zimin et al., 2014).

Current research of this genome indicates that approximately 82% of it consists of repetitive sequences (Wegrzyn et al., 2014). These repetitive sequences come in many forms

such as transposable elements, gene duplications, or tandem repeats and can range from a few copies to millions of copies. They arise from a multitude of biological mechanisms and can be useful for tracing back evolutionary heritage (Jurka, Kapitonov, Kohany, & Jurka, 2007). Plant genomes in general are known for having particularly high percentages of repeats. Computationally, genomic repeats and large genomes have a tendency to hinder genomic sequencing, and methods to find ways to enrich genomic libraries for regions associated with desirable phenotypic traits are an ongoing process (Treangen & Salzberg, 2011).

The genetic complexity of loblolly pine is in part what makes it a good candidate for breeding purposes. It has good genetic adaptability, with a wide natural range from southern Virginia down into Florida and across the Mississippi to the Lost Pines area of Texas (Zobel & Talbert, 1984). Therefore, it is also important to quantify the population structure of loblolly pine which may be associated with locally-adapted subgroups. A common practice is to use molecular markers, generally adaptively neutral ones, to determine the genetic F_{ST} values. F_{ST} values give a standardized measure of the genetic differentiation among populations for a genetic locus. If there is a large F_{ST} value between different populations, it could indicate trees have adapted to, or are adapting to, local climates and further investigation efforts should be made to discover the potential importance of these adaptive traits.

The yield of data per unit cost produced by high-throughput sequencing platforms is currently expanding at an exponential rate, allowing genotyping methods to become more cost-effective over time. Previously, sequencing cost kept pace with Moore's law of computing hardware, but starting in 2007 new technologies have triggered genome

sequencing prices to drop significantly faster. For example, sequencing the human genome, which once took several years to complete and cost millions dollars, now takes about 7 days and costs only a few thousand dollars (Stein, 2010). Some of the technologies that are making these reductions possible are known as next generation sequencing (NGS) tools. NGS technologies are not only lowering the cost of sequencing, but also increasing the amount of coverage available. However, NGS also has its drawbacks. It is known to be more error prone than other strategies, such as Sanger sequencing, which affects our ability to decide what is a true SNP and what is an error.

NGS has previously been used in genome-wide association studies (GWAS), in the discovery of SNPs, and in the construction of haplotype maps (Metzker, 2010). Restriction site-associate genomic DNA (RAD) was first demonstrated by Baird et al. (2008). In order to make sequencing higher throughput, a multiplexing barcoding system used with restriction enzymes (REs) was developed and is known as genotyping-by-sequencing (GBS). Barcodes are added in the adaptors just upstream of restriction enzyme (RE) sites after ligation, allowing multiple samples to be pooled together and later differentiated by their barcodes. This allows other down-stream library preparation steps such as polymerase chain reaction (PCR) amplification to be done in smaller reactions, thus reducing cost. This method was first employed in maize, a genetically diverse and large-genome species (2.3Gbp) and since has also been employed in several other species (Elshire et al., 2011).

This experiment will use the double enzyme approach to reduce the complexity of the loblolly pine genome. The double enzyme method was first performed successfully by Poland et al. (2012) and was used in the complex genomes of wheat and barley. They genotyped bi-parental populations to develop a de-novo reference map of identified SNPs

and tags. Their results indicated the potential for broad application in genomic-assisted plant breeding programs.

By reducing the complexity of the genome, the computational complexity of aligning the read fragments is reduced for both de-novo and reference alignments. There are a wide variety of open source computing platforms available for NGS analysis. This experiment will mainly use a program called STACKS. STACKS is a program created to specialize in the analysis of next-generation sequencing data from short read restriction enzyme genotyped libraries (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; Catchen et al., 2011). The program can be used de-novo or with comparison to a reference genome to identify thousands of SNPs markers to be used in an array of population genetic analyses and in reference genome assembly.

The reduction in the cost of DNA sequencing means the costs associated with sample collection, storage, and DNA library construction now comprise a larger fraction of the total cost of sequencing experiments. The simple act of collecting and storing samples from the field environment can be time consuming and expensive. Here a robust, cost effective, and high throughput method is developed for collecting tissue samples, extracting DNA, and conducting DNA sequencing of loblolly pine to identify genetic variation for potential breeding purposes.

2.2 Materials and Methods

2.2.1 Sample Material

1,562 samples of tree phloem were collected from the Plant Seed Source Selection Study (PSSSS) located in Screven County, Georgia. The PSSSS consists of 140 pollen mix families planted in 20 test locations representing much of the loblolly pine range. Height and

diameter at breast height (DBH) measurements were taken at this site for ages 4 and 8 years and were available for use in the analysis.

2.2.2 Tissue Sampling and Storage

Tissue samples were collected using 19mm diameter leather hole punches. The phloem was peeled off the bark disk and placed in a pre-labeled 2.0mL tube with 1mL of storage solution containing 10mM EDTA (pH 8.0) and 10mM sodium bisulfite dissolved in water. The tissue samples were stored at room temperature for up to 4 months before DNA was extracted.

2.2.3 DNA Extraction

One half of each phloem disk was used per individual extraction. The disks were smashed with the bottom of a test tube prior to grinding to increase yields. DNA extractions were performed using a modified protocol from the Canadian Center for DNA Barcoding (Ivanova, Dewaard, & Hebert, 2006) with the insect lysis buffer and the following parameters:

Tissue samples were ground in 96-sample deep-well plates from Fisher Scientific using a ceramic bead in the bottom of each well, followed by the tissue sample, and topped with a ceramic cylinder. 600µl of warmed Insect Lysis buffer was used per sample. Samples were ground in a Mixer Mill at a speed of 30 Hz for 6 minutes on each side. A 1.5 hour of incubation at 60°C followed, with shaking at 30 Hz for 1 minute at every 30 minute interval of the 1.5 hour incubation. 250µl of crude lysate was taken from each sample and allowed to bind with the Plant Binding Buffer for 15 minutes before transferring to the glass filter plates. Both Pall and Nunc glass filter plates were used successfully. A 15 minute spin after the final

wash step was done in place of the 1.5 hour drying in a pre-heated oven. Samples were eluted using a low salt solution and allowed to stand for 5 minutes before centrifugation.

2.2.4 GBS Library Preparations

DNA was quantified using a fluorescent dye-binding assay called AccuBlue (Biotium). DNA yields were lower than expected so three different concentration methods were used. Samples were either dried down to concentration in a large scale plate speed vacuum, dried-down individually in a single tube speed vacuum, or were pooled at different volumes (differential pooling) during the first plate pooling process. The restriction digest, ligations, and sample pooling were performed according to the published protocol (Poland et al., 2012) with the following modifications (See Figure 2.1-2.2 for detailed information) :

Approximately 700ng of DNA was digested with PstI and MspI restriction enzymes, and then ligation using T7 or T4 ligase was performed overnight (~18 hours). No 65°C kill step was performed. 96 of the designed 384 barcode adapters from the Poland et al., 2012 protocol were used. Ligated samples from each plate were pooled together using 5 µl of each ligation reaction (unless the sample started with less than 700ng; if so, the volume pooled was adjusted appropriately). After ligation, samples were size selected for 200-500base pair (BP) range using a PippinPrep instrument (Sage Science). Each pooled plate was separated into 6 PCR reactions. A total of 12 indexing primers were used. Three different primers per plate were used, allowing for a total of 4 plates to be pooled together. The PCR reactions were done using Q5 polymerase (New England Biolabs). After amplification, all plates were quantified using PicoGreen (Life Technologies) and the plates were normalized to the same concentration and pooled together, producing one library with up to 384 samples each (4 plates). A total of 4 pooled libraries were produced.

2.2.5 Sequencing

Each of the libraries were sequenced for 100bp paired end reads over 4 flowcell lanes (~96 samples/lane) on an Illumina HiSeq2000 using V3 chemistry. A total of 16 flowcell lanes were sequenced by the Beijing Genomics Institute service center in Philadelphia.

2.2.5 Processing of Illumina data

Data in each of the raw fastq.gz data files from the four lanes within a pool were split into the four plates present based on the index sequences of the adapters, using a Linux command line script. Then, the 96 samples in each plate were split based on their read1 barcodes then read1 and read2 data were merged together into one file. The reads were trimmed and filtered using the FASTX-toolkit quality filter program. Raw sequence read counts per sample were generated using command line tools and the fastq.gz files.

2.2.6 SNP discovery

The ustacks program from the STACKS package (v1.10) was used to align short sequence reads de-novo with a minimum depth coverage of 3 and maximum distance allowed between stacks of 4.The option of disabling calling haplotypes from secondary reads was performed. A catalog of alleles was compiled using the cstacks program, and the sstacks program was used to identify haplotypes in each individual sample and produce a haplotype map matrix. The matrix was filtered for any loci with greater than two haplotypes and a heatmap/dendrogram was produced using the R environment.

2.2.7 Populations-F_{ST} Values

For analysis by the STACKS populations program, the individual sample files were filtered to only use the top ten samples from each 96-well plate from the output of cstacks.

The sstacks program was used to identify haplotypes in each individual sample, sorting in each plate for only samples that gave more than 1000 reads. The populations program to find FST values was run with -r .5 -m 3 -p 3 -a 0.02 options.

2.2.8 Alignment to reference

Samples were aligned with the *P.taeda* draft genome assembly v.1.01 of the reference genome sequence, and individual bam files were produced. The SAMtools program (version 0.19) was used to process aligned sequences, and the Freebayes program (v.9.9.2) was used to call SNPs and create Variant Call Format (VCF) files of genotypes. A filtered VCF file was created by filtering to retain about half the samples with the least missing data, retaining loci with a read depth mean of 10. Insert fragment size distribution in each library was analyzed using the largest single-sample BAM file in each plate. To confirm this was a good representative, one plate was analyzed using all samples, and the insert size distribution compared to that estimated from the largest BAM file of that plate.

2.2.9 Association Testing

A RV (rare variant) test was performed using the filtered VCF file and known seed parent pedigree file with height measurements at age 8 for association testing of rare variant alleles.

2.3 Results

2.3.1 Quality Time and Cost

The adapted DNA extraction method showed that high quality DNA suitable for NGS can be obtained from a variety of tissue types including needles, megagametophytes, tissue cultures and cambium. Samples stored in the storage solution should be extracted at 3 days to

one month after tissue collection for best DNA quality, but will remain viable at room temperature up to approximately four months before the degradation of DNA quality is too great for further manipulation. These results are illustrated in Figure 2.3, showing DNA extracted from phloem stored in the modified storage solution at room temperature for 3 days, 4 months and 7 months. The quantity of DNA from similar weight/sized phloem disks was highly variable. Of the 1600 samples collected only 1471 had DNA yields high enough to be sequenced.

Transferring the samples into a 96 well plate format allowed for a robust high throughput process. The method permitted up to 192 samples to be processed in approximately 4 hours by a trained technician. The most time-consuming step in the extraction method was the pre-smashing of the tissue samples and the transfer to the 96 deep-well plates. DNA yields for samples that were not hand smashed before extraction were not consistently high enough to be continued in the process. The calculated cost of consumable supplies and reagents for DNA extraction and library preparation was about $1.12 per sample or $107.77 per plate. This cost per plate decreases when two plates are done in tandem.

2.3.2 Sequencing Reads

By using DNA barcoding and restriction enzyme digestion, a mean of 7.61 million reads of data per sample was produced using 16 Illumina HiSeq V3 flow cells lanes. Of the sequenced reads returned, there was a wide range in the distribution of the number of reads. A Kolmogorov-Smirnov test revealed a significant difference in the number of reads per sample in all pairwise comparisons, except the samples pooled differentially and samples dried individually (Table 2.1). A visual representation of these distributions can be seen in figure 2.4 as well as the distribution of reads among the 96-barcoded adapters used for each

individual sample (figure 2.5). Of the two different ligases used (T4 ligase and T7 ligase) a Kolmogorov-Smirnov test was also performed and a significant difference between the two was found. Upon inspection of the paired end reads returned, the read 2 files had generally poor quality scores and had to be trimmed to 70 base pairs, resulting in less data than expected.

Using the STACKS de-novo analysis pipeline, unique polymorphic SNP loci were detected in ~94.3% of the individual samples sequenced. Figure 2.6 displays a graph of SNP counts versus the number of reads per sample.

The STACKS program produced a set of matching loci files for each individual sample. In an attempt to visually inspect the loci present across similar individuals, a dendrogram/heatmap of all loci vs. individuals was produced (Figure 2.7). The image revealed unique patterns of absent and present loci across individuals. The lighter regions reveal where loci are missing while darker yellow/red is present. There are clear sections where loci are present in certain groups of individuals. This distinctive pattern suggests that systematic differences in library preparation or sequencing led to differences in the yield of SNP loci recovered in different libraries.

2.3.3 Populations

Using the STACKS populations program with a de-novo approach, only 688 samples passed the filtering restrictions. From those 688 samples, 3.8 million loci were available but only 813 had enough samples in common to be used for analysis. The results are displayed in Table 2.2. The highest amount of differentiation among the various seed source populations occurred between populations 4 and 2, and the lowest amount of differentiation occurred

between populations 1 and 2. A map displaying the regions can be seen in figure 2.8 and the distribution of F_{ST} values among the populations is shown in figure 2.9.

2.3.4 Alignment and Association Testing

Approximately 1.69 million SNPs were detected from aligning the reads to the available draft reference genome assembly. However, only 4.47% of the aligned SNPs were present in more than half the individual samples. Of the 50,172 currently annotated genes 1,887 contained SNP variants. During the analysis process, a large amount of data was discovered to be missing. This missing data caused major analysis problems, including the inability to produce kinship matrices based on the SNP genotype data, due to the intolerance of matrix methods for missing data.

The DNA fragments were sequenced from both ends and then mapped to the reference genome, so it was possible to infer the insert fragment size distribution of each sample library mapped. The largest binary alignment and mapping file (.BAM file) from each plate was used as representative of the plate. Mapped DNA insert size distributions were found to be variable among the different pools and plates (Figure 2.10). This was most likely due to a poor quality size selection technique. This error could have been due the machine or human error. Figure 2.11 shows all samples from plate 4. These samples were all size selected together on the same lane of a Pippin-Prep cassette giving a very consistent range.

Using a kinship matrix built with the known seed parent, a rare single variant association test (RV Test) was performed using grammar gamma statistics. Grammar gamma was originally proposed by Svishcheva (2012) as a less computationally exhaustive GWAS test when dealing with thousands of SNPs. It has been determined to have statistical power

similar to likelihood ratio test-based methods while scaling linearly with sample size (Eu-ahsunthornwattana et al., 2014; Svishcheva, Axenovich, Belonogova, van Duijn, & Aulchenko, 2012). The RV Test found 180 SNPs to be significantly associated with height at age 8 after the Bonferroni correction was applied. This result is based on the assumption that Linkage Disequilibrium (LD) is close to 0 and the SNPs are independent. Table 2.3 lists the p-values. An RV Test using a kinship matrix built from the genotypic information would have been preferable to one based on recorded family pedigrees, but was not a viable option due to the aforementioned problem with excessive missing data.

When grouping multiple GBS-tagged sites per scaffold within 10Kb distance to account for rare variants of low allele frequencies, 5,838 SNP groups were produced. These SNP groups can be used for further association analysis to achieve stronger associations as they account for the rare variants that would not be associated otherwise.

2.4 Discussion/Conclusions

2.4.1 DNA Quality and Cost Optimization

For the process described here, samples can be stored up to 4 months before sample degradation is too great, but extractions should be performed between 1-6 weeks from sample collection for optimal DNA quality and yields. Further research on optimization of this method is currently underway such as adding a chloroform extraction step for increased yield. To lower cost of reagents per sample, a lower grade of Guanidinium thiocyanate can be used without compromising DNA yields, or an alternative chaotropic agent could be used in the extraction buffer. In the comparison of DNA concentration strategies, drying samples all the way down in a low heated speed vacuum may have degraded the DNA, resulting in the wide range in number of reads per sample. Drying down samples with small volumes left

was more time consuming due to having to measure the remaining liquid for each sample before normalizing but gave the least amount of variation in number of reads between samples.

In library preparation, the use of T7 ligase showed a significant difference in reads when compared to T4, but this could have been due to the differences in how the samples were concentrated and pooled. In preliminary tests, T7 showed no visible difference in gel images (Figure 2.12). As T7 costs less than half T4 ligase, it appears that the cost benefits are worth pursuing. In order to get more even distribution of reads per sample, either better quantification steps need to be used before and after pooling, or a higher number of samples should be pooled together and spread out into all of the lanes of the flow cell. It has been shown that quantitative PCR (qPCR) is a more accurate method of quantification, because it only quantifies DNA fragments with the barcodes that will be amplified by the sequencer (Rohland & Reich, 2012). In addition, the samples can be run on a DNA base pair size detection instrument such as an Agilent TapeStation 2200 or a Bioanalyzer to get accurate average fragment sizes for the total quantification.

Quantification should occur at multiple steps for better accuracy. Firstly, as individual samples after extraction, secondly after pooling of the 96-individually barcoded samples, and finally after each pooled plate is size selected. Another method to help reduce the range of read distribution could be addition of a greater number of indexing primers. This would allow a larger number of samples to be pooled together and spread over all eight lanes of a flow cell, or even multiple flow cells. By putting the same number of samples on a flow cell this would reduce a number of variable factors influencing the number of reads per sample.

2.4.2 Sequencing Reads and Genetic Variability

The biggest problem with sequencing genomic DNA from an ancient lineage such as loblolly pine is the size and variability of its genome. The original goal of the experiment was to get enough coverage to create genetic kinship matrices based on genotypic data for breeding as well as generating a cost effective system to check accuracy in recorded seedling lineages. As ~82% of the pine genome is repetitive, it becomes difficult to get the same sequenced fragments over a large range of samples (Wegrzyn et al., 2014). The obtained sequence information for each sample was large but not consistent across enough individuals to form conclusions, even with the use of missing data imputation software. Looking at the clustered heat map (Figure 2.7), it clearly shows the extent of missing data, as well as a clear pattern of separation among the identified loci present in some individuals and not others. This leads to the question of how genetically variable individuals might be across different seed sources.

Finding distinct population differentiation values (F_{ST}) among loblolly pine populations can be challenging due to both its continuous range and long distance gene distance gene dispersal system of wind pollination. When F_{ST} values were calculated, a large number of loci were dropped from analysis due to missing data. However, 813 loci were retained which is still a decent number for F_{ST} calculations. The results indicated very small amounts of differences among the seven seed sources tested. The greatest difference was seen between a seed source population located in the northeast part of the native range of loblolly pine, and the seed source population from the humid Lower Gulf part of the range. These results are logical considering the large environmental and geographical differences between those two seed sources. It is important to mention that one seed source of the

loblolly pine natural range was not present in the original PSSSS study. That seed source is known as the Lost Pines of Texas and is located west of the Mississippi River. Another study done on loblolly pine population structure found similar population structures based on F_{ST} values using SNPs as markers. They found the largest amount of population differentiation to occur between the trees west of the Mississippi and the populations east of the Mississippi, with a smaller magnitude differentiation occurring between the northern and southern populations (Eckert et al., 2010).

2.4.3 Usage of the Reference Genome

At the time of this experiment's conception, the loblolly pine reference genome was not yet available and thus the experiment was designed for the analysis to be done de novo. Halfway through the analysis process, the published reference genome assembly became available thus allowing for further analysis of the data. The reference genome assembly (v1.01) contains scaffolds ranging in size from 70 bp up to 8,000kb. Mapping the reads to the reference genome allowed for the identification not just of SNPs that were associated with the height phenotype, but also their scaffold position, thus making follow up and validation experiments more plausible. One method that could be explored more thoroughly is the idea of grouping SNPs together within a certain base pair range. The scaffolds available are small enough relative to the entire genome that a recombination event would be unlikely to occur, while still being large enough to identify a subset with multiple SNPs per scaffold. This information could be used with the objective to solve missing data problems, or the group of SNPs discovered by the scaffold information could be used in association of rare variants that otherwise would not be detected, such as with the 5,838 unique SNPs groups found in this study.

Another benefit of having a reference genome was that it allowed the inference of the average sequence fragment read length. The suggestion of using a PippinPrep machine as a means for base pair fragment size selection for a more even distribution of reads was proposed in a de novo SNP discovery paper that had positive results (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). Upon a closer examination of the average library insert fragment sizes in this study, based on alignments to the reference genome, it was discovered that the insert fragment lengths varied greatly among the 16 independent libraries. However, when all of the samples in one plate were examined together there was seen to be in a very tight range (Figure 2.11). Since each pooled library prepared from a plate was size-selected on a PippenPrep instrument prior to pooling the libraries together, this suggests the size-selection on the PippinPrep instrument is not as reproducible in our experiment as was originally reported by Peterson et al (2012). This could have been due to a number of factors such as technician or mechanical error. To help avoid this factor of variability in the future, two rounds of size-selection on a PippinPrep instrument could be performed. The first round would be a large size cut range of 150-450bp at the plate level, and then a much narrower range of 300 +/- 50bp for the second round once all the plates are pooled together to focus on a specific set of restriction fragments used.

2.4.4 Further Steps

Ultimately, it appears that the major problem to overcome in implementing genetic markers in a cost effective, high throughput breeding program is developing a method to successfully reduce the variability of the sections of the genome isolated for sequence analysis among different individuals of loblolly pine.

The double enzyme digest was proven to be successful in maize which also has a large complex genome (2.3GB), but it is possible to create inbred lines in maize which can be used to help interpret missing data areas (Schnable et al., 2009). Loblolly pine is a highly heterozygous, outcrossing species that does not tolerate inbreeding very well. Studies have shown decreases in growth and quality traits for inbred progeny (Ford, McKeand, Jett, & Isik, 2015). One alternative to the double enzyme digest, such as the one performed in this experiment, could be a triple or even quadruple enzyme digestion of DNA fragments. This, paired with a size selection step, could reduce the large genome variability between samples enough to build kinship matrices to aid in association testing. The problem of missing data could also be solved by getting a greater amount of sequencing coverage per sample. This could be achieved by putting fewer samples per flow cell lane and thus getting a deeper coverage of each sample. However, by lowering number of samples per lane, the cost of sequencing each sample increases. Some alternatives to sequencing DNA directly include full RNA sequencing and the mapping of DNASE hypersensitive sites. Both are ways to discover genetic variability in specific subsets of the pine genome, and are discussed in Chapter 3.

2.5 Tables

Table 2.1: Comparison of four concentration methods. The P-values from Kolmogorov-Smirnov two-way comparison test shown for four different methods of concentrations used.

	Number of individuals	Not dried	Dried	Dried individually	T4 ligase
Number of individuals	1466	186	980	103	468
Dried	980	0.00024			
Dried Individually	103	3.60E-11	4.49E-07		
pooled	197	5.55E-16	1.68E-13	0.1182	
T7 ligase	998				0.002116

Table 2.2: Pairwise FST values of the seven PSSSS seed source regions. The regions are designated in Figure 2.8.

Regions	7	2	6	1	4	5
3	0.005235	0.005886	0.006278	0.005528	0.008592	0.005322
5	0.006695	0.007407	0.005853	0.00709	0.007754	
4	0.01098	0.013245	0.00654	0.011488		
1	0.004796	0.004743	0.008102			
6	0.008044	0.009311				
2	0.00491					

Table 2.3: List of P-value associations. 180 Chromosome scaffold IDs and their corresponding P-values from the RV Test for association to height at age 8.

Chromosome	P-Value	Chromosome	P-Value	Chromosome	P-Value
C25710560	1.38E-08	scaffold290639	3.68E-07	scaffold904324	2.06E-07
C29088176	7.29E-09	scaffold291742.3	2.12E-07	scaffold904568.2	1.04E-09
C30899172	1.50E-07	scaffold316511	1.96E-07	scaffold99282.2	4.15E-10
C30899172	6.41E-07	scaffold325073.2	5.25E-08	tscaffold1124	1.60E-07
C31083456	4.11E-08	scaffold344999.1	1.90E-07	tscaffold1274	1.08E-07
C31307240	1.00E-07	scaffold352463	2.10E-07	tscaffold1433	2.10E-07
C31589840	1.91E-08	scaffold3623	4.04E-07	tscaffold1433	4.18E-08
C31627720	2.74E-07	scaffold3623	4.07E-07	tscaffold1994	1.82E-07
C31627720	3.96E-07	scaffold369565	9.93E-09	tscaffold2199	2.14E-07
C31627720	4.26E-07	scaffold387017.1	1.82E-07	tscaffold235	2.24E-07
C31627720	1.04E-08	scaffold421845	2.46E-08	tscaffold2351	6.64E-07
C31627720	5.68E-10	scaffold431825	4.37E-07	tscaffold2418	1.97E-07
C32056646	5.31E-07	scaffold462106	7.97E-11	tscaffold245	2.12E-07
C32056678	8.35E-08	scaffold505226	2.80E-07	tscaffold2467	3.56E-08
C32056678	2.40E-08	scaffold534355.3	8.62E-09	tscaffold2677	7.17E-10
C32056678	1.33E-07	scaffold534355.3	9.07E-09	tscaffold2820	3.38E-07

Table 2.3 Continued

C32056678	1.18E-08	scaffold534355.3	1.51E-08	tscaffold2841	2.54E-08
C32056678	1.72E-08	scaffold534355.3	1.63E-07	tscaffold3121	1.59E-09
C32056678	6.72E-08	scaffold534355.3	6.80E-09	tscaffold3206	3.63E-11
C32056678	1.29E-07	scaffold534355.3	6.15E-09	tscaffold3212	1.80E-07
C32056678	1.64E-07	scaffold534355.3	6.52E-09	tscaffold3449	3.79E-11
C32077252	2.15E-09	scaffold554602	1.72E-07	tscaffold3751	1.70E-07
C32179806	8.77E-08	scaffold614158.1	2.59E-07	tscaffold3756	1.75E-07
C32204638	1.32E-08	scaffold617864	4.48E-10	tscaffold3821	3.05E-11
C32207922	4.95E-07	scaffold632842	2.29E-07	tscaffold4004	2.09E-07
C32420918	1.91E-07	scaffold634476	4.43E-07	tscaffold4019	5.32E-07
C32437858	3.97E-07	scaffold634476	9.12E-08	tscaffold4244	2.03E-07
C32462510	3.69E-09	scaffold64212	9.16E-09	tscaffold4360	1.91E-07
C32489810	1.73E-07	scaffold668959	3.07E-07	tscaffold4568	4.95E-07
C32546566	1.37E-08	scaffold690924	3.70E-07	tscaffold4773	5.25E-11
C32561210	1.75E-07	scaffold70602	2.34E-07	tscaffold4866	3.85E-08
C32564876	4.90E-10	scaffold709992	5.01E-07	tscaffold4915	3.68E-08
scaffold102550.3	6.11E-08	scaffold722884	3.14E-08	tscaffold5122	5.60E-07
scaffold102550.3	1.65E-09	scaffold725889.2	2.44E-07	tscaffold545	6.17E-09
scaffold10723	2.54E-10	scaffold728462	1.63E-07	tscaffold545	9.53E-09
scaffold123774.2	2.95E-07	scaffold739062.1	4.05E-11	tscaffold545	3.58E-09
scaffold140209.2	1.19E-08	scaffold76648.1	6.58E-09	tscaffold5514	1.42E-07
scaffold144909.2	1.86E-08	scaffold779694	1.50E-08	tscaffold5739	4.61E-08
scaffold158244.2	6.10E-07	scaffold823304	5.00E-07	tscaffold5991	5.10E-11
scaffold165922	7.94E-08	scaffold823579	3.05E-08	tscaffold6066	1.98E-07
scaffold165922	4.71E-07	scaffold82446.1	6.65E-07	tscaffold6075	8.81E-08
scaffold165922	5.11E-07	scaffold827872	1.79E-07	tscaffold6194	2.04E-07
scaffold170728	5.03E-09	scaffold838092	1.95E-07	tscaffold641	1.51E-07
scaffold176753	4.91E-07	scaffold845834	6.33E-07	tscaffold6660	5.47E-07
scaffold182294	1.54E-10	scaffold846213.1	5.61E-07	tscaffold6916	3.83E-11
scaffold182294	1.91E-10	scaffold851023	8.93E-08	tscaffold6924	3.33E-11
scaffold182294	8.44E-11	scaffold852101	1.96E-07	tscaffold7004	3.13E-08
scaffold182294	3.98E-11	scaffold859531	4.45E-11	tscaffold7004	3.07E-08
scaffold182294	1.72E-10	scaffold861095	3.52E-11	tscaffold7408	1.38E-08
scaffold185655.2	4.71E-07	scaffold861095	3.52E-11	tscaffold7735	4.37E-08
scaffold188.2	1.27E-07	scaffold868319.1	5.34E-11	tscaffold8140	4.26E-11
scaffold190152.2	3.74E-11	scaffold880868.2	6.94E-08	tscaffold8398	1.73E-07
scaffold212472	1.10E-07	scaffold884698.1	3.39E-11	tscaffold843	4.94E-11
scaffold212874	2.71E-09	scaffold890017	1.95E-07	tscaffold8443	2.27E-07
scaffold218949	5.81E-09	scaffold893600	4.36E-11	tscaffold8462	1.71E-07
scaffold221300	4.52E-11	scaffold901561.2	3.26E-08	tscaffold8469	1.44E-07
scaffold22142	1.95E-07	scaffold901631	1.91E-07	tscaffold8526	5.69E-09

Table 2.3 Continued

scaffold22989	3.50E-11	scaffold902596	3.02E-08	tscaffold8577	9.71E-08
scaffold268478	1.66E-08	scaffold903014	4.82E-08	tscaffold9153	1.96E-07
scaffold279993	4.15E-11	scaffold903253	5.71E-07	tscaffold92	4.56E-11

2.6 Figures

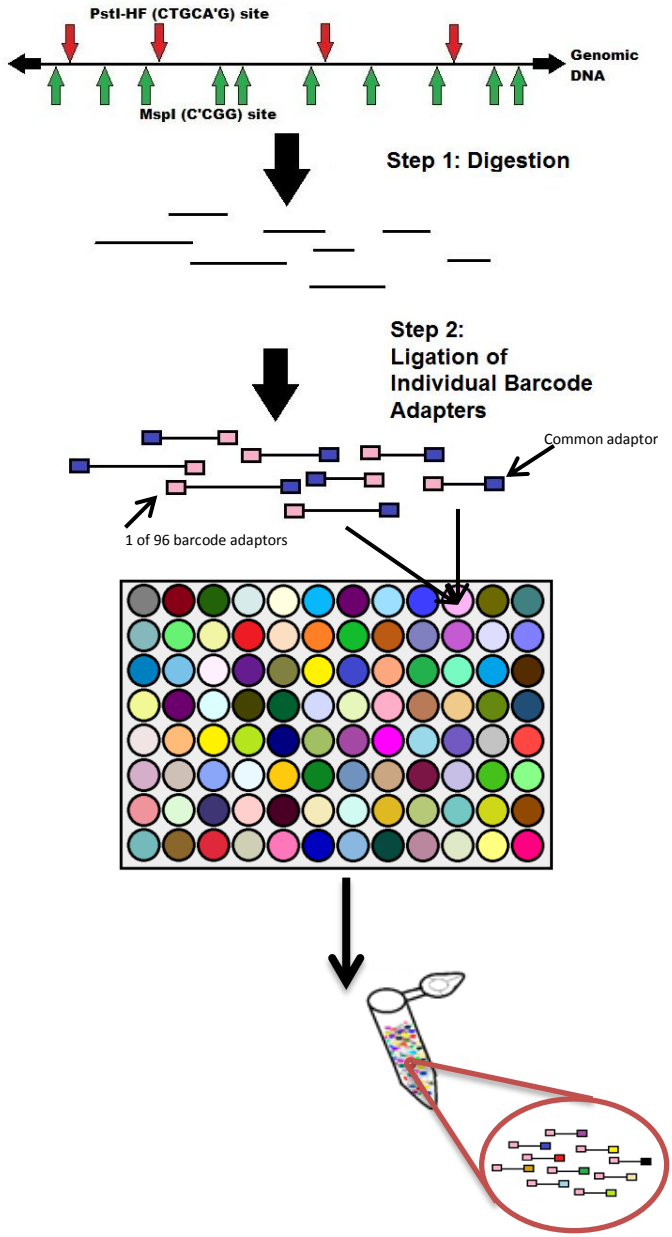

Figure 2.1: GBS Library Preparation. Step one illustrates restriction enzyme digestion of the DNA into small pieces. Common cutter enzyme MSPI was used as well as methylation sensitive rare enzyme PSTI. Once the digestion was complete, 96 distinct variable length barcodes were ligated onto the 5' end of the DNA fragment containing the PSTI cut site. A common DNA barcode was ligated onto the 3' end of the MSPI cut site. Once ligation was complete 96-individually tagged samples from each plate were pooled together into one tube.

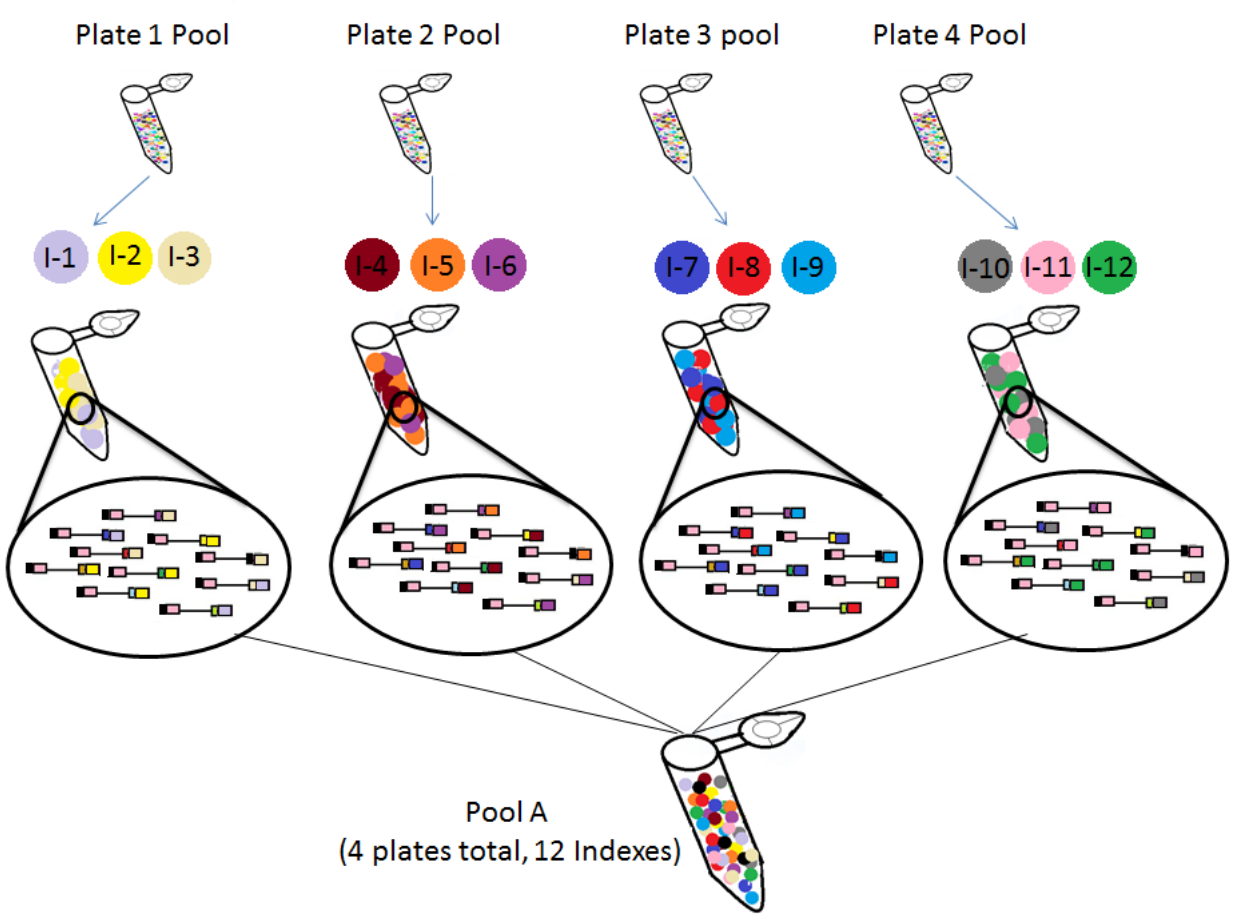

Figure 2.2: GBS Library Pooling. Once plates were pooled together, the DNA fragments were size selected on a Pippin-Prep and then PCR was performed. This image illustrates that 12 indexing primers were used to pool up to 4 plates together. The indexing primers were designed to only amplify fragments containing one of the 96-barcodes and one common barcode thus only amplifying fragments with a PSTI and MSPI site. Three different indexes were used per plate, after PCR 4 plates were pooled together to make one library. Four total libraries (16 plates) were produced

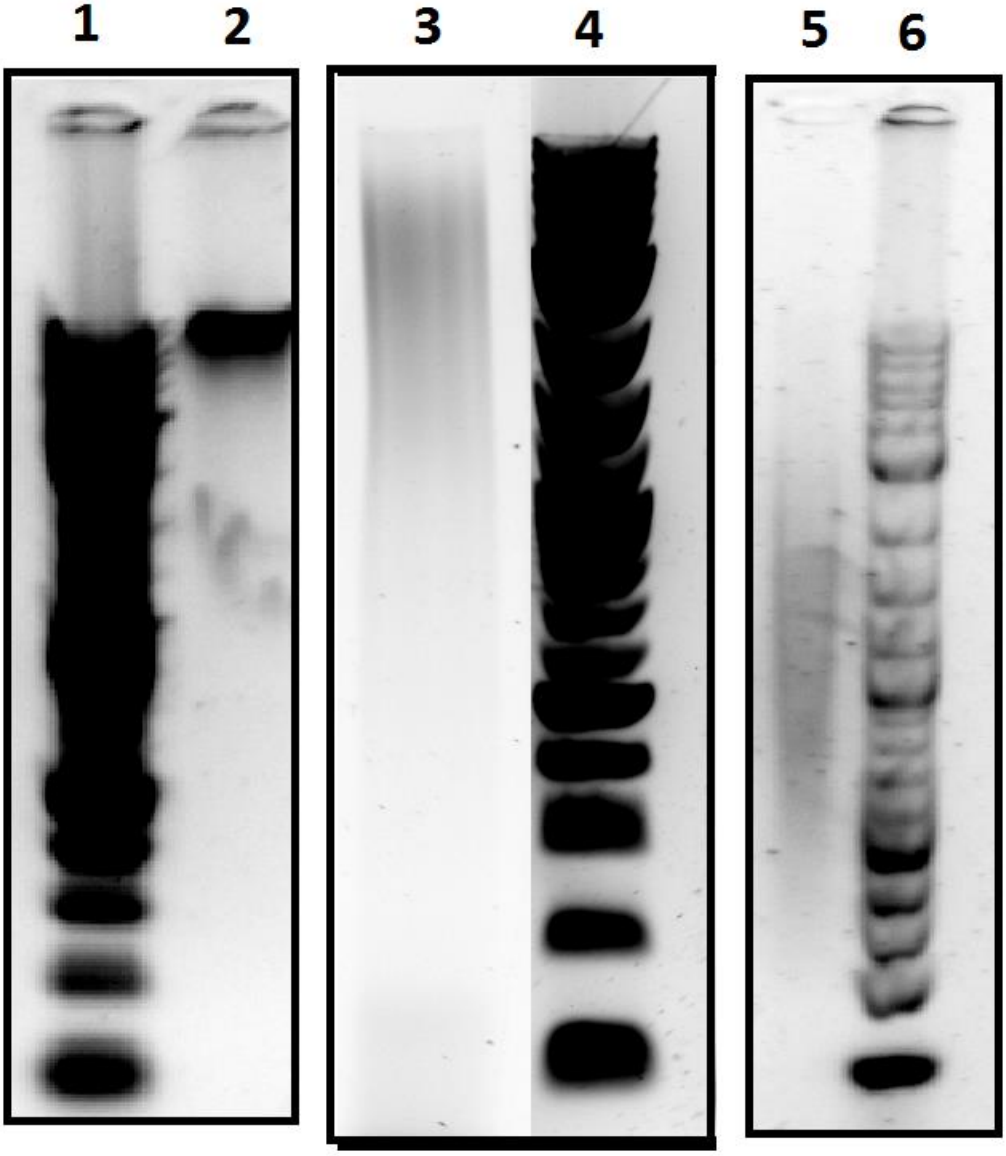

Figure 2.3: Storage Solution Degradation Over Time. 1% Agarose gels showing DNA extracted from cambium disk stored in modified water storage solution. Lane 1: 2-log DNA ladder, lane 2: DNA extracted after 3 days in storage solution, lane 3: DNA extracted after 4 months in storage solution, lane 4: 2-log ladder, lane 5 DNA extracted after 7 months in storage solution, lane 6: 2-log ladder.

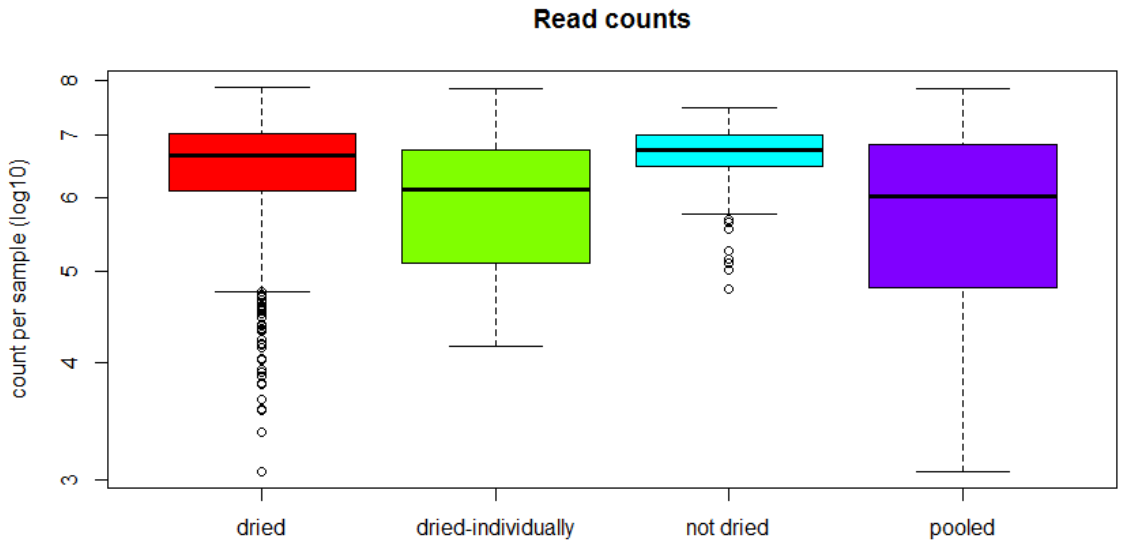

Figure 2.4: Read Counts by Concentration Strategies. Distribution of reads per individual by DNA concentration strategies used.

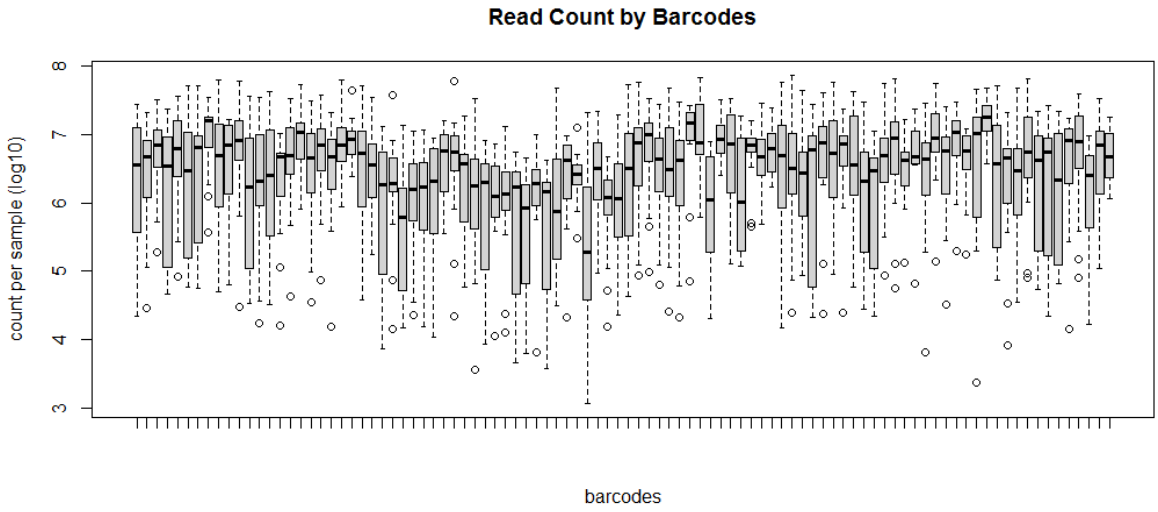

Figure 2.5: Read Count by Barcodes. Distribution of the number of reads of the 96 individual barcodes used across all 16 plates.

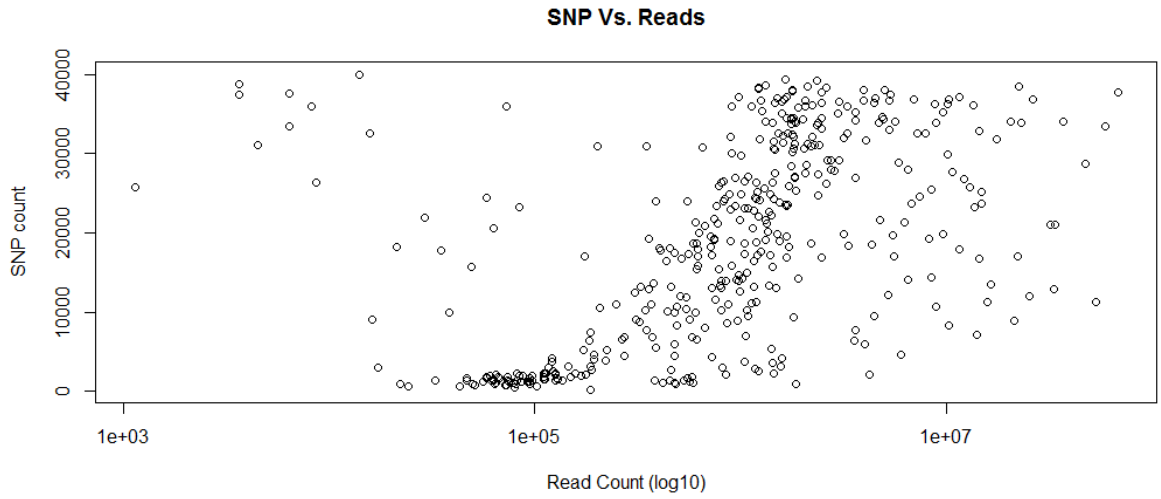

Figure 2.6: SNPs versus Reads. Distribution of the number of reads per sample versus number of counts per sample.

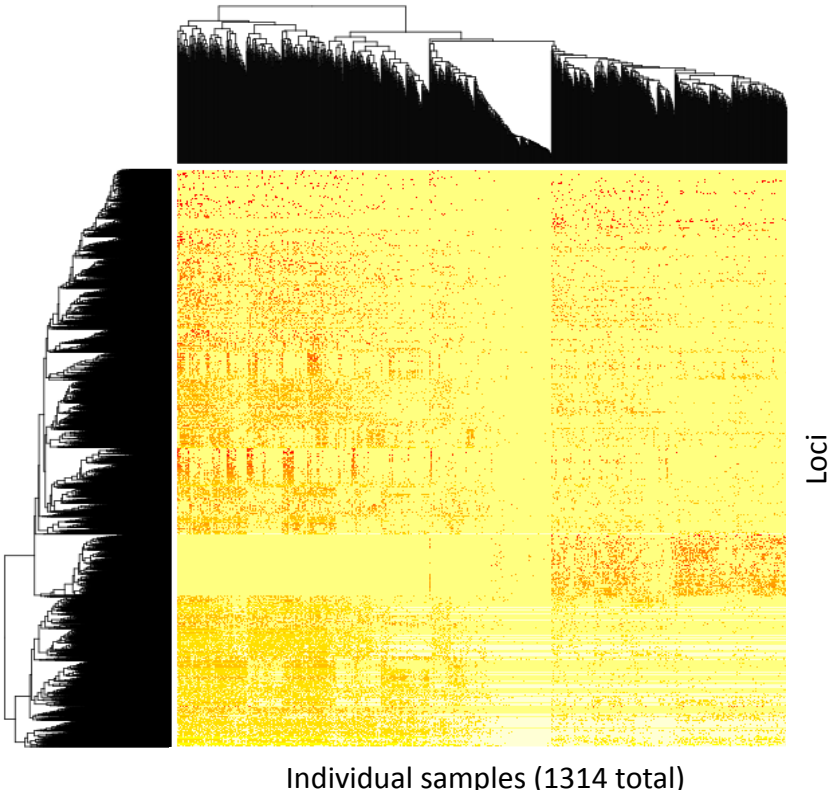

Figure 2.7: Dendrogram /Heatmap. Graphical display showing binary cluster of loci present/absence across all individuals sampled.

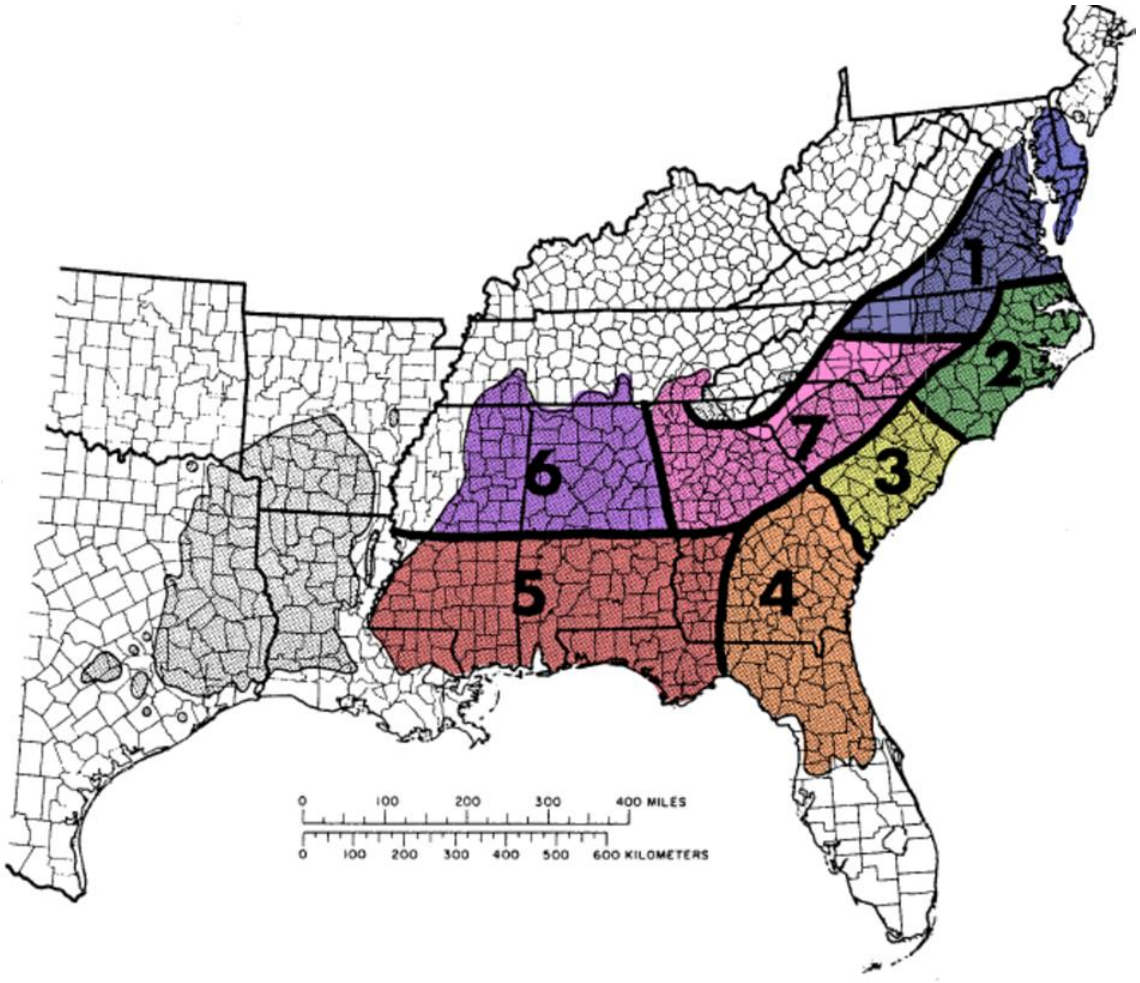

Figure 2.8: Seed Sources. Map displaying the 7 different seed sources used in the PSSSS trial.

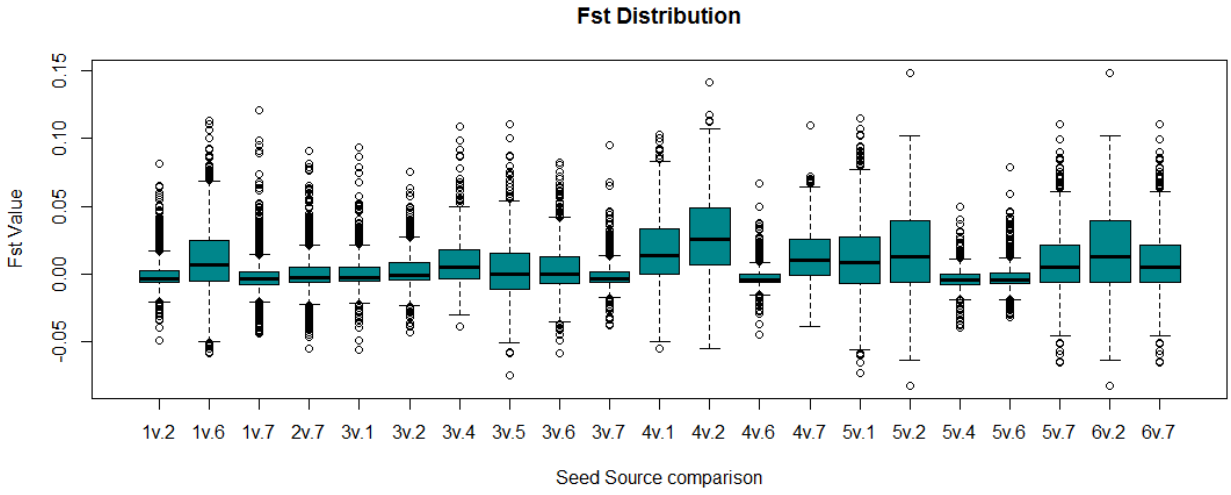

Figure 2.9: F_{ST} Distribution. Boxplot of the comparison of the pairwise F_{ST} values of the seven different seed sources in the PSSSS trial.

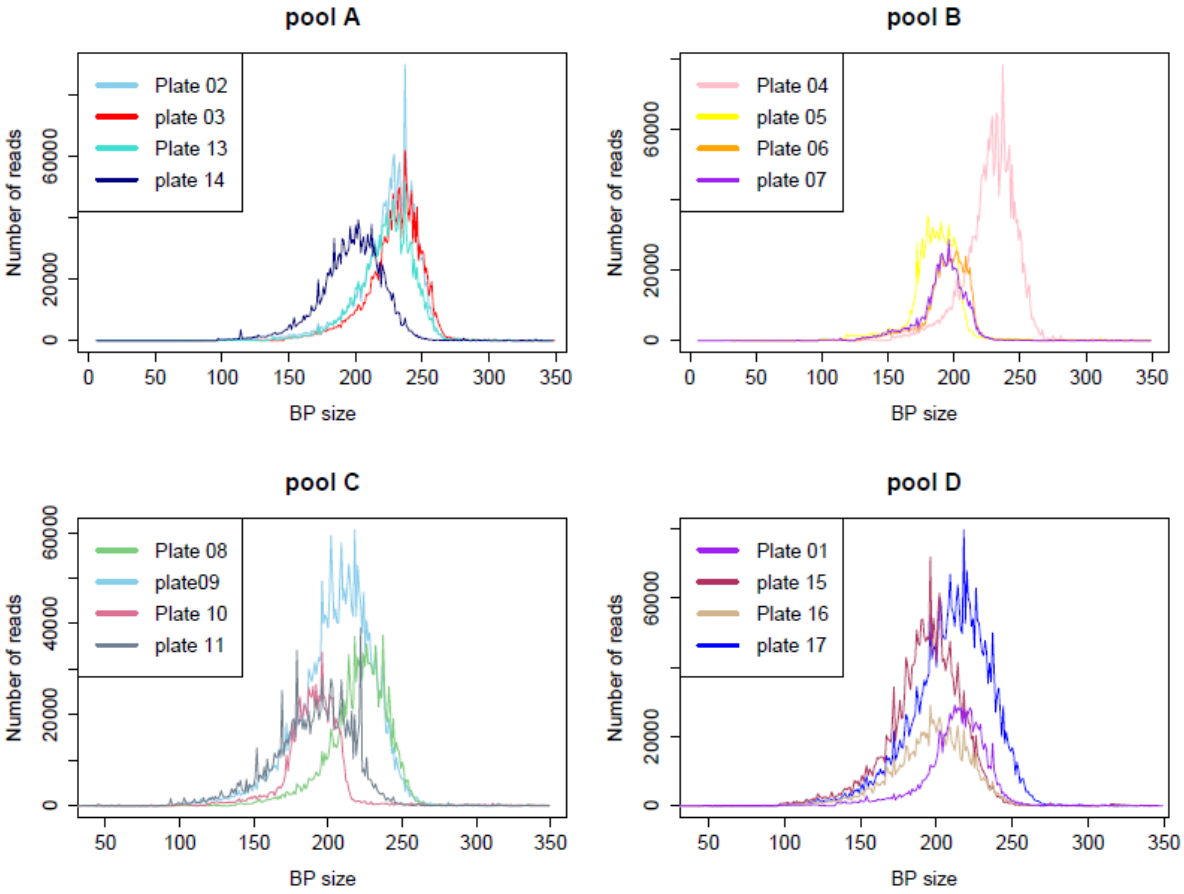

Figure 2.10: Distribution of Size Classes per Plate. Four pools with the largest .BAM file representing the average base pair size distributions for plate of samples.

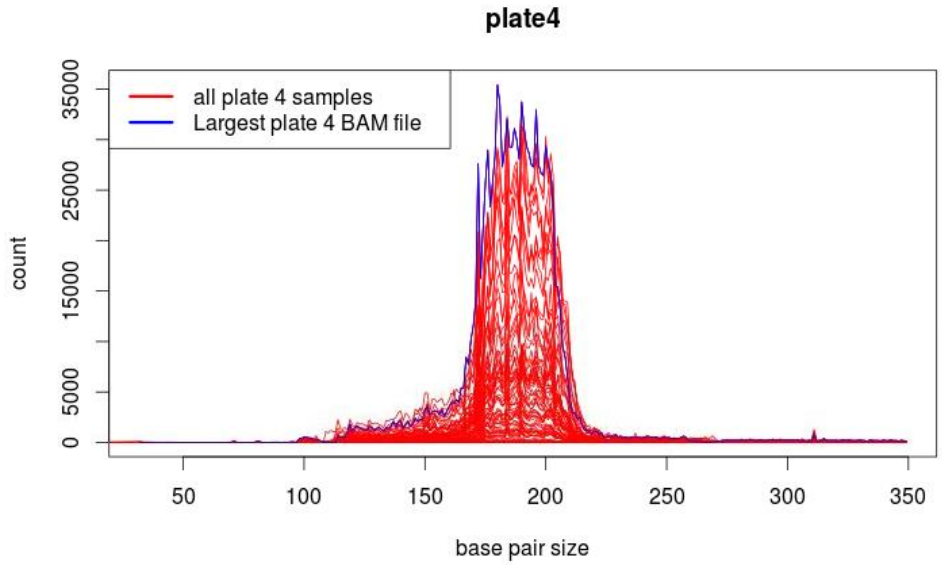

Figure 2.11: Distribution of Reads of All Samples in Plate 4. All size distributions of the 96 samples in plate 4 are represented with red lines from the mapped BAM files with the largest BAM file being represented in blue.

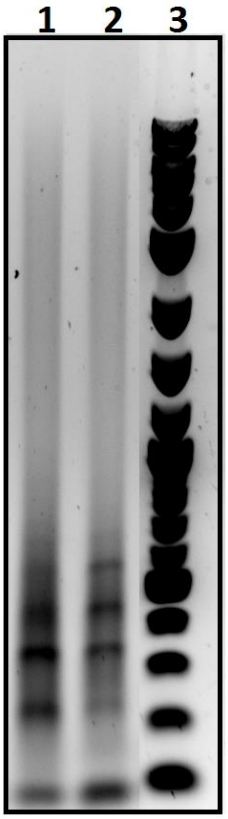

Figure 2.12: T7 Vs. T4 Ligase. 1% Agarose gel. Lane 1: Cambium extracted DNA after ligation with T4 ligase and 17 cycles of PCR, lane 2: Cambium extracted DNA after ligation with T7 ligase and 17 cycles of PCR, lane 3: 2-log ladder.

2.7 Literature Cited

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, *3*(10), e3376. http://doi.org/10.1371/journal.pone.0003376

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. http://doi.org/10.1111/mec.12354

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J. H., & De Koning, D.-J. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics*, *1*(3), 171–182. http://doi.org/10.1534/g3.111.000240

Dorman, K. W. (1976). *The Genetics and breeding of southern pines*. Washington: U.S Dept. of Agriculture, Forest Service: for sale by Supt. of Docs., U.S. Govt.

Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., Gonzalez-Martinez, S. C., & Neale, D. B. (2010). Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (Pinus taeda L., Pinaceae). *Genetics*, *185*(3), 969–982. http://doi.org/10.1534/genetics.110.115543

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, *6*(5), e19379. http://doi.org/10.1371/journal.pone.0019379

Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Wellcome Trust Case Control Consortium 2, Jeronimo, S. M. B., Blackwell, J. M., & Cordell, H. J. (2014). Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. *PLoS Genetics*, *10*(7), e1004445. http://doi.org/10.1371/journal.pgen.1004445

Ford, G. A., McKeand, S. E., Jett, J. B., & Isik, F. (2015). Effects of Inbreeding on Growth and Quality Traits in Loblolly Pine. *Forest Science*, *61*(3), 579–585. http://doi.org/10.5849/forsci.13-185

Ivanova, N. V., Dewaard, J. R., & Hebert, P. D. N. (2006). An inexpensive, automation-friendly protocol for recovering high-quality DNA: TECHNICAL NOTE. *Molecular Ecology Notes*, *6*(4), 998–1002. http://doi.org/10.1111/j.1471-8286.2006.01428.x

Jurka, J., Kapitonov, V. V., Kohany, O., & Jurka, M. V. (2007). Repetitive Sequences in Complex Genomes: Structure and Evolution. *Annual Review of Genomics and Human Genetics*, *8*(1), 241–259. http://doi.org/10.1146/annurev.genom.8.080706.092416

McKeand, S., Mullin, T., Byram, T., & White, T. (2003). Deployment of Genetically Improved Loblolly and Slash Pines in the South. *Journal of Forestry*, *101*(3), 32.

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

Neale, D. B., & Williams, C. G. (1991). Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. *Canadian Journal of Forest Research*, *21*(5), 545–554. http://doi.org/10.1139/x91-076

O'Brien, I. E. W., Smith, D. R., Gardner, R. C., & Murray, B. G. (1996). Flow cytometric determination of genome size in Pinus. *Plant Science*, *115*(1), 91–99. http://doi.org/10.1016/0168-9452(96)04356-7

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, *7*(5), e37135. http://doi.org/10.1371/journal.pone.0037135

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE*, *7*(2), e32253. http://doi.org/10.1371/journal.pone.0032253

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939–946. http://doi.org/10.1101/gr.128124.111

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Graves, T. A. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, *326*(5956), 1112–1115. http://doi.org/10.1126/science.1178534

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biology*, *11*(5), 207. http://doi.org/10.1186/gb-2010-11-5-207

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., & Aulchenko, Y. S. (2012). Rapid variance components–based method for whole-genome association analysis. *Nature Genetics*, *44*(10), 1166–1170. http://doi.org/10.1038/ng.2410

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*. http://doi.org/10.1038/nrg3117

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L.-S., Loopstra, C. A., Vasquez-Gross, H. A., … Neale, D. B. (2014). Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. *Genetics, 196*(3), 891–909. http://doi.org/10.1534/genetics.113.159996

Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marcais, G., … Langley, C. H. (2014). Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. *Genetics*, *196*(3), 875–890. http://doi.org/10.1534/genetics.113.159715

Zobel, B., & Talbert, J. (1984). *Applied Forest Tree Improvement*. United States of America: John Wiley & Sons.

CHAPTER 3

3.1 Introduction

Loblolly pine is an economically and environmentally important species. It comprises approximately 84% of the seedlings planted in the southern United States each year, and of that, 99% is from improved seed stock (McKeand et al., 2003). Tree breeders are interested in the discovery of sequence variants associated with desirable phenotypes, both for the forestry industry and for adaptability to varying climates. As the loblolly pine genome is approximately 22 Gb, and as 82% of it consist of repetitive sequences, finding the relevant variants is challenging (Wegrzyn et al., 2014). There have been many methods applied to similar species with large genomes in attempts to reduce the sections being sequenced to only transcriptionally active gene regions. RNA expression profiles are widely popular and have been used in studies for a number of species such as Arabidopsis, maize, and humans to name a few (Dittmar, Goodenbour, & Pan, 2006; Mach, 2011; Wunderlich, Groß-Hardt, & Schöffl, 2014). The use of methylation sensitive restriction enzymes has also been successfully used in RADseq and GBS experiments (Elshire et al., 2011; Peterson et al., 2012). Another method for discovering relevant variants is the mapping of DNase I sensitive and hypersensitive sites. This technique not only identifies transcriptionally active regions of the genome but can also discover many different types of genetic regulatory elements.

Inside a cell's nucleus, DNA is packed into various structure levels. Figure 3.1 shows the six main structures. In the 11nm fiber structure DNA is wound around a positively charged protein known as a histone. This structure is eventually more densely compacted into chromosomes. The small fragments of DNA between the histones are where DNase-hypersensitivity sites are theorized to reside. Here the DNA is physically accessible

(transcriptionally poised) to many kinds of transcription factors. The use of DNase sensitivity sites as a method of discovery for important genetic variants was originally developed by Harold Weintraub. He reasoned that if DNA in chromatin form is physically inaccessible to transcription factors in vivo, then it should also be inaccessible to endonucleases in vitro (Weintraub & Groudine 1976). DNase I is an endonuclease that readily cuts transcriptionally poised DNA that is in chromatin form. Weintraub's group showed that DNase I sensitivity is not confined to the transcribed and promoter regions of a gene but can be hundreds of thousands or even millions of base pairs away from the promoters they regulate. The theory of using DNase-hypersensitive sites as a way to discover relevant DNA variants is similar to the concept of RNA expression. Different tissues of an organism are actively transcribing different proteins, thus the DNA should be packed in accordance to which regions need to be transcriptionally active for a protein to be produced. This has been studied quite extensively in humans. Recent genome wide association studies (GWAS) have found 88% of the 5,654 SNPs detected were in non-coding DNA of unknown function rather than in a protein coding regions within genes. 76% of these variants in non-coding sequences are in DNase-hypersensitive sites or in Linkage Disequilibrium (LD) with DNase-hypersensitive sites, and this proportion appears to increase with increased strength of association with phenotype (Hindorff et al., 2009; Maurano et al., 2012).

A limitation in detection of DNase-hypersensitive sites in plant chromatin is in the extraction of intact nuclei from plant cell tissue types. This is due to the levels of polysaccharides, chloroplast, phenolic, and other organic compounds typically found in plant tissues that interfere with DNA extractions (Paterson, Brubaker, & Wendel, 1993). A plethora of buffer solutions have been developed to combat these issues when performing

high molecular weight extractions, but the biochemical diversity of plant species makes universal methods and buffers ineffective (Hein, Williamson, Russell, & Wayne, 2005). This experiment attempts to develop a robust and effective DNA nuclei isolation protocol to be used with various tissue types from loblolly pine, and to show that nuclei obtained from different tissue types have different patterns of accessible chromatin.

Based on these theories, a method was developed to isolate nuclei with DNA in its intact chromatin structure from loblolly pine and digest the chromosomal DNA to prepare sequencing libraries, with the goal of finding SNPs in LD with transcriptionally active genes in four distinct tissue types. The restriction enzyme MseI will be used instead of a nuclease as nucleases usually fragment DNA with ragged ends that then require the addition of an end repair step and a large amount of template DNA. In contrast, the MseI restriction enzyme leaves a 5' –TA overhanging end, allowing for a more efficient ligation and thus smaller sample sizes can be used.

3.2 Materials and Methods

3.2.1 Plant Material

All sample material was taken from the Schenck Forest in Raleigh, NC. Four distinct plant tissues including phloem, cambium, shoot tips, and needles were taken from three different loblolly pine trees, frozen in liquid nitrogen, and stored in -80^{o}C freezer until extractions were performed.

3.2.2 Nuclei Extraction

For the intact nuclei isolations, a series of different buffers were tested including the use of non-ionic detergents and Percoll filtration layers (Sikorskaite, Rajamäki, Baniulis,

Stanys, & Valkonen, 2013). MEB-buffer and MPDB-buffer from Hein et al., (2005) were found to be the most effective and used for this experiment. For best results all samples were kept on ice during extraction and filtering. Approximately 3-4 grams of plant tissue (Table 3.1 lists specific weights) were ground in liquid nitrogen to a fine powder using a mortar and pestle. 8mL of chilled MEB-buffer solution was transferred to the ground powder in the mortar and the solution was gently stirred every 2 minutes for a total of 12 minutes. The homogenate was filtered through two layers of Kimtech Science KimWipes, and then transferred to 2mL tubes with 1,500ul of filtered homogenate in each tube. Each 2mL tube was centrifuged at 1,500g for 4 minutes at 4^{o}C. The supernatant was decanted and the nuclei pellet retained. The pellet was re-suspended in 750ul of MPDB-buffer. The centrifugation wash step was repeated a total of 4 times. After the final spin the pellets were re-suspended in 20ul MPDB wash buffer. All pellets from the same sample were combined into a new 1.5mL tube.

3.2.3 DNA Extraction

One sample of genomic DNA was extracted using the protocol described in section *2.2.3 DNA Extractions*. This sample was used as a control of library purified DNA that has no chromatin structure, thus allowing for random digestion of DNA wherever the MseI restriction sites occur.

3.2.4 GBS Library Preparation and Sequencing

Suspensions of purified intact nuclei were digested with 1ul MSEI enzyme at 37^{o} C for 15 minutes on a rotating plate for continual agitation of the solution; table 3.2 lists exact volumes. The enzyme diffused through nuclear pores to find accessible sites in the chromatin structure to cut. The cut DNA fragments diffused out through the pores and were suspended

in the supernatant. After digestion, samples were incubated at 65^{o}C for 20 minutes to destroy the restriction enzyme activity. Samples were centrifuged at 12,000 rpm for 7 minutes to pellet the debris. DNA was recovered from the supernatant on a DNA Clean & Concentrator^{TM}-25 column (Zymo Research). All samples were quality checked on 1% agarose gel (Figure 3.2). Thirteen custom designed barcoded adapters were ligated overnight onto the digested DNA with the parameter found in Table 3.3-3.4. Fragments of 120-300bp range were size selected on a PippinPrep (Sage Science) instrument. A large fraction of each DNA sample was lost during size selection as evident from table 3.5, showing DNA concentrations before and after size selection. PCR reactions were performed using three reactions per sample type for all tissue types (Table 3.6-3.7) and quality checked on 1% agarose gel (Figure 3.3). Samples were size selected again on the Pippin-Prep for 190-370bp range to clean up any primer dimers and quantified on a Bioanalyzer. All samples were normalized and pooled together to create one library with 13 samples. Samples were sequenced over two lanes on an Illumina HiSeq2500 using V4 chemistry with 100BP paired end reads for approximately 29 million reads per sample.

3.2.5 Processing of Illumina Data

The Hi-Seq data were downloaded and reads were split based on their tagged barcoded sequences using the program Flexbar. Flexbar also allowed the 3' ends to be trimmed until a quality score of 20 was reach for each read sequence. Samples were aligned to the *P.taeda* assembly v.1.01 using the BWA alignment software V0.7.5. Aligned reads were summarized using SAMtools. The BEDtools program was used to convert .BAM files to BED format for sample comparisons. Sample comparisons were done using an open source multi-intersect function tool. These comparisons were filtered for read alignments

greater than 50 nucleotides. Comparisons and summaries were produced in the R environment.

3.3 Results and Discussion

3.3.1 Nuclei Extraction and Library Preparation

Overall, the nuclei extraction procedure developed allowed for high quality intact nuclei to be extracted from various loblolly pine tissue/organ types using low inputs of samples. The use of the common cutter enzyme MseI allowed for quick and easy ligation of adapters and barcodes for library preparation. When viewing the different tissue types after digestion on a 1% agarose gel (Figure 3.2), there are visibly different smear patterns for the different tissue types. The size selection step on the PippinPrep did account for a large quantity of sample loss as seen in Table 3.5, but still allowed high enough DNA yields for further processing and sequencing. The entire process from nuclei isolation to library preparation can be performed in approximately 3 days but is not ideal for high throughput. The manual grinding of tissue samples in liquid nitrogen with mortar and pestle is time consuming and labor intensive. An alternative method of grinding was attempted using ceramic beads and 1.5mL conical tubes in a Fast Prep grinder (Bio101), but proved unsuccessful due to liquid nitrogen shattering the tubes.

3.3.2 Mapping Reads and Comparisons

A total of 382 million reads were produced between the two Illumina lanes. Approximately 47 million unique scaffold positions were mapped to the reference genome when combining results from all tissue types. Figure 3.5 displays the comparison of all similar positions mapped among the different tissue types. Shoot tips and needles appear to

share the largest number of sites in common, which would be expected as they are very similar tissue types. The genomic DNA control returned the smallest amount of mapped reads compared to all tissue types, but the proportions of which each tissue type had in common was not widely varied. This would suggest that the digestion of DNA in chromatin form from the different tissue types returned specific mapped sites. Table 3.10 displays the total number of mapped sites per tissue type along with the number of sites that were found only in that tissue. When comparing mapped sites from the nuclei digestion it is important to remember that just because the site was physically accessible to the restriction enzyme while in chromatin form, that means that the region is only transcriptionally poised and not necessarily transcriptionally active. Further experiments would be needed for specific sites found to determine their relevance to the phenotypic traits of the trees.

3.3.3 Steps for the Future

From the 13 samples sequenced, a total of 4.4 million unique scaffolds mapped to the reference genome, or almost 1/3 of the known scaffolds available. The goal of this experiment was to find SNPs in regions that could be potentially associated with desired phenotypic traits through the identification of DNase-hypersensitive sites. The SNPs discovered thus far may provide a good launch point for further investigation but the regions discovered did not provide enough information to justify the cost of either a SNP chip assay or customized baits.

SNP chip assays have been used in numerous studies, usually in well studies species such as humans or cattle. These highly studied species allow for SNP chips designs to be relatively easy and many commercial variations are readily available (Ha, Freytag, & Bickeboeller, 2014; Wilkinson et al., 2011). For a species such as loblolly pine for which no

commercial chips yet exist, the decision as to which SNPs are important enough to be placed on the chip is a challenge both economically and functionally. A big advantage of the SNP chips arrays is the small amount of missing data produced between samples. If the SNP exists in the sample and it is on the chip, then it will be detected. The challenge of collecting nuclease hypersensitivity data from the dozens or even hundreds of trees necessary to discover which SNPs to be used in an assay would be too time consuming and economically expensive using this current method of nuclei isolation.

An alternative to the SNP assays would be the design of custom capture baits from some of the regions identified here. Baits are a cost-effective method used in next generation sequencing that targets specified regions of the genome (Samorodnitsky et al., 2015). However, as with the SNP chip, the absence of any indication for which identified regions would provide the most information is to too great to justify the cost of customized baits.

The amount of important data that can be discovered through the identification of DNase-hypersensitive sites is too great to disregard completely in pine trees. New methods are constantly being discovered, such as the one developed in Buenrostro et al. (2013). The method is a simple two-step process shown to capture information on open chromatin sites equal to that from DNase chromatin digestion. It is not only fast and easy, but also uses very low genomic inputs (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013). The discovery of SNP variants to be used in association testing is not limited to DNase-hypersensitive sites. Many methods such as RNA profiling, and enzyme digestion are all plausible. With the advancements in sequencing technologies and methodologies growing so rapidly the possibilities are endless.

3.4 Tables

Table 3.1: List of volumes/weights used per tissue type during nuclei isolations

Tissue type	Phloem	Xylem	Shoot tips	Needles
Tissue used (g)	3	4	3	3
re-suspension digested (ul)	200	150	150	200
Supernatant recovered (ul)	155	120	130	165
Clean up elution (ul)	30	30	30	30

Table 3.2: List of volumes/weights used per tissue type during digestion of the isolated nuclei

Digestion	DNA	Phloem	Xylem	Shoot tips	Needles
MseI (ul)	1	1	1	1	1
Cutsmart buffer (10x)	2.6	23	17.8	17.8	23
H_2O (ul)	0	7	.2	.2	7
DNA (ul)	22.4	200	160	160	200
Total Reaction Volume	26	230	178	178	230

Table 3.3: DNA ligation parameters

Ligation	ul
Master mix	12
Digested DNA	25
Common Adaptor (10uM)	1.5
Barcode Adaptor (10uM)	1.5
Total volume	40

Table 3.4: Ligation master-mix parameters

Ligation Master mix	ul
ATP	4
T7 ligase	0.33
Buffer B4	4
Water	3.67
Total	12

Table 3.5: Concentration of ligated DNA before and after size selection

Tissue	Tree	Nano drop conc (ng/ul)	Post Size Selection (pg/ul)
Shoot Tip	1	5.6	190.32
	2	4.6	42.6
	3	8.1	73.63
Needle	1	5.2	71.54
	2	8.4	35.05
	3	8.4	80.61
Pholem	1	11.8	69.29
	2	14	22.79
	3	17.4	62.81
Xylem	1	41.8	105
	2	14.2	54.65
	3	89.4	124.85

Table 3.6: PCR volume parameter

PCR	ul
Template DNA	11
Primer	1
Index Primer	1
Master Mix	12
Total Volume	25

Table 3.7: PCR master-mix volume parameters

PCR Master Mix	ul
Q5 buffer (5x)	5
DNTP's	0.5
Q5 polymerase	0.5
Water	6
Total Volume	12

Table 3.8: Comparison of unique mapped regions per tissue type. Pairwise comparison of the number of unique mapped regions in common with other tissue types across all trees. This information is also displayed in graph form in Figure 3.5.

Tissues	Xylem	Phloem	Needles	Genomic
Shoot Tips	12986606	12344281	15876862	5797902
Xylem	20612975	10622971	13108677	4832566
Phloem		19447808	12629333	4481694
Needles			26830868	5724187

Table 3.9: Comparison of number of mapped sites in each tissue type. This information is also displayed in graph form in figure 3.6

	Tree 1	Tree 2	Tree 3
phloem	15679431	13472374	14485492
Xylem	16425775	15529752	13798032
Needles	19723257	17214539	19660357
Shoot Tips	19866733	17780837	18662421

Table 3.10: Table comparison of mapped sites sums. Totals for each tissue type and mapped sites only occurring in each tissue type.

	Total Mapped Sites	Unique Mapped Sites
Shoot Tips	26914214	6199802
Xylem	20612975	3436016
Phloem	19447808	3051224
Needles	26830868	5913431
Genomic	9306688	1493078

3.5 Figures

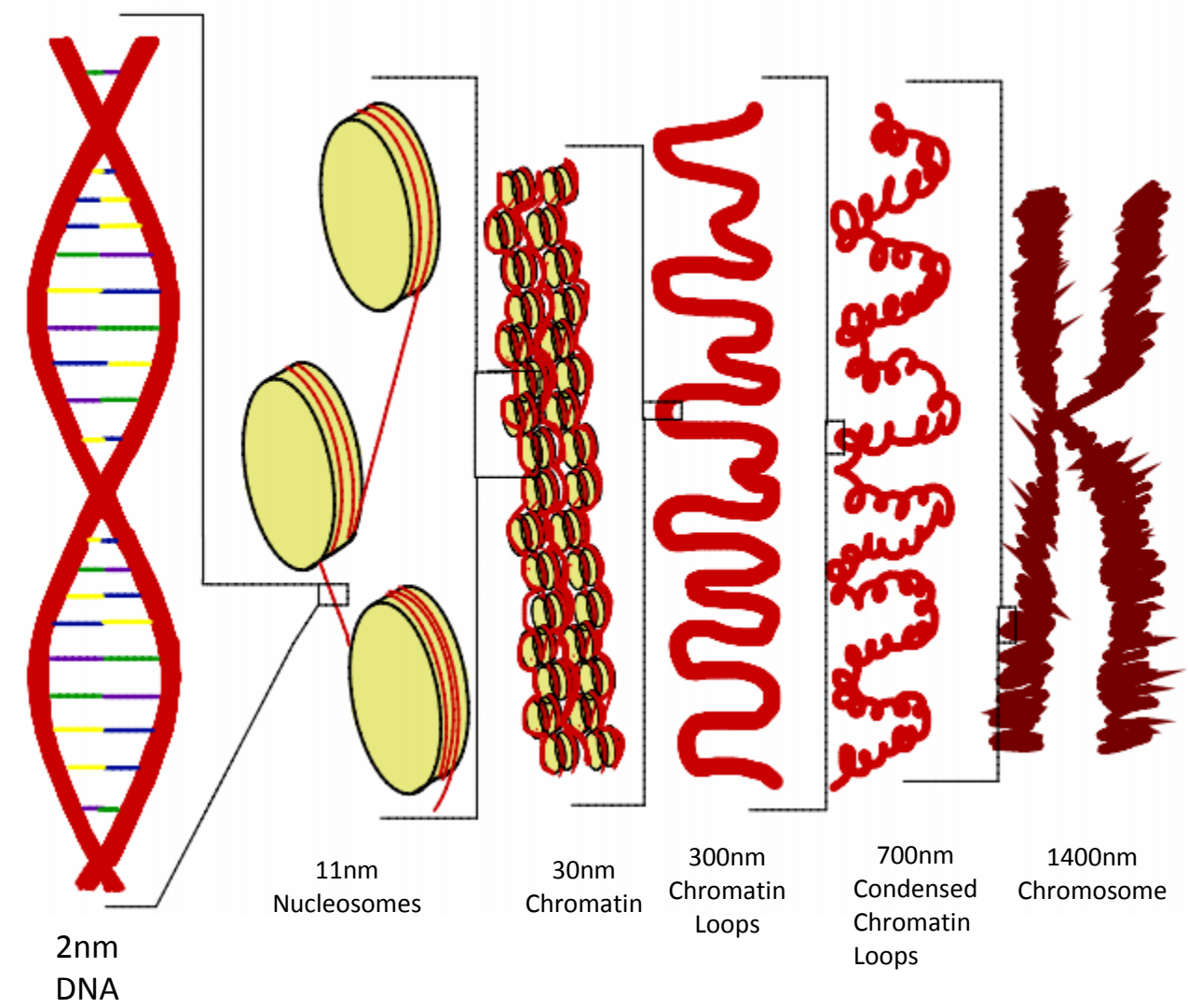

Figure 3.1: 6-Order Folding Structure of DNA.

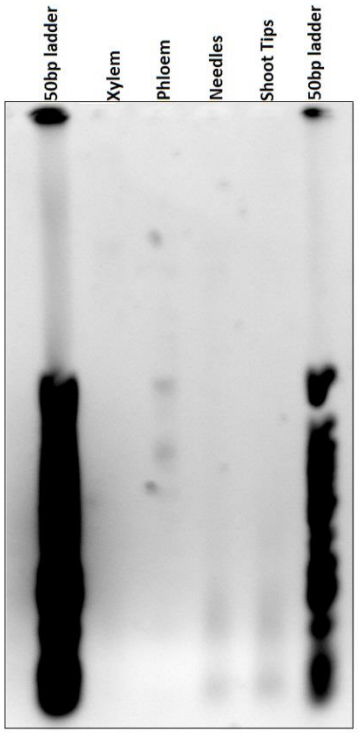

Figure 3.2: Digestion of all Tissue Types. 1% Agarose gel image of the nuclei digestion of the four plant tissue types.

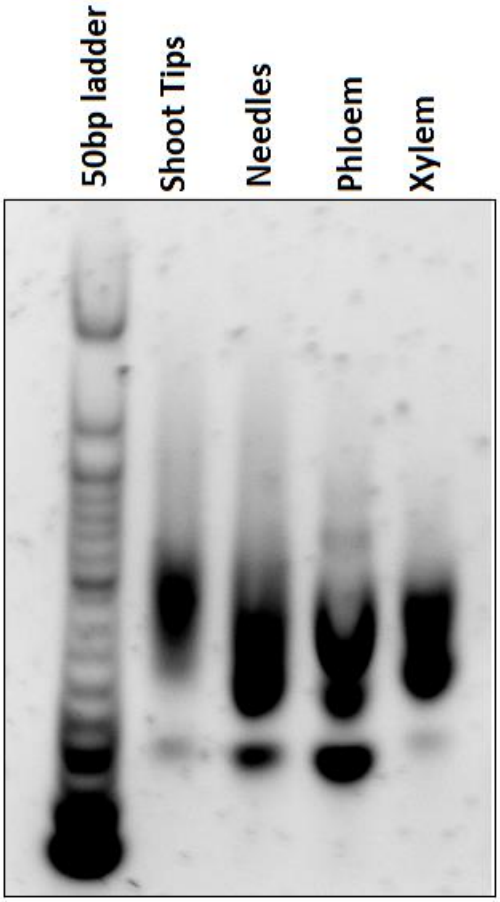

Figure 3.3: PCR Amplification of Four Tissue Types. 1% Agarose Gel image showing PCR amplification product of the 4 different tissue types.

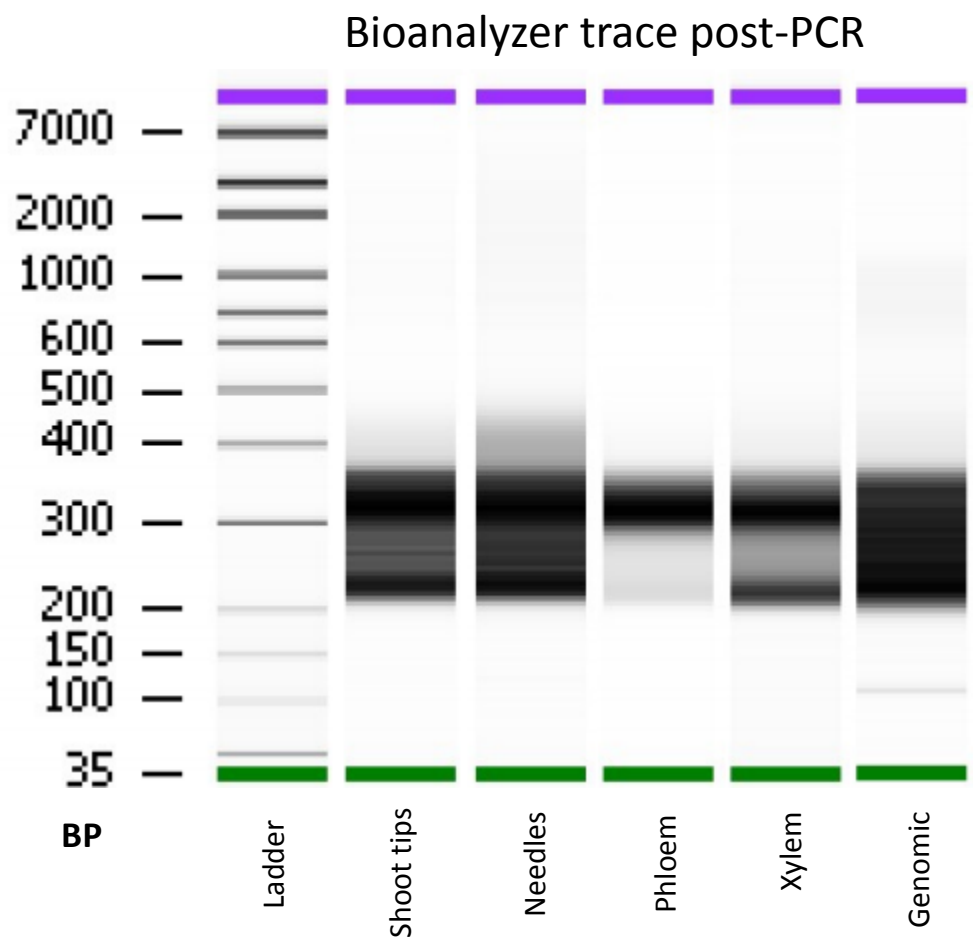

Figure 3.4: Bioanalyzer Results. Displaying the 4 tissue samples plus genomic DNA after PCR and size selection of 370-190 base pairs.

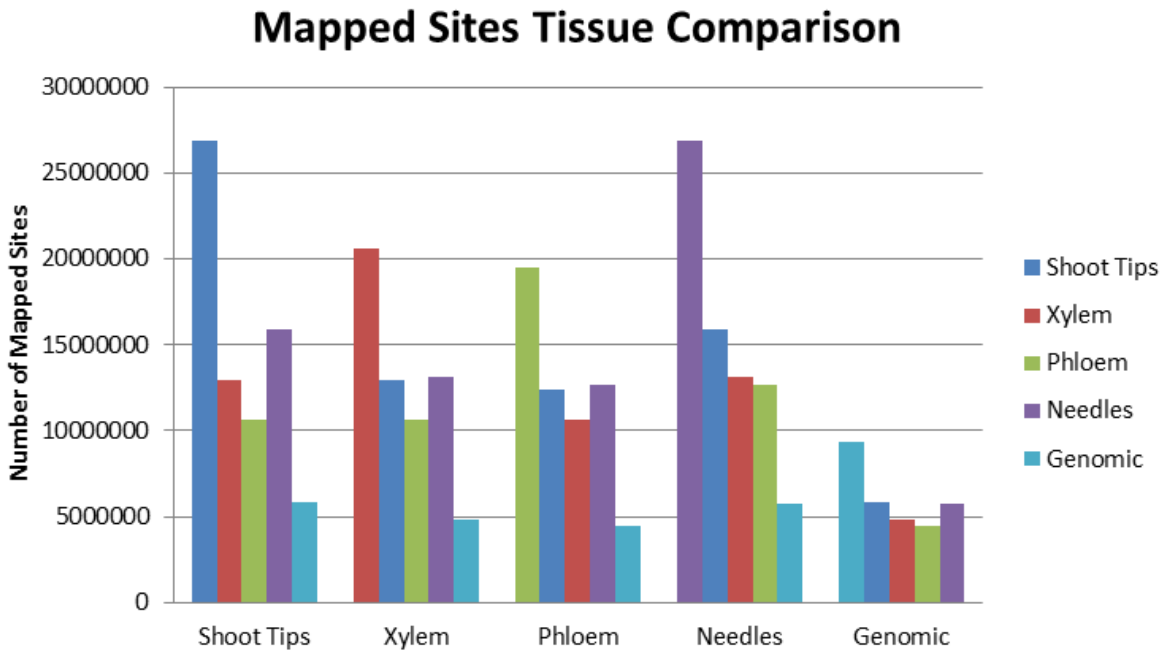

Figure 3.5: Mapped Sites Tissue Comparison. Comparison of mapped sites to the reference genome between tissue types present. Bars that are the same tissue as the tissue category represent all the unique mapped sites for that tissue.

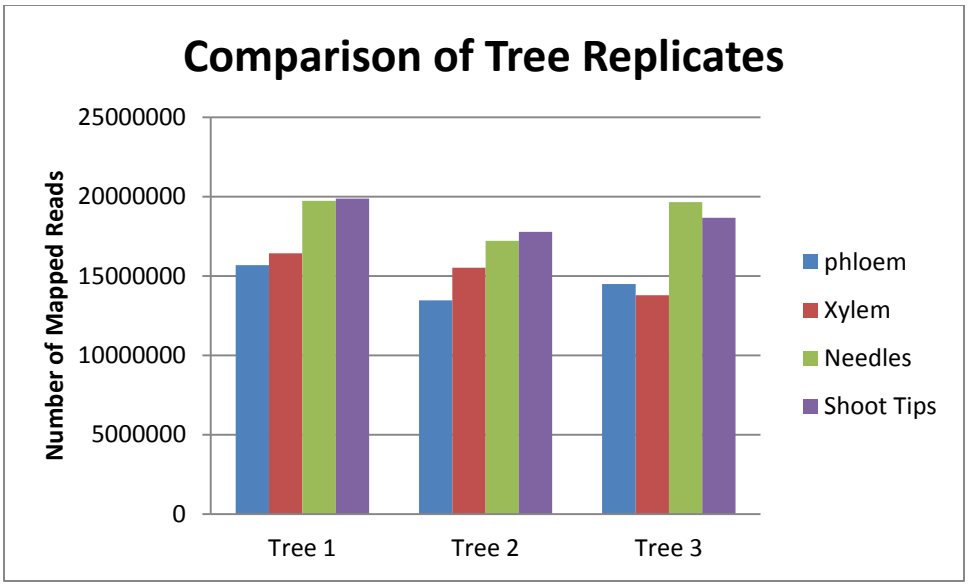

Figure 3.6: Comparison of Tree Replicates. Comparison of number of mapped sites of each tissue by tree replicate.

3.6 Literature Cited

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. http://doi.org/10.1038/nmeth.2688

Dittmar, K. A., Goodenbour, J. M., & Pan, T. (2006). Tissue-Specific Differences in Human Transfer RNA Expression. *PLoS Genetics*, *2*(12), e221. http://doi.org/10.1371/journal.pgen.0020221

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, *6*(5), e19379. http://doi.org/10.1371/journal.pone.0019379

Ha, N.-T., Freytag, S., & Bickeboeller, H. (2014). Coverage and efficiency in current SNP chips. *European Journal of Human Genetics*, *22*(9), 1124–1130. http://doi.org/10.1038/ejhg.2013.304

Hein, I., Williamson, S., Russell, J., & Wayne, P. (2005). Isolation of high molecular weight DNA suitable for BAC library construction from woody perennial soft-fruit species. *BioTechniques*, *38*, 69–71.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367. http://doi.org/10.1073/pnas.0903103106

Mach, J. (2011). Large-Scale RNA Sequencing to Identify Maize Genes with Parent-of-Origin Expression Effects. *The Plant Cell*, *23*(12), 4166–4166. http://doi.org/10.1105/tpc.111.231210

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, *337*(6099), 1190–1195. http://doi.org/10.1126/science.1222794

McKeand, S., Mullin, T., Byram, T., & White, T. (2003). Deployment of Genetically Improved Loblolly and Slash Pines in the South. *Journal of Forestry*, *101*(3), 32.

Paterson, A. H., Brubaker, C. L., & Wendel, J. F. (1993). A rapid method for extraction of cotton (Gossypium spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Molecular Biology Reporter*, *11*(2), 122–127.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, *7*(5), e37135. http://doi.org/10.1371/journal.pone.0037135

Samorodnitsky, E., Datta, J., Jewell, B. M., Hagopian, R., Miya, J., Wing, M. R., … Roychowdhury, S. (2015). Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing. *The Journal of Molecular Diagnostics*, *17*(1), 64–75. http://doi.org/10.1016/j.jmoldx.2014.09.009

Sikorskaite, S., Rajamäki, M.-L., Baniulis, D., Stanys, V., & Valkonen, J. P. (2013). Protocol: Optimised methodology for isolation of nuclei from leaves of species in the

Solanaceae and Rosaceae families. *Plant Methods*, *9*(1), 31. http://doi.org/10.1186/1746-4811-9-31

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L.-S., Loopstra, C. A., Vasquez-Gross, H. A., … Neale, D. B. (2014). Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. *Genetics*, *196*(3), 891–909. http://doi.org/10.1534/genetics.113.159996

Weintraub, H., & Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. *Science*, (193), 848–56.

Wilkinson, S., Wiener, P., Archibald, A. L., Law, A., Schnabel, R. D., McKay, S. D., … Ogden, R. (2011). Evaluation of approaches for identifying population informative markers from high density SNP Chips. *BMC Genetics*, *12*(1), 45. http://doi.org/10.1186/1471-2156-12-45

Wunderlich, M., Groß-Hardt, R., & Schöffl, F. (2014). Heat shock factor HSFB2a involved in gametophyte development of Arabidopsis thaliana and its expression is controlled by a heat-inducible long non-coding antisense RNA. *Plant Molecular Biology*, *85*(6), 541–550. http://doi.org/10.1007/s11103-014-0202-0

APPENDIX

4.1 Chapter 2 Detailed Sequencing Analysis Code

4.1.1- Filtering and Trimming Reads for Quality

```
FILE=$1
# input $FILE contains list of sample names, each followed by the barcode length in
#that file
exec 0<$FILE
while read -a line
do
 firstbase=$((${line[1]}+1)) # double (()) required for math operations with variables

  cat p01_${line[0]}.fq | awk -v FIRST=${firstbase} -f trimsplit2.awk | gzip >
p01${line[0]}.fq #| fastq_quality_filter -q 20 -p 95 | gzip > p01${line[0]}.fq.gz
# this line sends each sequence file to awk to have 70 nt of read1 (starting with the
PstI remnant) fused to 70 nt of read2
# starting with the MspI remnant, then sends the output to the FASTX-toolkit
#quality_filter program, which removes all reads that
# The filtered output is then compressed using gzip.
done
```

4.1.2- Split Data from Four Lanes of Pooled Plates into Separate Directories and Files

```
# PoolB files Example
# First create output directories for reads from each plate
mkdir plate4 plate5 plate6 plate7
# Run several parallel processes to split read files based on index sequences into
#output directories, name files according to GBS_barcode.pl format
zcat ../Br1_FCC21M0ACXX_s_1_fastq.gz | tee >(grep -EA3
'#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
plate4/FCC21M0ACXX_1_1.gz) >(grep -EA3 '#TGACCA|#ACAGTG|#GCCAAT' |
sed -e '/^--$/d' | gzip > plate5/FCC21M0ACXX_1_1.gz) >(grep -EA3
'#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
plate6/FCC21M0ACXX_1_1.gz) >(grep -EA3 '#TAGCTT|#GGCTAC|#CTTGTA' |
sed -e '/^--$/d' | gzip > plate7/FCC21M0ACXX_1_1.gz) > /dev/null &
        zcat ../Br1_FCC21WAACXX_s_8_fastq.gz | tee >(grep -EA3
        '#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
        plate4/FCC21WAACXX_8_1.gz) >(grep -EA3
        '#TGACCA|#ACAGTG|#GCCAAT' | sed -e '/^--$/d' | gzip >
        plate5/FCC21WAACXX_8_1.gz) >(grep -EA3
        '#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
        plate6/FCC21WAACXX_8_1.gz) >(grep -EA3
        '#TAGCTT|#GGCTAC|#CTTGTA' | sed -e '/^--$/d' | gzip >
        plate7/FCC21WAACXX_8_1.gz) > /dev/null &
zcat ../Br1_FCD23ETACXX_s_7_fastq.gz | tee >(grep -EA3
'#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
```

```
plate4/FCD23ETACXX_7_1.gz) >(grep -EA3 '#TGACCA|#ACAGTG|#GCCAAT' |
sed -e '/^--$/d' | gzip > plate5/FCD23ETACXX_7_1.gz) >(grep -EA3
'#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
plate6/FCD23ETACXX_7_1.gz) >(grep -EA3 '#TAGCTT|#GGCTAC|#CTTGTA' |
sed -e '/^--$/d' | gzip > plate7/FCD23ETACXX_7_1.gz) > /dev/null &
zcat ../Br1_FCD23ETACXX_s_8_fastq.gz | tee >(grep -EA3
'#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
plate4/FCD23ETACXX_8_1.gz) >(grep -EA3 '#TGACCA|#ACAGTG|#GCCAAT' |
sed -e '/^--$/d' | gzip > plate5/FCD23ETACXX_8_1.gz) >(grep -EA3
'#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
plate6/FCD23ETACXX_8_1.gz) >(grep -EA3 '#TAGCTT|#GGCTAC|#CTTGTA' |
sed -e '/^--$/d' | gzip > plate7/FCD23ETACXX_8_1.gz) > /dev/null &
zcat ../Br2_FCC21M0ACXX_s_1_fastq.gz | tee >(grep -EA3
'#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
plate4/FCC21M0ACXX_1_2.gz) >(grep -EA3 '#TGACCA|#ACAGTG|#GCCAAT' |
sed -e '/^--$/d' | gzip > plate5/FCC21M0ACXX_1_2.gz) >(grep -EA3
'#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
plate6/FCC21M0ACXX_1_2.gz) >(grep -EA3 '#TAGCTT|#GGCTAC|#CTTGTA' |
sed -e '/^--$/d' | gzip > plate7/FCC21M0ACXX_1_2.gz) > /dev/null &
zcat ../Br2_FCC21WAACXX_s_8_fastq.gz | tee >(grep -EA3
'#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
plate4/FCC21WAACXX_8_2.gz) >(grep -EA3 '#TGACCA|#ACAGTG|#GCCAAT' |
sed -e '/^--$/d' | gzip > plate5/FCC21WAACXX_8_2.gz) >(grep -EA3
'#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
plate6/FCC21WAACXX_8_2.gz) >(grep -EA3 '#TAGCTT|#GGCTAC|#CTTGTA' |
sed -e '/^--$/d' | gzip > plate7/FCC21WAACXX_8_2.gz) > /dev/null &
zcat ../Br2_FCD23ETACXX_s_7_fastq.gz | tee >(grep -EA3
'#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
plate4/FCD23ETACXX_7_2.gz) >(grep -EA3 '#TGACCA|#ACAGTG|#GCCAAT' |
sed -e '/^--$/d' | gzip > plate5/FCD23ETACXX_7_2.gz) >(grep -EA3
'#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
plate6/FCD23ETACXX_7_2.gz) >(grep -EA3 '#TAGCTT|#GGCTAC|#CTTGTA' |
sed -e '/^--$/d' | gzip > plate7/FCD23ETACXX_7_2.gz) > /dev/null
zcat ../Br2_FCD23ETACXX_s_8_fastq.gz | tee >(grep -EA3
'#ATCACG|#CGATGT|#TTAGGC' | sed -e '/^--$/d' | gzip >
plate4/FCD23ETACXX_8_2.gz) >(grep -EA3 '#TGACCA|#ACAGTG|#GCCAAT' |
sed -e '/^--$/d' | gzip > plate5/FCD23ETACXX_8_2.gz) >(grep -EA3
'#CAGATC|#ACTTGA|#GATCAG' | sed -e '/^--$/d' | gzip >
plate6/FCD23ETACXX_8_2.gz) >(grep -EA3 '#TAGCTT|#GGCTAC|#CTTGTA' |
sed -e '/^--$/d' | gzip > plate7/FCD23ETACXX_8_2.gz) > /dev/null
```

4.1.3- Create pool.key file in GBS_barcode.pl Format

```
 # Create first set of 96 samples, plus header line:
awk 'OFS="\t" {print $4,$3,"p","FCC21M0ACXX_1_1.gz"}' ../poolA/poolAr1.key |
head -97 > poolB.key
```

```
 # Reformat header line
sed -i
's/Sample\tBarcode\tp\tFCC21M0ACXX_1_1.gz/SampleName\tcode\tpaired\tfile/'
poolB.key
 # Create three more sets of 96 samples from other 3 lanes, using the same 96
barcodes and well IDs example of one:
tail -n +2 poolB.key | awk 'OFS="\t" {print $1,$2,$3,"FCC21WAACXX_8_1.gz"}'
>> poolB.key
```

4.1.4-A script to run GBS_barcode.pl

```
#perl script to split paired-end reads based on variable-length barcodes into individual
sample read1 and read2 files.
# To run GBS_barcode.pl on output from SplitPoolB.sh shell script:
#  1. Change to each output directory in turn
#  2. Make symbolic link to poolB.key in each directory.
#  3. Run GBS_barcode.pl script to split reads from 4 lanes into 96 individual sample
files,
#       using PstI enzyme file (final length 70 nt/read) and poolB.key file.
cd /media/3tb2/PSSSS/poolB/plate4
ln -s ../poolB.key poolB.key
~/software/perlscripts/GBS_barcode.pl poolB.key
~/software/perlscripts/GBS_barcodePstI_70nt.txt &
cd /media/3tb2/PSSSS/poolB/plate5
ln -s ../poolB.key poolB.key
~/software/perlscripts/GBS_barcode.pl poolB.key
~/software/perlscripts/GBS_barcodePstI_70nt.txt &
cd /media/3tb2/PSSSS/poolB/plate6
ln -s ../poolB.key poolB.key
~/software/perlscripts/GBS_barcode.pl poolB.key
~/software/perlscripts/GBS_barcodePstI_70nt.txt &
cd /media/3tb2/PSSSS/poolB/plate7
ln -s ../poolB.key poolB.key
~/software/perlscripts/GBS_barcode.pl poolB.key
~/software/perlscripts/GBS_barcodePstI_70nt.txt
 # wait until those processes are complete before moving on to the next
# Set up bash loop to compress read1 and read2 files, then merge into a single 140-nt
sequence for input to STACKS
# Run the loop on read1 and rea2 files in each of the four plate directories for a pool
of samples  example showing one pool.
cd /media/3tb2/PSSSS/poolB/plate4
for N in {A..H}
 do
   for M in {01..12}
      do
         gzip ${N}${M}_1.fastq
         gzip ${N}${M}_2.fastq
```

```
        ~/software/perlscripts/mergeGZFastqFiles.pl ${N}${M}_1.fastq.gz
${N}${M}_2.fastq.gz p04${N}${M}.fq.gz
     done
 done

 cd /media/3tb2/PSSSS/poolB/plate5
 for N in {A..H}
 do
   for M in {01..12}
      do
        gzip ${N}${M}_1.fastq
        gzip ${N}${M}_2.fastq
        ~/software/perlscripts/mergeGZFastqFiles.pl ${N}${M}_1.fastq.gz
${N}${M}_2.fastq.gz p05${N}${M}.fq.gz
     done
 done

cd /media/3tb2/PSSSS/poolB/plate6
for N in {A..H}
 do
   for M in {01..12}
      do
        gzip ${N}${M}_1.fastq
        gzip ${N}${M}_2.fastq
        ~/software/perlscripts/mergeGZFastqFiles.pl ${N}${M}_1.fastq.gz
${N}${M}_2.fastq.gz p06${N}${M}.fq.gz
     done
 done

 cd /media/3tb2/PSSSS/poolB/plate7
 for N in {A..H}
  do
   for M in {01..12}
      do
        gzip ${N}${M}_1.fastq
        gzip ${N}${M}_2.fastq
        ~/software/perlscripts/mergeGZFastqFiles.pl ${N}${M}_1.fastq.gz
${N}${M}_2.fastq.gz p07${N}${M}.fq.gz
     done
 done
```

4.1.5 Command line for Read Count Information

```
for dir in {Ar1,Br1,Cr1,Dr1}; do for file in
/media/3tb2/PSSSS/${dir}*_fastqc/fastqc_data.txt;
do sed -n '4,7p;8q' ${file} > seq.counts.txt ; done; done
```

4.1.6-Example of UStacks for One Plate

```
for let in {A..H}; do for num in {01..12}; do ustacks -t gzfastq -f
plate03/p03${let}${num}.fq.gz -d -r -o
plate03/stacks -i ${count} -m 3 -M 4 -p 4 -H; count=$( expr $count + 1 ); done; done
```

4.1.7-cstacks command

```
list1=`ls -l poolA/plate02/stacks/*.alleles.tsv | tr -s ' ' | tail -96 | awk '{print $9}' | sed
's/.alleles.tsv/ -s/' | tr "\n" " " |sed 's/-s $//'`
cstacks -b 3 -s ${list1} -o fullset > fullset/cstacks.out 2>&1
echo "plate02 done" >> fullset/cstacks.out
for file in
{poolA/plate03/stacks/,poolA/plate13/stacks/,poolA/plate14/stacks/,poolB/plate4/stac
ks/,poolB/plate5/stacks/,poolB/plate6/stacks/,poolB/plate7/stacks/,poolC/plate08/stac
ks/,poolC/plate09/stacks/,poolC/plate10/stacks/,poolC/plate11/stacks/,poolD/plate1/st
acks/,poolD/plate15/stacks/,poolD/plate16/stacks/,poolD/plate17/stacks/}
 do
  list1=`ls -l ${file}*.alleles.tsv | tr -s ' ' | tail -96 | awk '{print $9}' | sed 's/.alleles.tsv/
-s/' | tr "\n" " " | sed 's/-s $//'`
  cstacks -b 3 -s ${list1} --catalog fullset/batch_3.catalog >> fullset/cstacks.out 2>&1
  echo "${file} done" >> fullset/cstacks.out
 done
```

4.1.8-sstacks command

```
for dir in
{poolA/plate02/stacks/,poolA/plate03/stacks/,poolA/plate13/stacks/,poolA/plate14/st
acks/,poolB/plate4/stacks/,poolB/plate5/stacks/,poolB/plate6/stacks/,poolB/plate7/sta
cks/,poolC/plate08/stacks/,poolC/plate09/stacks/,poolC/plate10/stacks/,poolC/plate11
/stacks/,poolD/plate1/stacks/,poolD/plate15/stacks/,poolD/plate16/stacks/,poolD/plate
17/stacks/}
 do
  array=( `ls -l ${dir}*.alleles.tsv | tr -s ' ' | sort -nrk5,5 | awk '$5>0{print $9}' | sed
's/.alleles.tsv//' | tr "\n" " " ` )
  for file in "${array[@]}"
   do
    sstacks -b 3 -c fullset/batch_3 -s ${file} -o /media/sg3/pssss/sstacsfullout/ -p 6
   done
 done
```

4.1.9-cstacks Done for Population Calculations

```
#Creating catalog with only top ten samples per plate
list1=`ls -l poolA/plate02/stacks/*.alleles.tsv | tr -s ' ' | sort -nrk5,5 | head -10 | awk
'{print $9}' | sed 's/.alleles.tsv/ -s/' | tr "\n" " " | sed 's/-s $//'`
  cstacks -b 2 -s ${list1} -o topten > topten/cstacks.out 2>&1
echo "plate02 done" >> topten/cstacks.out
for file in
{poolA/plate03/stacks/,poolA/plate13/stacks/,poolA/plate14/stacks/,poolB/plate4/stac
ks/,poolB/plate5/stacks/,poolB/plate6/stacks/,poolB/plate7/stacks/,poolC/plate8/stack
```

```
s/,poolC/plate9/stacks/,poolC/plate10/stacks/,poolC/plate11/stacks/,poolD/plate1/stac
ks/,poolD/plate15/stacks/,poolD/plate16/stacks/,poolD/plate17/stacks/}
  do
   list1=`ls -l ${file}*.alleles.tsv | tr -s ' ' | sort -nrk5,5 | head -10 | awk '{print $9}' | sed
's/.alleles.tsv/ -s/' | tr "\n" " " | sed 's/-s $//'`
   cstacks -b 2 -s ${list1} --catalog topten/batch_2.catalog >> topten/cstacks.out 2>&1
   echo "${file} done" >> topten/cstacks.out
  done
SStacks done for population using only the topten catalog samples per plate
for dir in
{poolA/plate02/stacks/,poolA/plate03/stacks/,poolA/plate13/stacks/,poolA/plate14/st
acks/,poolB/plate4/stacks/,poolB/plate5/stacks/,poolB/plate6/stacks/,poolB/plate7/sta
cks/,poolC/plate8/stacks/,poolC/plate9/stacks/,poolC/plate10/stacks/,poolC/plate11/st
acks/,poolD/plate1/stacks/,poolD/plate15/stacks/,poolD/plate16/stacks/,poolD/plate17
/stacks/}
  do
   array=( `ls -l ${dir}*.alleles.tsv | tr -s ' ' | sort -nrk5,5 | awk '$5>1000{print $9}' |
sed 's/.alleles.tsv//' | tr "\n" " " ` )
   for file in "${array[@]}"
    do
     sstacks -b 2 -c topten/batch_2 -s ${file} -o ${dir} -p 6
    done
  done
```

4.1.10-Populations Command

```
populations -b 2 -P stacks_output/ -M popmap -r .5 -m 3 -p 3 -a 0.02 -t 20
```

4.1.11-Finding SNP Counts for Each match.tsv File

```
for pool in {A..D}; do for dir in pool${pool}/plate*;
do for file in ${dir}/stacks/*.matches.tsv; do awk -v file="$file"
'$6!="consensus"{s+=length($6)}END{print s,file}'
${file} >> SNPcounts.txt; done; done;done
```

4.1.13-Alignment to Reference Using BWA and SAMtools

```
#Was ran in loop for all pools
for X in poolA poolB poolC poolD; do bwa075 mem -t 6 -R
"@RG\tID:${X}1\tSM:${X}1" /media/seagate2/ptaeda/v101.scaffolds
samples/${X}1.r1.fq.gz samples/${X}1.r2.fq.gz \
| samtools view -SuF2308 - | samtools sort -l 5 -@ 6 -m 2G - bamfiles/${X}1pe; done
```

4.1.14- Filtering Half-Present Samples Using Freebayes in .VCF Format

```
# Merge bam files for each plate across all pools so freebayes can call SNPs
# Need a header.sam file that lists all scaffolds in @SQ lines, and all read groups
# (ie sample IDs) in @RG lines.
# First recover header with scaffold @SQ lines: example for one sample of a pool:
cd /media/seagate2/pssss/poolA/
```

```
samtools view -H p02bamfiles/A01.bam > header.sam
# This header.sam file already contains @RG tag for  rest of 96 samples:
for col in {02..12}; do echo -e "@RG\tID:p02A${col}\tSM:p02A${col}" >>
header.sam ; done
for row in {B..H}; do for col in {01..12}; do echo -e
"@RG\tID:p02${row}${col}\tSM:p02${row}${col}" >> header.sam ; done; done
# Similarly, add @RG tags to header for plates all other plates in the pool, and then
#all other pools
# To simplify calling samtools merge, create a new directory bamlinks with symlinks
#to bamfiles of all 16 plates
mkdir bamlinks
cd bamlinks
# The pool bamfiles don't contain plateID in filenames, but do contain it in RG tags
#on alignment lines and in header
# Example to change it in one plate
for row in {A..H}; do for col in {01..12}; do ln -s
../poolA/p02bamfiles/${row}${col}.bam p02${row}${col}.bam ; done; done
# From /media/seagate2/pssss directory, run samtools merge to merge all 1536 bam
#files into one, then index merged file:
 samtools merge -h poolA/header.sam pssss_merged.bam bamlinks/*.bam
 samtools index pssss_merged.bam
# Run freebayes on the new bam file:
~/software/freebayes/bin/freebayes -C 3 -p 2 --haplotype-length 75 -f
/media/seagate2/ptaeda/bowtie/v101.scaffolds.fa -q 20 -w --populations
/media/3tb2/pSSSS/popmap/populations_stacks.txt --report-all-haplotype-alleles --
prob-contamination 10e-3 -b pssss_merged.bam -v pssss_merged.vcf
-C = min read depth per sample to call genotype
-p = ploidy of samples
-f = reference genome
-q = min base quality for SNP locus
-w = disable HWE calculation and filtering
-j = use mapping quality of alleles when estimating data likelihoods
-= = calculate genotype quality (GQ) as marginal probability of genotypes and
incorporate in sample field
# Filter pssss_merged.vcf to recover SNPs with total depth (across all samples) >
7500 and NS>750, using vcflib vcffilter program:
/home/ross/software/vcflib/bin/vcffilter -f "DP > 7500 & NS > 750"
pssss_merged.vcf > half.present.vcf
# Use vcflib vcf2tsv to convert half.present.vcf to long matrix format for
# import into snpStats R package
~/software/vcflib/bin/vcf2tsv -n "NA" -g half.present.vcf | gzip > half.present.long.gz
# Combine scaffold and position to create unique SNP id for use in importing into
snpStats:
BEGIN{OFS="\t"}
{I=$1"_"$2
 N=$47
```

```
split($52,b,"/")
b[1]=="0" ? X=$4 : X=$5
b[2]=="0" ? Y=$4 : Y=$5
split($50,a,",")
if(a[1]==0 && a[2] >= a[3])
{
 C = -10*a[2]
}
else if(a[1]==0 && a[3] >= a[2])
{
 C = -10*a[3]
}
else if(a[3]==0 && a[2] >= a[1])
{
 C= -10*a[2]
}
else if(a[3]==0 && a[1] >= a[2])
{
 C= -10*a[1]
}
else if(a[2]==0 && a[1] >= a[3])
{
 C= -10*a[1]
}
else if(a[2]==0 && a[3] >= a[1])
{
 C= -10*a[3]
}
printf "%s\t%s\t%s\t%s\t%.2f\n",I,N,X,Y,C/100
}
```

4.1.15-Association Testing Using RVTest

```
rvtest --inVcf half.present.vcf --pheno all.pssss.ped --kinship all.pssss.kinship --out
all.pssss.kintest.out --single famScore,famLRT,famGrammarGamma
#all analysis of results was done using R environment
```

4.2 Chapter 3 Detailed Sequencing Analysis Code

4.2.1- Trim Adapters, Filter on Quality, and Split Reads into Barcode Sets Using Flexbar

```
# barcode trim mode will be left-tail
# adapter trim mode will be right tail
# Run two separate processes, one for each lane of data, looping across all subsets
for N in {01..47}; do flexbar -r NucleiDNA_NoIndex_L005_R1_0${N}.fastq.gz -p
NucleiDNA_NoIndex_L005_R2_0${N}.fastq.gz -f i1.8 -a adapters.fa -ao 16 -ae
RIGHT_TAIL -at 2 \
   -b barcodes.fa -be LEFT_TAIL -u 1 -t split/lane5_${N} -q 20 -m 75 -z GZ -s -n 4
&> lane5.log; done
```

```
# Merge files for each barcode across lanes and subsets to create single pair of files
#per sample.
# Run as 13 separate processes, in groups of four tissues per tree:
zcat split/lane5_[0-4][0-9]_barcode_A03_1.fastq.gz split/lane6_[0-5][0-
9]_barcode_A03_1.fastq.gz | gzip > samples/st1.r1.fq.gz &
zcat split/lane5_[0-4][0-9]_barcode_A03_2.fastq.gz split/lane6_[0-5][0-
9]_barcode_A03_2.fastq.gz | gzip > samples/st1.r2.fq.gz &
zcat split/lane5_[0-4][0-9]_barcode_A03_1_single.fastq.gz split/lane6_[0-5][0-
9]_barcode_A03_1_single.fastq.gz \
split/lane5_[0-4][0-9]_barcode_A03_2_single.fastq.gz split/lane6_[0-5][0-
9]_barcode_A03_2_single.fastq.gz  | gzip > samples/st1.single.fq.gz
```

4.2.2-Example of BWA Alignment for One Tissue Type

```
# Code was ran for all tissue types plus genomic controll
for X in ne ph xy; do bwa075 mem -t 6 -R "@RG\tID:${X}1\tSM:${X}1"
/media/seagate2/ptaeda/v101.scaffolds
samples/${X}1.r1.fq.gz samples/${X}1.r2.fq.gz \
| samtools view -SuF2308 - | samtools sort -l 5 -@ 6 -m 2G - bamfiles/${X}1pe; done
```

4.2.3- Use BEDtools to Convert BAM Files to BED Format for Multi-Intersect Function

```
# can be ran in parallel for each file, example of one:
/home/ross/software/bedtools/bin/bamToBed -i ne2pe.bam > ne2pe.bed
```

4.2.4-Making Multiple Comparisons

```
# Aaron Quinlan has a tutorial describing multi-intersect tool :
#https://github.com/arq5x/bedtools-protocols/blob/master/bedtools.md#bp6--
#measuring-dataset-similarity
# Use multi-intersect function to make table of comparisons of all 13 samples
~/software/bedtools/bin/bedtools multiinter -header -i ge0sort.bed ne1sort.bed
ne2sort.bed ne3sort.bed \
 ph1sort.bed ph2sort.bed ph3sort.bed xy1sort.bed xy2sort.bed xy3sort.bed st1sort.bed
st2sort.bed st3sort.bed \
-names ge0 ne1 ne2 ne3 ph1 ph2 ph3 xy1 xy2 xy3 st1 st2 st3 | gzip > all13.table.gz
#filter for alignments >= 50 nt,
# Sort by decreasing value of the number of samples (out of 13) in which each
#segment was detected.
zcat all13.table.gz | awk '$3-$2>=50{print $0}' | sort -T ./ -k4,4nr > all13.gt50sort.txt
# All summaries were done using standard R environment.
```