GENETIC APPROACHES TO IMPROVE ARCHITECTURAL TRAITS OF PIEDMONT AZALEA (*RHODODENDRON CANESCENS*) FOR LANDSCAPING

by

LAV KUMAR YADAV

(Under the Direction of H. Dayton Wilde)

ABSTRACT

Rhododendron canescens (Michaux) Sweet (Piedmont azalea) is a native azalea with potential in landscaping because of its adaptability, early flowering, and lace bug resistance. But use of *R. canescens* in urban settings is limited, however, because of its sparse branching and other architectural traits. The goal of this project was to determine whether genetic variation can be identified for the development of a more compact phenotype. Genotyping-by-sequencing was used to examine the genetic diversity and population structure of a *R. canescens* collection from across Georgia. Single nucleotide polymorphisms (SNPs) were identified. The SNP data supported the presence of three populations. Statistical results indicated that there was low genetic differentiation between the populations, but relatively high genetic diversity within populations. Next, the transcriptome of vegetative and reproductive tissues was sequenced to identify orthologs of genes known to control height and branching in multiple plant species. Long-read sequencing using PacBio Iso-Seq methods generated 24,244 full-length isoform sequences, of which 16,825 were annotated. We successfully identified orthologs of thirteen genes regulating plant architecture through regulatory factors and

phytohormone biosynthesis and signaling. Sequence data for these genes enabled RNA probes to be designed to capture the coding sequences from 216 *R. canescens* genotypes including one dwarf genotype. Variation in these genes among individuals was identified using capture sequencing. The structural and functional effect of these variants in protein function were predicted. Of the 69 variants, 16 SNPs were predicted to lead to deleterious missense mutations. These genetic variations could potentially be used to breed *R. canescens* with shorter stature or increased branching. The data analyzed using Discriminant Analysis of Principle Component (DAPC) helped us group azalea genotypes based on the variants. We found that for genes *MAX2, PHOR1* and *FLC*, the dwarf genotype was significantly different from the other genotypes. Our findings further support that the suspected dwarf genotype has unique mutations in the plant architecture related genes. In addition, transcriptome analysis identified 2,871 long non-coding RNA and 13,116 simple sequence repeats. The genetic resources developed in this study have applications in molecular breeding, comparative genomics, and evolutionary studies.

Keywords: Genotyping-by-sequencing, Single nucleotide polymorphisms, isoform, transcriptome, gibberellins, dwarf genotype.

GENETIC APPROACHES TO IMPROVE ARCHITECTURAL TRAITS OF PIEDMONT AZALEA (*RHODODENDRON CANESCENS*) FOR LANDSCAPING

by

LAV KUMAR YADAV

BS, Institute of Agriculture and Animal Sciences, 2013

MS, West Virginia State University, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

GENETIC APPROACHES TO IMPROVE ARCHITECTURAL TRAITS OF PIEDMONT AZALEA (*RHODODENDRON CANESCENS*) FOR LANDSCAPING

by

LAV KUMAR YADAV

Major Professor:	H. Dayton Wilde
Committee:	Scott A. Merkle
	Donglin Zhang
	Carol Robacker
	Ali Missaoui

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2021

DEDICATION

I would like to dedicate this thesis to my loving wife Krittika Tonapi, my parents Ram Kripal Yadav and Babita Singh, my brother Kush Kumar Yadav and my loving grandparents.

ACKNOWLEDGMENTS

It has been a wonderful journey that would not have been possible without many helping hands. I am very fortunate to have come across many amazing people and get my Ph.D. in Dr. Wilde's lab.

I would like to thank my major advisor, Dr. Dayton Wilde for his continued support and mentorship. It has been a very pleasing experience working with him. I am very grateful for his continuous guidance and encouragement during the course of my PhD. Thank you Dr. Wilde, for giving me room to grow and realize my potential here at UGA.

I would also like to thank my committee members Dr. Scott Merkle, Dr. Donglin Zhang, Dr. Carol Robacker and Dr. Ali Missaoui for their mentorship and continued guidance throughout my PhD. Thank you all for serving on my committee.

I would also like to thank Dr. Edward McAssey for his patience and guidance with bioinformatics work.

Thank you Dr. Hanieh Hadizadeh for helping with initial collection of Azalea samples. Dr. Heather Gladfelter (coolest lab mate ever), I loved working with you and collaborating with you on the Franklinia project. I will always be thankful to you for all your lab tips and for letting me borrow your solutions all the time. I will always cherish our time together be it our 'Snelling lunch' or 'drinks break'. Thank you for being a friend.

I would also like to thank John Doyle and Dr. Malladi for letting me borrow lab materials during the course of my Ph.D.

Thanks to the Department of Horticulture, Mary Jane, Vickie, and other administrative staff for always assisting me with a smile in administrative work.

Thank you to my wife Krittika Tonapi for being my support system during the ups and downs of my Ph.D. journey.

Thank you to all the people my path cross with during my PhD journey.

Thank you God for all the blessings!!!

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Piedmont azalea

Rhododendron canescens (Michaux) Sweet (Piedmont azalea) is native to the eastern United States from Texas to North Carolina. It flowers in spring with subsequent vegetative growth in the summer. Exotic azaleas are popular in landscape settings, but native azaleas are less common. Azaleas are believed to have migrated from Asia through the Bering Strait land bridge about 5-65 million years ago (Scharff, 1912). Because of this migration, very little variation is seen in chromosome number and morphology in Asian and North American populations. Therefore, any genetic variation present in the native azalea population must be due to geographical isolation and habitat preference (Kron et al., 1993).

Kron et al. (1993) used chloroplast DNA markers to distinguish *R. canescens* from the sympatric species *R. flammeum*. Population studies have also been carried out in azalea using RAPD (Random Amplification of Polymorphic DNA) and AFLP (Amplified Fragment Length Polymorphism) markers. These studies looked at the relatedness between *R. canescens* and other members of the subgenus *Pentathera* and found that interspecific hybridization and introgression exist between these species (Chappell et al., 2008). These studies helped us understand the evolutionary pedigree of native azalea. Introgression is the mixture of the genomes of two populations that phenotypically might look different but genotypically might be similar. If that is the case in *R. canescens*, it could complicate population studies based on DNA as we might get admixtures. This phenomenon would lead to significant bias when estimating inter- and

intraspecific diversity. GBS analysis of *R. canescens* will help us better understand how introgression has affected the population structure and diversity of *R. canescens*. Previously, the level of genetic diversity has been studied only interspecifically in azalea (Chappell et al., 2008; Scheiber et al., 2000). In this study I will use GBS to uncover the genetic diversity and population structure of naturally occurring native *R. canescens* in Georgia.

R. canescens is a hardy plant. It is one of the first to bloom in spring. *R. canescens* along with *R. flammeum* and *R. austrinum* blooms at least two weeks earlier than other azaleas and has very vibrant colors. This early blooming can be an advantage for pollinators early in the season. *R. canescens* also has resistance to lace bug, one of the major pests faced by azaleas (Braman et al., 1992). All of these characteristics make *R. canescens* a very good landscape and garden plant. However, *R. canescenss* are not widely used in gardening and landscape, in part due to their size and shape. *R. canescens* has indeterminate growth and a lanky appearance (Figure 1-1). Consumers desire more compact plants for their gardens as they are easier to manage and take less space to grow, but *R. canescens* lacks this characteristic. Here, we aim to understand the architectural complexity of the *R. canescens* through population analysis and exome capture through both horticultural and genetic approaches. Dwarf *R. canescens* was found by chance in the wild and this study aims to develop suitable vegetative propagation techniques to multiply these plants successfully for field trials. Natural variation in plant architecture exists in the wild population of *R. canescens*. I will also explore the plant architecture at the genetic level. I will try to find natural variation present in the wild that can be exploited to give *R. canescens* a determinate growth and bushy appearance. To exploit this natural variation, we first need to understand the genetic diversity and population structure. Population structure analysis will give us probable areas where we can collect samples to study gene diversity. Once equipped with this

information, we can analyze the genetic diversity present in the genes regulating plant architecture and use this knowledge to develop desirable *R. canescens*. Hence this study will lay the foundation on the basis of which other projects can be designed to develop *R. canescens* cultivars that are better suited for landscapes and gardens.

Figure 1-1: Typical architecture of *R. canescens*

CHAPTER 2

LITERATURE REVIEW

Genetic diversity and genotyping

Genetic diversity is one of the major tools for plant breeders and evolutionary biologists to help understand plant growth and development. Identification of genetic diversity heavily depends on genetic markers such as SSRs (Simple Sequence Repeats) and SNPs (Single Nucleotide Polymorphisms)(Kumar et al., 2012) . Genotyping-by-sequencing (GBS)(Bräutigam & Gowik, 2010; Metzker, 2010) is a cost-effective method used to identify a large number of SNP markers in the absence of a reference genome. This method was used in Piedmont azalea. GBS includes the following major steps: DNA extraction and digestion with one or two restriction enzymes, adapter ligation, PCR amplification, fragment size selection, library pooling, sequencing, data processing and SNP calling. GBS requires no prior sequencing information and is able to provide information about SNP discovery and genetic diversity following downstream analysis (Fu et al., 2014). In spite of these advantages, GBS studies are not devoid of limitations. GBS may have a large amount of missing data due to low coverage of the sequencing.

GBS provides a large amount of sequence data that needs further processing. Several bioinformatics tools are available to process this data. Depending on the resources available, the proper method needs to be selected. Some of the bioinformatics approaches available are: TASSEL (Glaubitz et al., 2014), UNEAK pipeline (Lu et al., 2013), STACKS (Catchen et al., 2013), GBS-SNP-CROP (Melo et al., 2016), etc. Some of the pipelines require a reference

genome. Although these pipelines are different, the final output of all the approaches tends to be similar. STRUCTURE analysis (Falush et al., 2003, 2007; Hubisz et al., 2009; Pritchard et al., 2000), principal component analysis (PCA), and phylogenetic trees are used as the common method to analyze GBS data. Genetic diversity and F statistics are calculated to better understand the genetic diversity and population structure of the plant under study. GBS is an easy method to extract high throughput information from any plant species for SNP discovery and genetic diversity analysis. GBS has allowed for the study of non-conventional crops and provides us with a plethora of data using restriction-based sequencing. This method does not bias any plant species depending upon the availability of reference genomes and helps us understand complex plant genomes. This technique serves as a major tool for marker assisted selection and GWAS (Genome Wide Association Studies), which are important techniques of crop development in the 21st century.

Genetic Control of Plant Architecture

Genes controlling two major components of architecture, height and branching was studied. Genes for regulatory factors and phytohormone biosynthesis and signaling play a major role in shaping plants. For example, an important gene regulating plant height is *TFL1* (*TERMINAL FLOWER 1*). *TFL1* inhibits the transition from vegetative to reproductive growth and *TFL1* mutants have determinate growth (Shannon & Meeks-Wagner, 1991). *TFL1* regulates a number of genes like *LEAFY (LFY*) and *APETALA1 (AP1)* (Ratcliffe et al., 1999). These genes initiate flowering and *TFL1* negatively regulates their expression, hence promoting vegetative growth. Plant architecture and reproductive phase are controlled by two genes, *TFL1* and *FLOWERING LOCUS T* (*FT*) (Wickland & Hanzawa, 2015). These two genes control the

growth and differentiation of plants across species. Both genes are similar in sequence (60% similarity) and have homology to phosphatidylethanolamine-binding proteins (PEBPs) (Karlgren et al., 2011). *FT* favors the shift of the plant from the vegetative phase to the reproductive phase whereas *TFL1* hinders this shift. The *SELF PRUNING (SP*) gene in tomato is a homolog of *TFL1* and regulates stem growth and development (Pnueli et al., 1998). In plants like rose and strawberry, the recessive allele of the *TFL1* homolog (*RoKSN* and *FvKSN*, respectively) when expressed, gives a short and determinate look to the plants (Iwata et al., 2012). This appearance is more appealing to consumers. In soybean, the null allele of the *TFL1* homolog, *Dt1*, promotes determinate growth and early flowering (Liu et al., 2010). This shows that *TFL1* is a major gene regulating plant architecture.

The gibberellin (GA) biosynthesis gene *GA20 oxidase* is of prime importance to plants as it regulates the level of GA in plants. *GA20ox* regulates activity like stem/petiole elongation, a more vertical growth habit, and flowering in plants (Rieu et al., 2008). In Arabidopsis there are five members in *GA20ox* family. Reverse genetic analysis of these genes showed that they express during vegetative and early reproductive development (Rieu et al., 2008). *GA20ox1* and *GA20ox2* are the genes that are most highly expressed during the vegetative state among the five in plant tissues (Phillips et al., 1995). Over expression studies of *GA20ox* genes have shown a decrease in the production of GA in plant tissue (Hisamatsu et al., 2005). *GA20ox1* is expressed in the stem. Mutant studies have shown reduced plant height in the Arabidopsis plant (Rieu et al., 2008). Other *GA20ox* gene family members are expressed in flowers, seeds, siliques etc. Overall, the *GA20ox* gene family members regulate the GA concentration in the tissues. They regulate a variety of plant growth and development activities like internode elongation, root growth and

leaf expansion, etc. (Rieu et al., 2008). This gene family is crucial in regulating plant growth and development which ultimately determines the architecture of any plant.

Deleterious alleles of Gibberellin insensitive dwarf 1 (*GID1c*) gene cause dwarfism in plants. Due to the two SNPs present in peach coding sequence, it produces a faulty gibberellic acid (Cantín et al., 2018b) that results in stunting of the plant. *PHOR1* is a transcription factor that regulates the growth of the plant. Knockout of *PHOR1* gives a semi-dwarf appearance to tomato plants and its overexpression shows prolific growth (Busov et al., 2008). Hence, *PHOR1* and *GID1c* will be ideal candidates to study for variation through capture sequencing.

Brassinosteroids (*BR*) are natural plant growth hormones that are very similar to animal steroids. They were first discovered in *Brassica napus* and are found across plants from monocots to dicots (Wang & Li, 2008). BRI1 has a role in internode elongation. It helps in the promotion of cell expansion and elongation. It also has a significant role in BR signaling. A membrane-localized LRR receptor-like kinase (RLK) is regulated by *BR1* which plays a key role in BR signaling (Clouse, 2002). When *BR1* was knocked out, it resulted in dwarfism in rice (Fujioka & Yokota, 2003). Some evidence also suggests that BR has a role in chilling and drought resistance (Clouse & Sasse, 1998). Hence, it is an ideal candidate to study architecture of azalea.

Branching is defined as the process of producing auxiliary shoots (Falush et al., 2007). Depending on the plant species, branching can be a desirable or undesirable trait. For example, branching is desired in rice as it produces a higher number of tillers, but in sugarcane, single stalks are desired to improve juice quality. Plant architecture is dependent upon several factors. In addition to genetic make-up, sunlight and nutrient levels may influence the architecture of the

plant (Falush et al., 2007). Due to advancements in modern genomics and gene mapping technologies, understanding plant architecture at the genetic level has become feasible.

BRANCHED1 (BRC1), a TCP transcription factor, is a negative regulator of axillary bud development in plants (Aguilar-Martínez et al., 2007). This transcription factor is closely related to teosinte branched1 (*tb1*) from maize (*Zea mays*) (Aguilar-Martínez et al., 2007). When *BRC1* is upregulated, the axillary bud remains dormant, preventing branching. It works similarly to *tb1* which promotes single stalk in maize (Doebley et al., 1997). When *BRC1* is down-regulated, axillary bud development can be seen. BRC1 plays a crucial role in regulating apical dominance in plants (Aguilar-Martínez et al., 2007). They are expressed in the apical meristems where they promote bud arrest (Kosugi & Ohashi, 1997, 2002). Mutants of these genes have shown enhanced shoot branching (Doebley et al., 1997; Takeda et al., 2003). *BRC1* is also transcriptionally regulated by environmental factors such as planting density. In high planting density, *BRC1* is upregulated, promoting apical dominance (Casal et al., 1986).

Plant architecture depends upon a number of factors. Environmental factors, nutritional availability and physiological response all act to shape plant architecture. All these factors together create a hormonal effect that regulates developmental response in plant shoots. Some of the major hormones playing a role in this phenomenon, such as cytokinins and strigolactone (SL), are produced in roots. SL inhibits shoot branching and controls the shoot via roots (Teichmann & Muhr, 2015). SL in the plant is regulated by *MAX* genes. These genes are expressed in roots and have been fully characterized in Arabadopsis (Booker et al., 2004). Mutagenized populations in Arabidopsis for *MAX1, MAX2* and *MAX3* have shown inhibited branching (Sorefan et al., 2003). MAX1 gene activates P450 cytochrome and MAX2 regulates F-box protein to synthesize SL (Booker et al., 2004; Sorefan et al., 2003). *MAX3* produces CCD7

carotenoid cleavage dioxygenase to complete the cycle in the production of SL (Booker et al., 2004). Rootstocks with mutations in these genes have been shown to lack SL in plants like poplar and willow (Ward et al., 2013). The orthologs of *MAX* genes are also present in a wide variety of plant species from rice to poplar, which makes it an ideal candidate to study in azalea. There is evidence that these genes play major roles in governing the architecture of many plant species.

Identification of Candidate Mutations

Discovering gene variants at a large scale and designing breeding pipelines has been a major challenge in crop improvement. Methods like TILLING (Henikoff et al., 2004) and EcoTilling (Gilchrist & Haughn, 2005) have been used in the past to detect these rare variants in large populations (Marroni et al., 2011; Raja et al., 2017). This technology, though cheap and useful in many plant species, comes with several drawbacks. Previous knowledge of DNA sequence is a must while employing these techniques for variant detection. Also, the efficiency of these techniques decreases when variation is high (Barkley & Wang, 2008), which is mostly the case when dealing with woody ornamentals. Next-generation sequencing (NGS) has given researchers access to large amounts of genomic data that can be used to screen a variety of populations and detect rare variants. There are both *de novo* and reference-based methods to detect these variants with great accuracy (Druley et al., 2009; Out et al., 2009). *R. canescens* does not have a reference genome; hence, a *de novo* based approach has to be employed to identify genetic variations in genes regulating plant architecture.

Transcriptome analysis-based next-generation sequencing (NGS) technology is a powerful and economical way to obtain genetic information on a large scale and has been widely

used to uncover genes involved in various functions. Although NGS technology like RNA-seq (Wang et al., 2009) is widely used for this approach, short sequence reads generated by this method create assembly and annotation problems (Sharmaö et al., 2018). Pacbio Isoform sequencing (Iso-seq) solves this problem and yields longer reads, thus providing more full-length transcripts and direct evidence of the structural variation of isoforms (Rhoads & Au, 2015). It is a proven technology which helps create transcriptome databases and has been successfully used in plant species, such as *Zea mays* (Wang et al., 2016), *Sorghum bicolor* (Abdel-Ghany et al., 2016), *Arabidopsis thaliana* (Zhu et al., 2017), and *Fragaria* × *ananassa* (Li et al., 2017) for various downstream analyses.

Detection of polymorphisms can aid in a variety of projects from genotyping to trait development. The recent breakthroughs in NGS have made possible the development of a sequence-based approach for variant detection. But the cost and time for whole-genome sequencing is still high, which has created room for hybridization-based approaches in unison with high-throughput sequencing that focus on targeted capture sequencing. Capture-Seq is a technology that allows us to target a specific part of the genome and study its sequence. Exome capture sequencing is designed to enrich exons of genes using hybridization probes present in the genome. Targeted capture sequencing is cost-effective and provides higher sequencing depth over whole-genome sequencing. This helps us detect the rare functional mutations present in the genic region that directly affect a trait in a plant. Plants with a high level of genetic diversity can be sequenced and have their variants called with high confidence. Capture sequencing is a cost-effective method to identify natural genetic variations in genes of species with large genomes or no reference genome (Grover et al., 2012; Müller et al., 2015). However, It is not clear what challenges it may pose in sequencing a small fraction of the genome and how intronic sequences

may affect this targeted sequencing (Gnirke et al., 2009; Zhou & Holliday, 2012). Exon capture has yielded promising results in species like poplar, eucalyptus and loblolly pine (Pavy et al., 2016). SNP markers identified in targeted regions can be of interest in breeding programs and could be developed as significant genetic markers. Targeted capture sequencing has also successfully identified variation among flowering-time regulatory genes in *Brassica napus* (Schiessl et al., 2014) and candidate mutations for cuticle wax in rice (Kim & Tai, 2019). Targeted exon sequencing also helps to study the speciation and evolution of targeted genes associated with genotypic and phenotypic variations (McNally et al., 2009).

In this study, we selected a total of fifteen genes regulating plant architecture for exome enrichment and sequencing. The hybridization probes were designed on the sequences identified by Iso-seq study. This enrichment helped us study these genes across 216 azalea genotypes and uncover any genetic variation present. It also helped us identify rare mutants present in our gene of interest that can further be used in breeding pipelines.

CHAPTER 3

GENETIC DIVERSITY AND POPULATION STRUCTURE OF *RHODODENDRON CANESCENS*, A NATIVE AZALEA FOR URBAN LANDSCAPING[1]

[1] L. K. Yadav, E. V. McAssey, and H. D. Wilde. 2019. *HortScience.* 54.4: 647-651.
Reprinted here with permission of publisher.

Abstract

Rhododendron canescens is a deciduous azalea native to the southeastern United States that is used in landscaping due to its ornamental qualities. A genotyping-by-sequencing approach was taken to characterize the genetic structure and diversity of a *R. canescens* germplasm collection. Single nucleotide polymorphisms (SNPs) were identified by two software platforms, STACKS and GBS-SNP-CROP. Three distinct *R. canescens* populations were detected by STRUCTURE analysis with GBS-SNP-CROP data, while two populations were distinguished using STACKS data. Principal component analysis with data from both SNP pipelines supported the presence of three populations. Statistical results indicated that there was low genetic differentiation between the populations, but relatively high genetic diversity within populations. The inbreeding coefficient of the *R. canescens* accessions was low, which would be expected with an outcrossing species. These results suggest that there may be a significant level of gene flow between populations of *R. canescens*.

Keywords: genotyping-by-sequencing, ornamental plant, Piedmont azalea, single nucleotide polymorphism

Introduction

Rhododendron canescens (Michaux) Sweet is a deciduous shrub commonly known as Piedmont azalea or sweet azalea. It is a diploid species (2n=26) that has a native range in the southern US from Texas to North Carolina. *R. canescens* is a member of the *Rhododendron* section *Pentanthera*, a group of interfertile species that generally maintain species identity through habitat preference and flowering time. Twelve species of section *Pentanthera* are native to regions of the southern US. Where *R. canescens* populations overlap with other native azalea species, hybridization and introgression have been demonstrated (King et al. 1977; Kron et al. 1993).

R. canescens is of value as an ornamental landscaping plant due to its showy, scented flowers, wide geographic distribution, and lace bug resistance (Galle et al. 1967; Wang et al. 1998). It is one of the first native azaleas to bloom and it could benefit native pollinators in urban landscapes in early spring (Mader et al. 2011). Cultivars of *R. canescens* and *R. canescens* hybrids are available currently for a niche market. *R. canescens* has architectural characteristics that may limit more widespread use in urban settings, namely an open growth habit and height up to five meters. We are interested in looking for genetic variation in wild germplasm that could be used to develop a more compact phenotype for landscaping.

Genetic analysis of species in *Rhododendron* section *Pentanthera* has been conducted using DNA sequences of the internal transcribed spacer (ITS) region of rRNA genes (Scheiber et al. 2000) and the chloroplast *matK* and *trnK* intron region (Kurashige et al. 2001). These studies found little sequence variation among species of the section *Pentanthera*, suggesting that they are a closely related group. Low genetic diversity among section *Pentanthera* species was further observed in a study with several hundred AFLP (amplified fragment length polymorphism)

markers (Chappell et al. 2008). In that study, AFLP analysis was also used to investigate variation between and within populations of individual azalea species, including *R. canescens*. In contrast to the low level of variation found between populations and species, a high level of variation was observed within populations. Variation within *R. canescens* populations may be due in part to hybridization and introgression from other native azalea species (Chappell et al. 2008).

We obtained leaf samples from 290 *R. canescens* genotypes, with a long-term objective of screening for variation in genes known to control plant height and branching. We first investigated the genetic diversity of a representative subset of this collection using genotyping-by-sequencing (GBS). Through GBS analysis, thousands of SNP (single nucleotide polymorphism) markers can be generated and used to examine genetic diversity in non-model species (Peterson et al. 2014). We used SNP genotypes to characterize the genetic structure and diversity of the *R. canescens* germplasm collection.

Materials and Methods

Plant material collection

Young leaves were collected from 247 *R. canescens* plants from 18 sites across Georgia, primarily within the Piedmont ecoregion (Omernik and Griffith 2014). Plants sampled at each site were at least 10 m apart and the species was confirmed based on floral characteristics as described by Kron (1993). The GPS coordinates of the plants were recorded, and the samples were frozen and stored at -80^{o}C until further use. In addition, silica-dried leaves from 43 *R. canescens* plants were received from collaborators in Georgia and northern Florida. The locations of the accessions used for GBS analysis are shown in Supplemental Table 1.

GBS library preparation and sequencing

Approximately 150 mg of frozen leaf tissue was ground using a TissueLyser bead mill (Qiagen) and DNA was isolated using an E.Z.N.A. HP Plant DNA kit (Omega Bio-Tek), following the manufacturers' protocols. The DNA quantity was measured with a Qubit 2.0 (Invitrogen) using a Qubit dsDNA HS assay kit. DNA quality was determined by analysis with a NanoDrop 8000 (Thermo Scientific) and electrophoresis through 0.8% agarose. A subset of 96 samples were chosen for GBS that were of high DNA quality and representative of 16 collection sites in Georgia and acquisitions from two collaborators.

DNA samples (250 ng) were digested with *Msp1* and *Pst1* at 37^{o}C for 2 hours in a 96-well plate. Barcoded *PstI* adapters and *MspI* Y-adapters were ligated to the digested DNA fragments, as described in Qi et al. (2018). Small DNA fragments (<400bp) were eliminated using a Mag-Bind RxnPure Plus kit (Omega Bio-Tek). PCR was conducted with each sample individually using a barcode-specific forward primer and a common adapter-specific reverse primer using the following conditions: 95^{O}C for 30s, then 16 cycles of 95^{O}C for 30s, 62^{O}C for 20s, 68^{O}C for 15s, followed by 68^{O}C for 5 min. Following cleanup with a Mag-Bind RxnPure Plus kit, PCR products were quantified by SYBR green fluorometry on a plate reader. The PCR products of the 96 samples were pooled (5 ng each) and the library was quantified using a Qubit 2.0. The GBS library was sequenced with an Illumina Nextseq 500 mid output flow cell by the Georgia Genomics Facility (Athens, GA), generating single-end reads of 150 bp in length.

Sequence data processing

FastQC (Leggett et al., 2013) was used to determine the quality of the sequence data. The sequence data were then processed using two software packages, STACKS v.1.44 (Catchen et

al., 2013) and GBS-SNP-CROP (Melo et al., 2016). For STACKS, the raw sequence reads were filtered and trimmed to the length of 115 bp. STACKS analyses were performed using the following pipeline: process_radtags – ustacks – cstacks – sstack - populations for diploid species, with 0.05 minor allele frequency. This generated a VCF file of the SNP matrix and initial population statistics.

For GBS-SNP-CROP analysis, raw GBS data were parsed to remove barcode sequences and cut sites and then trimmed using Trimmomatic (Bolger et al., 2014) to a uniform length (115 bp). The minimum phred score was set to 20 and the sliding window to 4 bp. Sequence reads were aligned using the Burrows-Wheeler Alignment tool (Li & Durbin, 2009) to a mock reference developed from *R. canescens* accession DA09. The binary matrix was generated and parsed using SAMtools (H. Li et al., 2009) in the downstream steps. Using the default settings for the diploid crop, SNP master matrix was generated followed by SNP calling.

Analysis of SNP Data

The ancestral population clusters of *R. canescens* were established with the admixture model of STRUCTURE (Falush et al., 2003, 2007; Hubisz et al., 2009; Pritchard et al., 2000) three to ten parallel Markov chains with a burn-in of 100,000 iterations and a run length of 1,000,000 iterations following the burn-in. The STRUCTURE Harvester program was used to determine natural logarithms of probability data (LnP(K)) and the ΔK. STRUCTURE PLOT version 2.0 was used to create visual structure charts (Ramasamy et al., 2014). Principal component analysis (PCA) was conducted in R version 3.5.1 using the *PCAdapt package* (Duforet-Frebourg et al., 2014; Duforet-Frebourg et al., 2015; Luu et al., 2017) with 87 accessions that met threshold requirements. A weighted neighbor-joining tree was created in

DARwin v6.0 using the default settings and data imported from the Genepop output file of STACKS.

The variant call file was used to manually make a numeric file with 0, 1, and 2 representing, respectively, homozygous reference alleles, heterozygous alleles, and homozygous alternate alleles. Nei statistics (Nei & Roychoudhury, 1974) were calculated using R software (version 3.5.1; 2018-07-02) to estimate the genetic distance among the 87 accessions and the three population clusters determined by the PCA analysis. R was used to calculate all the population statistics using HierFstat (De Meeûs & Goudet, 2007) and adegenet (Jombart, 2008a) function. This included gene diversity (D_{ST}) and corrected gene diversity (D_{STP}) among individuals, the overall gene diversity (H_T) and corrected gene diversity (H_{TP}) among populations, the fixation index (F_{ST}) and corrected (F_{STP}) based on population, and the inbreeding coefficient (F_{IS}). The overall observed heterozygosity (H_O) and genetic diversity (H_S) within population was estimated based on mean allele frequency. G_{ST}, the proportion of species genetic diversity in relation to among-population variation, was calculated as $1-(H_S/H_T)$.

Results

GBS sequence data

Genetic analysis was conducted with *R. canescens* samples collected from 16 sites in Georgia (Fig. 3-1). Single-end sequencing of a GBS library of 96 *R. canescens* accessions yielded 167,783,620 sequence reads. FastQC analysis indicated that the raw sequences were of good quality, with an average length ranging from 120-135 bp and an average GC content of 46%. The read depth count (Fig. 3-2) indicated an even coverage of the *R. canescens* genome. After filtering, 57% of the sequences were retained as high-quality reads.

SNP identification and analysis

SNPs were identified and analyzed from *R. canescens* accessions with high quality sequences and no missing data using STACKS and GBS-SNP-CROP software tools. A total of 3955 high quality SNPs were called by STACKS from 91 accessions. These polymorphic sites comprised 0.85% of the total loci examined. The observed heterozygosity and homozygosity present in our GBS data were 0.1958 and 0.8042, respectively. This matched the expected heterozygosity and homozygosity of 0.2557 and 0.7443 respectively. The inbreeding coefficient F_{IS} of the *R. canescens* lines was 0.2437, indicative of an outcrossing population. All the statistics were calculated for variant sites.

In contrast to the *de novo* analysis of GBS data using STACKS, a reference-based analysis was conducted with GBS-SNP-CROP software. Because the *R. canescens* genome has not been sequenced, a mock reference was developed using GBS data of one of the accessions (DA09) and all other sample reads were aligned to this reference. After filtering, 3185 high quality SNPs were called by GBS-SNP-CROP from 96 accessions.

Population structure analysis

The genetic structure of the *R. canescens* collection was examined through a STRUCTURE analysis of SNP data. Using GBS-SNP-CROP data, three populations (P1-P3) were identified based on LnP(K) variance and delta K value (Figs. 3-3 A and B). When STRUCTURE was conducted with STACKS data, two *R. canescens* populations were identified (Fig. 3-3C). The larger group consisted of 52 azalea genotypes and a smaller group had 39 genotypes. In light of PCA results below, the smaller group was re-examined by STRUCTURE, but there was no further differentiation (data not shown).

Principle component analysis

Genetic relationships within the *R. canescens* collection were further examined by PCA of SNP data. Three major clusters were observed among the 87 accessions using SNPs identified by STACKS analysis (Fig. 3-4) or GBS-SNP-CROP analysis (not shown). The first and second principal components explained approximately 12% of variance. The largest cluster contained samples from the central Georgia Piedmont, with the second largest containing samples from the western Georgia Piedmont region (Fig. 3-1). The third cluster was composed of samples from southern Georgia, northern Florida, and *R. canescens* samples of unspecified provenance from collaborators.

Population statistics

Genetic diversity was analyzed within and between the *R. canescens* populations (Table 3-1). The average heterozygosity within the species (H_T) was 0.25 and the average heterozygosity within populations (H_S) was 0.24. The proportion of species genetic diversity attributed to variation among populations (G_{ST}) was calculated to be 4.0%, a low value indicating significant gene flow within the species. Relatively low values were also observed for gene diversity among individuals (D_{ST}) and the fixation index (F_{ST}). The low F_{ST} indicates that there was no significant differentiation between the populations and is consistent with G_{ST} results.

The inbreeding coefficient (F_{IS}) of the population under study is 0.2163, a low value that would be expected with an outcrossing species. F_{IS} was positive, indicating that individuals within the population are more related than expected under a random mating model. Nei genetic diversity and genetic gain was also estimated among the 87 *R. canescens* accessions and between the population clusters identified by PCA. This analysis found more variation among the

individuals than between the three population clusters (Supplemental Table 3-2), in agreement with previous results.

Conclusions

A genotyping-by-sequencing approach was taken to characterize the genetic structure and diversity of a *Rhododendron canescens* germplasm collection. SNPs were identified by two software platforms, STACKS and GBS-SNP-CROP, and a genome-wide genetic variant file was developed. The genetic variation present in the germplasm collection was examined by STRUCTURE, a model-based Bayesian analysis, and PCA, a distance-based method. Taken together, these analyses indicated that there three population clusters present among the accessions analyzed.

The STRUCTURE results varied depending on whether SNP data from the GBS-SNP-CROP or the STACKS pipeline were used. STRUCTURE with GBS-SNP-CROP and STACKS data identified three and two population clusters, respectively. These different outcomes may be due to the fact that STACKS analysis involved *de novo* assembly, whereas GBS-SNP-CROP was reference genome-based. The largest population cluster of 48 accessions (P1) was the same for both methods, but the remaining accessions were partitioned by STRUCTURE into two clusters of 22 accessions (P2) and 17 accessions (P3) when using GBS-SNP-CROP data. PCA of SNP data from GBS-SNP-CROP or STACKS supported the presence of three population clusters.

An analysis of *R. canescens* genetic diversity was included in a prior study (Chappell et al. 2008) that examined four *R. canescens* populations (6 accessions each) with AFLP markers. Similar to that study, our investigation found a low G_{ST} value, indicating that the proportion of diversity between populations was low, while the proportion of diversity within populations was

high. Minimal differentiation between *R. canescens* populations was also indicated by the low F_{ST} value. Chappell et al. (2008) suggested that this may be the result of gene flow between populations due to insect pollination. *R. canescens* is known to be pollinated by bumblebees, adrenid bees, butterflies, and hummingbirds. Populations P1 and P2 are geographically close, whereas accessions of P3 had more diverse origins. Introgression from other species of section *Pentanthera* may also have played a role in similarity found between populations.

Genetic markers, including SNPs and cpDNA loci, have been used to examine the genetic structure of a Japanese evergreen azalea species, *R. indicum* (Yoichi et al. 2018). SNPs were identified by multiplexed ISSR genotyping-by-sequencing (MIG-seq). Two genetically distinct lineages were detected, both of which had DNA introgressed from geographically close populations of *R. kaempferi*. MIG-seq was also used to investigate rhododendron plants in a hybrid zone between two natural varieties of *R. japonoheptamerum* (Tamaki et al. 2016). SNP analysis distinguished the varieties and their hybrids and provided the basis for estimating that hybridization occurred 0.4 million years ago. In our investigation, GBS provided a cost-efficient means of generating SNP markers for genetic characterization an *R. canescens* germplasm collection. The high level of genetic diversity found within this collection indicates that screening for allelic variation in genes controlling architecture could be a viable approach to accelerate the breeding of plants with improved form.

Acknowledgments: This work was supported by Specialty Crop Block Grant 16-SCBGP-GA-0010, the Azalea Society of America, and USDA NIFA Hatch project GEO00755. We thank Matthew Chappell for advice and Carol Robacker, Ron Miller, and Charles Andrew for providing azalea material.

References

Bolger, A.M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Catchen J., P.A. Hohenlohe, S. Bassham, A. Amores, and W.A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol., 22, 3124–3140.

Chappell M., C. Robacker, and T.M. Jenkins. 2008. Genetic diversity of seven deciduous azalea species (Rhododendron spp. section Pentanthera) native to the eastern United States. J. Amer. Soc. Hort. Sci. 133: 374–382.

De Meeûs T. and J. Goudet. 2007. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. Infect. Genet. Evol. 7: 731–735

Duforet-Frebourg N., E. Bazin, and M.G.B Blum. 2014. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. Mol. Biol. Evol. 31: 2483-2495.

Galle F.C. 1967. Native and some introduced azaleas for southern gardens: kinds and culture. Amer. Hort. Mag. 46:13–23

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405

King B.L. 1977. Flavonoid analysis of hybridization in Rhododendron sect. Pentanthera (Ericaceae). Syst. Bot. 2: 14-27.

Kron K.A. 1993. A revision of *Rhododendron* section *Pentanthera*. Edinburgh J. Bot. 50: 249-364

Kron K.A., L.M. Gawen, and M.W. Chase.1993. Evidence for introgression in azaleas (Rhododendron; Ericaceae): Chloroplast DNA and morphological variation in a hybrid swarm on Stone Mountain, Georgia. Amer. J. Bot. 80:1095–1099.

Kurashige Y., J.I. Etoh, T. Handa, K. Takayanagi, and T. Yukawa. 2001. Sectional relationships in the genus Rhododendron (Ericaceae): evidence from matK and trnK intron sequences. Plant Syst. Evol. 228:1–14.

Leggett R.M., R.H. Ramirez-Gonzalez, B.J. Clavijo, D. Waite, and R.P. Davey. 2013. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front. Genet. 4: 288

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

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079.

Mader E., M Shepherd, M. Vaughan, S. Hoffman-Black, and G. LeBuhn. 2011. Attracting Native Pollinators: Protecting North America's Bees and Butterflies. Storey Publishing, North Adams, MA.

Melo A.T., R. Bartaula, and I Hale. 2016. GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC Bioinform. 17: 1

Nei M., and A.K.Roychoudhury. 1974. Sampling variances of heterozygosity and genetic distance. Genetics 76: 379-390.

Omernik J.M., and G.E. Griffith. 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. Environ. Mgt. 54:1249-1266.

Peterson G.W., Y. Dong, C. C. Horbach, and Y.B. Fu. 2014. Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. Diversity 6: 665-680

Pritchard J.K., M Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Ramasamy R.K., S. Ramasamy, B.B.Bindroo, and V.G. Naik. 2014. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. SpringerPlus, 3, 431.

Qi P., D. Gimode, D. Saha, S. Schröder, D. Chakraborty, X. Wang, M.M. Dida, R.L. Malmberg, and K.M. Devos. 2018. UGbS-Flex, a novel bioinformatics pipeline for imputation-free SNP discovery in polyploids without a reference genome: finger millet as a case study. BMC Plant Biol.18:117. https://doi.org/10.1186/s12870-018-1316-3

Scheiber S.M., R.L. Jarret, C.D. Robacker, and M. Newman. 2000. Genetic relationships within Rhododendron L. section Pentanthera G. Don based on sequences of the internal transcribed spacer (ITS) region. Scientia Hort. 85:123–135.

Tamaki I., W. Yoichi, Y. Matsuki, Y. Suyama, and M. Mizuno. 2017. Inconsistency between morphological traits and ancestry of individuals in the hybrid zone between two *Rhododendron japonoheptamerum* varieties revealed by a genotyping-by-sequencing approach. Tree Genet. Genomes 13: 4.

Wang Y., C.D. Robacker, and S.K. Braman.1998. Identification of resistance to azalea lace bugs among deciduous azalea taxa. J. Amer. Soc. Hort. Sci. 123: 592-597

Yoichi W., I. Kawamata, Y. Matsuki, Y. Suyama, K. Uehara, and M. Ito. 2018. Phylogeographic analysis suggests two origins for the riparian azalea Rhododendron indicum (L.) Sweet. Heredity 121:594–604

Table 3-1. Genetic diversity statistics of STACKS data

H_O	H_S	H_T	D_{ST}	H_{TP}	D_{STP}	F_{ST}	F_{STP}	F_{IS}	D_{EST}
0.1933	0.2452	0.2572	0.0120	0.2631	0.0180	0.0465	0.0682	0.2117	0.0238

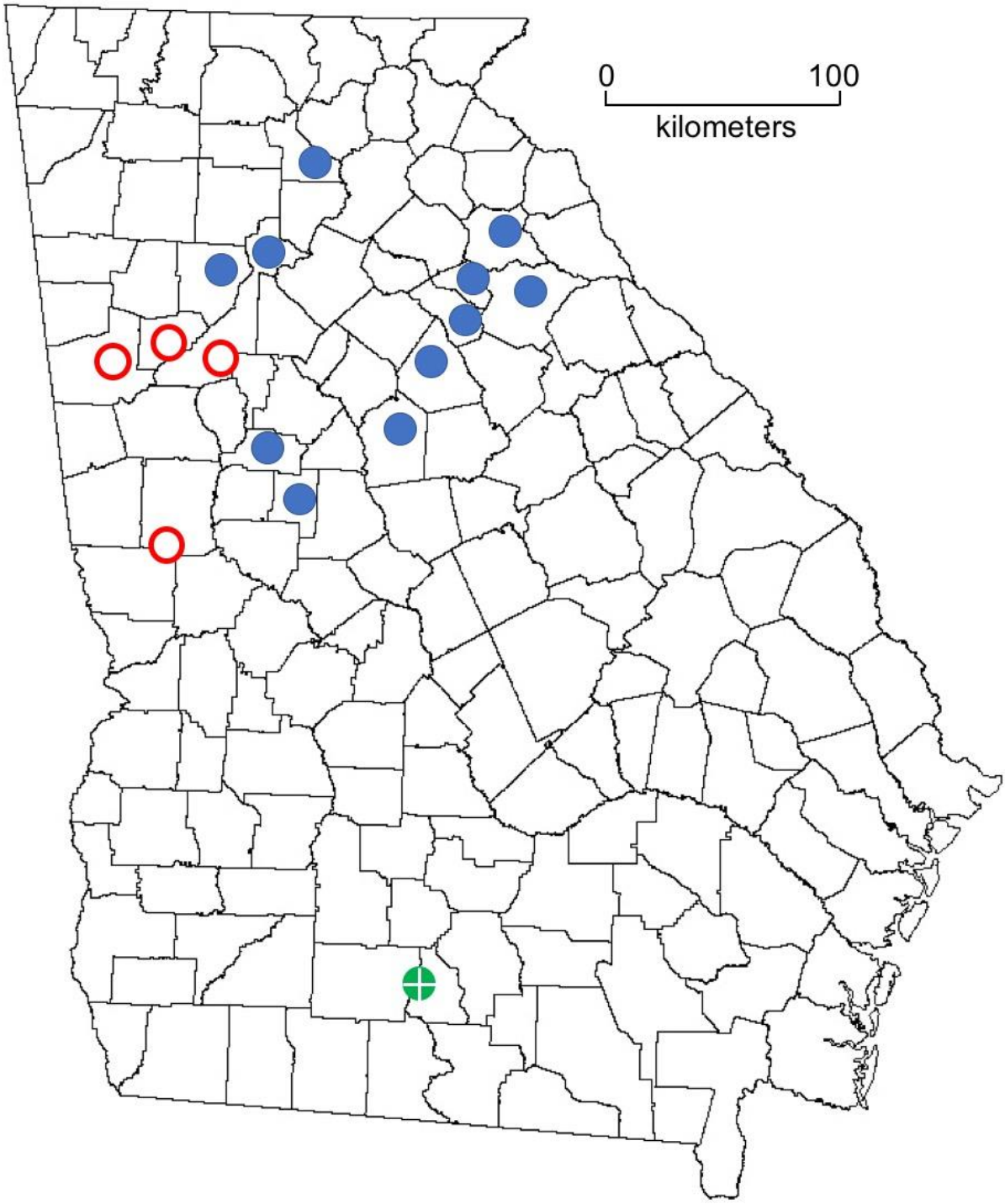


Figure 3-1. Sites of *R. canescens* collection in Georgia. Leaf samples from at least 4 plants were collected per site. P1, filled blue circles; P2, open red circles; P3, cross-marked green circle

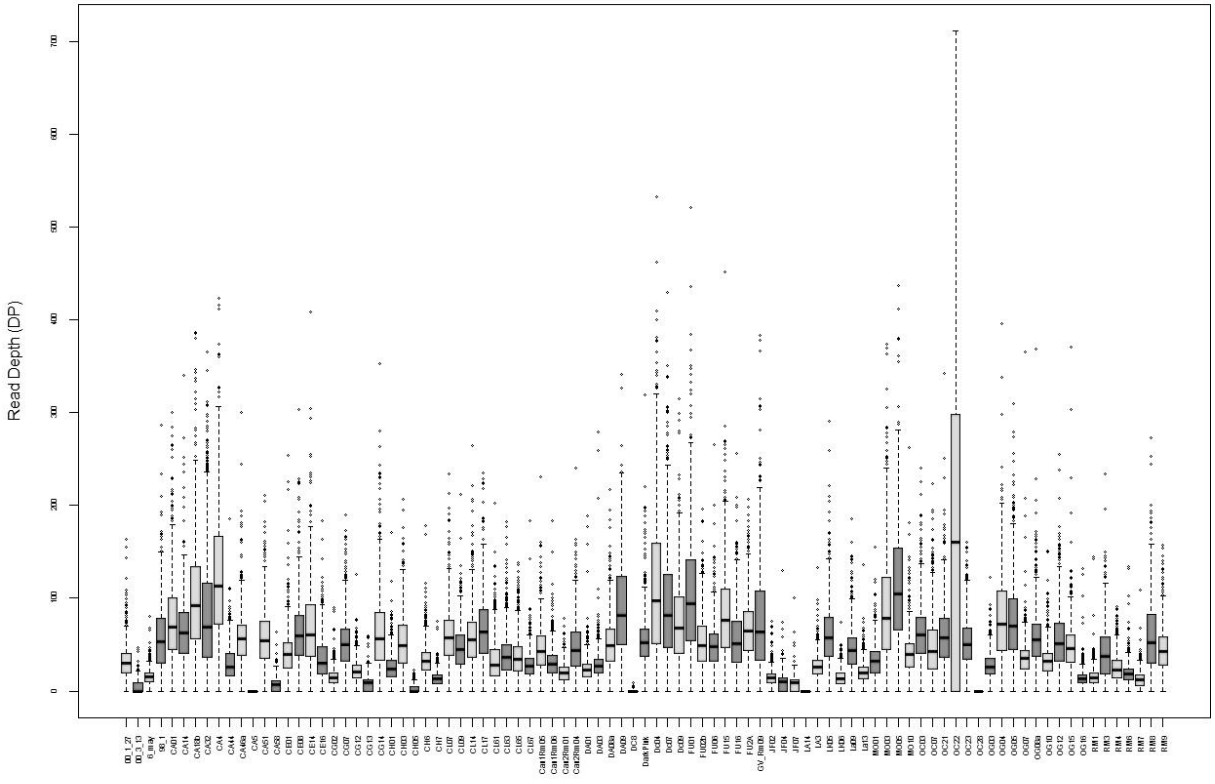


Figure 3-2. Read depth of sequencing data

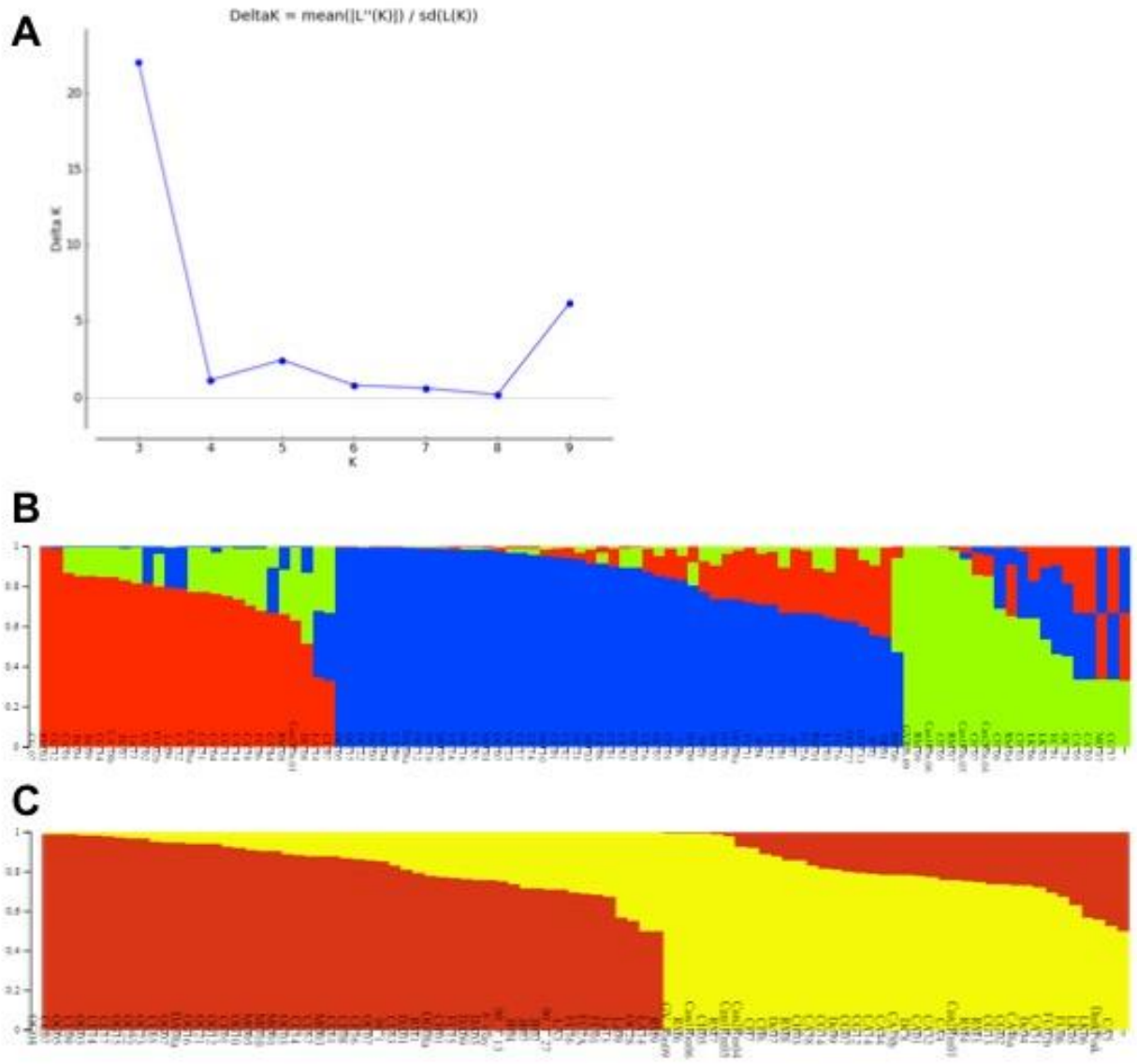


Figure 3-3. STRUCTURE analysis. A. Support for 3 optimal clusters based on delta K estimates from GBS-SNP-CROP data. B. STRUCTURE results with GBS-SNP-CROP data. C. STRUCTURE results with STACKS data

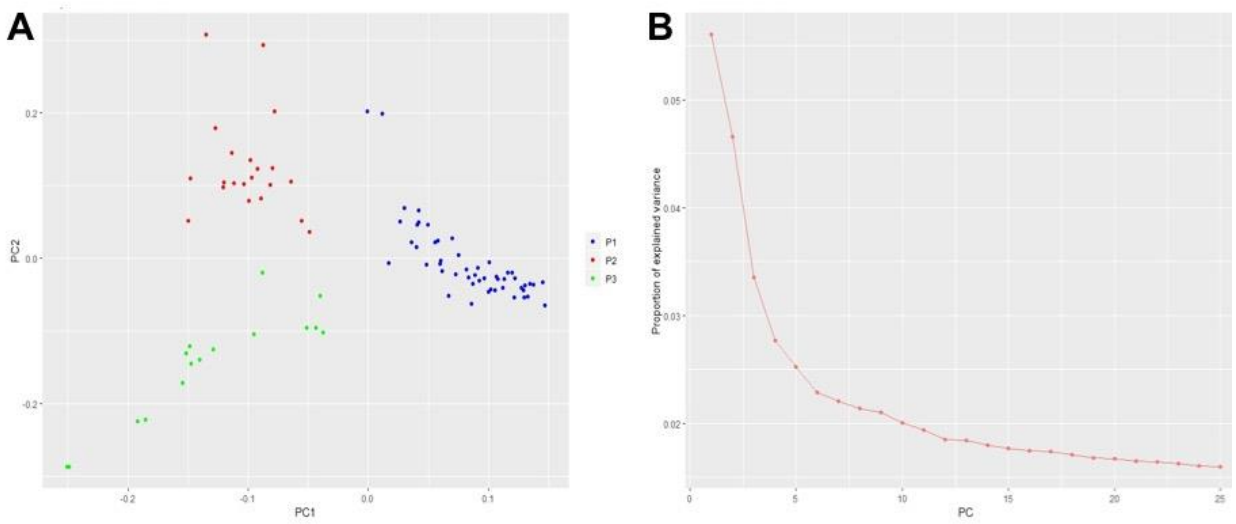
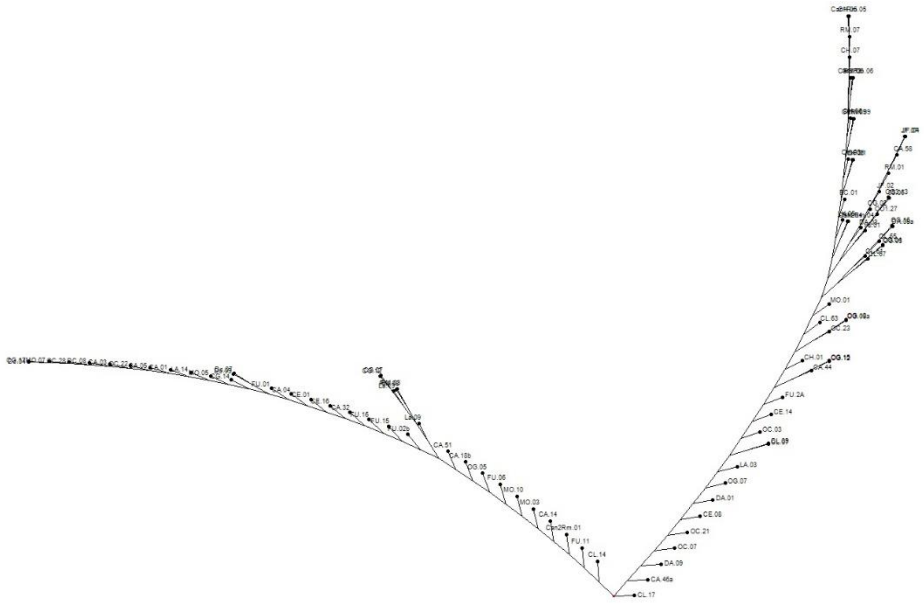


Figure 3-4. Principal component analysis. A. PCA with SNP data from STACKS pipeline. Populations P1, P2, and P3 as defined by STACKS analysis. B. Proportion of variance explained by principal components



Supplementary Figure 3-1. Phylogenetic relationships of *R. canescens* accessions. A weighted neighbor-joining phylogenetic tree created in DARwin v6.0.

CHAPTER 4

LONG READ ISOFORM SEQUENCING OF *RHODODENDRON CANESCENS* AND *DE NOVO* ASSEMBLY OF FULL-LENGTH TRANSCRIPTOME TO IDENTIFY GENES REGULATING PLANT ARCHITECTURE[1]

[1] L. K. Yadav, and H. D. Wilde. To be submitted to *Tree Genetics & Genomes*.

Abstract

Rhododendron canescens (Michaux) Sweet is a deciduous azalea from the southeastern United States that is used as an ornamental landscaping plant. The transcriptome of *R. canescens* vegetative and reproductive tissues was characterized as part of an investigation into the control of its architecture. Transcriptome data were obtained by long-read sequencing and analysis using PacBio Iso-Seq methods. The analysis generated 24,244 full-length isoform sequences, of which 16,825 were annotated. Gene ontology classification grouped these unigenes into three clusters: cellular component (12,826), molecular function (13,685), and biological processes (13,102). We identified orthologs of 13 genes that have been shown to control height or branching across multiple plant species. These included genes for regulatory factors and components of phytohormone biosynthesis and signaling pathways. Sequence data from *R. canescens* orthologs of these architecture genes will enable their genetic variation to be investigated within a large *R. canescens* population by exon capture and sequencing. In addition to gene sequence data, bioinformatic analysis of long read sequences identified 2871 long non-coding RNA (lncRNA) sequences and13,116 simple sequence repeats (SSRs). This data adds to the genetic resources available for *Rhododendron*, which include at least fourteen species with sequenced genomes or transcriptomes. This information has potential applications in *Rhododendron* comparative genomics, evolutionary studies, and molecular breeding.

Key words: azalea, transcriptome, architecture, long non-coding RNA

Introduction

Rhododendron canescens (Michaux) Sweet is a deciduous azalea and one of over 1000 species of *Rhododendron* (Chamberlain et al. 1996). It is a diploid species (2n = 26) with a native range across the southern United States from Texas to North Carolina. *R. canescens* is used as an ornamental landscaping plant due to its showy, scented flowers that bloom in early spring. In addition to its wide adaptability and floral display, it has greater resistance to the azalea lace bug (*Stephanitis pyrioides*) than other deciduous azaleas (Wang et al. 1998). The use of *R. canescens* in urban settings, however, may be limited by architectural characteristics such as an open growth habit and height up to five meters. We are interested in identifying genes involved in determining *R. canescens* architecture, as the first step towards developing a more compact phenotype for landscaping.

Significant progress has been made towards understanding the genetic control of plant stature and form (Wang and Li 2008; Busov et al 2011, Teichmann and Muir 2015, Hill and Hollender 2019). Our objective was to identify *R. canescens* orthologs of genes that have been shown to control height or branching across multiple plant species. Genes that encode components of phytohormone biosynthesis and signaling pathways, together with regulatory genes, play a major role in determining plant architecture. For example, the "Green Revolution" genes *GA20 OXIDASE* (*GA20ox*) and *GA INSENSITIVE* (*GAI*) are involved in gibberellin biosynthesis and signaling, respectively (Hedden 2002), and they have been found to have similar roles regulating height in different plant species. The gibberellin biosynthesis gene *GA2 OXIDASE* (*GA2ox*; Biemelt et al. 2004; Braun et al. 2019) and signaling genes *GA-INSENSITIVE DWARF1* (*GID1*; Liang et al. 2014, Cantin et al. 2018) and *PHOTOPERIOD*

RESPONSIVE1 (*PHOR1*; Amador et al. 2001; Zawaski et al. 2012) have also been shown to control plant height.

Deficiencies in the biosynthesis or signaling pathways of the brassinosteroid phytohormones have been shown to reduce plant stature (Wang and Li 2008). *DWARF4* (*DWF4*) encodes an enzyme for a rate-limiting step in brassinosteroid biosynthesis and its loss-of-function results in a dwarf phenotype (Choe et al. 2001; Sakamoto et al. 2006). Mutations in *BRASSINOSTEROID INSENSITIVE1* (*BRI1*), a gene involved in signal transduction, can lead to dwarfing in dicots and monocots (Yamamuro et al. 2000; Clouse 2002). Alternatively, plant stature can be modified by changes in genes that regulate shoot meristem determination, including *TERMINAL FLOWER 1* (*TFL1*), *FLOWERING LOCUS T* (*FT*), and *FLOWERING LOCUS C* (*FLC*) (Schiessl et al. 2014; Moraes et al. 2019).

Branching patterns are another major component of plant architecture. Axillary bud development is inhibited by a transcription factor encoded by *BRANCHED1* (*BRC1*), mutations of which can alter branch density (Teichmann and Muhr 2015). The phytohormone strigolactone (SL) plays a major role in the control of branching in dicots and monocots (Gomez-Roldan et al. 2008; Umehara et al. 2008). Plant architecture is altered by variants of SL biosynthesis genes *MORE AXILLARY BRANCHING1* (*MAX1*), *MAX3*, and *MAX4* and the SL signaling gene *MAX2* (Teichmann and Muhr 2015). Sequence data from *R. canescens* orthologs of these architecture genes would enable their genetic variation to be investigated within a large *R. canescens* population by exon capture and sequencing (*e.g*. Schiessl et al. 2014; Kim and Tai 2019).

Genetic and genomic resources for *Rhododendron* species have increased significantly since the publication of the draft genome of *R. delavayi* (Zhang et al. 2017). Sequencing and analysis of the genomes of *R. williamsianum* and *R. simsii* have identified two whole-genome

duplications that are shared with other members of the Ericaceae (Sosa et al. 2019; Yang et al. 2020). These are the only genomes in the genus that have been sequenced to date, due in part to the genome size of diploid *Rhododendron* species, which have an estimated 2C value of 1.2-1.9 pg (Jones et al. 2007). Transcriptomes, however, have been sequenced from ten *Rhododendron* species representing the subgenera *Pentathera, Tsutsusi*, *Hymenanthes*, *Rhododendron*, and *Azaleastrum*. These studies involved the characterization of gene expression across the whole plant (Xing et al. 2017; Cheng et al. 2018; Jia et al. 2020), as well as more focused research to identify molecular markers (Zhang et al. 2017; Xing et al. 2017; Choudhary et al. 2018; Li et al. 2018) and genes expressed during heat stress (Fang et al. 2017; Zhao et al. 2018), drought stress (Wang et al. 2020), flower development (Wang et al. 2018), flower color (Xiao et al. 2018), and secondary metabolism (Zhao and Zhu 2020).

Molecular studies of *R. canescens* to date have used genetic markers to examine relationships within the subgenus *Pentathera* (Kron et al. 1993; Scheiber et al. 2000; Kurashige et al. 2001) and genetic diversity within *R. canescens* populations (Chappell et al. 2008; Yadav et al. 2019). We have expanded the genetic resources for *R. canescens* by sequencing the transcriptome of vegetative (leaf, internode, terminal and lateral bud) and reproductive (flower, fruit) tissues. PacBio single-molecule real-time (SMRT) sequencing and analysis was used to assemble the full-length transcriptome without a reference genome. The *R. canescens* transcriptome was characterized and putative orthologs of 13 of the 15 targeted genes were identified. In addition to gene sequence data, bioinformatic analysis of long read sequences generated other genomic resources including simple sequence repeat (SSR) and long non-coding RNA (lncRNA) sequences.

Materials and Methods

Collection of tissue samples and RNA extraction

Young leaves, mature leaves, internodes, terminal buds, axillary buds, flowers, and siliques were collected from a single *R. canescens* plant located in Athens, Georgia (33.95° N, 83.36° W). Plant tissues were flash frozen in liquid nitrogen and stored in -80°C for future analysis. Total RNA was isolated from the six tissues individually using the CTAB (cetyltrimethyl ammonium bromide) extraction procedure of Vashisth et al. (2011). After treatment with RNase-free DNase (New England Biolabs Inc.), the quantity and quality of the RNA was determined using a NanoDrop 8000 (Thermo Scientific, Rockford, IL). RNA from the different tissues was pooled in equal quantity in a 50 µl volume. The RIN (RNA Integrity Number) and concentration of the pooled RNA sample were determined with an Agilent 2100 Bioanalyzer.

cDNA library preparation and PacBio sequencing

cDNA synthesis and isoform sequencing (IsoSeq) were conducted at the Georgia Genomic Bioinformatics Core facility in Athens, Georgia. The SMARTer PCR cDNA Synthesis kit (Clontech) was used to generate full-length cDNA and add adapters (Gonzalez-Garay 2016). The cDNA transcripts were divided into four groups based on size: 1-2 Kb, 2-3 Kb, 3-6 Kb, and 6-10 Kb. Each group was PCR-amplified using KAPA HiFi DNA polymerase (Roche) following the manufacturer's protocol. After PCR, the cDNA was circularized using PacBio reagents and sequenced on a PacBio Sequel II platform.

Iso-Seq data polish and full length iso-form identification

Reads were extracted with the read of inserts (ROI) component of the RS IsoSeq protocol (SMRT Analysis 3.0). Full-length and non-full length reads were distinguished and only the full-

length, non-chimeric reads were used in downstream analysis. The full-length reads were clustered using the isoform-level clustering (ICE) and the clustered reads were further processed into high and low quality polished reads. Both high and low quality polished reads were merged and used in 'cd-hit-est' from the Cd-HITv4.6 package with the following parameters: -c 0.99 –G 0 –aL 0.00 –aS 0.99 –AS 30 -M 0 –d 0 – (Li & Godzik, 2006). Non-redundant transcripts were then analyzed using the COding GENome reconstruction Tool (Cogent v6.0, https://github.com/Magdoll/Cogent). Cogent utilizes K-mer values of the transcripts to calculate pairwise distances and cluster the transcripts using similarity. These clusters were further dissolved into one or multiple unique transcripts. The error-corrected non-redundant transcripts from cd-hit were used by Cogent to partition the iso-forms into gene families and provide additional filtering to increase the quality of the transcript sequences.

Iso-forms functional annotation

Non-redundant isoforms were subjected to BLASTX against the NR database (Altschul et al., 1997). All the six reading frames were blasted against the protein database to identify orthologues in other plants. The significant threshold set for a BLAST hit was E-value ≤ 1e-5. The results from BLASTX were annotated using BLAST2GO v5.0 (Conesa et al., 2005) and were further used for GO term identification. The isoforms were clustered into three classes, namely: molecular function, cellular component and biological processes (Götz et al., 2008). The GO terms identified were classified using WEGO software (http://wego.genomics.org.cn/cgi-bin/wego/index.pl). The assembled unigenes were further annotated using KEGG pathways platform KEGG Automatic Annotation Server (KAAS) (http://www.genome.jp/tools/kaas/) (Moriya et al., 2007). The Bi-directional Best Hit (BBH) method was used to obtain KEGG Orthology (KO) assignment.

Identification of target genes, lncRNA, and SSR markers

Transcriptome data were used to query the NCBI database to identify putative architectural genes based on amino acid similarity. The isoforms annotation was done and, based on the annotation results, isoforms were filtered to retain sequences related to genes regulating plant architecture. Based on our filtered results, fifteen genes were targeted: *GA20ox, GA2ox, GAI, GID1c, PHOR1, DWF4, BRI1, TFL1, FT, FLC, BRC1, MAX1, MAX2, MAX3* and *MAX4.* The unigene sequences of the selected genes were converted into amino acid sequences and analyzed using the ORFfinder program of NCBI (https://www.ncbi.nlm.nih.gov/orffinder/). The protein sequence was further analyzed for its structure and compared to known domains of the gene of interest using the SMART mode (http://smart.emblheidelbergde/index2.cgi). These analyses helped us to narrow down transcriptome sequences to be used for Capture-Seq. LncRNAs were detected using CNIT software (Guo et al., 2019; Sun et al., 2013) and PLEK VM software (Li et al., 2014). CNIT identifies lncRNA using sequence similarity of protein database from plants including rice, arabidopsis, and soybean. PLEK uses the k-mer value of the sequence and predicts lncRNA. PLEK is more sensitive then CNIT in lncRNA identification. Simple Sequence Repeat (SSR) loci were identified using MISA software (http://pgrc.ipk-gatersleben.de/misa/; version: 1.0) (Beier et al. 2017). The parameters set to identify SSRs were as follows: di-nucleotide repeats of more than 6 repeats, tri-nucleotide, tetra-nucleotide, penta-nucleotide and hexa-nucleotide of more than 5 repeats.

Results

R. canescens transcriptome analysis using PacBio Iso-seq

The full-length transcriptome of six *R. canescens* tissues was sequenced using PacBio Iso-seq, yielding 16 Gb of data. The raw data were deposited in the NCBI database and can be accessed in the BioSample accession SAMN17098567. The SMRT Iso-Seq pipeline was used to analyze the raw data, producing 47,289 high quality (HQ) and 519 low quality (LQ) transcript sequences (Fig. 4-1). The total length of these sequences was 65,500,237 bp, with a mean length of 1386 bp and an N50 length of 1746 bp (Table 4-1). Although PacBio yields longer sequence, it has a higher error rate and does not account for redundant sequences. Therefore, Cogent and Cd-hit were employed to polish and collapse the HQ transcript sequences. A total of 24,244 unigenes were obtained with a total length of 36,290,184 bp and a mean length of 1497 bp and N50 of 1851 bp. The length distribution of unigenes is shown in Fig. 4-2. The results demonstrate that PacBio Iso-Seq is an efficient way to generate a high-quality, full-length transcriptome for a species without a reference genome.

Functional annotation and classification

The unigenes were aligned to public gene and protein databases. The isoforms were compared to the NR database with BLASTX (E-value $\leq$ 1e – 5), which identified 22,580 unigenes. With blast2go, 16,825 genes were annotated and 39,613 GO terms were assigned (Fig. 4-3). The GO term annotations were further classified into three categories: biological processes (13,102), molecular function (13,685), and cellular components (12,826). Cellular and metabolic processes were the most abundant sub-categories in biological processes. Within molecular functions, catalyzing and binding activity were the major sub-categories. Cell and cell parts were the major sub-categories identified by GO terms in the cellular component.

KEGG pathway analysis was conducted as an alternative approach to identify unigenes. The unigenes were annotated using KAAS and categorized into different functional groups based on their Enzyme Commission (EC) numbers (Fig. 4-4). A total of 3,666 unigenes were annotated using this method. Forty-two percent of the annotated sequences (1466 unigenes) belonged to the Metabolism category, with Carbohydrate (314 unigenes) and Amino acid metabolism (267 unigenes) being major sub-groups. Signal transduction (404 unigenes) and Translation (336 unigenes) were the largest sub-groups overall. The percentage of unigenes in other categories were Genetic Information Processing (23%), Environmental Information Processing (12%), Cellular Processes (14%), and Organismal Systems (8%).

SSR Discovery

The unigenes were mined for SSRs using MISA software and 13,116 SSRs were identified (Table 4-2). The most common repeats were di-nucleotide (80%) and tri-nucleotide (19%) motifs. Tetra-nucleotide (0.81%), penta-nucleotide (0.30%), and hexa-nucleotide (0.34%) repeats were also found, as well as other SSRs in a very low number. The di-nucleotide AG/CT was the motif found at the highest frequency (73.6 %), followed by the tri-nucleotide AAG/CCT at 4.7 % (Fig. 4-5).

LncRNA identification in full-length transcriptome

LncRNA is a type of RNA of more than 200 bp that is not translated into protein. The unigenes identified by PacBio Iso-Seq were subjected to CNIT and PLEK, which are sequence-based and k-mer-based identification tools, respectively. CNIT identified 2871 lncRNAs (Fig. 4-6) and PLEK identified 11,084 lncRNAs (Fig. 4-7). Most of the sequences identified by CNIT and PLEK were between 200-700 bp and 500-1400 bp, respectively. The difference between the two outcomes is due to the methodology of the two programs. CNIT uses sequence comparison

to genomes such as rice, arabidopsis, and soybean, whereas PLEK uses the K-mer value of the sequence and predicts lncRNA. These results are consistent with the annotation results by blast2go. While most of the identified lncRNA have unknown function, some of them are annotated, but need further characterization.

Genes associated with Plant Architecture and Branching

Plant height and branching are quantitative traits regulated by number of genes. We selected 15 target genes that play a role in shaping plant architecture across multiple species. BLASTX comparison of the *R. canescens* Iso-Seq database with the NCBI database identified 237 potential transcripts from 13 of the 15 targeted genes. Multiple unigenes were found for each of the following target genes: *GA2ox, GAI, GID1c, PHOR1, FLC, DWF4, BRI1, BRC1, MAX2, MAX3* and *MAX4* (Table 4-3). One unigene each was identified for *GA20ox* and *TFL1* and no unigenes were found for *FT* or *MAX1*. The open reading frames and amino acid sequence of these transcripts were determined. The degree of sequence similarity of these genes was above 90% at the protein level compared to the ORF sequences of unigenes. Furthermore, SMART analysis showed that the protein domains and structure of these unigenes were similar to the known orthologs of these genes.

Discussion

Through long-read sequencing and analysis using PacBio Iso-Seq methods, we developed a *R. canescens* transcriptome database with annotation. This analysis generated 24,244 full-length isoform sequences, of which 16,825 were annotated. This information adds to the genetic resources available for *Rhododendron*, which includes at least fourteen species with sequenced genomes or transcriptomes. The *R. lapponicum* transcriptome (roots, stems, leaves, flower) was

also analyzed by PacBio Iso-Seq (Jia *et al.* 2020), while the other *Rhododendron* transcriptomes were shot-gun sequenced on Illumina platforms. Errors in assembly of *R. canescens* and *R. lapponicum* transcripts were minimized by the long-read sequencing of full-length transcripts. In the *R. lapponicum* study, of the 75,002 isoform sequences obtained, 71,386 were annotated. The difference in the number of transcripts identified in *R. canescens* and *R. lapponicum* is likely due to the inclusion of root tissue in the *R. lapponicum* study, as well as different transcript filtering stringencies. In the sequenced genomes of *R. williamsianum* and *R. simsii*, there were 23,559 and 34,170 genes predicted, respectively (Sosa et al. 2019; Yang et al. 2020).

The full-length sequence of transcripts facilitated their annotation in the absence of reference genome. Transcriptome data from six *R. canescens* tissues (leaves, internodes, terminal buds, axillary buds, flowers, siliques) were analyzed by BLASTX and Blast2GO. Blast2GO identified 39,613 GO terms in the *R. canescens* database, which is approximately the same number of GO terms (33,653) assigned in the *R. lapponicum* transcriptome study (Jia *et al.* 2020). There were a number of *R. canescens* isoforms (5755) that could not be annotated, potentially indicating genes that are unique to *R. canescens.* The characterization of the *R. canescens* transcriptome enabled the identification of 13 of our target genes.

Transcriptome studies of *Rhododendron* species have led to a better understanding of the genetic control of flower color and development (Xiao et al. 2018; Wang et al. 2018), secondary metabolism (Zhao and Zhu 2020), and abiotic stress response (Fang et al. 2017; Zhao et al. 2018; Wang et al. 2020). The *R. canescens* study presented here is the first to focus on genes involved in the control of plant height and branching patterns. We identified 237 transcripts for specific regulatory proteins, phytohormone biosynthesis enzymes, and signal transduction proteins. The number of transcripts per gene ranged from one for *TFL1* to 78 transcripts for *PHOR1*. BLASTP

analysis revealed that the sequences had high similarity (E-value < 1e-111) to orthologs of several plant species in NCBI database. The identity of the transcript sequences was further confirmed at the protein domain level using ORF and SMART analysis.

R. canescens-specific sequence information is needed to examine variation in architecture genes within a large *R. canescens* population (216 genotypes). *R. canescens* transcript data enables the design of tiled RNA probes to capture exons of these genes for sequencing and analysis. A similar capture/sequence strategy identified variation in orthologs of 29 regulatory flowering-time genes among four diverse *Brassica napus* accessions (Schiessl et al. 2014). The *Brassica* study detected non-synonymous SNPs, copy number variation, and presence-absence variation in the target genes. More recently, a mutation underlying a rice wax mutant was identified through capture/sequence analysis of exons from 321 genes in 12 rice genotypes (Kim and Tai 2019).

In addition to the characterization of protein-encoding transcripts, the analysis of the *R. canescens* transcriptome identified SSRs and lncRNAs. SSRs are polymorphic loci of 1-6 bp DNA repeats that are used as molecular markers in applications such as population genetics, genetic mapping and genome-wide association studies ((Bruni et al., 2012; Varshney et al., 2005). Analysis of *R. canescens* transcripts found 13,116 genic SSRs, with the di-nucleotide motif AG/TC being the most prevalent (74%). Previous transcriptome analyses have identified genic SSRs in *R. latoucheae, R. rex, R. molle, R. longipedicellatum, R. arboretum, R. hybridum*, and *R. lapponicum* (Xing et al. 2017, Zhang et al. 2017, Xiao et al. 2018, Li et al. 2018, Choudhary et al. 2018, Cheng et al. 2018, Zhao and Zhu 2020, Jia et al. 2020). Genic SSR markers have been found that are transferable from one *Rhododendron* species to others in different subgenera (Zhang et al. 2017; Xing et al. 2017; Li et al. 2018).

LncRNAs have been found to regulate transcription patterns and protein activity at the transcriptional, post-transcriptional, and epigenetic levels (Wu et al. 2020, Sanchita et al. 2020). Sequence similarity and k-mer techniques were used to identify lncRNA in the *R. canescens* transcriptome. A sequence similarity screen using CNIT software ((Guo et al., 2019; Sun et al., 2013) detected 2871 lncRNA. In contrast, the k-mer based technique PLEK found a much larger number of lncRNA (11,084). Within other *Rhododendron* species, lncRNAs were identified in a screen of the *R. lapponicum* transcriptome, which found 2,011 lncRNAs (Jia et al. 2020). This suggests that the number of *R. canescens* lncRNAs detected by sequence similarity may be more accurate. LncRNA is a relatively new field of study in plant genomics and the availability of these sequences could be useful in the characterization of their role in *Rhododendron*.

In conclusion, we have expanded the genetic resources for *R. canescens* to include a full-length transcriptome, gene annotation, SSR information, and a list of potential lncRNAs. This data has potential applications in *Rhododendron* comparative genomics, evolutionary studies, and molecular breeding. For *R. canescens*, it enabled the identification of orthologs of genes involved in determining plant architecture.

Data Archiving Statement: All sequence data supporting this work are available in the NCBI SRA database under the project ID: PRJNA685847

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402

Amador V, Monte E, García-Martínez JL, Prat S (2001) Gibberellins signal nuclear import of PHOR1, a photoperiod-responsive protein with homology to *Drosophila armadillo*. Cell 106:343-354

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583-2585

Biemelt S, Tschiersch H, Sonnewald U (2004) Impact of altered gibberellin metabolism on biomass accumulation, lignin biosynthesis, and photosynthesis in transgenic tobacco plants. Plant Phys 135: 254-265

Braun EM, Tsvetkova N, Rotter B, Siekmann D, Schwefel K, Krezdorn N, Plieske J, Winter P, Melz G, Voylokov AV, Hackauf B (2019) Gene expression profiling and fine mapping identifies a gibberellin 2-oxidase gene co-segregating with the dominant dwarfing gene ddw1 in rye (Secale cereale L.). Frontiers Plant Science 10:857

Busov VB, Brunner AM, Strauss SH (2008) Genes for control of plant stature and form. New Phytol 177:589-607

Cantín CM, Arús P, Eduardo I (2018) Identification of a new allele of the Dw gene causing brachytic dwarfing in peach. BMC Res. Notes 11:1-5

Chamberlain D, Hyam R, Argent G, Fairweather G, Walter KS (1996) The genus *Rhododendron*: its classification and synonymy. Royal Botanic Garden, Edinburgh

Chappell M, Robacker C, Jenkins TM (2008) Genetic diversity of seven deciduous azalea species (*Rhododendron spp. section Pentanthera*) native to the eastern United States. J Am Soc Hortic Sci 133:374-382

Cheng S, Zong Y, Chen M, Wang J, Liao M, Liu F (2018) De novo assembly and characterization of *Rhododendron hybridum* hort.(Ericaceae) global transcriptome using Illumina sequencing. Pak J Bot 50:757-761

Choe S, Fujioka S, Noguchi T, Takatsuto S, Yoshida S, Feldmann KA (2001) Overexpression of DWARF4 in the brassinosteroid biosynthetic pathway results in increased vegetative growth and seed yield in Arabidopsis. Plant J 26:573-582

Choudhary S, Thakur S, Najar RA, Majeed A, Singh A, Bhardwaj P (2018) Transcriptome characterization and screening of molecular markers in ecologically important Himalayan species (*Rhododendron arboreum*). Genome 61:417-428

Clouse SD (2002) Brassinosteroid signal transduction: clarifying the pathway from ligand perception to gene expression. Mol. cell 10:973-982

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674-3676

Fang L, Tong J, Dong Y, Xu D, Mao J, Zhou Y (2017) De novo RNA sequencing transcriptome of *Rhododendron obtusum* identified the early heat response genes involved in the transcriptional regulation of photosynthesis. PloS one 12:e0186376

Gonzalez-Garay ML (2016) Introduction to isoform sequencing using pacific biosciences technology (Iso-Seq). In: Wu J. (ed) Transcriptomics and Gene Regulation. Translational Bioinformatics, vol 9. Springer, Dordrecht, pp 141-160

Götz S, García-Gómez JM, Terol J et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420-3435

Guo J-C, Fang S-S, Wu Y et al. (2019). CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. Nucleic Acids Res 47:W516-W522

Hedden P (2003) The genes of the Green Revolution. TRENDS in Genetics 19:5-9

Hill Jr. JL, Hollender CA (2019) Branching out: new insights into the genetic regulation of shoot architecture in trees. Curr. Opin. Plant Biol. 47:73-80

Jia X, Tang L, Mei X, Liu H, Luo H, Deng Y, Su J (2020) Single-molecule long-read sequencing of the full-length transcriptome of *Rhododendron lapponicum L.* Sci Rep 10:1-11

Kim H, Tai TH (2019) Identifying a candidate mutation underlying a reduced cuticle wax mutant of rice using targeted exon capture and sequencing. Plant Breed Biotechnol 7:1-11

Kron KA, Gawen LM, Chase MW (1993) Evidence for introgression in azaleas (Rhododendron; Ericaceae): Chloroplast DNA and morphological variation in a hybrid swarm on Stone Mountain, Georgia. Am. J. Bot. 80:1095-1099

Kurashige Y, Etoh JI, Handa T, Takayanagi K, Yukawa T (2001) Sectional relationships in the genus *Rhododendron* (Ericaceae): evidence from matK and trnK intron sequences. Plant Syst. Evol 228:1-14

Li A, Zhang, J, Zhou Z (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics 15:311.

Li T, Liu X, Li Z, Wan Y, Liu X, Ma H (2018) Development of novel EST-SSR markers for *Rhododendron longipedicellatum* (Ericaceae) and cross-amplification in two congeners. APPS. 6:e01162

Li W, Godzik, A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659

Liang YC, Reid MS, Jiang CZ (2014) Controlling plant architecture by manipulation of gibberellic acid signalling in petunia. Hortic. Res. 1:1-6

Moraes TS, Dornelas MC, Martinelli AP (2019) FT/TFL1: Calibrating plant architecture. Front. Plant Sci. 10:97

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182-W185

Sakamoto T, Morinaka Y, Ohnishi T, Sunohara H, Fujioka S, Ueguchi-Tanaka M, Mizutani M, Sakata K, Takatsuto S, Yoshida S, Tanaka H (2006) Erect leaves caused by brassinosteroid deficiency increase biomass production and grain yield in rice. Nat. Biotechnol 24:105-109

Sanchita, Trivedi PK, Asif MH (2020) Updates on plant long non-coding RNAs (lncRNAs): the regulatory components. Plant Cell, Tissue Organ Cult. 140:259–269

Scheiber SM, Jarret RL, Robacker CD, Newman M (2000) Genetic relationships within *Rhododendron L. section Pentanthera* G. Don based on sequences of the internal transcribed spacer (ITS) region. Sci. Hortic 85:123-135

Schiessl S, Samans B, Hüttel B, Reinhard R, Snowdon RJ (2014) Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species *Brassica napus*. Front. Plant Sci. 5:404

Soza VL, Lindsley D, Waalkes A, Ramage E, Patwardhan RP, Burton, J.N., Adey, A, Kumar, A, Qiu, R, Shendure J, Hall B (2019) The Rhododendron genome and chromosomal organization provide insight into shared whole-genome duplications across the heath family (Ericaceae). Genome Biol. Evol. 11:3353-3371

Sun L, Luo H, Bu D et al. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res 41:e166-e16

Teichmann T, Muhr M (2015) Shaping plant architecture. Front. Plant Sci. 6:233

Umehara M, Hanada A, Yoshida S, Akiyama K, Arite T, Takeda-Kamiya N, Magome H, Kamiya Y, Shirasu K, Yoneyama K, Kyozuka J (2008) Inhibition of shoot branching by new terpenoid plant hormones. Nature 455:195-200

Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. Trends Biotechnol., *23*(1), 48-55.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23:48-55

Vashisth T, Johnson LK, Malladi A (2011) An efficient RNA isolation procedure and identification of reference genes for normalization of gene expression in blueberry. Plant Cell Rep. 30:2167-2176

Wang S, Li Z, Jin W, Fang Y, Yang Q, Xiang J (2018) Transcriptome analysis and identification of genes associated with flower development in *Rhododendron pulchrum* Sweet (Ericaceae). Gene 679:108-118

Wang Y, Li J (2008) Molecular basis of plant architecture. Annu Rev Plant Biol 59:253-279

Wang Y, Robacker CD, Braman SK (1998) Identification of resistance to azalea lace bug among deciduous azalea taxa. J Am Soc Hortic Sci 123:592-597

Wu L, Liu S, Qi H, Cai H, Xu M (2020) Research progress on plant long non-coding RNA. Plants 9:408

Xiao Z, Su J, Sun X, Li C, He L, Cheng S, Liu X (2018) De novo transcriptome analysis of *Rhododendron molle* G. Don flowers by Illumina sequencing. Genes & genomics 40:591-601

Xing W, Liao J, Cai M, Xia Q, Liu Y, Zeng W, Jin X (2017) De novo assembly of transcriptome from *Rhododendron latoucheae* Franch. using Illumina sequencing and development of new EST-SSR markers for genetic diversity analysis in Rhododendron. Tree Genet. Genomes 13:53

Yadav LK, McAssey EV, Wilde HD (2019) Genetic diversity and population structure of *Rhododendron canescens*, a native azalea for urban landscaping. HortScience 54:647-651

Yamamuro C, Ihara Y, Wu X, Noguchi T, Fujioka S, Takatsuto S, Ashikari M, Kitano H, Matsuoka M (2000) Loss of function of a rice brassinosteroid insensitive1 homolog prevents internode elongation and bending of the lamina joint. The Plant Cell 12:1591-1605

Yang FS, Nie S, Liu H, Shi TL, Tian XC, Zhou SS, Bao YT, Jia KH, Guo JF, Zhao W, An N (2020) Chromosome-level genome assembly of a parent species of widely cultivated azaleas. Nat. Commun. 11:1-13

Zawaski C, Ma C, Strauss SH, French D, Meilan R, Busov VB (2012) PHOTOPERIOD RESPONSE 1 (PHOR1)-like genes regulate shoot/root growth, starch accumulation, and wood formation in Populus. J. Exp. Bot. 63:5623-5634

Zhang Y, Zhang X, Wang Y-H, Shen S-K (2017) De novo assembly of transcriptome and development of novel EST-SSR markers in *Rhododendron rex* Lévl. through illumina sequencing. Front. Plant Sci. 8:1664

Zhou G, Zhu P (2020) De novo transcriptome sequencing of *Rhododendron molle* and identification of genes involved in the biosynthesis of secondary metabolites. BMC Plant Biol. 20:414

Table 4-1. Metrics of assembled transcripts and unigenes

Nucleotide Properties	***Transcripts***	***Unigenes***
Number of Sequences	47,289	24,244
Largest Sequence	7,713	7713
Shortest Sequence	53	81
No. of Sequences > 500 nt	41,574 (87.9%)	22,078 (91.1%)
No. of Sequences > 1000 nt	30,301 (64.1%)	16,752 (69.1%)
Mean Sequence Size	1,385	1,497
N50 Length	1,746	1,851
Sequence %A	26.80	26.97
Sequence %T	28.05	28.12
Sequence %G	24.08	23.76
Sequence %C	21.07	21.16
Total Nucleotides	**65,500,237**	**36,290,184**

Table 4-2. Composition of SSRs identified through MISA

Repeats	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**>18**	**Total**	**%**
Di-	-	1794	1313	1145	976	790	655	545	502	420	415	380	302	297	922	10456	79.72
Tri-	1364	600	241	115	57	29	19	14	8	9	5	1	-	3	3	2468	18.82
Tetra-	82	16	3	5	1	-	-	-	-	-	-	-	-	-	-	107	0.82
Penta-	31	7	2	-	-	-	-	-	-	-	-	-	-	-	-	40	0.30
Hexa-	35	4	6	-	-	-	-	-	-	-	-	-	-	-	-	45	0.34

Table 4-3. Identification of *R. canescens* architecture genes

Target gene	Unigenes identified	Length[a] (nt)	Conserved protein domain[b]	BLASTp: greatest similarity	
				E-value	GenBank accession (species)
GA20ox	1	1606	gibberellin 20 oxidase 2	0	PSR93386.1 [*Actinidia chinensis*
GA2ox	5	1324	gibberellin 2-beta-dioxygenase 2	0	NP_001313008.1 [*Nicotiana tabacum*]
GAI	23	3048	GRAS family TF	0	EOY24632.1 [*Theobroma cacao*]
GID1c	48	3003	alpha/beta-hydrolases superfamily	0	EOY30167.1 [*Theobroma cacao*]
PHOR1	78	6557	RING/U-box superfamily protein	0	EOY34643.1 [*Theobroma cacao*]
DWF4	5	2414	cytochrome P450 78A5	0	EOY19585.1 [*Theobroma cacao*]
BRI1	9	4442	serine/threonine-protein phosphatase	0	PSR85447.1 [*Actinidia chinensis*]
TFL1	1	914	PEBP family protein	2.53e-111	ART91289.1 [*Vaccinium corymbosum*]
FT	0	-	-	-	-
FLC	26	1128	K-box region/MADS-box TF	6.04e-157	XP_027096723.1 [*Coffea arabic*
BRC1	32	2297	TCP family TF	0	GAV63848.1 [*Cephalotus follicularis*]
MAX1	0	-	-	-	-
MAX2	17	4129	F-box/LRR-repeat protein	0	PSR93622.1 [*Actinidia chinensis*]
MAX 3/4	8	2152	carotenoid oxygenase family	0	APO15143.1[*Camellia sinensis*]

[a]longest unigene with lowest E-value

[b]SMART Domain Analysis

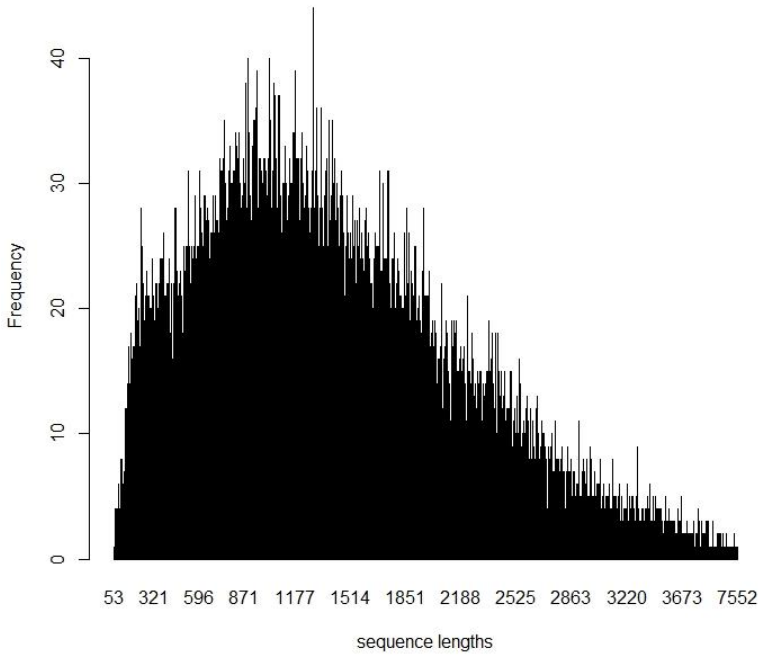


Figure 4-1. Size distribution of Iso-Seq transcripts. Sequence length in base pairs

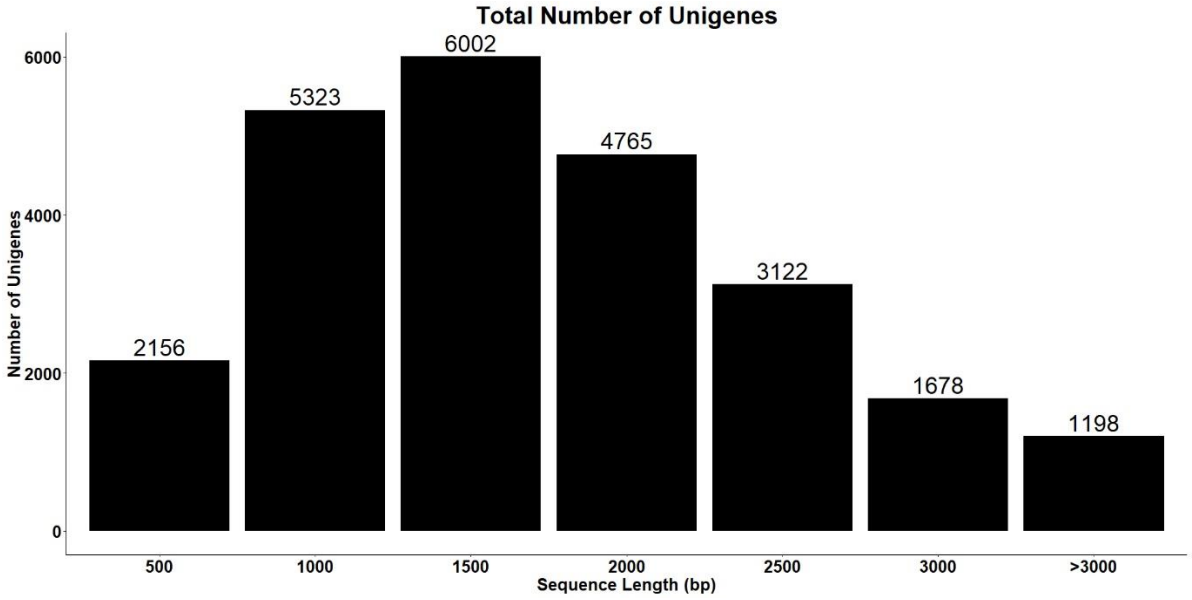


Figure 4-2. Size distribution of assembled unigenes

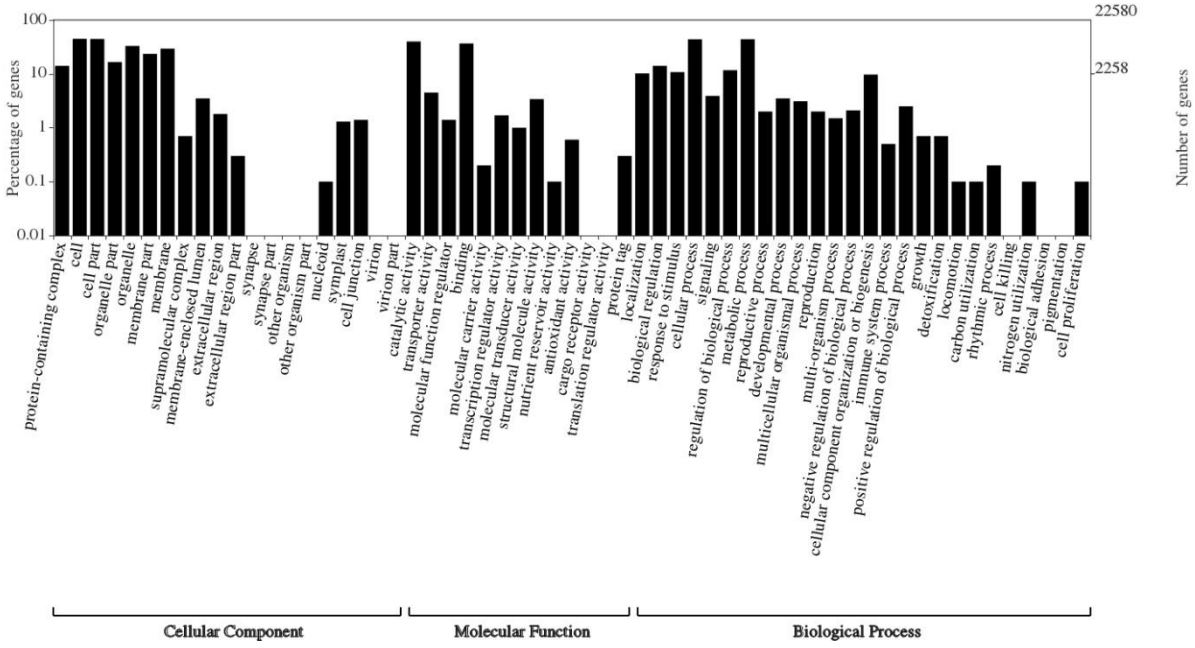


Figure 4-3. Gene ontology classification of unigenes

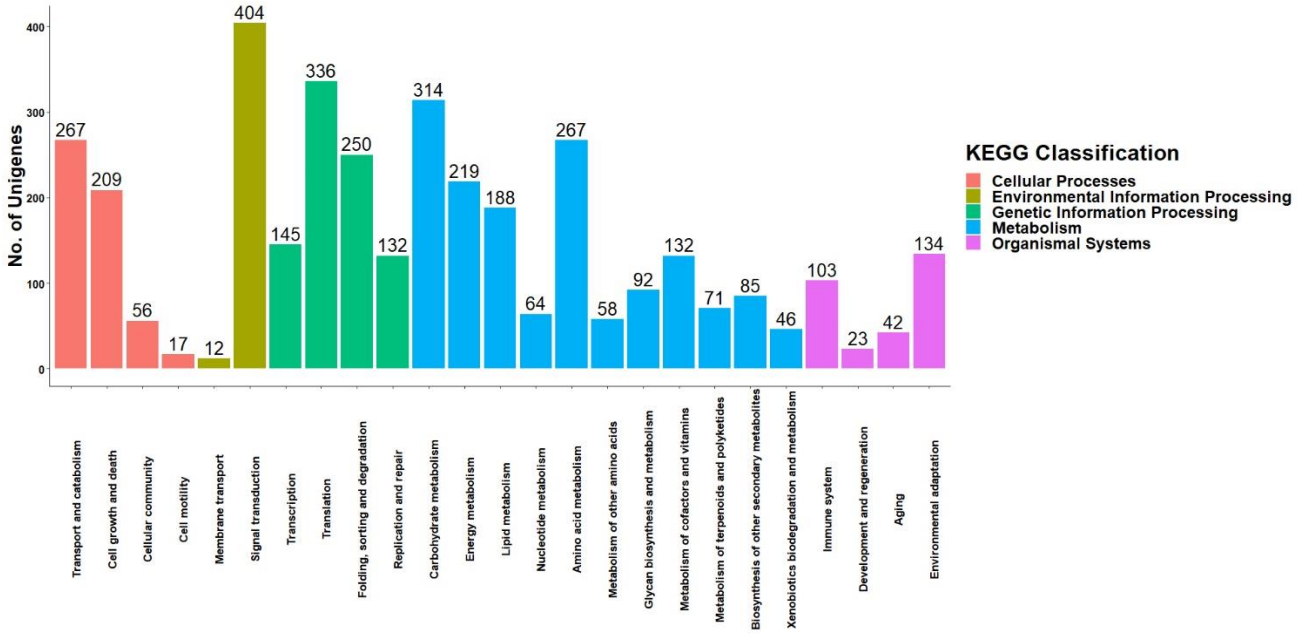


Figure 4-4. Functional classification of unigenes by KEGG pathway analysis

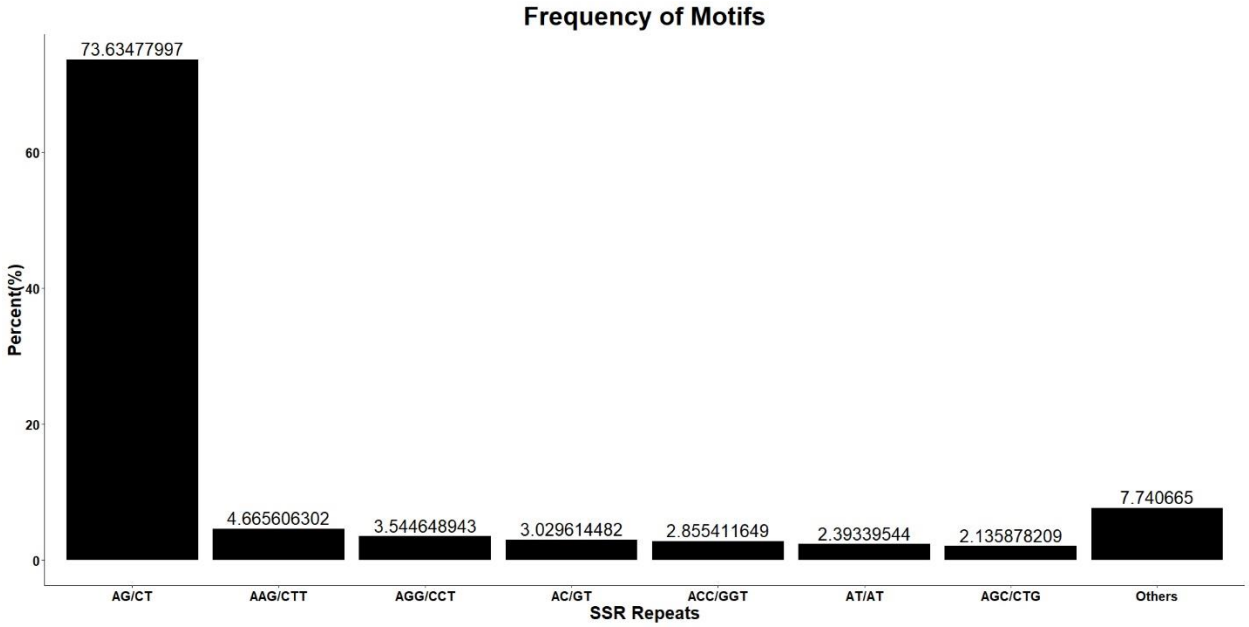


Figure 4-5. Frequency distribution of SSR motifs.

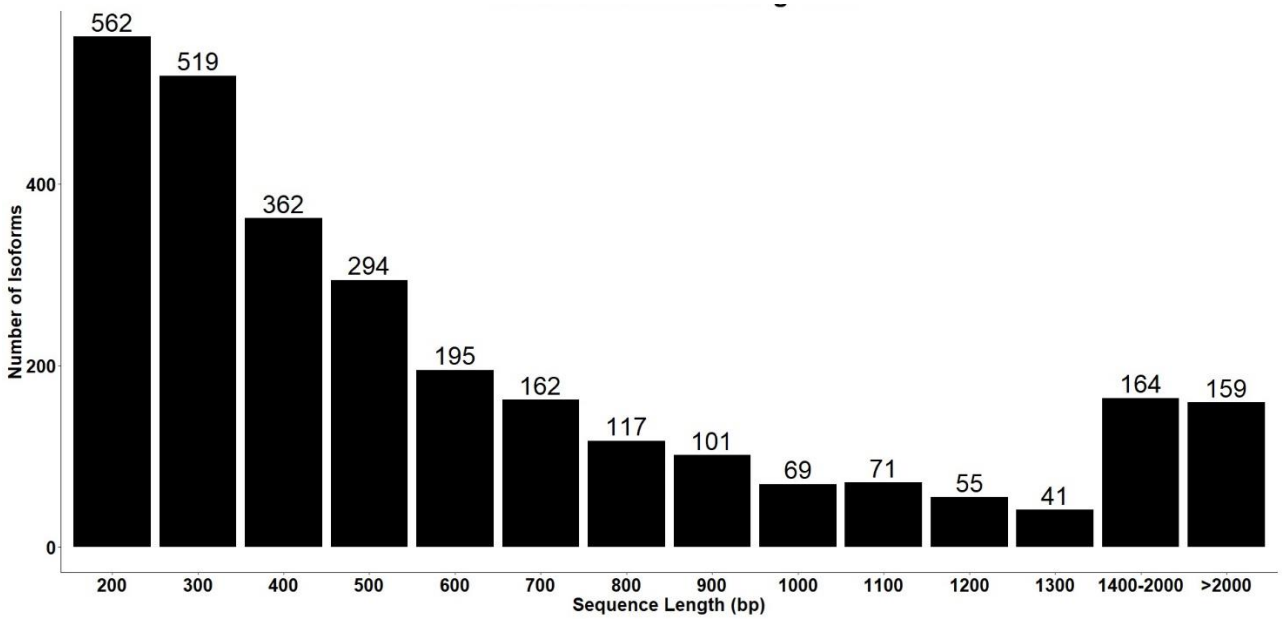


Figure 4-6. Size distribution of lncRNA sequences identified by CNIT analysis

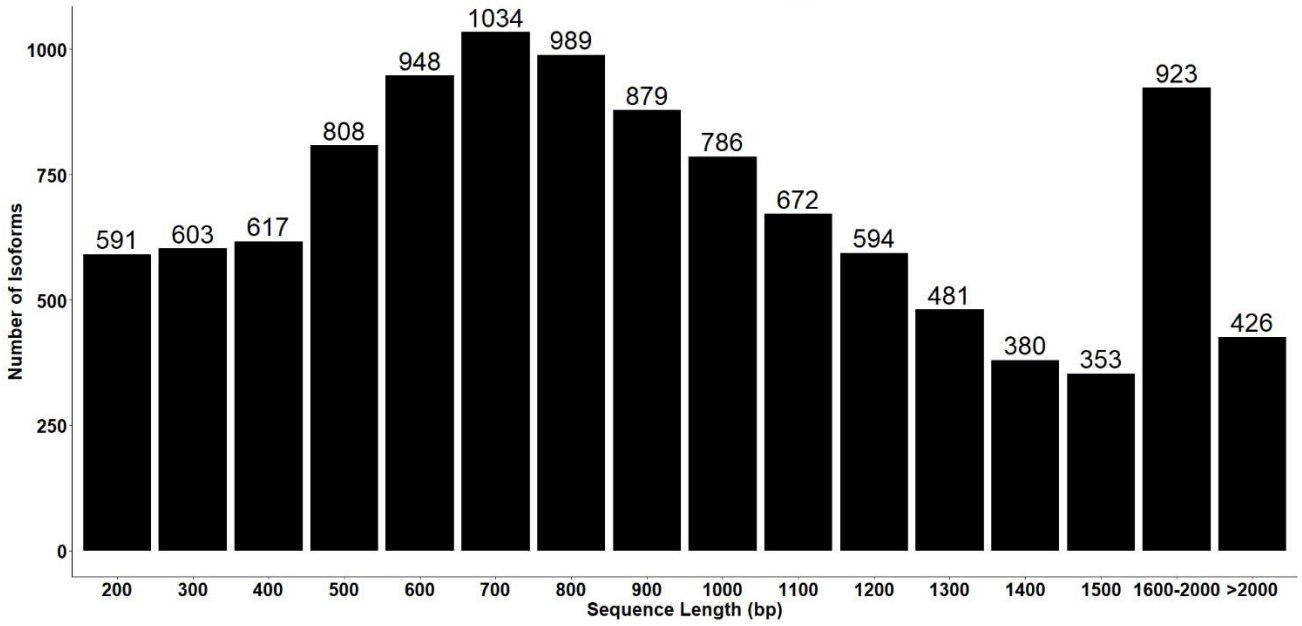


Figure 4-7. Size distribution of lncRNA sequences identified by PLEK analysis

CHAPTER 5

IDENTIFICATION AND BIOINFORMATIC CHARACTERIZATION OF NATURAL VARIANTS IN ARCHITECTURE GENES OF *RHODODENDRON CANESCENS*[1]

[1] L. K. Yadav, and H. D. Wilde. To be submitted to *Tree Genetics & Genomes*.

Abstract

Piedmont azalea (*Rhododendron canescens*) is the most common native azalea in the southeast US, ranging from Texas to North Carolina. It is of interest as a landscaping plant because of its adaptability, lace bug resistance, and early flowering. The use of *R. canescens* in urban settings, however, may be limited by architectural characteristics such as an open growth habit and height up to five meters. The goal of this project was to determine whether genetic variation can be identified for the development of a more compact phenotype. Sequence data from *R. canescens* orthologs of 13 architecture genes were investigated for genetic variation within a large *R. canescens* population (216 genotypes) by exon capture and sequencing. The 216 genotypes also included one suspected dwarf genotype. Variants were called on these genes using Iso-seq database as a reference. We were able to identify 69 high-quality SNPs in 7 plant architecture related genes namely: *BRC1, BRI1, Max2, G1D1c, PHOR1, GAI* and *FLC*. The SNP dataset had nucleotide diversity (π) ranging from 0.0002 to 0.2. Tajima's D value ranged from 4.2 to -0.2, indicating low levels of both low and high-frequency polymorphisms. The impact of the SNPs on protein function was examined using sequence- and structure-based models and 12 missense mutations were identified as deleterious. The 216 genotypes were then geographically divided into three groups based on our previous GBS study. Further, the discriminant analysis of 216 genotypes was able to separate the dwarf genotype from the other 215 genotypes. Based on the SNPs, we found some genotypes similar to the dwarf for the *PHOR1* and *FLC* genes. The biggest separation was seen for *MAX2* gene, for which the dwarf genotype was clearly separated from all three clusters. The *MAX2* gene regulates growth and branching, and the variants identified here might be responsible for dwarfing of this genotype, but further functional analysis is warranted. The mutations "F268S", "P203L" and "P203L" in the *MAX2* gene are supported by

both sequence and structure-based approaches to have deleterious effects. The identification of *R. canescens* accessions with architecture gene mutations can be used to choose potential parents for breeding more compact plants.

Keywords: plant architecture, exon capture, dwarf, Iso-seq, Tajima's D, compact phenotype.

Introduction

Azaleas are diverse ornamental species that are known for their cultural and economic importance around the world. *Rhododendron canescens* (Piedmont Azalea) is a deciduous diploid (2n = 26) that is native to the southeastern USA. *R. canescens* is less popular as a landscaping plant than exotic azalea cultivars that, among other things, have a more compact phenotype. Breeding and development of landscaping plants has been mainly focused on exotic species; hence native azaleas are lagging behind. The development of new azalea cultivars has been carried out by classical breeding approaches that involve long timelines. Molecular breeding provides an opportunity to accelerate the production of native azalea cultivars by identifying genotypes with novel alleles in genes for traits such as height and branching patterns.

Significant progress has been made towards understanding the genetic control of plant stature and form (Busov et al., 2008; Hill Jr & Hollender, 2019; Teichmann & Muhr, 2015; Wang & Li, 2008). Our objective was to identify *R. canescens* orthologs of genes that have been shown to control height or branching across multiple plant species. Genes that encode components of phytohormone biosynthesis and signaling pathways, together with regulatory genes, play a major role in determining plant architecture. For example, the "Green Revolution" genes *GA20 OXIDASE* (*GA20ox*) and *GA INSENSITIVE* (*GAI*) are involved in gibberellin biosynthesis and signaling, respectively (Hedden et al., 2002), and they have been found to have similar roles regulating height in different plant species. The gibberellin biosynthesis gene *GA2 OXIDASE* (*GA2ox)* (Biemelt et al., 2004; Braun et al., 2019) and signaling genes *GA-INSENSITIVE DWARF1* (*GID1*) (Cantín et al., 2018a; Liang et al., 2014) and *PHOTOPERIOD RESPONSIVE1* (*PHOR1)* (Amador et al., 2001; Zawaski et al., 2012) have also been shown to independently control plant height.

Deficiencies in the biosynthesis or signaling pathways of the brassinosteroid phytohormones have been shown to reduce plant stature (Wang & Li, 2008). *DWARF4* (*DWF4*) encodes an enzyme for a rate-limiting step in brassinosteroid biosynthesis and its loss-of-function results in a dwarf phenotype (Choe et al., 2001; Sakamoto et al., 2006). Mutations in *BRASSINOSTEROID INSENSITIVE1* (*BRI1*), a gene involved in signal transduction, can lead to dwarfing in dicots and monocots (Clouse, 2002; Yamamuro et al., 2000). Alternatively, plant stature can be modified by changes in genes that regulate shoot meristem determination, including *TERMINAL FLOWER 1* (*TFL1*), *FLOWERING LOCUS T* (*FT*), and *FLOWERING LOCUS C* (*FLC*) (Moraes et al., 2019; Schiessl et al., 2014). Branching patterns are another major component of plant architecture. Axillary bud development is inhibited by a transcription factor encoded by *BRANCHED1* (*BRC1*), mutations of which can alter branch density (Teichmann & Muhr, 2015). The phytohormone strigolactone (SL) plays a major role in the control of branching in dicots and monocots (Gomez-Roldan et al., 2008; Umehara et al., 2008). Plant architecture is altered by variants of SL biosynthesis genes *MORE AXILLARY BRANCHING1* (*MAX1*), *MAX3,* and *MAX4* and the SL signaling gene *MAX2* (Hedden et al., 2002; Teichmann & Muhr, 2015).

Next generation sequencing-based approaches enable the large-scale screening of populations for rare functional variants in specific genes. For example, EcoTILLING-by sequencing of 768 *Populus nigra* accessions identified non-synonymous single nucleotide polymorphisms (nsSNPs) in five lignin biosynthesis genes (Marroni et al., 2011). A natural mutation in *HCT1* (*hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase1*) was subsequently used for the breeding of poplars with altered lignin composition (Vanholme et al., 2013). Exon capture and sequencing (Capture-Seq) is an alternative approach for high-

throughput discovery of variation in target genes. Whereas EcoTILLING-by-sequencing uses PCR products from target genes, Capture-Seq examines genomic DNA fragments that bind to specific RNA probes. Capture-Seq was used, for example, to screen for nsSNPs in 29 regulatory flowering-time genes of diverse *Brassica napus* accessions (Schiessl et al., 2014). We investigated Capture-Seq as a means to find variation in genes involved in the control of the architecture of *R. canescens.*

To design RNA probes for Capture-Seq, sequence data from *R. canescens* orthologs of these architecture genes was needed. Because the *R. canescens* genome (2C=1.7 pg;(Jones et al., 2007) has not been sequenced, we obtained this information by sequencing the transcriptome (Chapter 3). We have also used genotyping-by-sequencing to examine the genetic diversity and population structure of *R. canescens* genotypes collected across Georgia (Yadav et al., 2019). This study distinguished three subpopulations based on their SNP profiles.

SNPs identified by Capture-Seq in *R. canescens* architecture genes can affect protein function if they are located in coding regions and result in an amino acid change. The number of nsSNPs (non-synonymous) present in a population is low compared to synonymous SNPs. It is therefore critical to identify these SNPs in genes regulating economically important traits. A small subset of these nsSNPs are deleterious and only 20-30% of these substitutions cause a changes in function (Ng & Henikoff, 2006). Therefore, identified SNPs must be filtered for neutral SNPs. The nsSNPs create loss of function of the protein, either by changing the active residue (sequence-based) or by destabilizing the structure of the protein (structure-based) (Bromberg & Rost, 2009). Structural changes of nsSNPs can be evaluated by calculating Gibbs unfolding free energy difference between the wild type and mutant protein, i.e., $\Delta\Delta G = \Delta G_{wildtype} - \Delta G_{mutant}$ (Compiani & Capriotti, 2013). Sequence based methods are less

accurate then structural based methods but are better at providing functional disruption and evolutionary information (Fariselli et al., 2015).

To our knowledge, this study is the first use of targeted capture sequencing in a woody ornamental species to access genetic variation. Here, we present an exome capture study for *R. canescens*. Thirteen genes regulating plant architecture were selected for probe design and hybridization. The genic database was created by Isoform sequencing of *R. canescens* through long read sequencing. We detected the SNPs present in the exons of these genes and further classified them based on their ability to disrupt protein function. We were able to study the genetic diversity present in the wild population of *R. canescens* and compare it to a suspected naturally existing dwarf. Furthermore, we were able to establish the usefulness of targeted capture sequencing technology for studying natural variants in woody ornamentals. The availability of this NGS resource and nsSNPs information will help design breeding pipelines to develop a compact *R. canescens* cultivar. Hence, the objective of this study was: 1) to access the genetic polymorphism present in plant architecture genes of Piedmont azalea genotypes; and 2) to identify genetic variants associated with disruption of protein structure and function.

Materials and Methods

Plant Material and DNA Extraction

Young leaf samples were collected from 216 *R. canescens* genotypes for the capture sequencing study. *R. canescens* samples were collected from 14 counties in Georgia namely: Lamar, Jasper Douglas, Oglethorpe, Morgan, Carrol, Cobb, Oconee, Clarke, Madison, Spalding, Fulton, Dawson and Cook. Some samples were obtained from Callaway Gardens and private collectors. The suspected dwarf genotype sample was from Santa Rosa County, Florida. All the

plant tissues collected were frozen in liquid nitrogen and stored at -80^{0}C. DNA extraction was performed on frozen tissue using the OMEGA E.Z.N.A. HP Plant DNA Kit following the manufacturer's protocol. DNA quality was checked in a NanoDrop spectrophotometer, and a 260/280 ratio of 1.8-2.0 with a minimum concentration of 50 ng/µl were applied as cutoffs for acceptable extractions. The extracted DNA (minimum quantity: 1000 ng) was sent to RAPiD Genomics (Gainesville, FL) for capture sequencing.

Probe design and Capture-Seq

Orthologs of 11 genes (*GA20ox, GA2ox, GAI, GID1c, PHOR1, DWF4, BRI1, FLC, BRC1, MAX2, MAX4*) were identified from a transcriptome assembly developed from young leaves, mature leaves, internodes, terminal buds, axillary buds, flowers, and siliques of a single *R. canescens* plant. RNA probes of approximately 120 bases were designed by RAPiD Genomics to complement the coding regions of the target genes. Probe coverage of exonic regions was over 96% for all target genes, except for *RcFLC*, which was 86% covered. DNA libraries from the 216 *R. canescens* genomic samples were constructed by RAPiD Genomics. Capture-Seq was carried out by hybridizing DNA libraries with biotinylated probes and the DNA-probe hybrids were isolated by binding to magnetic streptavidin-coated beads. The captured DNA fragments were then eluted from the beads and sequenced with an Illumina HiSeq 2000 DNA analyzer.

Variant Calling

The dataset generated by capture sequencing is pair-end sequencing library. These samples are 2 × 250 nt paired-end libraries. The sequenced Illumina data was filtered and demultiplexed using Trimmomatic (Bolger et al., 2014) program. The FASTA sequences were trimmed and then split into their respective groups of pair end using this program. The next step was to align these filtered reads. The pair-end sequences were matched with their corresponding

pairs. The sequences which lost their partners during the filtering step were treated as single-end reads. Once grouping the sequences was completed, these sequences were aligned to the sequences used to design the probes. BWA-MEM v0.7.17 (Li, 2013) was used to align these raw reads. SAMtools v.1.8 (H. Li et al., 2009) was deployed for sorting of the BAM files. All mapping results were subjected to variant calling via BCFtools (Danecek et al., 2016). The variants generated were further filtered with VCFtools using default settings. Variants with more than 30% missing data were excluded from the study.

Genetic Diversity Analysis and Discriminant analysis of principal components (DAPC)

Variant data was used to study the genetic diversity and relationship between the genotypes. The nucleotide diversity (π) and heterozygosity were estimated using VCFtools (Danecek et al., 2011). The normalized measure of the difference between the observed and expected nucleotide diversity, known as Tajima's D, was also computed in VCFtools. Next, the population cluster based on STRUCTURE (Pritchard et al., 2010) was used perform DAPC analysis on the SNP dataset. DAPC was performed on each gene for 216 genotypes. Three eigenvalues were retained for each DAPC analysis. The DAPC analysis was done in R using Tydalverse (Wickham et al., 2019) and Adegenet (Jombart, 2008b; Jombart & Ahmed, 2011) packages.

Effect of Mutation on Protein

SNPs identified by variant analysis leading to non-synonymous amino acid substitution were further analyzed in two ways. Firstly, sequence-based analysis was done to evaluate the effect of substitution on protein function. The effect of mutations was evaluated using the computational tools PROVEAN (Choi & Chan, 2015), SIFT (Ng & Henikoff, 2003), PredictSNP (Bendl et al., 2014), Polyphen-2 (Adzhubei et al., 2010), Panther (Thomas et al., 2003), MutPred

(B. Li et al., 2009) and MUpro (Cheng et al., 2006) using default parameters (Table 5-1). Based on the majority of the results predicted by these programs, mutation effect was determined.

Secondly, protein structure was analyzed for its stability using Folding Energy Analysis. Gibbs free energy of unfolding (ΔΔG) between wild type and variant proteins, using structure information helped to determine its stability. It can be calculated using the following equation: ΔΔGstability = [G(folded:WT)−G(folded:AAS)] (Teng et al., 2010). Stabilizing mutants were classified in the range: $-0.5 < \Delta\Delta G > 0.5$. Any mutation with a ΔΔG value beyond this range was considered unstable. INPS-MD (Savojardo et al., 2016) and I-MUTANT (Capriotti et al., 2005) were used to predict ΔΔG value for protein structures.

Results

Variant Analysis and Nucleotide Diversity

R. canescens orthologs were identified in the Iso-seq transcriptome database for 10 of the 13 architecture-related genes. Orthologs of *TFL1*, *Max1*, and *Max3* were not found, perhaps because they were not expressed at sufficient levels in the tissues or time points sampled. Probes for the remaining ten genes captured corresponding DNA fragments from 216 *R. canescens* genotypes. Sequence capture was used to characterize genetic variation in the exon regions. Illumina sequencing generated short sequence reads that were then aligned to Iso-seq database and variants were called. Out of the 10 genes, we were successfully able to call variants in 7 genes. Variants that were genotyped in 50% of individuals, a minimum quality score of 30, and a minor allele count of 3 were kept for further analysis. Minor allele frequency (MAF) ranged from 0.05 to 0.50, with a mean of 0.4 (Figure 5-1). In total, we were able to identify 69 high quality SNPs in 7 plant architecture related genes, namely: *BRC1, BRI1, Max2, G1D1c, PHOR1,*

GAI and *FLC*. Single-nucleotide polymorphisms were considerably more frequent than indels. We estimated the nucleotide diversity (π) of the data set and the diversity ranged from 0.0002 to 0.2 (Table 5-2). BRI1 gene had the highest nucleotide diversity at 0.2. Next, we estimated statistical test for neutrality (Tajima's D test) on the variants. Tajima's D estimates the normalized measure of the difference between the observed (π) and expected (θ) nucleotide diversity. The result showed significant variation from neutrality, with most values being positive. Tajima's D value ranged from 4.2 to -0.2. The fact that the majority of the values were positive indicated low levels of both low and high-frequency polymorphisms. We also calculated the ratio of non-synonymous to synonymous nucleotide diversity of SNPs. For the *BRC1* and *G1D1c* genes, no synonymous SNPs were identified. Hence, we could not calculate a ratio of non-synonymous to synonymous nucleotide diversity for them.

Amino Acid Substitution and Protein Stability Analysis

Analysis of the SNPs was done on the exon regions of the seven genes. We successfully identified 69 SNPs in 216 genotypes. Most of the SNPs resulted in synonymous substitutions. Of the nsSNPs, a large percentage had a neutral effect on protein function and structure. The effect of amino acid substitution was analyzed by approaches based on either protein sequence or structure (Table 5-3). Sequence-based analysis compared our sequence to protein databases to predict the effect of amino acid substitution. The consequence of substitution was assessed by the software programs PROVEAN, MuPro, MuPred and PredictSNP, using a combination of tools including SIFT, Polyphen and Panther. The structure based approach utilizes Gibbs free energy ($\Delta\Delta G$) to determine protein stability. The range of DDG values for a Neutral mutation classification is: -0.5=<DDG=<0.5. Any value outside this range makes the protein structure unstable and may create loss of function for the protein sequence.

For the *RcGAI* gene we identified a "R218K" substitution which is predicted to have a neutral effect but seems to have a structural stability issue, as the ΔΔG value was -1.22. The two mutations identified for *RcG1D1c* gene have deleterious effects and this is further supported by stability analysis. In the *RcBRC1* gene, all the mutations had neutral effects and the *RcBRI1* gene mutations were predicted to have deleterious effects. The two genes (*RcFLC* and *RcMAX2*) that had interesting results for DPCA analysis had a majority of deleterious effects. We identified 8 deleterious mutations in these genes which might be responsible for dwarfing of the genotype (Table 5-2). The mutations "F268S", "P203L" and "P233L" in the *RcMAX2* gene are supported by both sequence and structure based approaches to have deleterious effects. "A204F "mutation in the *RcPHOR1* gene and three mutations in the *FLC* gene (T166R, G44E and G113E) have deleterious effects. This finding creates a useful starting point for functional analysis of these genes and validates their role in natural dwarfing of *R. canescens*.

Discriminant Principal Component Analysis of Plant Architecture Genes

Genetic diversity analysis of *R. canescens* previously identified three population clusters in our collection zone. We used this information, and based on the geography, we divided our 216 *R. canescens* genotypes into three clusters. We also had a potential dwarf included in our study which was not included into any clusters. Discriminant analysis was done on 216 genotypes for all the variants (Figure 5-2). The axes represent the first two linear discriminants (LD). Circumferences surround each group, and small solid dots represent individual clones. Linear discriminant (LD1) did not clearly separate the three STRUCTURE clusters but LD2 was able to separate the dwarf genotype from other 216 genotypes in the study.

Next, we performed DPCA analysis for the seven genes under study. Variants called for *BRC1, BRI1, GAI* and *G1D1c* genes did not separate the clusters significantly, the reason being the low level of variants present in the exon regions of these genes. *PHOR1* and *FLC* genes showed some separation between the clusters (Figure 5-3 and 5-4). LD2 for these genes were able to separate Cluster 1 and Cluster 2 from Cluster 3 and the Dwarf. Cluster 3 and the Dwarf separated together. This might be due to geographic similarity as Cluster 3 has genotypes collected from southern Georgia and Dwarf was collected from northern Florida.

The biggest separation was seen for the *MAX2* gene (Figure 5-5). LD2 analysis clearly separated all three clusters from the dwarf genotype. The *MAX2* gene is predominantly involved in photo-morphogenesis of plants, regulating growth and branching (Shen et al., 2007). The variants identified here for *MAX2* might be responsible for dwarfing of this genotype, but further functional analysis is warranted to validate this hypothesis. This study also helped us identify potential genotypes similar to the dwarf for the *FLC* and *PHOR1* genes, which can be used to develop the breeding pipelines in the future.

Discussion

Capture sequencing of targeted genomic regions serves as a powerful tool to identify variants, especially in the coding regions. Identification of natural variants and developing markers that can be used for a breeding pipeline are the major goals of SNP detection. As most ornamental species do not have a reference genome, variant detection and marker development poses a challenge. In this study, we used Capture-Seq to identify variants in genes regulating plant architecture.

We identified mutations that are predicted to have a significant impact on the function of proteins encoded by *GID1c, PHOR1, BRI1, FLC, and MAX2*. Sequence-based and stability-based models were used to determine the effects of missense mutations. Sequence based approaches, in which the novel amino acid residue is compared to a database of wildtype sequences, identified 16 deleterious mutations among the five genes. Next, the stability of these mutations was analyzed by calculating the ΔΔG stability change, which can change certain protein functions. The unfolding energy generated by mutation can destabilize the protein, which ultimately leads to loss of function. A ΔΔG energy outside the range $-0.5 > \Delta\Delta G < 0.5$, generates a large amount of energy that can lead to destabilization of the protein (Table 2). It is known that all functionally disruptive mutations affect structural stability, but not all structurally unstable proteins have a loss of function (DePristo et al., 2005; Steward et al., 2003). We see this in our results, as many mutations have high ΔΔG but have neutral affect. For example, the "F242S" mutation in the *MAX2* gene has a neutral effect, but its ΔΔG is on the higher side (-2.89 kcal/mol). The reason for this anomaly may be due to the requirement of higher ΔΔG energy by some proteins for function disruption. Proteins also have domains; mutations in these domains are more subjected to loss of stability. Mutation in the inter-domain region result results in a stable protein but may cause loss of function (Qu et al., 1997). Therefore, functional validation of these mutations is necessary.

Capture sequencing has been successfully used in plant research (Cronn et al., 2012; Neves et al., 2013), but the lack of a reference genome for many plant species is a constraint. Exon capture sequencing is a cost-effective method to identify natural genetic variations in genes of species with large genomes or no reference genome (Grover et al., 2012; Müller et al., 2015). Exon capture has yielded promising results in species like poplar, eucalyptus and loblolly pine

(Pavy et al., 2016). Probe design and lower capture efficiency have been major challenges in using capture sequencing as a prominent tool for variant calling. Apart from lower efficiency, not knowing the exon boundaries decreases sequencing depth, which could affect SNP detection (Bundock et al., 2012; Zhou & Holliday, 2012). Here, we customized our approach and generated an Iso-seq database containing transcript sequences of our gene of interest. This gave us the transcript sequences needed to design probes.

This is the first study in *R. canescens* to investigate the allelic diversity of specific genes. *R. canescens* is cross pollinating species, resulting in a homogenizing effect on the population (Heuertz et al., 2006). Yadav et al. (2019) have previously found that heterozygosity between populations is lower but higher within the population clusters. This is in line with our findings in the current study, as nucleotide diversity (π) for the seven genes is on the lower side. Cross pollination leads to recombination, but it does not necessarily lead to mutation; hence we detect low nucleotide diversity even in non-synonymous and synonymous nucleotide diversity. The observed lower genetic diversity may be mainly due to being a cross pollinating species, but also might be due to a population bottleneck event in the past (Liu & Burke, 2006).

Out of the seven genes analyzed, we did not detect any synonymous substitution for *BRC1* and *G1D1c*. For the rest of the genes, the ratio of non-synonymous to synonymous nucleotide diversity (Πns/ Πs) was estimated. All the genes except *PHOR1* had a low Πns/ Πs ratio. This shows that these genes were undergoing normal purifying selection. *PHOR1* had a significantly higher Πns/ Πs ratio (12.05). The *PHOR1* gene regulates shoot and root growth, starch accumulation, and wood formation in *Populus* (Zawaski et al., 2012). This significantly higher ratio suggests that *PHOR1* gene may have undergone divergent selection pressure due to abiotic effects on plant growth and development (Milla et al., 2003). Tajima's D was calculated

for the seven genes under study. Most of the values generated were positive, highlighting the fact that there is low level of rare variants present in the population and these gene have gone through balancing selection (Simonsen et al., 1995). This is in line with our finding of low nucleotide diversity, as the *R. canescens* population is going through a homogenizing or balancing phase. This result also supports our previous finding that low level of heterozygosity is present in *R. canescens* subpopulations.

Our previous genetic diversity study of *R. canescens* identified three population clusters in Georgia, USA. This differentiation correlated with geography, with the first cluster (P1) found in northeastern GA (east of Atlanta, GA), the second cluster (P2) found in northwestern GA (west of Atlanta, GA) and the third cluster (P3) found in southern Georgia (Yadav et al., 2019). Using this information, our capture sequencing population (215 genotypes) was divided into three clusters. A suspected naturally occurring dwarf genotype was also introduced in the study. DPCA analysis was conducted on the genotypes. Overall, no clear separation was seen between the three population clusters based on the variants of seven genes. However, the dwarf genotype had a clear separation from the three clusters. Further analysis revealed that, for the genes *MAX2*, *PHOR1* and *FLC*, the dwarf genotype is significantly different form the rest. The alleles present in the dwarf loci for these genes were majority of alternate kind. This finding further supports that suspected dwarf genotype has unique mutations in its plant architecture related genes.

Conclusion

In conclusion, here we present a customized exon capture sequencing technique to identify rare variants in the targeted genes of a genome. We were successfully able to design

probes and identify SNPs for seven architecture related genes using our *R. canescens* Iso-seq database. In 216 genotypes, we identified 69 high-quality SNPs. SNPs discovered here highlighted that the *R. canescens* population is going through a balancing effect. It also helped us to validate our previous finding that there is low genetic diversity among the population clusters. It also helped us to establish the fact that the suspected dwarf genotype is indeed different from the other 215 genotypes and can be potentially used in breeding for compact consumer friendly plants. The amino acid mutation study also helped us understand the effect of mutations on protein function and stability. This study helped us establish capture sequencing as a promising tool for variant detection in species without a reference genome and helped us identify mutations in seven architecture related genes that can be used in breeding pipelines for *R. canescens.*

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods, 7*(4), 248-249.

Amador, V., Monte, E., García-Martínez, J.-L., & Prat, S. (2001). Gibberellins signal nuclear import of PHOR1, a photoperiod-responsive protein with homology to Drosophila armadillo. *Cell, 106*(3), 343-354.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., & Damborsky, J. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol, 10*(1), e1003440.

Biemelt, S., Tschiersch, H., & Sonnewald, U. (2004). Impact of altered gibberellin metabolism on biomass accumulation, lignin biosynthesis, and photosynthesis in transgenic tobacco plants. *Plant Physiology, 135*(1), 254-265.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Braun, E.-M., Tsvetkova, N., Rotter, B., Siekmann, D., Schwefel, K., Krezdorn, N., Plieske, J., Winter, P., Melz, G., & Voylokov, A. V. (2019). Gene expression profiling and fine mapping identifies a gibberellin 2-oxidase gene co-segregating with the dominant dwarfing gene Ddw1 in rye (Secale cereale L.). *Frontiers in plant science, 10*, 857.

Bromberg, Y., & Rost, B. (2009). Correlating protein function and stability through the analysis of single amino acid substitutions. *BMC bioinformatics, 10*(8), 1-9.

Bundock, P. C., Casu, R. E., & Henry, R. J. (2012). Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid crop plant. *Plant biotechnology journal, 10*(6), 657-667.

Busov, V. B., Brunner, A. M., & Strauss, S. H. (2008). Genes for control of plant stature and form. *New Phytologist, 177*(3), 589-607.

Cantín, C. M., Arús, P., & Eduardo, I. (2018). Identification of a new allele of the Dw gene causing brachytic dwarfing in peach. *BMC research notes, 11*(1), 1-5.

Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research, 33*(suppl_2), W306-W310.

Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics, 62*(4), 1125-1132.

Choe, S., Fujioka, S., Noguchi, T., Takatsuto, S., Yoshida, S., & Feldmann, K. A. (2001). Overexpression of DWARF4 in the brassinosteroid biosynthetic pathway results in increased vegetative growth and seed yield in Arabidopsis. *The Plant Journal, 26*(6), 573-582.

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics, 31*(16), 2745-2747.

Clouse, S. D. (2002). Brassinosteroid signal transduction: clarifying the pathway from ligand perception to gene expression. *Molecular cell, 10*(5), 973-982.

Compiani, M., & Capriotti, E. (2013). Computational and theoretical methods for protein folding. *Biochemistry, 52*(48), 8601-8624.

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., & Udall, J. (2012). Targeted enrichment strategies for next-generation plant biology. *American journal of botany*, *99*(2), 291-311.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., & Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

DePristo, M. A., Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*, *6*(9), 678-687.

Fariselli, P., Martelli, P. L., Savojardo, C., & Casadio, R. (2015). INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics, 31*(17), 2816-2821.

Gomez-Roldan, V., Fermas, S., Brewer, P. B., Puech-Pagès, V., Dun, E. A., Pillot, J.-P., Letisse, F., Matusova, R., Danoun, S., & Portais, J.-C. (2008). Strigolactone inhibition of shoot branching. *Nature, 455*(7210), 189-194.

Grover, C. E., Salmon, A., & Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis1. *American journal of botany*, *99*(2), 312-319.

Hedden, P., Phillips, A. L., Rojas, M. C., Carrera, E., & Tudzynski, B. (2002). Gibberellin biosynthesis in plants and fungi: a case of convergent evolution? *Journal of Plant Growth Regulation, 20*(4), 319-331.

Heuertz, M., De Paoli, E., Källman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., & Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]. *Genetics, 174*(4), 2095-2105.

Hill Jr, J. L., & Hollender, C. A. (2019). Branching out: new insights into the genetic regulation of shoot architecture in trees. *Current opinion in plant biology, 47*, 73-80.

Jombart, T. (2008). *Analyses multivariées de marqueurs génétiques: développements méthodologiques, applications et extensions* Lyon 1].

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics, 27*(21), 3070-3071.

Jones, J. R., Ranney, T. G., Lynch, N. P., & Krebs, S. L. (2007). Ploidy levels and relative genome sizes of diverse species, hybrids, and cultivars of Rhododendron. *J. Amer. Rhododendron Soc, 61*(4), 220-227.

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D., & Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics, 25*(21), 2744-2750.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Liang, Y.-C., Reid, M. S., & Jiang, C.-Z. (2014). Controlling plant architecture by manipulation of gibberellic acid signalling in petunia. *Horticulture research, 1*(1), 1-6.

Liu, A., & Burke, J. M. (2006). Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics, 173*(1), 321-330.

Marroni, F., Pinosio, S., Di Centa, E., Jurman, I., Boerjan, W., Felice, N., Cattonaro, F., & Morgante, M. (2011). Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: towards next-generation Ecotilling. *The Plant Journal, 67*(4), 736-745.

Milla, M. A. R., Maurer, A., Huete, A. R., & Gustafson, J. P. (2003). Glutathione peroxidase genes in Arabidopsis are ubiquitous and regulated by abiotic stresses through diverse signaling pathways. *The Plant Journal, 36*(5), 602-615.

Moraes, T. S., Dornelas, M. C., & Martinelli, A. P. (2019). FT/TFL1: Calibrating plant architecture. *Frontiers in plant science, 10*, 97.

Müller, T., Freund, F., Wildhagen, H., & Schmid, K. J. (2015). Targeted re-sequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection. *Tree genetics & genomes, 11*(1), 816.

Neves, L. G., Davis, J. M., Barbazuk, W. B., & Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant Journal, 75*(1), 146-156.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research, 31*(13), 3812-3814.

Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet., 7*, 61-80.

Pavy, N., Gagnon, F., Deschênes, A., Boyle, B., Beaulieu, J., & Bousquet, J. (2016). Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (Picea mariana). *Molecular ecology resources, 16*(2), 588-598.

Pritchard, J. K., Wen, W., & Falush, D. (2010). Documentation for STRUCTURE software: Version 2. *University of Chicago, Chicago, IL*.

Qu, C., Akanuma, S., Moriyama, H., Tanaka, N., & Oshima, T. (1997). A mutation at the interface between domains causes rearrangement of domains in 3-isopropylmalate dehydrogenase. *Protein engineering, 10*(1), 45-52.

Sakamoto, T., Morinaka, Y., Ohnishi, T., Sunohara, H., Fujioka, S., Ueguchi-Tanaka, M., Mizutani, M., Sakata, K., Takatsuto, S., & Yoshida, S. (2006). Erect leaves caused by brassinosteroid deficiency increase biomass production and grain yield in rice. *Nature biotechnology, 24*(1), 105-109.

Savojardo, C., Fariselli, P., Martelli, P. L., & Casadio, R. (2016). INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics, 32*(16), 2542-2544.

Schiessl, S., Samans, B., Hüttel, B., Reinhard, R., & Snowdon, R. J. (2014). Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species Brassica napus. *Frontiers in plant science, 5*, 404.

Shen, H., Luong, P., & Huq, E. (2007). The F-box protein MAX2 functions as a positive regulator of photomorphogenesis in Arabidopsis. *Plant Physiology, 145*(4), 1471-1483.

Simonsen, K. L., Churchill, G. A., & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics, 141*(1), 413-429.

Steward, R. E., MacArthur, M. W., Laskowski, R. A., & Thornton, J. M. (2003). Molecular basis of inherited diseases: a structural perspective. *TRENDS in Genetics, 19*(9), 505-513.

Teichmann, T., & Muhr, M. (2015). Shaping plant architecture. *Frontiers in plant science, 6*, 233.

Teng, S., Srivastava, A. K., & Wang, L. (2010). Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC genomics, 11(2), 1-8.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research, 13*(9), 2129-2141.

Umehara, M., Hanada, A., Yoshida, S., Akiyama, K., Arite, T., Takeda-Kamiya, N., Magome, H., Kamiya, Y., Shirasu, K., & Yoneyama, K. (2008). Inhibition of shoot branching by new terpenoid plant hormones. *Nature, 455*(7210), 195-200.

Vanholme, B., Cesarino, I., Goeminne, G., Kim, H., Marroni, F., Van Acker, R., Vanholme, R., Morreel, K., Ivens, B., & Pinosio, S. (2013). Breeding with rare defective alleles

(BRDA): a natural P opulus nigra HCT mutant with modified lignin as a case study. *New Phytologist*, *198*(3), 765-776.

Wang, Y., & Li, J. (2008). Molecular basis of plant architecture. *Annu. Rev. Plant Biol.*, *59*, 253-279.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Yadav, L. K., McAssey, E. V., & Wilde, H. D. (2019). Genetic Diversity and Population Structure of Rhododendron canescens, a Native Azalea for Urban Landscaping. *HortScience*, *54*(4), 647-651.

Yamamuro, C., Ihara, Y., Wu, X., Noguchi, T., Fujioka, S., Takatsuto, S., Ashikari, M., Kitano, H., & Matsuoka, M. (2000). Loss of function of a rice brassinosteroid insensitive1 homolog prevents internode elongation and bending of the lamina joint. *The Plant Cell*, *12*(9), 1591-1605.

Zawaski, C., Ma, C., Strauss, S. H., French, D., Meilan, R., & Busov, V. B. (2012). PHOTOPERIOD RESPONSE 1 (PHOR1)-like genes regulate shoot/root growth, starch accumulation, and wood formation in Populus. *Journal of experimental botany*, *63*(15), 5623-5634.

Zhou, L., & Holliday, J. A. (2012). Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. *BMC genomics*, *13*(1), 1-12.

Table 5-1. Effect of Mutation Analysis.

Genes	Mutations	Predict SNP	MAPP	PhD-SNP	PolyPhen-1	PolyPhen-2	SIFT	SNAP	PANTHER	PROVEAN Score
GAI	R218K	83%	-	83%	-	-	-	-	-	0.000
GID1C	R177K	83%		89%	-	-	-	-	-	-3.000
	Q577T	87%		77%	-	-	-	-	-	-6.000
Max2	F268S	83%	-	58%	-	-	-	-	-	-8.000
	S84N	60%		72%	-	-	-	62%	-	1.000
	T171N	83%		78%	-	-	-	77%	-	-0.250
	F172L	83%		72%	-	-	-	77%	-	0.000
	R175H	83%		68%	-	-	-	77%	-	1.500
	H188Q	83%		83%	-	-	-	67%	-	0.250
	P203L	83%		78%	-	-	-	67%	-	-3.000
	V209L	83%		83%	-	-	-	61%	-	0.500
	P210S	83%		83%	-	-	-	61%	-	-0.500
	R213H	83%		72%	-	-	-	50%	-	0.000
	S226P	83%		89%	-	-	-	58%	-	-0.250
	L230I	83%		89%	-	-	-	67%	-	-0.250
	P233L	83%		58%	-	-	-	58%	-	-2.500
	F242S	83%		55%	-	-	-	58%	-	-0.750
	L244F	61%		83%	-	-	-	56%	-	0.750
PHOR1	A204F	51%	76%	58%	74%	60%	53%	50%	87%	-2.273
	R209S	83%	64%	68%	67%	64%	76%	77%	65%	-0.113
	D213V	63%	84%	68%	67%	43%	66%	50%	61%	-6.122
	V257D	87%	-	68%	-	-	-	85%	-	-5.000

	T166R	87%	-	68%	-	-	-	-	-	-6.000
FLC	G44E	51%	-	68%	-	-	-	72%	-	-8.000
	G48R	55%	-	83%	-	-	79%	62%	64%	-0.025
	G113E	51%	-	66%	-	-	-	85%	-	-8.000
	S364Y	83%	-	72%	-	-	-	-	-	-6.000
	S373T	83%	-	72%	-	-	-	-	-	-3.000
	I376M	83%	-	72%	-	-	-	-	-	-3.000
	E395K	87%	-	58%	-	-	-	-	-	-4.000
BRI1	L404S	83%	-	51%	-	-	-	-	-	-6.000
	S409L	87%	-	59%	-	-	-	-	-	-6.000
	L12F	83%	77%	68%	67%	68%	81%	50%	-	0.573
	F24L	74%	80%	89%	67%	87%	82%	56%	-	1.650
	E169G	74%	68%	72%	67%	87%	70%	56%	69%	-1.636
	A211T	83%	-	66%	-	-	-	55%	-	-0.100
	T212S	83%	-	83%	-	-	-	77%	-	-0.317
BRC1	H228Y	60%	-	89%	-	-	-	62%	-	-0.392
	F233S	83%	-	66%	-	-	-	71%	-	1.137
	F238S	83%	-	68%	-	-	-	71%	-	0.783

Table 5-2: Estimates of nucleotide diversity and Tajima's D in seven genes involved in Plant Architecture.

Genes	Nucleotide Diversity (Π)	Tajima's D	Nucleotide Diversity (Π) nsSNP	Tajima's D nsSNP	Nucleotide Diversity (Π) sSNP	Tajima's D sSNP	Π_{ns}/Π_{s}
BRC1	2.95E-05	4.01591	2.952E-05	4.01591	na	Na	na
BRI1	0.00034	6.57639	0.0002207	4.92829	0.000119	4.83833	1.849451
FLC	0.000318	9.226841	5.016E-05	3.366871	0.000268	6.895775	0.187422
G1D1c	4.19E-05	0.890158	4.185E-05	0.890158	na	Na	Na
GAI	5.57E-06	-0.03439	5.572E-06	-0.03439	1.49851e-05	1.52823	3.72E-01
Max2	0.000391	5.74113	0.0003063	5.24991	8.47E-05	3.12978	3.61731
PHOR1	6.35E-05	1.318365	5.868E-05	1.927931	4.87E-06	-0.96034	12.05367

Table 5-3: Protein Function and Stability Analysis.

Genes	Mutations	Mutation Effect	Protein Stability (ΔΔG)
GAI	R218K	Neutral	-1.22166
GID1C	R177K	Neutral Deleterious	-1.22166
	Q577T		-0.611821
Max2	F268S	Deleterious	-2.91933
	S84N	Neutral	-0.590824
	T171N	Neutral	-0.919964
	F172L	Neutral	-1.70714
	R175H	Neutral	-1.08482
	H188Q	Neutral	-0.877581
	P203L	Deleterious	-1.07188
	V209L	Neutral	-0.985348
	P210S	Neutral	-1.03979
	R213H	Neutral	-1.08482
	S226P	Neutral	-0.969785
	L230I	Neutral	-0.542999

	P233L	Deleterious	-1.07188
	F242S	Neutral	-2.88525
	L244F	Neutral	-0.808395
PHOR1	A204F	Deleterious	-0.447127
	R209S	Neutral	-0.274325
	D213V	Deleterious	-0.0159666
	V257D	Neutral	-2.70332
FLC	T166R	Deleterious	-0.466094
	G44E	Deleterious	-1.04517
	G48R	Neutral	0.178356
	G113E	Deleterious	-1.09303
BRI1	S364Y	Neutral	-0.478397
	S373T	Neutral	-0.221758
	I376M	Neutral	-0.837455
	E395K	Deleterious	-0.677756
	L404S	Neutral	-2.70222
	S409L	Deleterious	-0.900298
	L12F	Neutral	na
	F24L	Neutral	na
	E169G	Neutral	na

BRC1	A211T	Neutral	na
	T212S	Neutral	na
	H228Y	Neutral	na
	F233S	Neutral	na
	F238S	Neutral	na

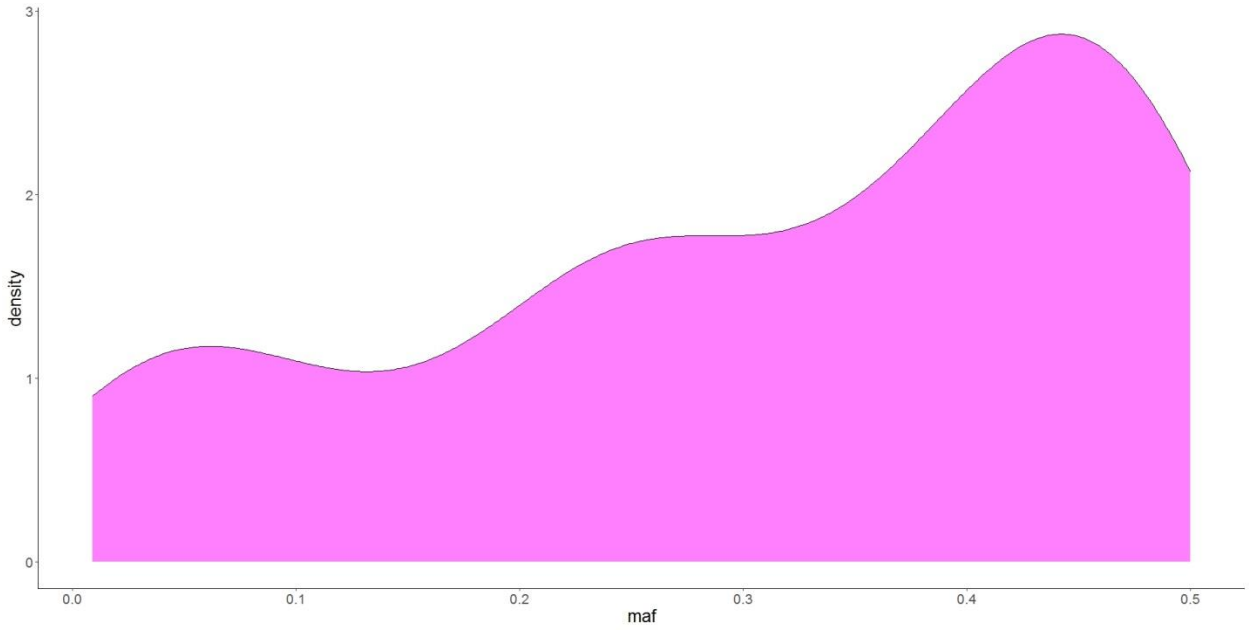


Figure 5-1: Distribution of minor allele frequency (MAF) of 216 *R. canescens* genotypes.

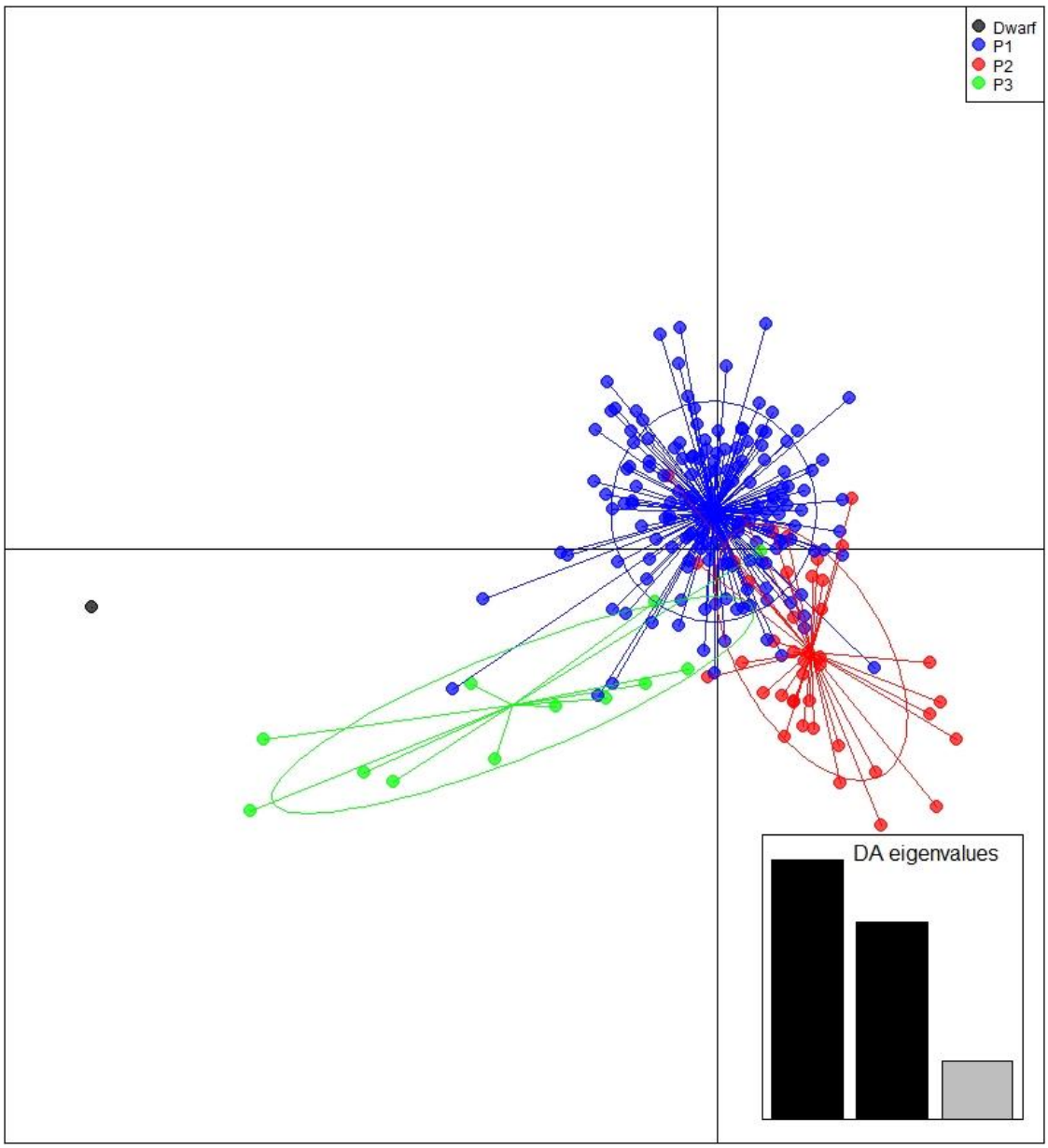


Figure 5-2: DPCA analysis of all variants

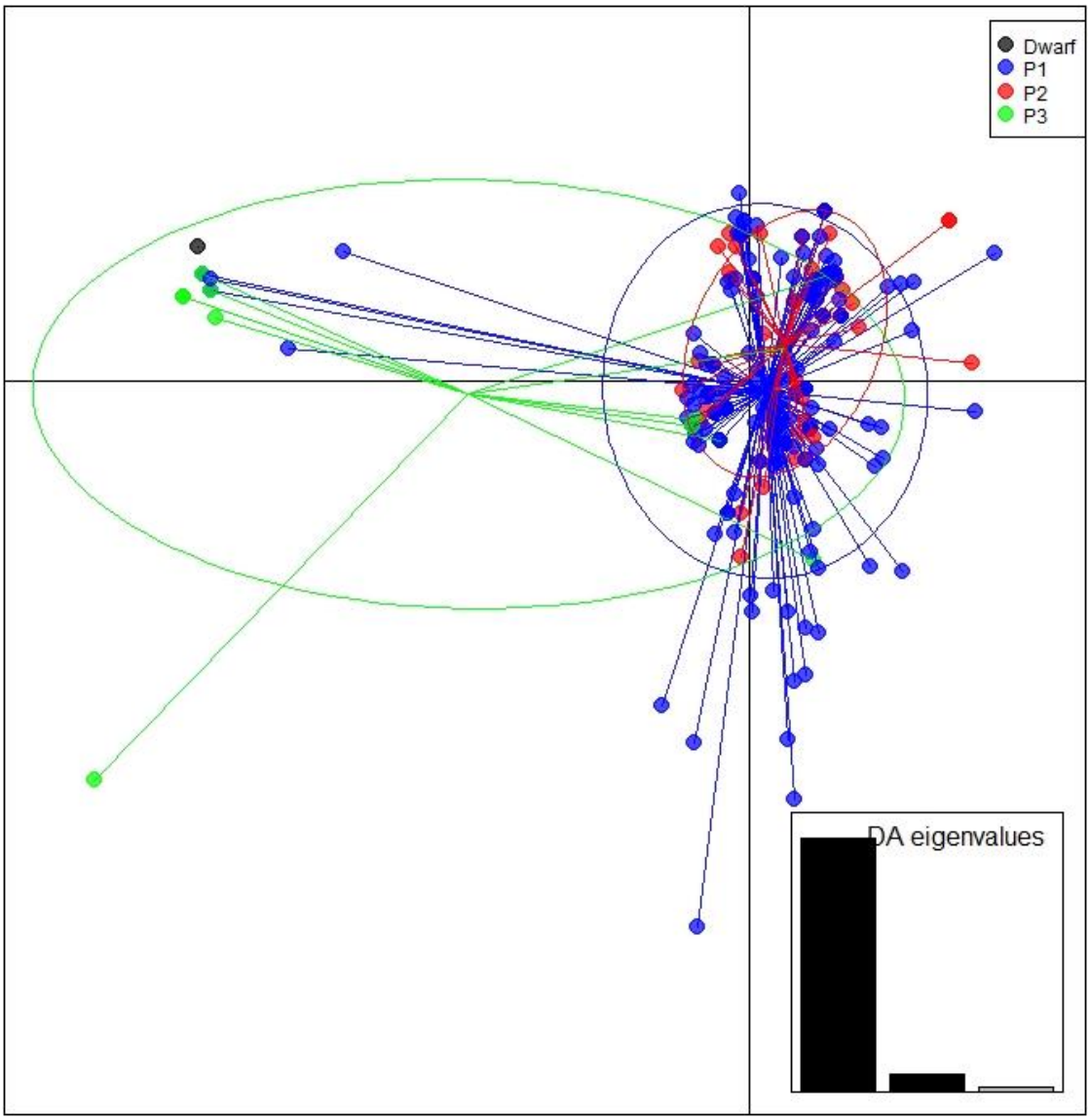


Figure 5-3: DPCA analysis of *PHOR1* gene.

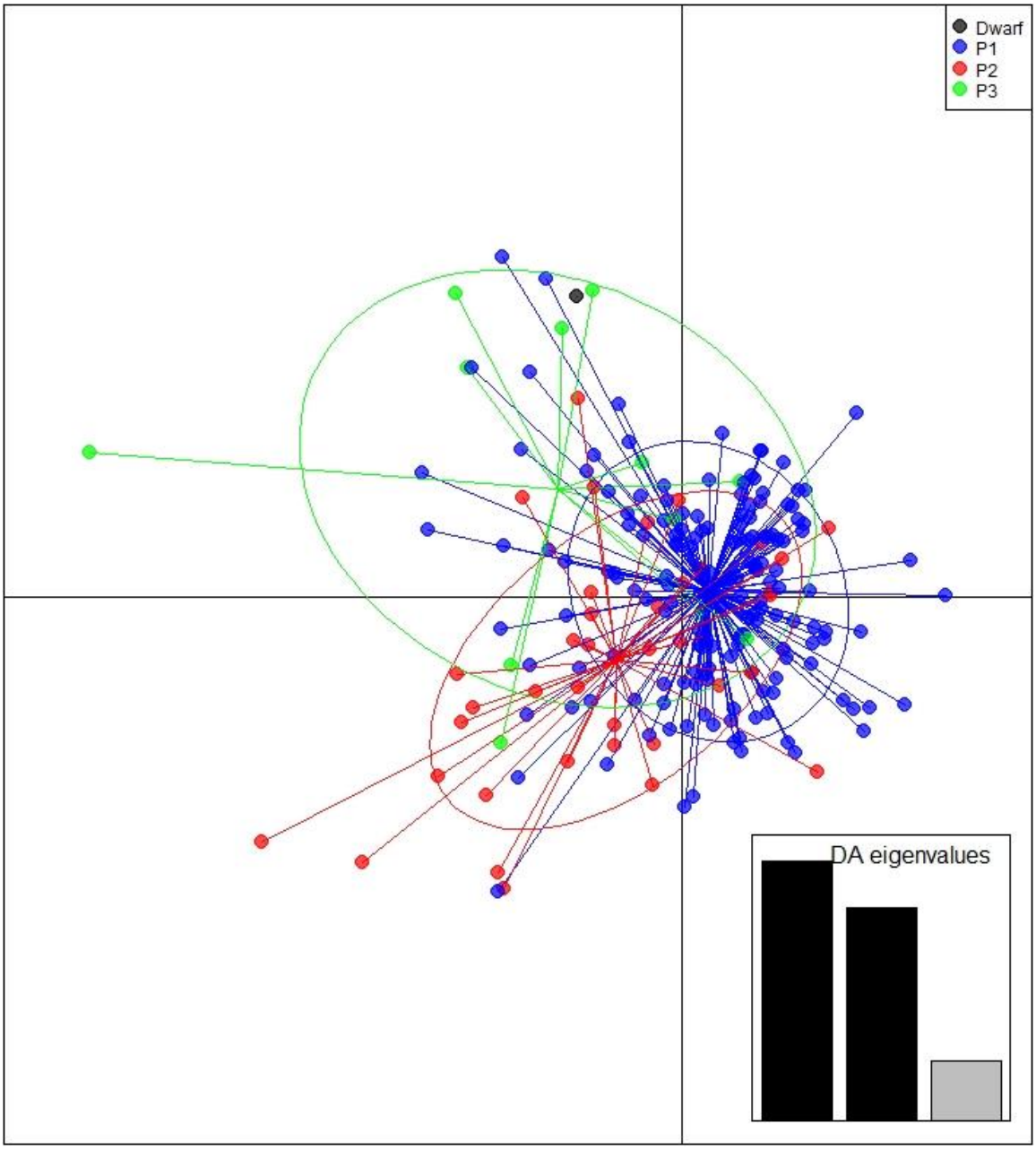


Figure 5-4: DPCA analysis of *FLC* gene.

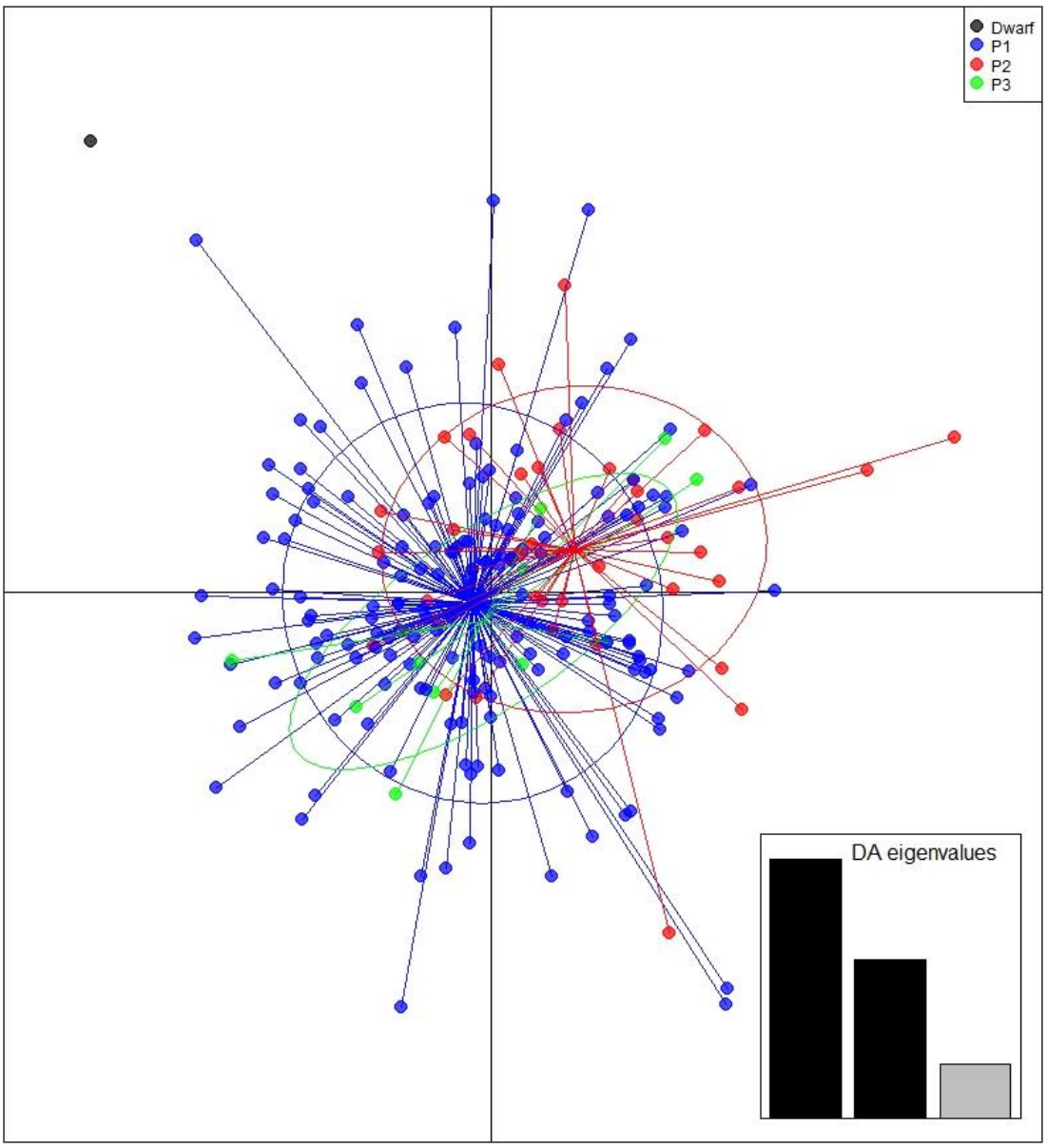


Figure 5-5: DPCA analysis of *MAX2* gene.

CHAPTER 6

CONCLUSIONS

R. canescens GBS Analysis

A genotyping-by-sequencing approach was taken to characterize the genetic structure and diversity of a *Rhododendron canescens* germplasm collection. SNPs were identified by two software platforms, STACKS and GBS-SNP-CROP, and a genome-wide genetic variant file was developed. The genetic variation present in the germplasm collection was examined by STRUCTURE, a model-based Bayesian analysis, and PCA, a distance-based method. Taken together, these analyses indicated that there were three population clusters present among the accessions analyzed.

The STRUCTURE results varied depending on whether SNP data from the GBS-SNP-CROP or the STACKS pipeline were used. STRUCTURE with GBS-SNP-CROP and STACKS data identified three and two population clusters, respectively. These different outcomes may be due to the fact that STACKS analysis involved *de novo* assembly, whereas GBS-SNP-CROP was reference genome-based. The largest population cluster of 48 accessions (P1) was the same for both methods, but the remaining accessions were partitioned by STRUCTURE into two clusters of 22 accessions (P2) and 17 accessions (P3) when using GBS-SNP-CROP data. PCA of SNP data from GBS-SNP-CROP or STACKS supported the presence of three population clusters.

An analysis of *R. canescens* genetic diversity was included in a prior study (Chappell et al. 2008) that examined four *R. canescens* populations (6 accessions each) with AFLP markers.

Similar to that study, our investigation found a low G_{ST} value, indicating that the proportion of diversity between populations was low, while the proportion of diversity within populations was high. Minimal differentiation between *R. canescens* populations was also indicated by the low F_{ST} value. Chappell et al. (2008) suggested that this may be the result of gene flow between populations due to insect pollination. *R. canescens* is known to be pollinated by bumblebees, adrenid bees, butterflies, and hummingbirds. Populations P1 and P2 are geographically close, whereas accessions of P3 had more diverse origins. Introgression from other species of section *Pentanthera* may also have played a role in similarity found among populations.

Genetic markers, including SNPs and cpDNA loci, have been used to examine the genetic structure of a Japanese evergreen azalea species, *R. indicum* (Yoichi et al. 2018). SNPs were identified by multiplexed ISSR genotyping-by-sequencing (MIG-seq). Two genetically distinct lineages were detected, both of which had DNA introgressed from geographically close populations of *R. kaempferi*. MIG-seq was also used to investigate rhododendron plants in a hybrid zone between two natural varieties of *R. japonoheptamerum* (Tamaki et al. 2016). SNP analysis distinguished the varieties and their hybrids and provided the basis for estimating that hybridization occurred 0.4 million years ago. In our investigation, GBS provided a cost-efficient means of generating SNP markers for genetic characterization an *R. canescens* germplasm collection. The high level of genetic diversity found within this collection indicates that screening for allelic variation in genes controlling architecture could be a viable approach to accelerate the breeding of plants with improved form.

Isoform Sequencing of *R. Canescens*

In this study, we were able to generate 24,244 full-length isoform sequences, of which we were able to successfully annotate 16,825 genes. We annotated and characterized these genes into three major categories using BLASTX and blast2go: biological processes (13,102), molecular function (13,685) and cellular components (12,826). We were also successfully able to generate 13,116 SSRs with di-nucleotide motif bring the most abundant (79%). The motif AG/TC accounted for 74% of the total SSRs found. For our plant architecture study, we were successfully able to identify 257 transcripts coding for enzymes related to our 13 target genes. Further analysis of these hits revealed that they were homologues of these genes across several species in NCBI database. In conclusion, through our work we have developed a publicly available computational transcriptomic resource for *R. canescens* containing different data types including full-length transcriptome, annotation, SSRs information and a list of potential LncRNA. We were able to identify transcripts regulating plant architecture for probe design, which was the major objective of this sequencing project. This research will also not only provide a robust pipeline to analyze the Iso-seq data in the absence of a reference genome but will also create the genomic resources that can directly be used in the development of *Rhododendron* sp. through molecular breeding, capture sequencing and genetic studies.

Capture Sequencing of *R. canescens*

Here we present a customized exon capture sequencing technique to identify rare variants in the targeted genes of a genome. We were successfully able to design probes and identify SNPs for seven architecture related genes using our *R. canescens* Iso-seq database. In 216 genotypes, we identified 69 high quality SNPs. SNPs discovered here indicated that the *R. canescens*

population is going through balancing effect. It also helped us to validate our previous finding that there is low genetic diversity among the population clusters. It also helped us to establish the fact that the suspected dwarf genotype is indeed different from the other 215 genotypes and can be potentially used in breeding for compact consumer friendly plants. The amino acid mutation study also helped us understand the effect of mutations on protein function and stability. This study helped us establish capture sequencing as a promising tool for variant detection in species without a reference genome and helped us identify mutations in seven architecture related genes that can be used in breeding pipelines for *R. canescens.*

References

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., & Reddy, A. S. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature communications, 7*(1), 1-11.

Aguilar-Martínez, J. A., Poza-Carrión, C., & Cubas, P. (2007). Arabidopsis BRANCHED1 acts as an integrator of branching signals within axillary buds. *The Plant Cell, 19*(2), 458-472.

Barkley, N., & Wang, M. (2008). Application of TILLING and EcoTILLING as reverse genetic approaches to elucidate the function of genes in plants and animals. *Current genomics, 9*(4), 212-226.

Booker, J., Auldridge, M., Wills, S., McCarty, D., Klee, H., & Leyser, O. (2004). MAX3/CCD7 is a carotenoid cleavage dioxygenase required for the synthesis of a novel plant signaling molecule. *Current biology, 14*(14), 1232-1238.

Bräutigam, A., & Gowik, U. (2010). What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology, 12*(6), 831-841.

Busov, V. B., Brunner, A. M., & Strauss, S. H. (2008). Genes for control of plant stature and form. *New Phytologist, 177*(3), 589-607.

Cantín, C. M., Arús, P., & Eduardo, I. (2018). Identification of a new allele of the Dw gene causing brachytic dwarfing in peach. *BMC research notes, 11*(1), 386.

Casal, J., Sanchez, R., & Deregibus, V. (1986). The effect of plant density on tillering: the involvement of R/FR ratio and the proportion of radiation intercepted per plant. *Environmental and Experimental Botany, 26*(4), 365-371.

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.

Chappell, M., Robacker, C., & Jenkins, T. M. (2008). Genetic diversity of seven deciduous azalea species (Rhododendron spp. section Pentanthera) native to the eastern United States. *Journal of the American society for horticultural science, 133*(3), 374-382.

Clouse, S. D. (2002). Brassinosteroid signal transduction: clarifying the pathway from ligand perception to gene expression. *Molecular cell, 10*(5), 973-982.

Clouse, S. D., & Sasse, J. M. (1998). Brassinosteroids: essential regulators of plant growth and development. *Annual review of plant biology, 49*(1), 427-451.

Doebley, J., Stec, A., & Hubbard, L. (1997). The evolution of apical dominance in maize. *Nature, 386*(6624), 485.

Druley, T. E., Vallania, F. L., Wegner, D. J., Varley, K. E., Knowles, O. L., Bonds, J. A., Robison, S. W., Doniger, S. W., Hamvas, A., & Cole, F. S. (2009). Quantification of rare allelic variants from pooled genomic DNA. *Nature methods, 6*(4), 263-265.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics, 164*(4), 1567-1587.

Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Resources, 7*(4), 574-578.

Fu, Y.-B., Cheng, B., & Peterson, G. W. (2014). Genetic diversity analysis of yellow mustard (Sinapis alba L.) germplasm based on genotyping by sequencing. *Genetic resources and crop evolution, 61*(3), 579-594.

Fujioka, S., & Yokota, T. (2003). Biosynthesis and metabolism of brassinosteroids. *Annual review of plant biology, 54*(1), 137-164.

Galle, F. (1967). Native and Some Introduced Azaleas for Southern Gardens-Kinds and Culture. *AMERICAN HORTICULTURAL MAGAZINE, 46*(1), 13-&.

Gilchrist, E. J., & Haughn, G. W. (2005). TILLING without a plough: a new method with applications for reverse genetics. *Current opinion in plant biology, 8*(2), 211-215.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS one, 9*(2), e90346.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., & Russ, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology, 27*(2), 182-189.

Grover, C. E., Salmon, A., & Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis1. *American journal of botany, 99*(2), 312-319.

Henikoff, S., Till, B. J., & Comai, L. (2004). TILLING. Traditional mutagenesis meets functional genomics. *Plant Physiology, 135*(2), 630-636.

Hisamatsu, T., King, R. W., Helliwell, C. A., & Koshioka, M. (2005). The involvement of gibberellin 20-oxidase genes in phytochrome-regulated petiole elongation of Arabidopsis. *Plant Physiology, 138*(2), 1106-1116.

Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources, 9*(5), 1322-1332.

Iwata, H., Gaston, A., Remay, A., Thouroude, T., Jeauffre, J., Kawamura, K., Oyant, L. H. S., Araki, T., Denoyes, B., & Foucher, F. (2012). The TFL1 homologue KSN is a regulator of continuous flowering in rose and strawberry. *The Plant Journal, 69*(1), 116-125.

Karlgren, A., Gyllenstrand, N., Källman, T., Sundström, J. F., Moore, D., Lascoux, M., & Lagercrantz, U. (2011). Evolution of the PEBP gene family in plants: functional diversification in seed plant evolution. *Plant Physiology*, pp. 111.176206.

Kim, H., & Tai, T. H. (2019). Identifying a candidate mutation underlying a reduced cuticle wax mutant of rice using targeted exon capture and sequencing. *Plant Breeding and Biotechnology*, *7*(1), 1-11.

Kosugi, S., & Ohashi, Y. (1997). PCF1 and PCF2 specifically bind to cis elements in the rice proliferating cell nuclear antigen gene. *The Plant Cell*, *9*(9), 1607-1619.

Kosugi, S., & Ohashi, Y. (2002). DNA binding and dimerization specificity and potential targets for the TCP protein family. *The Plant Journal, 30*(3), 337-348.

Kron, K. A., Gawen, L. M., & Chase, M. W. (1993). Evidence for introgression in azaleas (Rhododendron; Ericaceae): Chloroplast DNA and morphological variation in a hybrid swarm on Stone Mountain, Georgia. *American Journal of Botany, 80*(9), 1095-1099.

Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. *International journal of plant genomics, 2012*.

Li, Y., Dai, C., Hu, C., Liu, Z., & Kang, C. (2017). Global identification of alternative splicing via comparative analysis of SMRT-and Illumina-based RNA-seq in strawberry. *The Plant Journal*, *90*(1), 164-176.

Liu, B., Watanabe, S., Uchiyama, T., Kong, F., Kanazawa, A., Xia, Z., Nagamatsu, A., Arai, M., Yamada, T., & Kitamura, K. (2010). The soybean stem growth habit gene Dt1 is an ortholog of Arabidopsis TERMINAL FLOWER1. *Plant Physiology, 153*(1), 198-210.

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., Buckler, E. S., & Costich, D. E. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS genetics, 9*(1), e1003215.

Marroni, F., Pinosio, S., Di Centa, E., Jurman, I., Boerjan, W., Felice, N., Cattonaro, F., & Morgante, M. (2011). Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: towards next-generation Ecotilling. *The Plant Journal, 67*(4), 736-745.

McNally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., Zeller, G., Clark, R. M., Hoen, D. R., & Bureau, T. E. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences, 106*(30), 12273-12278.

Melo, A. T., Bartaula, R., & Hale, I. (2016). GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC bioinformatics, 17*(1), 29.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics, 11*(1), 31.

Müller, T., Freund, F., Wildhagen, H., & Schmid, K. J. (2015). Targeted re-sequencing of five Douglas-fir provenances reveals population structure and putative target genes of positive selection. *Tree genetics & genomes, 11*(1), 816.

Out, A. A., van Minderhout, I. J., Goeman, J. J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P. E., & Tops, C. M. (2009). Deep sequencing to reveal new variants in pooled DNA samples. *Human mutation, 30*(12), 1703-1712.

Pavy, N., Gagnon, F., Deschênes, A., Boyle, B., Beaulieu, J., & Bousquet, J. (2016). Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (Picea mariana). *Molecular ecology resources, 16*(2), 588-598.

Phillips, A. L., Ward, D. A., Uknes, S., Appleford, N. E., Lange, T., Huttly, A. K., Gaskin, P., Graebe, J. E., & Hedden, P. (1995). Isolation and expression of three gibberellin 20-oxidase cDNA clones from Arabidopsis. *Plant Physiology, 108*(3), 1049-1057.

Pnueli, L., Carmel-Goren, L., Hareven, D., Gutfinger, T., Alvarez, J., Ganal, M., Zamir, D., & Lifschitz, E. (1998). The SELF-PRUNING gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of CEN and TFL1. *Development, 125*(11), 1979-1989.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Raja, R. B., Agasimani, S., Jaiswal, S., Thiruvengadam, V., Sabariappan, R., Chibbar, R. N., & Ram, S. G. (2017). EcoTILLING by sequencing reveals polymorphisms in genes encoding starch synthases that are associated with low glycemic response in rice. *BMC plant biology, 17*(1), 1-13.

Ratcliffe, O. J., Bradley, D. J., & Coen, E. S. (1999). Separation of shoot and floral identity in Arabidopsis. *Development, 126*(6), 1109-1120.

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics, 13*(5), 278-289.

Rieu, I., Ruiz-Rivero, O., Fernandez-Garcia, N., Griffiths, J., Powers, S. J., Gong, F., Linhartova, T., Eriksson, S., Nilsson, O., & Thomas, S. G. (2008). The gibberellin biosynthetic genes AtGA20ox1 and AtGA20ox2 act, partially redundantly, to promote growth and development throughout the Arabidopsis life cycle. *The Plant Journal, 53*(3), 488-504.

Scharff, R. F. (1912). *Distribution and origin of life in America*. Macmillan.

Scheiber, S., Jarret, R., Robacker, C. D., & Newman, M. (2000). Genetic relationships within Rhododendron L. section Pentanthera G. Don based on sequences of the internal transcribed spacer (ITS) region. *Scientia Horticulturae, 85*(1-2), 123-135.

Schiessl, S., Samans, B., Hüttel, B., Reinhard, R., & Snowdon, R. J. (2014). Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species Brassica napus. *Frontiers in plant science, 5*, 404.

Shannon, S., & Meeks-Wagner, D. R. (1991). A mutation in the Arabidopsis TFL1 gene affects inflorescence meristem development. *The Plant Cell, 3*(9), 877-892.

Sharmaö, T. R., Devanna, B. N., Kiran, K., Singh, P. K., Arora, K., Öain, P., Tiwari, I. Ó., Dubey, H., Saklani, B. K., & Kumari, Ó. (2018). Status and Prospects of Nextógeneration Sequencing Technologies in Crop Plants.

Sorefan, K., Booker, J., Haurogné, K., Goussot, M., Bainbridge, K., Foo, E., Chatfield, S., Ward, S., Beveridge, C., & Rameau, C. (2003). MAX4 and RMS1 are orthologous dioxygenase-like genes that regulate shoot branching in Arabidopsis and pea. *Genes & development, 17*(12), 1469-1474.

Takeda, T., Suwa, Y., Suzuki, M., Kitano, H., Ueguchi-Tanaka, M., Ashikari, M., Matsuoka, M., & Ueguchi, C. (2003). The OsTB1 gene negatively regulates lateral branching in rice. *The Plant Journal, 33*(3), 513-520.

Teichmann, T., & Muhr, M. (2015). Shaping plant architecture. *Frontiers in plant science, 6*, 233.

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C., & Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature communications, 7*(1), 1-13.

Wang, Y., & Li, J. (2008). Molecular basis of plant architecture. *Annu. Rev. Plant Biol., 59*, 253-279.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics, 10*(1), 57-63.

Ward, S. P., Salmon, J., Hanley, S. J., Karp, A., & Leyser, O. (2013). Using Arabidopsis to study shoot branching in biomass willow. *Plant Physiology, 162*(2), 800-811.

Wickland, D. P., & Hanzawa, Y. (2015). The FLOWERING LOCUS T/TERMINAL FLOWER 1 gene family: functional evolution and molecular mechanisms. *Molecular plant, 8*(7), 983-997.

Zhou, L., & Holliday, J. A. (2012). Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. *BMC genomics, 13*(1), 1-12.

Zhu, F. Y., Chen, M. X., Ye, N. H., Shi, L., Ma, K. L., Yang, J. F., Cao, Y. Y., Zhang, Y., Yoshida, T., & Fernie, A. R. (2017). Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. *The Plant Journal, 91*(3), 518-533.

APPENDIX

A. PROPAGATION OF DWARF PIEDMONT AZALEAS THROUGH ROOTED CUTTINGS AND FIELD TEST ESTABLISHMENT

Introduction

Piedmont azalea (*Rhododendron canescens*) is the most widespread native azalea in the southeast US, ranging from Texas to North Carolina. It is of interest as a landscaping plant because of its adaptability, lace bug resistance, and early flowering. During the course of our sample collection for GBS, we were contacted by an azalea expert who had discovered some naturally growing dwarf Piedmont azaleas near Pensacola, Florida. We obtained samples from Florida in order to do field testing of the suspected dwarf plant through vegetative propagation techniques (Figure A-1). This plant could be a potential genetic resource and was used in our plant architecture study through capture sequencing. Depending on the field trial results, these dwarf plants could be recommended for landscape gardening.

Vegetative propagation is a widely used technique to propagate ornamental plants. It is fairly easy technique and can be used to clone the mother plant. Vegetative propagation ensures that the clone will have all the characteristics present in the mother plant. These plants will flower faster compared to seed plantings. There are various methods for azalea vegetative propagation, namely cuttings and layering, etc. For our research, we focused on cuttings. Cuttings are a widely used method to propagate woody or herbaceous plants. They are vegetative parts cut from the mother plant that can regenerate an entire new plant. They can be derived from stems, leaves or roots. Azalea is mainly propagated through seeds and rhizomes but cuttings are

also used as a means of propagation. We used the cutting method of propagation in dwarf Piedmont azalea because the reason for the dwarf trait is unknown. Also, it is the safest way to vegetatively propagate without doing harm to the mother plant.

Materials and Methods

Stem cuttings were taken from a dwarf Piedmont azalea that originated in the Black Water Forest Wildlife Area near Milton, Florida. Our initial attempt to root the cuttings was unsuccessful, so we obtained a plant propagated from the dwarf at Superior Trees nursery (Lee, FL). An azalea rooting protocol from Dr. Carol Robacker was followed. Cuttings were taken at a 45° angle at a leaf node, where the meristematic tissue is present. Cutting at a 45^0 angle near the meristem increased the chance of survival (Figure A-2). It was taken in the early hours of the day during the spring and soaked in mancozeb (Subdue Maxx - fungicide) for 5 min followed by a 24-hour soak in 50 ppm K-IBA. Cuttings were inserted into a coarse medium,either 50-50% peat perlite or 1/4-3/8 inch composted bark at a depth of 2 inches and placed on a mist bench with 40% shade. Plants were misted with water for 10 seconds every six minutes until roots appeared. After roots emerged, the mist setting was changed to 6 seconds every 10 minutes for about 10 days and then to 6 seconds every 20 minutes for the next 10 days. The cuttings were then removed from the mist chamber and placed under 60% shade for the remainder of the season and hand watered as needed. Subdue Maxx and Heritage (azoxystrobin) were applied at 14-day intervals at the lowest label rate. Cuttings were grown in the greenhouse for two years, with dormancy induced over the winters. In March of 2021, the cuttings were planted in a field trial (Figure A-3).

The cuttings were planted at UGA Horticulture farm, Watkinsville, GA. They were planted in six blocks (Table A-1). Each block contained six dwarf plants from rooted cuttings and one control (normal-sizes) Piedmont azalea plant from seelimg.

Results

The height and width of the azaleas planted in the field test were measured three weeks after planting (Table A-2). The average height and width of the dwarf plants was 10.5 in (26.67 cm) and 7.2 (18.28 cm) in respectively. Compared to dwarfs, the average height and width of control is 18.5 in (46.99 cm) and 10 in (25.4 cm) respectively. The height and the width of these dwarf plants will be measured at the end of the growing season for three years in a row and compared to the control. If the height and width of the dwarf plant is consistently smaller then the control, this dwarf genotype will be recommended for commercial use.

Table A-1: Map of the field trial.

Horticulture Farm			Hog Mountain Rd.
A1 A2 A3 A4 A5 A6 Con1	B1 B2 B3 B4 B5 B6 Con2	C1 C2 C3 C4 C5 C6 Con3	
D1 D2 D3 D4 D5 D6 Con4	E1 E2 E3 E4 E5 E6 Con5	F1 F2 F3 F4 F5 F6 Con6	

Table A-2: height and width of Piedmont azalea plants in field Trial.

Row	Height (inches)	Width (inches)	Row	Height (inches)	Width (inches)	Row	Height (inches)	Width (inches)
A1	12	8	**B1**	12	4	**C1**	10	6
A2	10	6.5	**B2**	7	8	**C2**	11	9
A3	11	9.5	**B3**	10	8	**C3**	12	8
A4	10	9	**B4**	11	6	**C4**	10	7
A5	10.5	7.5	**B5**	13.5	9	**C5**	11	9
A6	13	8.5	**B6**	8	6	**C6**	9	8
Control	14	6	**Control**	24	6	**Control**	13.5	12

Row	Height (inches)	Width (inches)	Row	Height (inches)	Width (inches)	Row	Height (inches)	Width (inches)
D1	8.5	8	**E1**	8	5.5	**F1**	8	7
D2	10	6.5	**E2**	7	7	**F2**	11.5	8.5
D3	13.5	6	**E3**	10	6	**F3**	8	6
D4	9.5	7	**E4**	12	6.5	**F4**	11.5	5.5
D5	8.5	6	**E5**	13	8	**F5**	10.5	7.5
D6	10	7	**E6**	12.5	7.5	**F6**	14	7
Control	20	14	**Control**	23	12	**Control**	16.5	10

Figure A-1: Dwarf *Rhododendron canescens*

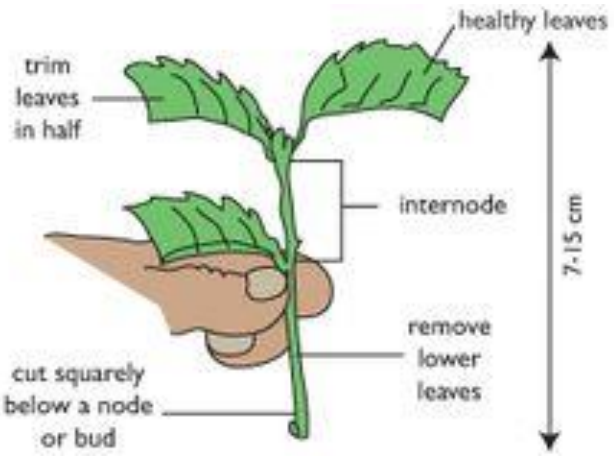


Figure A-2: Diagram of the cutting

Figure A-3: Field Trial Outlay (March, 2021)