2021-08-01

Analyzing genetic variation of the tussock cottongrass among ecotypes along a latitudinal gradient through transcriptomics

Carmen Webster
University of Texas at El Paso

ANALYZING GENETIC VARIATION OF THE TUSSOCK COTTONGRASS AMONG ECOTYPES ALONG A LATITUDINAL GRADIENT THROUGH TRANSCRIPTOMICS

CARMEN PAIGE WEBSTER

Master's Program in Biological Sciences

APPROVED:

Michael L. Moody, Ph.D., Chair

Jonathon E. Mohl, Ph.D.

Kelly S. Ramirez, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

ANALYZING GENETIC VARIATION OF THE TUSSOCK COTTONGRASS AMONG ECOTYPES ALONG A LATITUDINAL GRADIENT THROUGH TRANSCRIPTOMICS

by

CARMEN WEBSTER, B.A.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Biological Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2021

Acknowledgements

I would like to acknowledge and show my immense gratitude to my thesis committee, without whom, this research could not have been completed. To Dr. Michael Moody, for his patience in teaching me about plant genetics and his entirely admirable mentorship throughout this graduate degree. To Dr. Jonathon Mohl, for his willingness to teach me the infinite uses of computer programing with biological data and encouragement while writing my thesis. To Dr. Kelly Ramirez, for her guidance with organizing professional scientific writing and her apt eagerness to help and encourage in any other way needed. Moreover, I'd like to thank Ana Betancourt with the UTEP Border Biomedical Research Center (BBRC) Genomics Analysis Core Facility, for her capability of conducting sequence orders through the COVID-19 pandemic. I'd also like to give special thanks to Elizabeth Stunz, my lab mate in the Plant Evolution lab, who helped me sharpen my wet lab techniques and learn the *E. vaginatum* study site in northern Alaska. To all mentioned above, it was a true pleasure to have been able to learn from you, and words cannot express how grateful I am for each and every one of you.

Abstract

The Arctic is seeing some of the most extreme effects of climate change that induce environmental pressures, including warmer temperatures and longer growing seasons. Due to this, taxa may need to adapt or migrate in order to survive. The long-lived tussock cottongrass, *Eriophorum vaginatum*, is a foundation species in the Arctic, and little is currently known about the genetic constraints that could be playing a role in how this species will respond to the changing climate. Specific gene families that play an important role in signaling genetic pathways related to plant phenology and response to environmental stress are likely to be a key component to the performance of *E. vaginatum* under climate change in the Arctic. The purpose of this study was to investigate the genomics of adaptation, emphasizing the Phytochrome gene family and "Response to Stress" genes. Sanger sequencing was utilized to investigate evidence for selection among the Phytochrome gene family (PHYTA, PHYTB, and PHYTC) along a latitudinal gradient in northern Alaska. Analyses using Bayesian gene tree construction and nonsynonymous and synonymous (K_A/K_S) mutation rates showed that these genes are likely not under selection in relation to North/South ecotypes, but there is allelic variation in these genes and some evidence that is associated with specific populations. The *E. vaginatum* transcriptome, the program SciRoKo 3.4, and several Python scripts were used to identify genes that play a role in stress response and identify SSRs and SNPs associated with these genes for genetic marker development. Primers will be developed for these genetic markers to be used to examine the potential for ecotypic variation with stress response in future selection studies of *E. vaginatum*.

Table of Contents

Acknowledgements iv

Abstract v

Table of Contents vi

List of Tables ix

List of Figures x

Chapter 1: Genetic variation of Phytochrome genes in *Eriophorum vaginatum* along an Arctic latitudinal gradient 1

1.1 Introduction 1

1.1.1 Local Adaptations in Relation to Climate Change 2

1.1.2 Phenology and Fitness in Relation to Climate Change 3

1.1.3 Phytochromes and Light Regime 5

1.1.4 Chapter Aim and Hypothesis 6

1.2 Methods 7

1.2.1 Study Area and Study Organism 7

1.2.2 Sampling 7

1.2.3 Gene Assembly and Primer Design 9

1.2.4 Ecotype Sampling 10

1.2.5 DNA Extractions, PCR, and Sequencing 11

1.2.6 Data Analysis: Gene Trees 13

1.3 Results 15

1.3.1 Phytochrome Gene Primers 15

1.3.2 Sequence Data 17

1.3.3 Gene Trees 17

1.3.4 K_A/K_S Analysis 20

1.3.5 Sliding Window Analysis 21

1.4 Discussion 23

1.4.1 Phytochrome Genetic Variation 23

1.4.2 Future Directions 27

Chapter 2: Genetic Marker Identification in Genes Related to Stress Response 28
2.1 Introduction 28
2.1.1 Genetic Markers 28
2.1.2 Genetic Differentiation 28
2.1.3 Simple Sequence Repeats (SSRs) 29
2.1.4 Single Nucleotide Polymorphisms (SNPs) 30
2.1.5 Genetic Markers in the Arctic Foundation Species *Eriophorum vaginatum* 31
2.1.6 Response to Stress GO Term 31
2.1.7 Chapter Aims 32
2.2 Methods 33
2.2.1 Transcriptome Sampling 34
2.2.2 Isolation of "Response to Stress" Genes 35
2.2.3 Search Parameters for SSRs associated with RTS genes 36
2.2.4 Translations, Alignments, and creation of BED file 37
2.2.5 SNP Identification in RTS Genes 40
2.2.6 Primer Design 40
2.3 Results 41
2.3.1 SSR Detection in RTS Genes 41
2.3.2 Variable SSRs in Transcriptome Samples 42
2.3.3 SNP Detection in RTS Genes 42
2.3.4 Primer Design 43
2.4 Discussion 45
2.4.1 RTS Gene Identification 45
2.4.3 SSRs and SNPs 46
2.4.5 Future Research 49

References 50

Appendix 62
A. Python Scripts 62
A1. Chapter 1 Preliminary Script 62
A2. Chapter 1 SNP Script (C1S) 64
A3. Chapter 2 Translate Script (C2T) 68

Vita 73

List of Tables

Table 1.1 Sampling sites for Chapter 1. The northern ecotypes include PB, CP, SG, TL, and CH, treeline occurs between CH and CF, and the southern ecotypes include CF, GO, NN, EL, and EC. The first subset of populations includes EC, NN, CF, AT, TL, and PB..... 8

Table 1.2 Primers developed for Phytochrome A, B and C genes. Included are the gene ID (transcript) following Mohl et al. (2020), \\ gene length, and primer pairs to amplify the coding region. F = forward primer, R = reverse primer. Due to the length of the genes, coding region sizes, and the occurrence of multiple reading frames found, multiple primers were designed to amplify 600-800 bp regions for all reading frames. Primers in Bold type indicate those that provided high quality sequence data with polymorphisms..... 16

Table 1.3 Results from the Ka/Ks analysis conducted in DNASP 6 for each Phytochrome gene region. 20

Table 1.4 Results from the sliding window analysis conducted in DNASP 6 for each Phytochrome gene region. 21

Table 2.1 Displays location data for each *E. vaginatum* ecotype used in this study. Plants from each of these sites were transplanted into a common garden at the Toolik Field Station (*) in 2012 and 2013 (Mohl et al, 2020)..... 33

Table 2.2 Displays 44 RTS transcripts, or 36 RTS genes, with SSRs or SNPs among ecotypes with gene description, and the types of genetic markers they contain. SSRs are categorized by whether or not they occur in the coding regions. The number of SNPs within each RTS transcript are noted in the last column. 44

List of Figures

Figure 1.1 ArcGIS map displaying latitudinal gradient and locations of ecotypes in northern Alaska (Stunz et al., In Revision). Tree line is represented with dashed black line, blue stars designate reciprocal transplant gardens from previous studies (Bennington et al. 2012; Mohl et al. 2020), and circles depict the populations of study for my first chapter. Orange circles are the reciprocal transplant gardens and yellow are ecotypes. 3

Figure 1.2 Conceptual image describing the rationale behind Chapter 1, starting broad with a full latitudinal gradient and two major ecosystems, and ending with mutations that are likely to change gene function and be important for adaptations for *E. vaginatum*. 6

Figure 1.3 The general outline of the in-house script that was written to parse variable Phytochrome genes from the transcriptome. Starting with a FASTA file containing all the Phytochrome genes from the transcriptome and a VCF file containing mutations (Single Nucleotide Polymorphisms or SNPs) from the samples in the transcriptome, a for loop (portion of image in blue) was applied to identify mutations in all samples. Sample depth (SD) and allele depth (AD) were adjusted to target Phytochromes with only the most prominent mutations. 9

Figure 1.4 ArcGIS map displaying latitudinal gradient and locations of ecotypes in northern Alaska (Stunz et al., In Revision). Tree line is represented with dashed black line, blue stars designate reciprocal transplant gardens from previous studies (Bennington et al. 2012; Mohl et al. 2020), and yellow circles show the ecotypes of study for my first chapter, red circles display the new populations of focus with the given number of samples. 10

Figure 1.5 Possible Sanger sequencing results: (A) shows a poor-quality sequence with multiple nucleotide peaks at one base pair position, meaning that the nucleotide that was called for that position is likely inaccurate. (B) Polymorphism is highlighted in yellow, two alleles, one with a 'T' the other an 'A' is signified with a 'W'. (C) Poor quality sequence where individual nucleotide peaks can be identified, however the sequence quality is too poor to clearly identify polymorphisms. (D) High quality sequence data alignment with a SNP mutation (highlighted in yellow). 13

Figure 1.6 Bayesian gene trees resulting from analysis of the three Phytochrome genes using Geneious 10.0.9. Branch values are posterior probabilities 19

Figure 1.7 Two types of mutations that are present in the Phytochrome genes from populations of *E. vaginatum* along the latitudinal gradient. (A) synonymous, or silent mutation in Phytochrome C, where the variable base pair does not cause a change in the amino acid coded for. (B) a non-synonymous mutation in Phytochrome A, where the variable base pair causes a change in the amino acid 20

Figure 1.8 Data visualization of polymorphic sites in each Phytochrome gene, images created in DNASP 6. 22

Figure 2.1 Chapter 2 flow chart, starting broad with the *E. vaginatum* transcriptome (Mohl et al., 2020), focusing in on the genes of interest, locating genetic markers that could be variable among ecotypes with different environmental pressures, and preparing to determine their utility for future selection studies. 32

Figure 2.2 ArcGIS map displaying the locations of ecotypes on the northern Alaska latitudinal gradient used in developing the *E. vaginatum* transcriptome (Mohl et al., 2020). The dark purple circles denote ecotypes used in the transcriptome and the light purple circle is the location of the transplant garden containing all ecotypes. 33

Figure 2.3 Venn diagram showing the number and overlap of unigenes expressed among the 5 ecotypes utilized for the transcriptome of *E. vaginatum* on the ambient temperature day (13.8°C) in the Toolik Field Station common garden (Mohl et al. 2020). 35

Figure 2.4 Histogram of Gene Ontology classification for *E. vaginatum* unigenes showing the overall percentage of unigenes by their GO term and divided into 3 functional groups. Red bars are 26.6°C day and grey bars are 13.8°C day (Mohl et al. 2020). 35

Figure 2.5 Flow chart for Python script developed to identify SSRs associated with the RTS genes and their coding regions 37

Figure 2.6 Examples of different alignments of RTS transcripts that had multiple SSRs using Clustal Omega. (A) RTS gene that had a long region of repeats, both Perfect and Imperfect. (B) RTS gene that contained multiple SSRs motifs within. (C) RTS gene where a perfect and imperfect SSR were recognized. 39

Chapter 1: Genetic variation of Phytochrome genes in *Eriophorum vaginatum* along an Arctic latitudinal gradient

1.1 INTRODUCTION

Earth's global surface temperature has increased by 0.75°C over the past century, with most warming occurring in the past five decades (Stocker et al., 2013). Making up roughly 14% of the earth's land surface, the arctic ecosystem is currently facing increasingly dramatic effects of climate change, with predictive models estimating up to an 11°C increase by the end of the 21st century (Krinner et al., 2013). Local adaptation is a mechanism that takes place when a specific population of a species evolves to be better adapted to its local environment than other members of the same species that live in other environments or locations. If arctic plants are locally adapted, they will need to migrate or adapt in order to survive the changing climate. However, both will be a challenge within this short time scale, especially for long-lived organisms with little genetic turnover. Little is currently known about the genetics of local adaptations in arctic plants but understanding the mechanisms of genetic constraint can provide a better understanding of how organisms may respond.

The tussock cottongrass (*Eriophorum vaginatum*; Cyperaceae) is a foundation species and dominant plant of the moist acidic tundra of northern Alaska. It may face challenges with migration and new recruitment under climate change as there are ecotypes with some level of homesite adaptations across their range (Bennington et al., 2012; Curasi et al., 2019) and display low rates of seedling establishment, as they are long-lived (>100 years) with low turnover (Fetcher & Shaver, 1982). They may also face competition with other plants, such as shrubs (e.g. *Betula nana*) that could have a better ability to migrate (Curasi et al., 2019). The decline of tussock cottongrass in the warming Arctic could lead to dramatic effects on ecosystem CO_2 flux

responses (Oberbauer et al., 2007). A better understanding of the degree of local adaptations of *E. vaginatum* could be ascertained through understanding the underlying genetic variation in genes that would respond to changing environmental pressures and provide a starting point to understanding the future potential of the species to persist and compete in the new Arctic.

1.1.1 Local Adaptations in Relation to Climate Change

Local adaptations occur due to selective pressures related to environmental variables such as light, temperature, and predation across a species range leading to genetic differentiation. Long term ecological studies show that *E. vaginatum* ecotypes have local adaptations in different parts of their range in northern Alaska that can convey homesite advantage, a particularly important characteristic when discussing long-term fitness of a widespread species under climate change (Bennington et al., 2012; Parker, Tang, Clark, Moody, & Fetcher, 2017). Strong adaptation to local climates could leave arctic plants vulnerable to rapid climate change (Mcgraw et al., 2015). Due to the local adaptation of *E. vaginatum* along a latitudinal gradient (pictured in Figure 1.1) in Alaska, regional populations are described as ecotypes (Bennington et al., 2012; Souther, Fetcher, Fowler, Shaver, & McGraw, 2014). Not all ecotypes recognized in the long-term ecological studies (e.g., Bennington et al. 2012) were recovered in large scale population genomic studies (Stunz et al. In Revision) instead, population genomic markers recognized broader structure among plants of the region with a division between plants north and south of treeline and one population (Eagle Creek) unique from all others. However, transcriptome studies support variation among ecotypes in differential expression of genes (DEG) related to stress response when under heat stress (Mohl, Fetcher, Stunz, Tang, & Moody, 2020) and to a lesser extent among genes related to metabolic processes even when not under stress. Therefore,

there is potential for adaptations that are related to broad ecosystem variation between the Tundra Biome north of treeline and Taiga Biome south of treeline as well as local adaptation for homesite. A better understanding of the genetic mechanisms that drive local adaptations can give insight to the long-term fitness of an organism under climate change (Elmer & Meyer, 2011; Pavey, Bernatchez, Aubin-Horth, & Landry, 2012) and provide clarity of whether local adaptations are 'hard-wired' (based on genetic differences) to their environment.

1.1.2 Phenology and Fitness in Relation to Climate Change

Some aspects of fitness in plants can be measured through phenology. Plant phenology is the vegetative or reproductive life cycle events, usually in response to seasonal variation that can be influenced by environmental

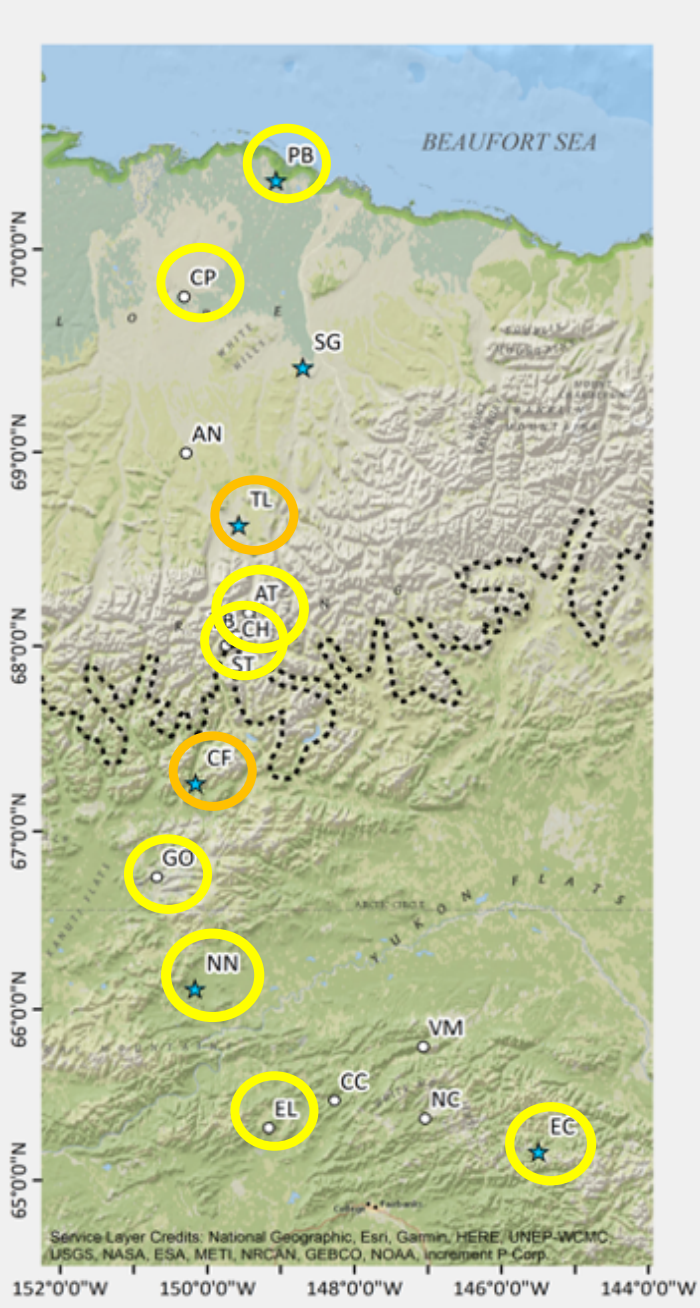


Figure 1.1 ArcGIS map displaying latitudinal gradient and locations of ecotypes in northern Alaska (Stunz et al., In Revision). Tree line is represented with dashed black line, blue stars designate reciprocal transplant gardens from previous studies (Bennington et al. 2012; Mohl et al. 2020), and circles depict the populations of study for my first chapter. Orange circles are the reciprocal transplant gardens and yellow are ecotypes.

pressures. Local adaptations in plant phenology are often related to the seasonal timing of ecological events such as flowering and senescence (Chapin, Shaver, Giblin, Nadelhoffer, & Laundre, 1995; Cleland, Chuine, Menzel, Mooney, & Schwartz, 2007). Phenological processes are likely under genetic control for *E. vaginatum* and other arctic plants. Over time, some plants have adapted to produce and drop leaves based on timing of snowmelt at the beginning of the season (Borner, Kielland, & Walker, 2013; Chapin et al., 1995; Parker et al., 2017). Recent evidence shows that different ecotypes of *E. vaginatum* retain phenological character traits from their homesites when moved. For example, when plants were moved to reciprocal transplant gardens along a latitudinal gradient in the Alaskan Arctic (see Figure 1.1) they retained their homesite leaf senescence timing (Parker et al., 2017). Meaning that southern plants underwent leaf senescence later than plants originating from northern populations no matter which garden they were planted in, utilizing the same senescence timing as if they were still in their homesite (Parker et al., 2017). Therefore, different ecotypes of *E. vaginatum* retain phenological character traits from their homesite when moved. This means that if this characteristic is genetically 'hardwired' in northern ecotypes, they will likely be unable to take advantage of the warmer temperatures and longer growing season already found with climate change in the Arctic (Parker et al., 2017). Phenology response is usually related to Plant Phytochrome genes, differential light receptivity (Ding & Nilsson, 2016; Schmitt, Dudley, & Pigliucci, 1999) and signaling to transcription factors (Kudoh, 2016) that will be discussed further below. Understanding specific mechanisms behind local adaptation and phenological plasticity in terms of genetic differentiation and the changing climate with *E. vaginatum* (Parker et al., 2021) can help us discover the extent of genetically 'hardwired' factors that play a role in adaptations with environmental pressures.

1.1.3 Phytochromes and Light Regime

While there appears to be a clear difference in phenology among ecotypes of *E. vaginatum* (Parker et al., 2017), a knowledge gap resides in identifying which genetic constraints control the timing of senescence in *E. vaginatum* with response to growing season length. Gene families are variants of similar genes that are historically the result of gene or chromosomal duplication, and these genes often have similar functions (Henikoff et al., 1997). Phytochromes are a family of genes that respond to light quantity and quality and have an important role in signaling the genetic pathways related to plant phenology (Halliday & Davis, 2016). They are the leading class of photoreceptors that regulate multiple developmental processes including interpreting photoperiodic signals, appearing in three different gene forms: PHYA, PHYB, and PHYC. These three Phytochrome gene forms have previously displayed playing important roles in photoperiodic responses and environmental senescence of vegetation (Chen et al., 2014; Lin, 2000; Schippers, 2015). Biologically, Phytochrome B plays an inhibitory role in floral initiation (Lin, 2000; Mockler, Guo, Yang, Duong, & Lin, 1999) and Phytochrome C is the least understood member of the Phytochrome family, but it has displayed involvement in transcriptional regulation of both photoperiod and clock genes, as well as a distinct role in the regulation of flowering time (Chen et al., 2014). Phytochromes function by fluctuating between two isomeric forms that respond to red light (Pr), occurring from 650-670nm and far-red light (Pfr) which occurs from 705-740nm. As the different spectrums of light fluctuate, Phytochromes respond to the light ratios. Direct sunlight has a red to far-red (R:FR) light ratio of roughly 1, in the temperate zone of the Arctic the R:FR light ratio changes from 1.1 to 0.8-0.9 (Holmes & Smith, 1977). As light quality changes under canopies and shrub encroachment, the Phytochromes must adjust accordingly. Furthermore, it is likely that Phytochromes could act as

temperature sensors as well, playing an important role for plant performance (Halliday & Davis, 2016). Due to the ecotype locations along the latitudinal gradient in northern Alaska, sensitivity to changes in R:FR light ratio may influence (Parker et al., 2017) the variability in the Phytochrome gene family of photoreceptors, and these variations are likely to be a key genetic component for the performance of *E. vaginatum* in its local environments under climate change in the Arctic. If there are missense mutations in these genes it could relate to functional changes, and this would be particularly pertinent if the changes correspond to North or South broad ecotypes or are isolated to specific populations.

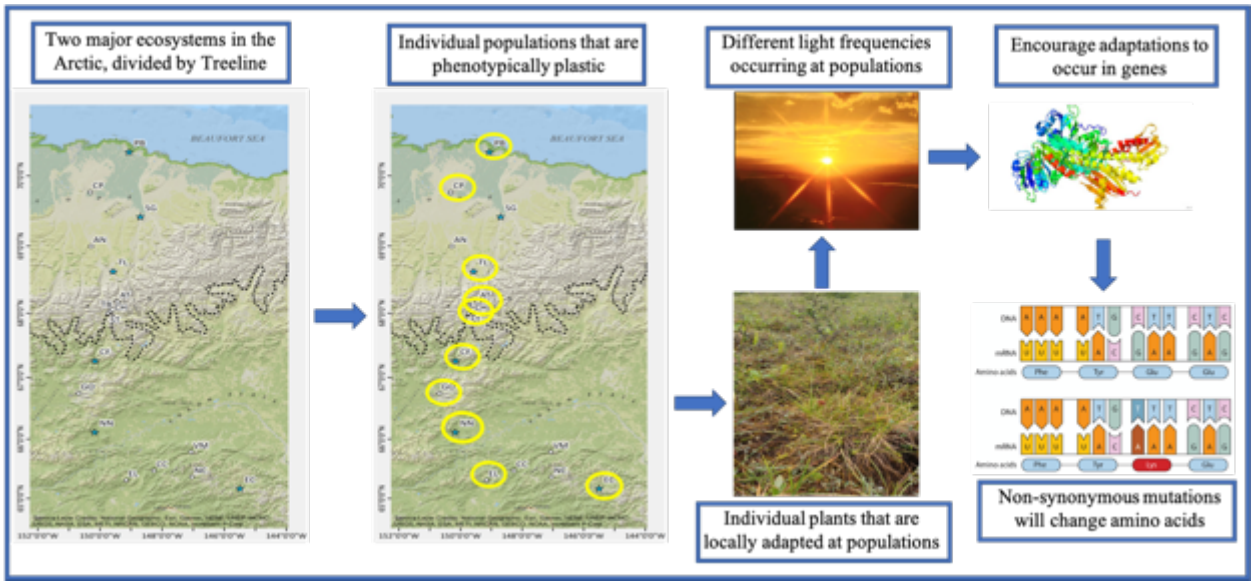


Figure 1.2 Conceptual image describing the rationale behind Chapter 1, starting broad with a full latitudinal gradient and two major ecosystems, and ending with mutations that are likely to change gene function and be important for adaptations for *E. vaginatum*.

1.1.4 Chapter Aim and Hypothesis

Here, the aim of this work is twofold: (1) To identify, isolate and create primers to amplify Phytochrome genes of *E. vaginatum*, and (2) to identify if there is ecotype specific variation in Phytochrome genes that could be related to adaptation for *E. vaginatum* in the

Arctic. This will be addressed by examining variation within members of the Phytochrome gene family between *E. vaginatum* ecotypes from north and south of treeline along a latitudinal gradient in the Alaskan Arctic. *Eriophorum vaginatum* has shown differences in phenology that are correlated with ecotype location (Parker et al., 2017). Given that Phytochrome genes have been found to be directly linked to phenology (Hill & Li, 2016; Hyles, Bloomfield, Hunt, Trethowan, & Trevaskis, 2020; Zhao et al., 2014), I hypothesize that there will be variation in these Phytochrome genes that alter amino acids between ecotypes and correlates with phenological differences between plants from these regions.

1.2 METHODS

1.2.1 Study Area and Study Organism

The study area covers a latitudinal gradient located in northern Alaska, beginning just north of Fairbanks and covering roughly 426 km between northern and central Alaska (Figure 1.1). The treeline, which represents the division of northern and southern ecotypes occurs on the southern slope of the Brooks Range (Figure 1.1). North of treeline is Tundra ecosystem where *E. vaginatum* is a dominant species and found in continuous population; South of treeline is Taiga ecosystem where *E. vaginatum* populations are interspersed in patches among continuous spruce forest. This study area has been utilized to research phenological, ecological, molecular, and environmental effects of the changing climate since the early 1970s (Hobbie & Kling, 2014).

1.2.2 Sampling

Eriophorum vaginatum samples used in this study originate from north of tree line (Prudhoe Bay, Coastal Plain, Toolik Lake, Atigun, and Chandalar), and south of tree line (Coldfoot, Gobbler's Knob, No Name Creek, Elliott Highway, and Eagle Creek) (Figure 1.1) and were taken during the summers of 2015 and 2017. Leaf material was taken from 30 individual

plants and immediately dried in silica gel for subsequent DNA extraction (Table 1.1). Here forward, these leaf sample DNA extractions will be referred to as "DNAs".

Table 1.1 Sampling sites for Chapter 1. The northern ecotypes include PB, CP, SG, TL, and CH, treeline occurs between CH and CF, and the southern ecotypes include CF, GO, NN, EL, and EC. The first subset of populations includes EC, NN, CF, AT, TL, and PB.

Site	Latitude (N), Longitude (W)	Elevation (m)	Vegetation Type
Eagle Creek (EC)	65.4332°, -145.5118°	771	MAT
Elliott Highway (EL)	65.3081°, -149.1230°	720	MAT
No Name Creek (NN)	66.1171°, -150.1676°	167	Tussock bog
Gobbler's Knob (GO)	66.7459°, -150.6862°	520	Muskeg
Coldfoot (CF)	67.2631°, -150.1591°	321	Muskeg
Chandalar (CH)	68.0518°, -149.6115°	968	MAT
Atigun (AT)	68.1730°, -149.4392°	1,063	MAT
Toolik Lake (TL)	68.6292°, -149.5778°	758	MAT
Coastal Plain (CP)	68.9945°, -150.2871°	173	MAT
Prudhoe Bay (PB)	70.3270°, -149.0645°	8	MAT

1.2.3 Gene Assembly and Primer Design

A preliminary in-house script (Chapter 1 Preliminary Script, see appendix A1 for full script) was written using the computer programming language Python to ensure general mutations occurred within the Phytochrome genes in the *E. vaginatum* transcriptome. Once mutations were confirmed, a more in-depth in-house Python script (Chapter 1 SNP Script, see appendix A2 for full script) was developed to locate and build individual assemblies of the 3 Phytochrome genes (Phytochrome A, B, and C) all approximately 4,000 bps in length, from the transcriptome (Table 1.2) (Mohl et al., 2020). A part of the script was also dedicated to parsing out the Phytochrome genes with a sample depth of 8 and allele depth of 60, this was done by calculating the Variant Allele Frequency (VAF) and identifying if there was variation in the

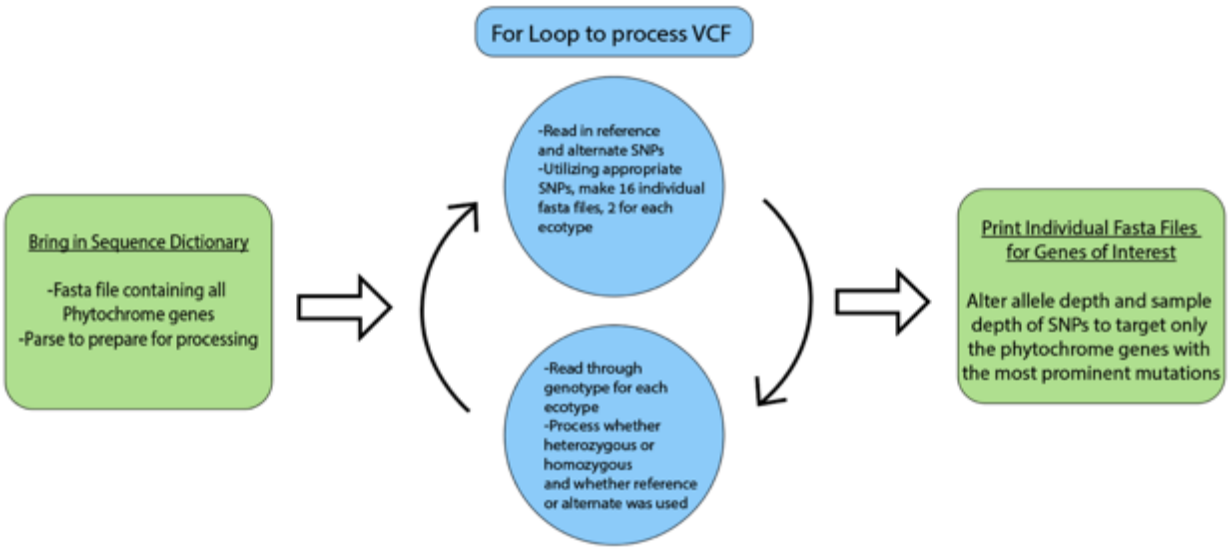


Figure 1.3 The general outline of the in-house script that was written to parse variable Phytochrome genes from the transcriptome. Starting with a FASTA file containing all the Phytochrome genes from the transcriptome and a VCF file containing mutations (Single Nucleotide Polymorphisms or SNPs) from the samples in the transcriptome, a for loop (portion of image in blue) was applied to identify mutations in all samples. Sample depth (SD) and allele depth (AD) were adjusted to target Phytochromes with only the most prominent mutations.

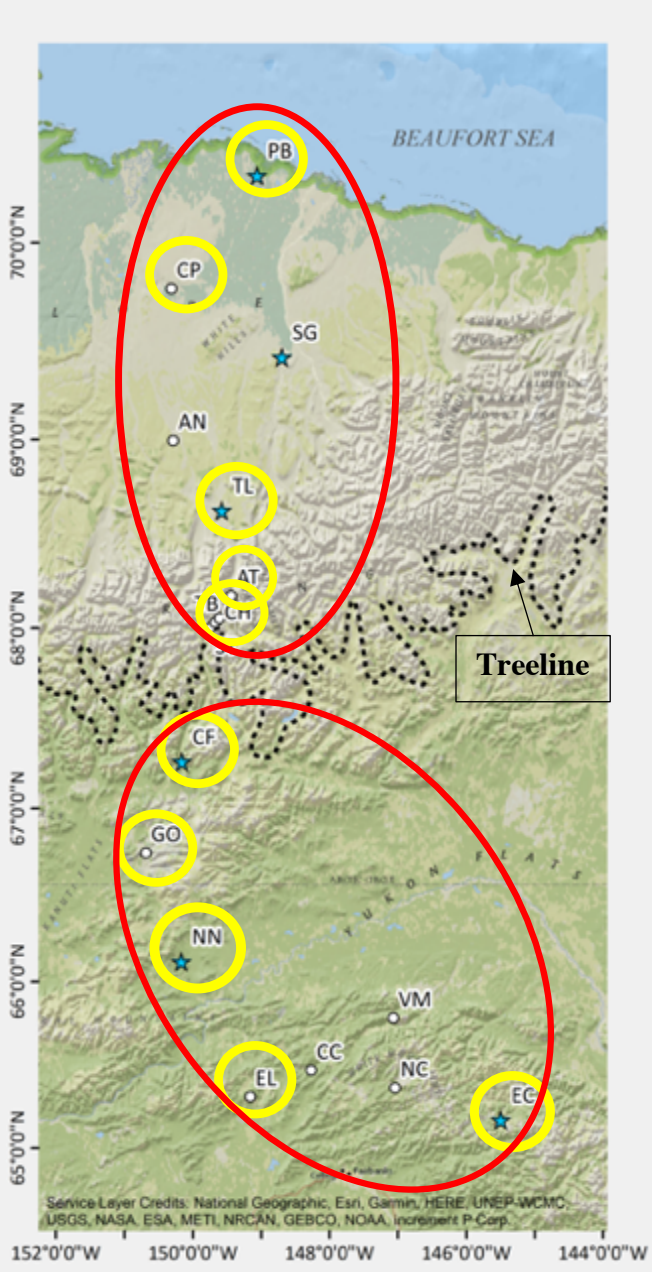


Figure 1.4 ArcGIS map displaying latitudinal gradient and locations of ecotypes in northern Alaska (Stunz et al., In Revision). Tree line is represented with dashed black line, blue stars designate reciprocal transplant gardens from previous studies (Bennington et al. 2012; Mohl et al. 2020), and yellow circles show the ecotypes of study for my first chapter, red circles display the new populations of focus with the given number of samples.

genes among 5 ecotypes used (Mohl et al., 2020) (Figure 1.3). The VAF is important for assessing the different alleles present at a mutation.

The Phytochrome genes were translated using Expasy Translate (web.expasy.org/translate/) and primers were designed to amplify the open reading frames of Phytochrome coding regions using the program Geneious 10.0.9 (Kearse et al., 2012). To design these primers, the following parameters were used: Primer Size of 18bp to 24bp; Temperature melting point of 52°C to 58°C; GC content of 40% to 60%; and Target length of 600-800bps. Primers were screened for hairpins and primer dimers.

1.2.4 Ecotype Sampling

Initially, the goal was to sequence 10 DNAs from 10 individual ecotypes, however due to unprecedented events, time constraints, and budgeting allotment, sequencing was conducted with 4 DNAs

from all 10 ecotypes in the latitudinal gradient. Population genomics along the same latitudinal gradient used for this study identified genetic structure at tree line (north vs south; Figure 1.4) (Stunz et al., In Revision), consistent with ecological specialization of ecotypes. Given these results, our approach is to sample 4 DNAs from 10 populations (5 north/ 5 south) to look at genetic variation between the distinct ecotypes. With this approach, 20 samples were sequenced from the northern populations and 20 samples were sequenced from the southern populations for each of three Phytochrome gene regions.

1.2.5 DNA Extractions, PCR, and Sequencing

Genomic DNA extractions from the 10 sampled populations were conducted from 50mg of dried leaf tissue (Stunz et al, In Revision) using the CTAB method (Doyle & Doyle, 1987). The DNA concentrations were then quantified using the Qubit dsDNA BR Assay Kit (Invitrogen) and Qubit 3.0 Fluorometer (Thermo Fisher Scientific), then stored in a -20°C freezer for later use.

The polymerase chain reaction (PCR) was used to amplify all Phytochrome gene regions with the custom designed primers (Table 1.2) using 2 individuals each from a subset of 4 populations (PB, TL, GO, and EC; Table 1.1) to determine if there was variation among populations in the Phytochrome genes (Table 1.2). Gene regions that provided good sequence data from the initial examination and were variable, were included for a more inclusive sampling of 4 individuals from each of the 10 populations.

PCR was performed with 14μl of Master Mix containing 6.8μl ddH_2O, 1.5μl 10x Buffer, 2.5μl $MgCl_2$, 2.3μl dNTPs, 0.4μl primers, 0.2μl taq and 2μl DNA template. PCR was performed in an Eppendorf Mastercycler Pro thermocycler. The cycle reaction was 94°C 5 min. followed by 30 cycles (94°C 45 s, 52°C or 56°C 45 s, 72°C 45 s) with a final extension 72°C for 10 min. All

unincorporated dNTPs and primers were removed using EXOSAP-IT (Applied Biosystems). The PCR products were sequenced using the same primers used for PCR and BigDye terminator chemistry (Applied Biosystems, Waltham, Massachusetts, USA) on an ABI 3730 × DNA sequencer at the UTEP Border Biomedical Research Center (BBRC) Genomics Analysis Core Facility.

Raw sequence data was uploaded to Sequencher® 4.0.5 to align, edit, and determine sequence quality and presence of sequence polymorphisms among individual accessions. If a polymorphism (Figure 1.5B) was found in the sequence data among the initial 8 samples, sequence data was collected from the other populations and accessions (as described above). Once the additional data was included, final sequence alignments were exported as aligned FASTA files. These FASTA files were then uploaded into the software DNASP 6 (Rozas et al., 2017) to phase sequences with polymorphisms to identify haplotypes. Phasing sequence data with polymorphisms provide a means of identifying and extracting the likely alleles for each accession.

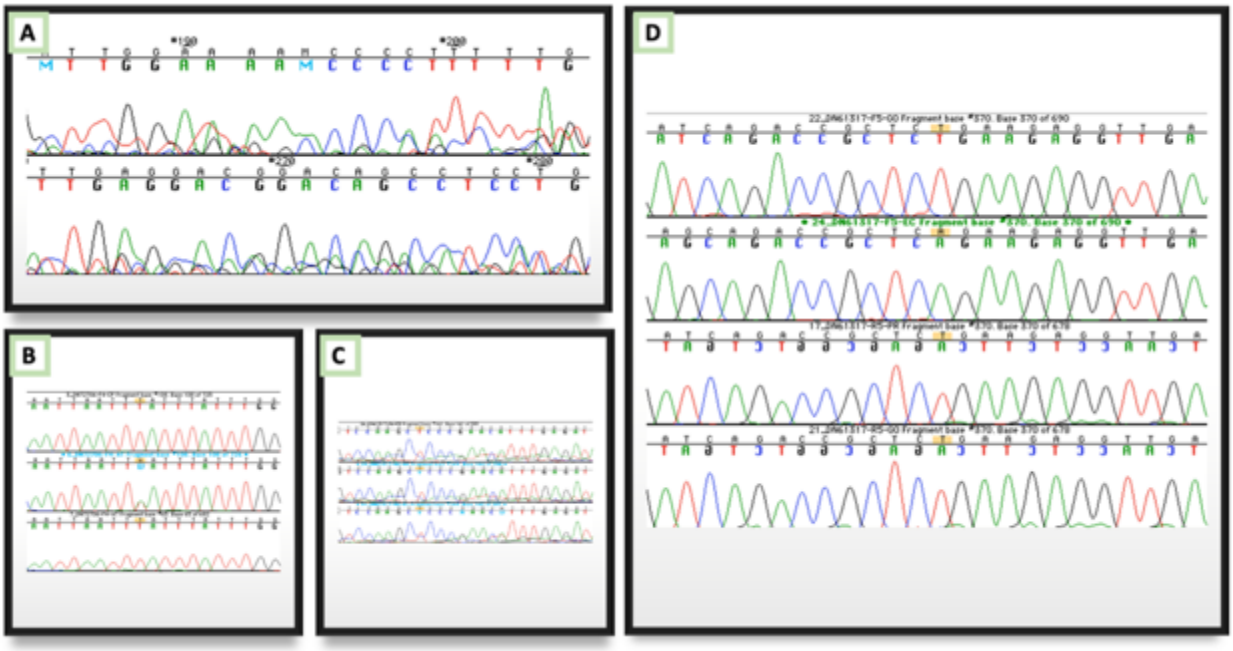


Figure 1.5 Possible Sanger sequencing results: (A) shows a poor-quality sequence with multiple nucleotide peaks at one base pair position, meaning that the nucleotide that was called for that position is likely inaccurate. (B) Polymorphism is highlighted in yellow, two alleles, one with a 'T' the other an 'A' is signified with a 'W'. (C) Poor quality sequence where individual nucleotide peaks can be identified, however the sequence quality is too poor to clearly identify polymorphisms. (D) High quality sequence data alignment with a SNP mutation (highlighted in yellow).

1.2.6 Data Analysis: Gene Trees

FASTA files with the phased sequence data was uploaded into Geneious 10.0.9 (Kearse et al., 2012) to construct gene trees using MrBayes (Huelsenbeck & Ronquist, 2001) and to align sequence exons. Aligned sequences of each Phytochrome gene were first analyzed for best fit model for Bayesian analyses using the Aikike Information (AIC) in the program JMODELTEST 2 (Darriba, Taboada, Doallo, & Posada, 2012) accessed through the web portal CIPRES (Miller, Pfeiffer, & Schwartz, 2010). In Geneious 10.0.9 (Kearse et al., 2012), the plugin "MrBayes" (Huelsenbeck & Ronquist, 2001) was then used to construct gene trees using the best fit model for each Phytochrome gene region. Phased sequence data that was aligned to a reference sequence for each Phytochrome region was then exported as an aligned FASTA file.

1.2.7 Data Analysis: K_A/K_s Analysis

Analyses were conducted to look for evidence that supports selective sweeps, these occur when a beneficial mutation increases in frequency and may eventually become fixed in a population or subset of populations. Selective sweep signatures can be identified when there is reduced allelic variation with one allele becoming more established in one population (or subset of populations), compared to overall expected distribution of that allele in all populations. They can also be identified at individual genes if specific non-synonymous mutations occur more frequently than would be expected at random among a set of populations or compared to other genes.

In order to identify polymorphisms that could potentially change gene function, I followed methods similar to Mattila et al. (2016) to analyze the sequences for evidence that supports selective sweeps. First, the coding region was identified for each data set uploaded in DNASP 6 (Rozas et al., 2017). The signature of selection was examined for each gene region by determining complete measures of nucleotide diversity, site specific variation, and K_A/K_s ratios across all accessions, among populations and north/south ecotypes within DNASP 6 (Rozas et al., 2017). Nucleotide diversity is a measure of the average proportion of nucleotide differences per site of the given samples, DNASP 6 (Rozas et al., 2017) used the total number of differences, sequence sizes, and varying numbers of pairwise comparisons to calculate this across all samples. The K_A/K_S ratio test is a measure of non-synonymous to synonymous mutations, the command estimates the number of non-synonymous substitutions per non-synonymous site and synonymous mutations per synonymous site between two sequences (Rozas et al., 2017). The analysis yields the average number of nucleotide differences per site between two sequences (or nucleotide diversity) using pairwise comparison (Rozas et al., 2017). Once the K_A/K_S statistic is

calculated, if the ratio is greater than 1, there is a higher likelihood for selection, if the statistic yields a number less than 1, it's likely that this gene or sequence data is not under selection (Rozas et al., 2017).

The DNA Polymorphism (sliding window) analysis in DNASP 6 was conducted to identify regions of the Phytochrome genes where mutations are occurring, and to visualize these mutations in the form of a line graph. With this analysis, each window of ~25bp (though this can vary) is considered a site, and as the window "slides" across the selected region of sequence data, the number of mutations is noted (Rozas et al., 2017). For this research, a window size of 25 bps and a window length of 100 sites was utilized.

1.3 RESULTS

1.3.1 Phytochrome Gene Primers

After identifying the reading frames for each of the Phytochrome genes, 16 primer pairs were designed to amplify the coding regions for each gene (Table 1.2). Due to the size of the Phytochrome gene coding regions, multiple primer pairs were needed to amplify them in their entirety. Primer pairs amplify 600bp to 800bp regions of the reading frame, and primer pairs were designed to cover the entire coding region. When multiple primer pairs for one coding region were designed, they had at least 20bp of overlap to avoid losing parts of the coding region during amplification (Table 1.2).

High quality sequence data that contained polymorphisms was attained from only some primer pairs for each gene. Initial sequencing revealed that of the Phytochrome A primers, there were two regions (F2-R2, F5-R5) with high quality sequence data and the occurrence of polymorphisms. For Phytochrome B, two primer pairs (F4-R4, 2_F1-2_R1) resulted in high quality sequence data, but only F4-R4 had polymorphisms. Phytochrome C had three primers

pairs that provided high quality sequence data, two had polymorphisms (F1-R1, F2-R2), and one lacked polymorphism (F3-R3). Due to the results of the initial sequencing run, only those primer pairs amplifying regions that returned high-quality sequence data with the occurrence of polymorphisms were used to attain further data.

Table 1.2 Primers developed for Phytochrome A, B and C genes. Included are the gene ID (transcript) following Mohl et al. (2020), \ \ gene length, and primer pairs to amplify the coding region. F = forward primer, R = reverse primer. Due to the length of the genes, coding region sizes, and the occurrence of multiple reading frames found, multiple primers were designed to amplify 600-800 bp regions for all reading frames. Primers in Bold type indicate those that provided high quality sequence data with polymorphisms.

Gene Name	Transcript	Gene Length (bp)	Primer Name	Sequence (5'-3')
Phytochrome A (PHYTA)	DN61317_c0_g1_i1	4077	F1	ATGAAAGAGAAGGAGGAGCTG
			R1	AAGGTAAGAATGATCTGCGA
			F2	**CTTTCACGATGACGACCAT**
			R2	**GTCAACCAGATAGCAATTTG**
			F3	ATGGATTTAGTGAAGTGCGA
			R3	CGATCTTCTGATTCCAACCAT
			F4	GTTACAAGTGAGATGGTGAGG
			R4	CCTGTCATAGCCGTGTTTA
			F5	**CAAAACCCTAACCCTCTGAT**
			R5	**CCGCTATCATTATCACATCCC**
			F6	ATTGATATTGGCTCGAGATGC
			R6	TGTCTACAAGAAGGTCCTCAT
Phytochrome B (PHYTB)	DN72706_c1_g1_i3	4377	F1	ACGATAGGGTGATGGTTTACA
			R1	GCAAGATGAGAGCCATTCAAT
			F2	TATGTTCTACCATGGCAGGTA
			R2	ATCTTCTTCTCCTCTCAGTGC
			F3	TTTAAGGAGTCTGAGGAGGTC
			R3	CTGTCATCTCCAAAAGTGAGT
			F4	**GGAATTGGCCTATCTTTGTCA**
			R4	**CTCCATGTTGCAGTCATTTTG**
			2_F1	CCTCACAATTCAATCACAAACC
			2_R1	ACAACTACAAGTACCAGTTCG
Phytochrome C (PHYTC)	DN67535_c0_g1_i1	4472	**RC_F1**	**TCTGGTTATGGACATGGAG**
			RC_R1	**CAGAATGTGCCTCTCCTTTG**
			RC_F2	**AGTTCCTAGTGCAAGTCTTTG**
			RC_R2	**TAGCCGAACCATCTCATTAGT**
			RC_F3	CGTCTGATGATGCAAGAAGAA
			RC_R3	GAGATGGCAGTGTTCAGAAT
			RC_F4	GTCAGGCATGTGTATAGAGTG
			RC_R4	TCATTCAAATCTGAGGACTGC
			RC_F5	TCTGGTCAAGATGTCGAAAAG
			RC_R5	TCTCTAAACCACTCTTAGCCA

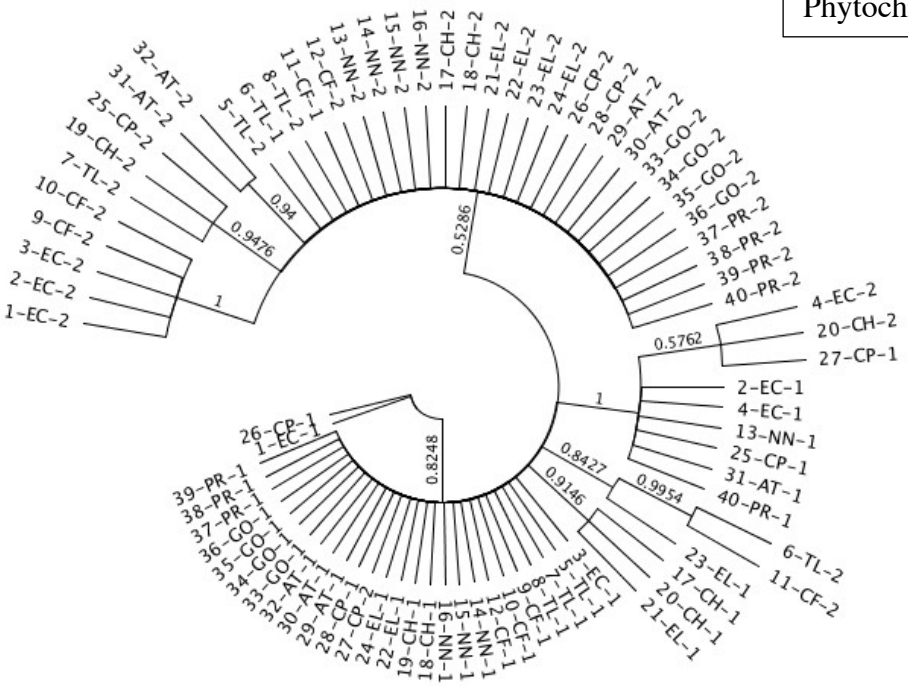
1.3.2 Sequence Data

Sequence data from Phytochrome A included two regions, here forward referred to as PHYTA1 and PHYTA2. PHYTA1 (F5-R5) had 702 total bases (included a 145 bp intron region) located at positions 924 bp to 1627 bp of the reference sequence, and PHYTA2 (F2-R2) has 548 total bases located at position 2575 bp to 3122 bp of the reference sequence. Partial sequence from Phytochrome B included one region, here forward referred to as PHYTB. PHYTB had 722 total bases (included a 234 bp intron region) located at positions 3624bp to 4349bp of the reference sequence. Partial sequence from Phytochrome C included two regions, here forward referred to as PHYTC1 and PHYTC2. PHYTC1 (F2-R2) had 648 total bases located at position 2953bp to 3600bp of the reference sequence, and PHYTC2 (F1-R1) with 682 total bases located at position 3580bp to 4261bp of which all were exons. Due to Phytochrome A and C having two different sections of sequence data from the coding regions, sequences were concatenated in Sequencher® 4.0.5 before further analyses were conducted. Once sequences from Phytochrome A and C were concatenated and reading frames with no stop codons were identified, protein sequences were blasted with the NCBI database verifying their identity as Phytochrome genes.

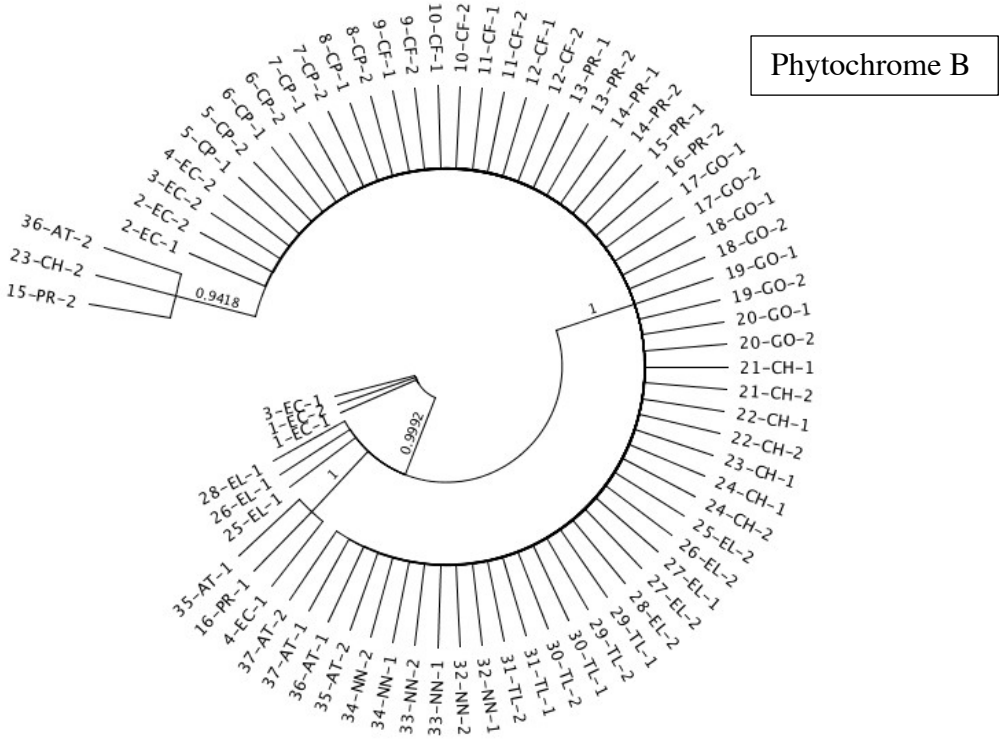
1.3.3 Gene Trees

JMODELTEST2 (Darriba et al., 2012) was used to determine the best fit model for Bayesian analysis for each data set resulting in the following models used: PHYTA; JC+I, PHYTB; F81, and PHYTC; K80+I. Each of these models were used to create gene trees using Bayesian analysis (Figure 1.6). The resulting Bayesian gene trees did not support divergence between northern and southern populations, though there is strong evidence for allelic variation (Figure 1.6).

Phytochrome A



Phytochrome B



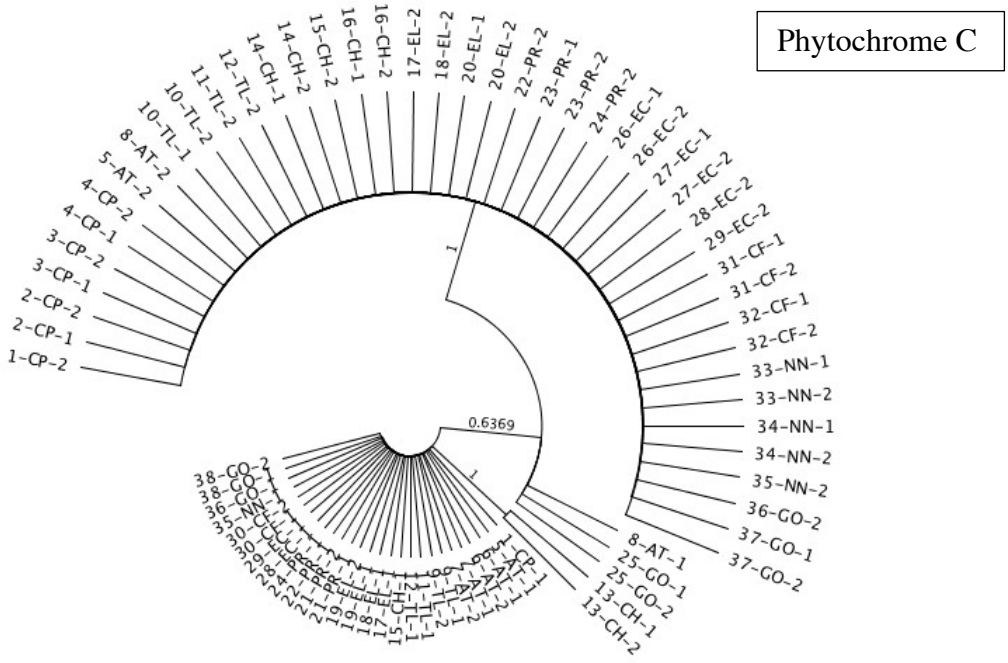


Figure 1.6 Bayesian gene trees resulting from analysis of the three Phytochrome genes using Geneious 10.0.9. Branch values are posterior probabilities.

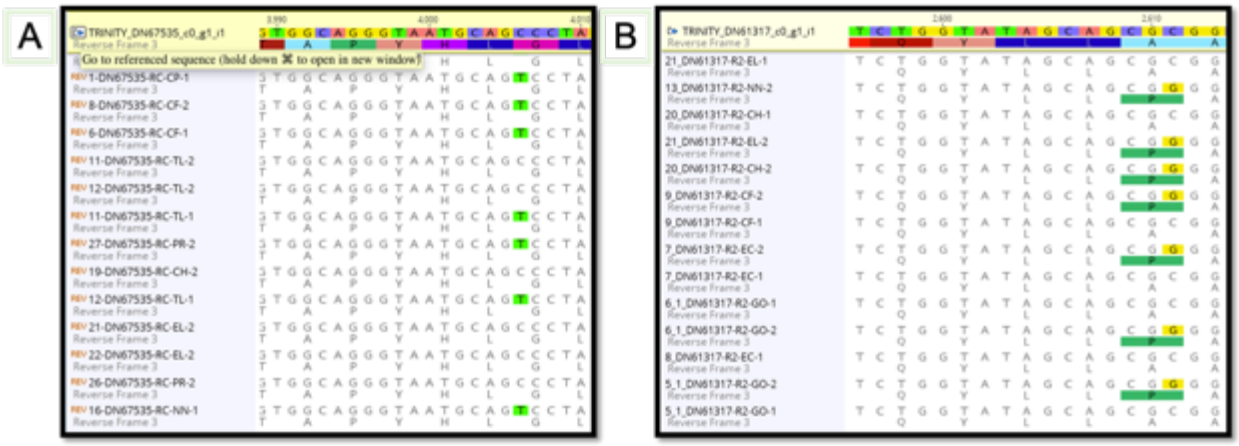


Figure 1.7 Two types of mutations that are present in the Phytochrome genes from populations of *E. vaginatum* along the latitudinal gradient. (A) synonymous, or silent mutation in Phytochrome C, where the variable base pair does not cause a change in the amino acid coded for. (B) a non-synonymous mutation in Phytochrome A, where the variable base pair causes a change in the amino acid.

1.3.4 K_A/K_S Analysis

Sliding window analysis DNASP 6 (Rozas et al., 2017) was used to measure the K_A/K_s ratio per sequence data set. Phytochrome A received a K_A/K_s statistic of 0.6 (Table 1.3). Phytochrome B received a K_A/K_s statistic of 0.67 (Table 1.3). Lastly Phytochrome C, received a K_A/K_s statistic of 0.3 (Table 1.3).

Table 1.3 Results from the Ka/Ks analysis conducted in DNASP 6 for each Phytochrome gene region.

Nonsynoymous/Synonymous Mutation Analysis Results							
Phytochrome	Theta		Pi, Jukes and Cantor		Number of Mutations		Ka/Ks Statistic
	Ka	Ks	Ka	Ks	Ka	Ks	
A	0.003	0.005	0.002	0.003	12	6	0.6
B	0.002	0.003	0.0004	0.001	2	1	0.67
C	0.0009	0.003	0.002	0.007	3	3	0.3

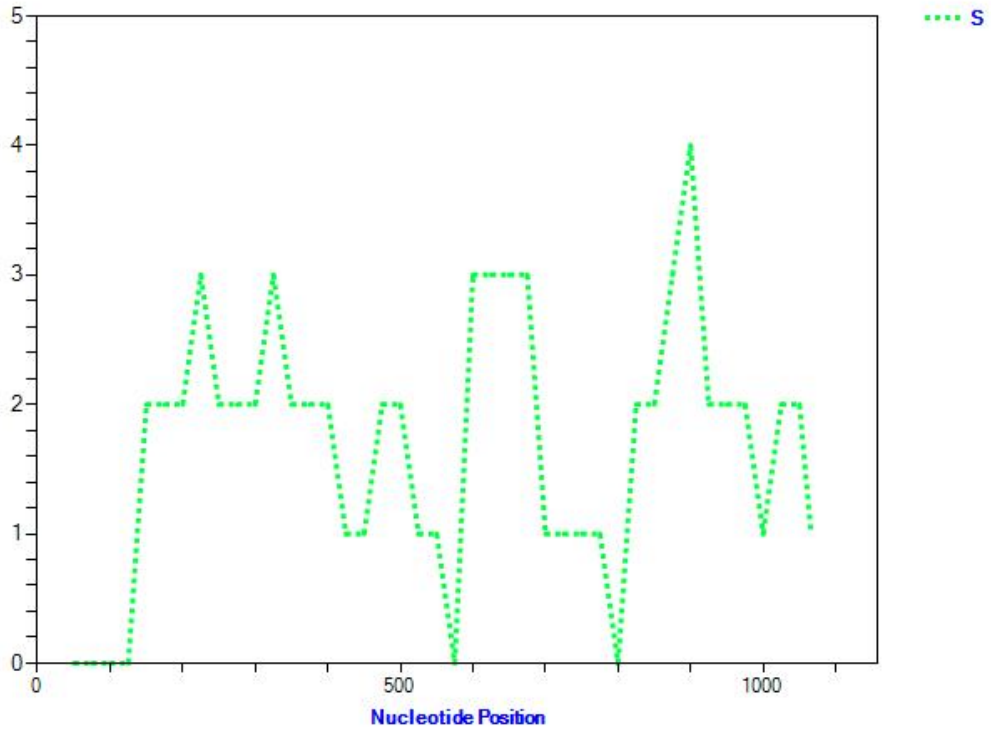
1.3.5 Sliding Window Analysis

The sliding window results displayed that Phytochrome A had 18 polymorphic sites, 16 haplotypes, and a nucleotide diversity of 0.0021, Phytochrome B had 3 polymorphic sites, 4 haplotypes, and a nucleotide diversity of 0.0005, and Phytochrome C had 21 polymorphic sites, 7 haplotypes, and a nucleotide diversity of 0.0036 (Table 1.4).

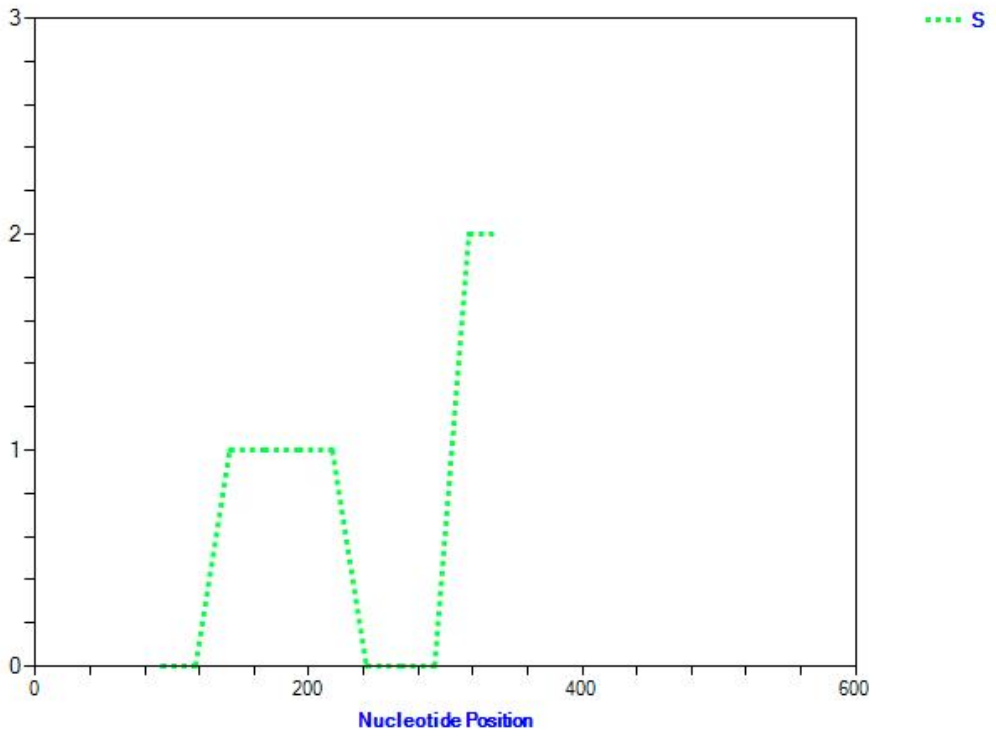
Table 1.4 Results from the sliding window analysis conducted in DNASP 6 for each Phytochrome gene region.

Sliding Window Results			
Phytochrome	Polymorphic Sites	Haplotypes	Nucleotide Diversity
A	18	16	0.00209
B	3	4	0.00055
C	21	7	0.00361

Phytochrome A



Phytochrome B



Phytochrome C

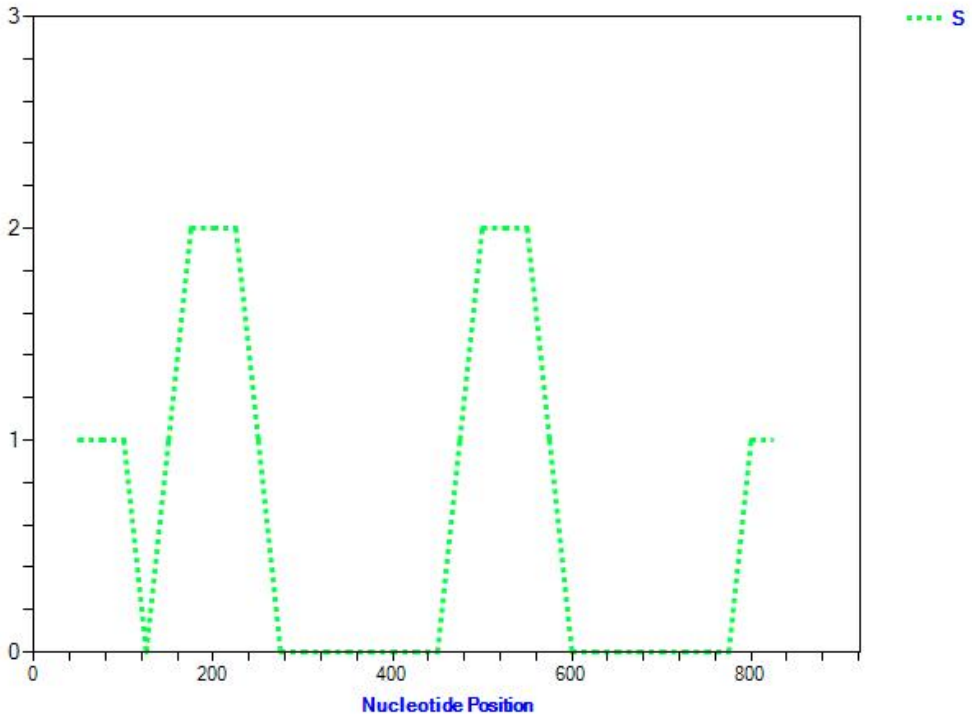


Figure 1.8 Data visualization of polymorphic sites in each Phytochrome gene, images created in DNASP 6.

1.4 DISCUSSION

Eriophorum vaginatum has phenological variation related to leaf senescence timing that is consistent between northern and southern ecotypes in the Alaskan Arctic (Parker et al., 2017). There is also evidence of genetic structure separating north and south ecotypes at treeline that is supported by population genomics (Stunz et al., In Revision). With the warming of the Arctic changing the length of growing season, there is a knowledge gap that resides in identifying the potential genetic constraints for adaption among ecotypes. Due to the role that the Phytochrome gene family plays in signaling the genetic pathways related to plant phenology (Halliday & Davis, 2016), and the recognized phenological differences among ecotypes of *E. vaginatum* in the Alaskan Arctic (Parker et al., 2017; Stunz et al., In Revision), it was hypothesized that Phytochrome genes of *E. vaginatum* would have genetic variation in a pattern consistent with northern and southern ecotypes. The overarching goal of this chapter was to determine if there were sequence polymorphisms within the Phytochrome genes of *E. vaginatum* and if found, determine if they alter amino acids, which could change gene function. The ultimate goal for this work was to determine if there were potentially adaptive changes in these genes associated with ecotypes north and south of treeline.

1.4.1 Phytochrome Genetic Variation

The C1S script (Figure 1.3) was used to identify Phytochrome genes and determine if any had SNPs among the populations sampled by Mohl et al. (2020). Sequence sampling across 10 populations identified variation in Phytochrome A, B, and C coding regions. However, gene tree construction for each Phytochrome gene did not provide evidence of allelic divergence that was consistent between northern and southern ecotypes along the latitudinal gradient, despite there being a strong display for allelic variation among ecotypes (Figure 1.6).

This could be due to the occurrence of multiple gene copies for some of the Phytochrome genes that the designed primers (Table 1.2) found during PCR, these Phytochrome genes (A, B, and C) have all displayed duplication previously in other plant genomes (Sheehan, Farmer, & Brutnell, 2004). There could be true allelic variation among these genes, yet no evidence of selection with consistencies among northern and southern ecotypes. The results of the K_A/K_S analysis suggest that these Phytochrome genes are likely not under selection (Table 1.3) despite displaying strong evidence for allelic variation (Figure 1.6). Results of the K_A/K_S analysis (Table 1.3) did not show a higher ratio of K_A (nonsynonymous) mutations to (synonymous) mutations, when correcting for total ratio of potential K_A and K_S type mutations for the Phytochrome genes. For gene duplication to lead to retention of multiple copies it would be expected that selection would lead to alternative function for these genes (Senetar & McCann, 2005). Given the lack of a signal of selection via the K_A/K_S analysis there isn't evidence for this. However further analysis of amino acid changes and whole genome analysis could provide evidence that the pattern we see is due to multiple copies of all phytochrome gene in *E. vaginatum*.

Additionally, the data do not show strong evidence for both copies of genes in a majority of accessions as would be expected if multiple copies of a gene were present (Table 1.4). While many heterozygotes are detected, homozygotes for an apparently more dominant allele are present in many accessions. Phytochrome A does not display any alleles found as homozygotes for the alternative nucleotide, only heterozygotes, suggesting that these are true mutations detected in this gene, rather than supporting the presence of multiple gene copies. However, this also may just be due to artifacts of PCR sometimes not capturing one gene copy or another and

sometimes preferring one copy over the other. In the latter case this may be due to primer design favoring the gene copy the primers are designed for.

Duplication events have resulted in multiple copies of Phytochrome genes. For example, there is genetic redundancy of Phytochrome A in other plant genomes, these multiple homologs are likely a result of ancient tetraploidization in ancestral lineage (Liu et al., 2008; Sheehan et al., 2004). If there are multiple copies of Phytochrome A in the *E. vaginatum* genome and the multiple copies are functional then this could explain the Phytochrome A results, these primers might have found a different copy of the gene. Biologically, Phytochrome A is involved throughout the whole life cycle of angiosperm plants including but not limited to light promotion of germination in seeds, shade perception, and resetting of circadian rhythms (Casal, Sanchez, & Yanovsky, 1997), though its most prominent role lies in promoting flowering (Bagnall et al., 1995; Lin, 2000). Phytochrome A, B and C did not yield statistics that indicate they are under selection (Table 1.3). In the closely related grass (Poaceae) lineage, the maize genome has multiple copies of Phytochrome A, B, and C (Sheehan et al., 2004). Due to a relatively recent polypoidization event in the maize genome (Gaut, 2001), and the conservation of sequence of the Phytochrome homeologs, there is likely functional redundancy of this gene family that was maintained during evolution of modern maize (Sheehan et al., 2004).

Recovering high quality Phytochrome B sequence data proved to be particularly problematic (though there were issues with the other genes as well). Sanger sequencing was utilized for this research due to its fast and cost-effective nature, and the relatively small sample size. Additionally, Sanger sequencing is an effective approach when looking at variant screening in single genes when working with a low number of samples (Schuster, 2008). However, this method can be problematic when multiple copies of the targeted sequence are found or there are

multiple alleles that have length differences based on indels or SSRs, as the multiple copies can be sequenced simultaneously and lead to poor quality sequencing results. The portion of Phytochrome B that was sequenced here had a number of introns, which are regions that indels and SSRs are more commonly found and could have led to sequencing issues. Within the transcriptome, multiple Phytochrome B transcripts were identified (Mohl et al. 2020), but it was unclear if this was due to multiple versions of the gene or variations of portions of the sequence assembled. In any case, given the Sanger sequencing issues encountered, using a Next Generation Sequencing (NGS) approach with targeted amplicon sequencing will be the preferred approach for *E. vaginatum*. Taking an NGS approach in the future will allow for attaining high quality sequence for regions of poor quality with the Sanger approach and assessing if multiple copies are found of each gene or if there is allelic variation. The NGS approach will allow better discovery power to identify novel variants, (König et al., 2015; Shendure & Ji, 2008), but is an approach beyond the budget available for this project.

For future selection studies including the Phytochrome gene family, an approach using reference genes that are not associated with phenology to serve as controls when running comparative analyses, will also be required. This approach was not taken during this study due to time constraints imposed by UTEP closure during 2020. These genes could be selected for neutrality and provide general mutation rates across the genome for analyses concerning natural selection, polymorphisms, and divergence data among the 10 ecotypes using likelihood HKA (Hudson-Kreitman-Aguadé) with a multilocus approach considering the reference genes could also be utilized. This could give valuable insight by conducting statistical analyses for allelic variations that are not the same across all populations for a specific locus and provide

assessments as to which loci contribute to drift under neutrality using the reference loci not related to phenology or under selection (Wright & Charlesworth, 2004).

1.4.2 Future Directions

To continue this work, primarily, it would be best to increase the number of samples sequenced from each ecotype. Sample collection from the field has already occurred, but more DNA extractions would need to be conducted in order to do this. Inclusion of more sequence data would allow for more in-depth statistical analyses to be conducted, such as the usage of reference genes and maximum likelihood HKA. The goal for number of samples acquired should be at least 10 DNAs from all 10 ecotypes spanning across the latitudinal gradient. Furthermore, given the evidence of variation in *E. vaginatum* phenology along the northern Alaskan latitudinal gradient (Parker et al. 2017), more gene families related to phenology need to be examined for selection.

Chapter 2: Genetic Marker Identification in Genes Related to Stress Response

2.1 Introduction

2.1.1 Genetic Markers

Genetic differentiation can be identified through observed variation in allelic frequencies, or DNA sequence variations at a given gene (Zhang & Hewitt, 2003). One way to recognize genetic differentiation and different allelic patterns is through the use of genetic markers. Genetic markers are powerful tools that can be used to link phenotypic and genotypic variation in organisms (Varshney, Graner, & Sorrells, 2005). Given the evidence of phenotypic variation in *E. vaginatum* (Parker et al., 2017) along the latitudinal gradient in northern Alaska, more insight is needed on possible changes in genetic markers to infer molecular response to environmental pressures that come with the changing climate in the Arctic. Two frequently used genetic markers are microsatellites (or simple sequence repeats (SSRs)) and single nucleotide polymorphisms (SNPs).

2.1.2 Genetic Differentiation

Genetic differentiation within a species can be uncovered by examining allelic frequencies, or DNA sequence variations at a given gene (Zhang & Hewitt, 2003), among populations. The variation can be driven by evolutionary factors including mutations, gene flow, and natural selection. Understanding genetic differentiation and allelic frequencies among and within populations gives insight to the potential for a taxon to evolve with environmental pressures and how populations may have evolved in the past. Methods for detection of significant genetic differentiation and variation among populations depend on many factors (Waples & Gaggiotti, 2006) including the type of genetic marker being studied. In this work I

chose to utilize two different types of genetic markers, SNPs, due to their ease of detection and SSRs due to their high levels of polymorphism (Zhang & Hewitt, 2003).

2.1.3 Simple Sequence Repeats (SSRs)

SSRs occur when segments of DNA are repeated anywhere from five to fifty times or more and can be found within either coding or noncoding regions of an organism's genome (Kalia, Rai, Kalia, Singh, & Dhawan, 2011). Plants are rich in dinucleotide AT repeats (Kalia et al., 2011; Morgante & Olivieri, 1993), and dinucleotide repeats are most common in many species, but these repeats are more frequent in non-coding regions rather than coding regions (Wang, Weber, Zhong, & Tanksley, 1994; Zane, Bargelloni, & Patarnello, 2002). Previous studies have suggested that in plants, AT repeats are more common to CG repeats and show more variation (Merritt, Culley, Avanesyan, Stokes, & Brzyski, 2015; Morgante & Olivieri, 1993). The AT dinucleotide repeats are generally favored for use due to their higher levels of variation, which is likely due to ease of mutation through DNA slippage during replication (Chakraborty, Kimmel, Stivers, Davison, & Deka, 1997; Levinson & Gutman, 1987; Merritt et al., 2015). Transcriptomic studies that include searches for SSRs find that the most abundant repeats in coding regions are trinucleotide (Han et al., 2018; Pramod, Perkins, & Welch, 2014), as they can occur without shifting reading frames as opposed to other repeats. Trinucleotide SSRs are found in both coding and noncoding regions of plants but have been found to occur more abundantly (nearly twice as often) in coding regions, most likely due to a result of positive selection for single amino acid stretches (Li, Korol, Fahima, Beiles, & Nevo, 2002). Smaller and larger motif repeats (such as dinucleotide and tetranucleotide) are more likely to be distributed in 5'UTRs and 3'UTRs as they would cause frameshifts if found in coding regions (Pramod et al., 2014).

Polymorphic SSR markers are uniquely valuable for genomic studies of adaptation and population structure due to their abundance and uniformity of genome coverage, their frequent association with expressed sequence tags (ESTs) (Kalia et al., 2011) and with functional genes in sequenced transcriptomes (Hodel et al., 2016). ESTs are small sequences of DNA (roughly 200-500 bps long) that are developed by sequencing one or both ends of an expressed gene, however fully sequenced transcriptomes are now more commonly associated with SSR discovery (Hodel et al., 2016). Previously, SSRs have been identified in publicly available EST projects and gene sequences using several tools that evaluate a single sequence at a time such as BLASTN tools and SSRfinder (Kalia et al., 2011; Scott et al., 2000; Temnykh et al., 2000; Varshney et al., 2005). Other tools are now available that will identify SSRs across entire genomes and transcriptomes, such as IMEx, SciRoKo (Kofler, Schlötterer, & Lelley, 2007; Mudunuri & Nagarajaram, 2007), and MISA (MIcroSAtellite) (Hodel et al., 2016). Additionally, several scripts in Perl and Python have been used to recognize SSR patterns in genomic sequence studies (Labbé, Murat, Morin, Le Tacon, & Martin, 2011; Varshney et al., 2005). The utilization of SSRs from transcriptomes that are related to functional genes can identify genetic variances in different populations that could be related to adaptation. These aspects of SSRs make them particularly useful for examining gene flow and/or selection patterns in natural populations (Kalia et al., 2011; Provan, Powell, & Hollingsworth, 2001).

2.1.4 Single Nucleotide Polymorphisms (SNPs)

SNPs are defined as a variation in a single nucleotide of DNA sequence that occurs throughout the genome, including both coding and non-coding regions. SNPs are commonly used today as a genetic marker to identify loci under selection in natural populations (Rellstab, Zoller, Tedder, Gugerli, & Fischer, 2013; Wessinger, Kelly, Jiang, Rausher, & Hileman, 2018)

or for development of sustainable agricultural crops (Jain, Darshan, & B. S. Ahloowalia, 2010; Varshney, Mahendar, Aggarwal, & Börner, 2007). The identification of SNP among populations in or associated with transcribed (coding) regions can help us further understand the effects of selection on population structure and gene flow in various model and non-model organisms (Emanuelli et al., 2013; van Inghelandt, Melchinger, Lebreton, & Stich, 2010).

2.1.5 Genetic Markers in the Arctic Foundation Species *Eriophorum vaginatum*

Due to the occurrence of SSRs and SNPs in and associated with coding regions, a transcriptome wide development of genetic markers can be used to further our understanding of ecotype specific genetic adaptations in *E. vaginatum*. Due to the variation uncovered among ecotypes in *E. vaginatum* with environmental response (Bennington et al. 2012; Mohl et al., 2020), I hope to discover patterns of variation in genetic markers related to functional genes involved in environmental stress response due to local adaptation and homesite advantage.

2.1.6 Response to Stress GO Term

When the *E. vaginatum* transcriptome was sequenced (Mohl et al., 2020), a GO (Gene Ontology) enrichment analysis was also conducted to classify genes by function. The enrichment analysis first classifies genes into three main domains: biological process, molecular function, and cellular component. The domain cellular component includes genes incorporated with the cell or the extracellular environment, molecular function includes genes with elemental activities at the molecular level, and biological process includes genes with operations or sets of molecular events pertinent to the functioning of living units (Ashburner et al., 2000). Within those three domains, genes are further classified by more specific functions and provided a "GO Term" or ID (example: **ID:** GO:0000016 **Name:** Lactase Activity **Ontology:** Molecular Function). The GO identifier "Response to Stress" is defined as any gene that plays a role in processes that

result in a change in state or activity of a cell or an organism as a result of disturbance in organismal or cellular homeostasis, typically due to exogenous factors such as temperature and humidity (Ashburner et al., 2000).

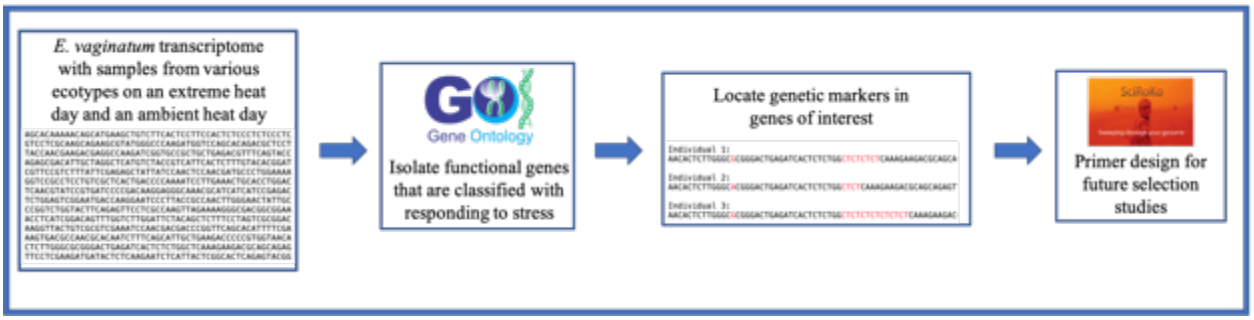


Figure 2.1 Chapter 2 flow chart, starting broad with the *E. vaginatum* transcriptome (Mohl et al., 2020), focusing in on the genes of interest, locating genetic markers that could be variable among ecotypes with different environmental pressures, and preparing to determine their utility for future selection studies.

2.1.7 Chapter Aims

The overarching goal of this work is to identify genetic markers that can be used to explore ecotype specific variation that could be related to adaptation for *E. vaginatum* in the Arctic. There are 4 goals for this research (1) Develop and modify Python scripts to isolate genes and genetic markers of interest; (2) Identify genes that are classified within the subgroup "Response to Stress" in the "Biological Process" group when using GO terms for *E. vaginatum*. These genes will be most likely to have function in response to climate variation found along the arctic latitudinal gradient; (3) SSRs and SNPs will be identified in "Response to Stress" genes that can be used in association studies for selection to environmental stressors across *E. vaginatum* ecotypes; and (4) Design primers for these genetic markers to use for future selection studies that will examine variation among ecotypes of *E. vaginatum* exposed to different environmental stressors.

2.2 METHODS

Site	Latitude (N), Longitude (W)	Elevation (m)	# Accessions
Eagle Creek (EC)	65.4332°, -145.5118°	771	3
Coldfoot (CF)	67.2631°, -150.1591°	321	6
Toolik Lake (TL)*	68.6292°, -149.5778°	758	6
Sagwon (SG)	69.4244°, -148.6976°	299	6
Prudhoe Bay (PB)	70.3270°, -149.0645°	8	3

Table 2.1 Location data for each *E. vaginatum* ecotype used in this study. Plants from each of these sites were transplanted into a common garden at the Toolik Field Station (*) in 2012 and 2013 (Mohl et al, 2020).

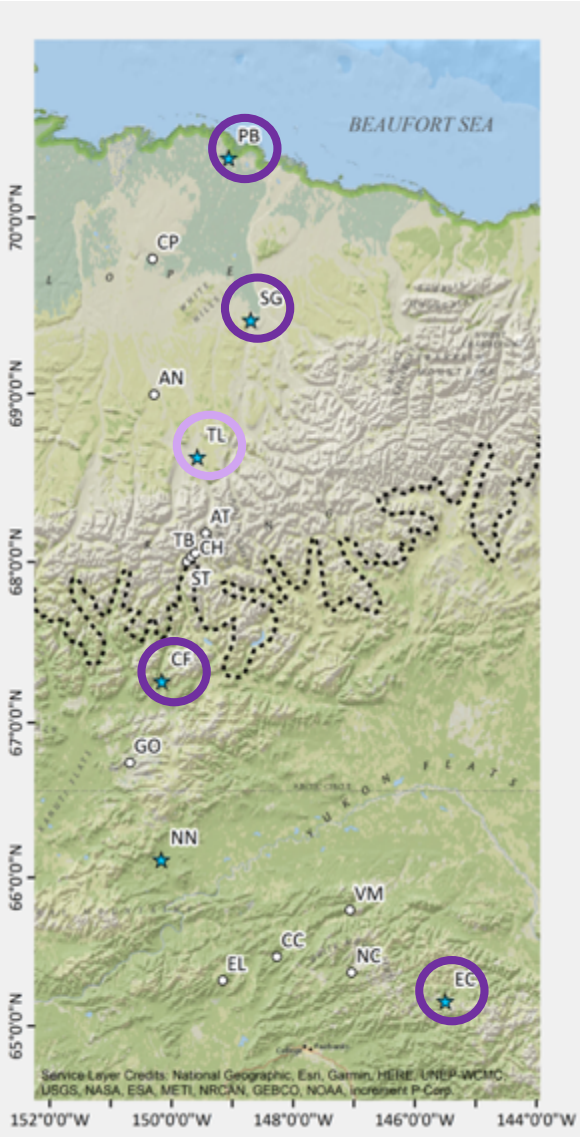


Figure 2.2 ArcGIS map displaying the locations of ecotypes on the northern Alaska latitudinal gradient used in developing the *E. vaginatum* transcriptome (Mohl et al., 2020). The dark purple circles denote ecotypes used in the transcriptome and the light purple circle is the location of the transplant garden containing all ecotypes.

2.2.1 Transcriptome Sampling

This research used the transcriptome sequence data for *E. vaginatum* (Mohl et al., 2020) that was derived from five populations (EC, CF, TL, SG, and PB; Figure 2.2, Table 2.1) and represents ecotypes from north (TL, SG, PB) and south (EC, CF) of treeline in the Alaskan Arctic, which were transplanted into a central common garden located at Toolik Lake (Table 2.1, Figure 2.2). Three accessions each were collected and pooled from all ecotypes in July of 2016 on an ambient day (13.8°C) and from only CF, TL, and SG on an extreme heat day (26.6°C) at the Toolik field station (Mohl et al., 2020). In total, representing 3 accessions each from EC and PB and 6 accessions each from CF, TL, and SG.

The transcriptome contains 182,744 transcripts that could be utilized for identifying SNPs and SSRs with 23,132 that were present in all ecotypes (Figure 2.3: Mohl et al., 2020). There were 124,150 transcripts assigned Gene Ontology (GO) classification terms resulting in a total of 286,156 GO terms recognized, 93,296 were assigned as biological processes and 207 of these were categorized as "Response to Stress" (RTS), (Mohl et al., 2020). The focus of this study is to identify SSRs and SNPs that will likely vary among ecotypes, to do this, genes associated with the GO term for "Response to Stress" in biological processes will be the primary targets (see Figures 2.3 and 2.4).

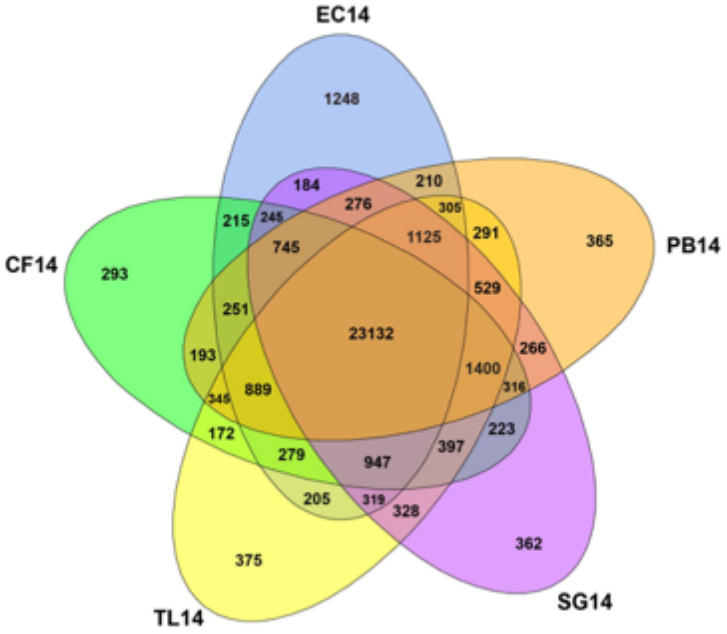


Figure 2.3 Venn diagram showing the number and overlap of unigenes expressed among the 5 ecotypes utilized for the transcriptome of *E. vaginatum* on the ambient temperature day (13.8°C) in the Toolik Field Station common garden (Mohl et al. 2020).

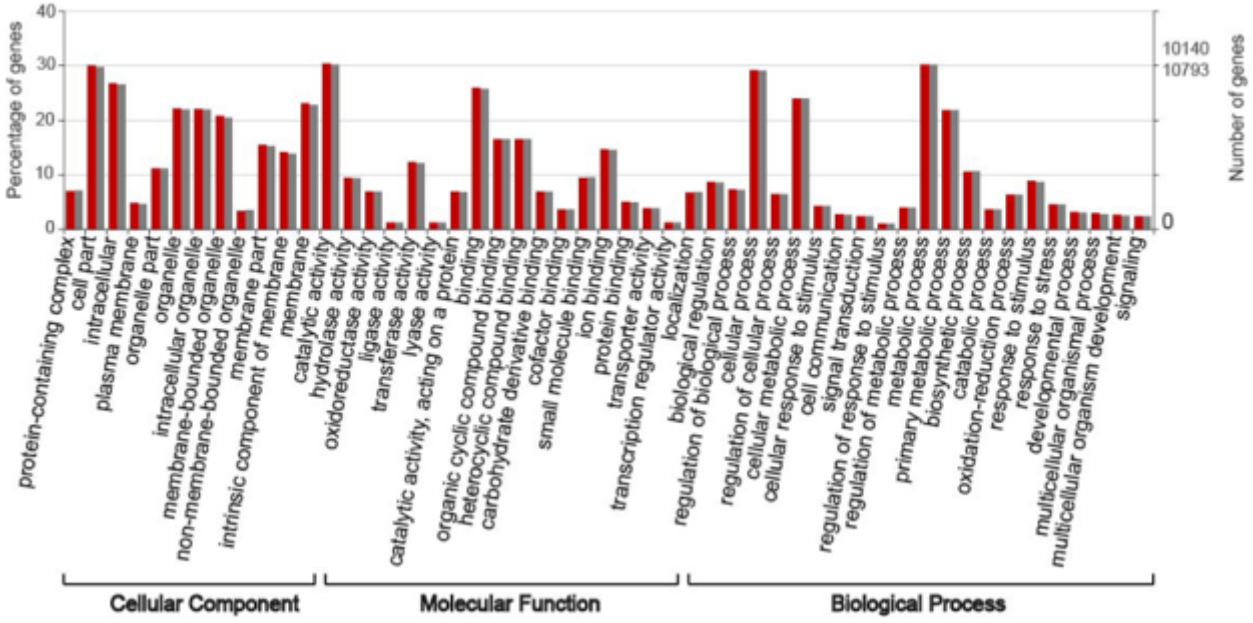


Figure 2.4 Histogram of Gene Ontology classification for *E. vaginatum* unigenes showing the overall percentage of unigenes by their GO term and divided into 3 functional groups. Red bars are 26.6°C day and grey bars are 13.8°C day (Mohl et al. 2020).

2.2.2 Isolation of "Response to Stress" Genes

A previously written in-house Python script (Mohl Script 1), here forward referred to as MS1, was utilized for this thesis and used to isolate the gene identifiers that are associated with the Gene Ontology term GO:0006950, which is the Response to Stress (RTS) biological process from the full transcriptome list of gene identifiers, this script was modified to isolate these genes of interest. This script utilizes arguments corresponding with the *E. vaginatum* Data Matrix file which contains expression counts and gene IDs and the Web Gene Ontology Annotation (WEGO) file which contains GO terms with corresponding genes. Once the gene identifiers for RTS genes were isolated, another previously written in-house Python script (Mohl Script 2), here forward referred to as MS2, was modified to use the list of RTS gene identifiers to extract the RTS sequences from the full transcriptome and write these to a FASTA file.

2.2.3 Search Parameters for SSRs associated with RTS genes

Once the RTS transcript sequences were isolated and placed into a separate FASTA file, the software SciRoKo 3.4 (Kofler et al., 2007) was used to create an SSR dataset focusing on di-, tri-, tetra-, and pentanucleotide Perfect and Imperfect SSRs.

Search parameters aligned closely with those given in Honig et al. (2017) and included a minimum of 6, 4, 4, and 4 repeat motifs for di-, tri-, tetra-, and pentanucleotide repeats respectively. However, due to limited setting parameters, and new updates, the program SciRoKo 3.4 (Kofler et al., 2007) was run twice (once for Imperfect SSRs and once for Perfect SSRs) to include a broad span of SSR motifs. The parameters for Perfect SSRs were any short sequence repeats (mononucleotide to hexanucleotide) that had a repeat length of at least 4 nucleotides. The parameters for the Imperfect SSRs include: (1) a minimum required score of 4, which refers to the total nucleotide length of the entire SSR (example: a dinucleotide SSR needs

to have a length of 8 and repeated at least 4 times); (2) a mismatch penalty of 5, referring to the number of nucleotide substitutions within the SSR that are allowed to occur; (3) SSR seed minimum length of 8. which is the length (bps) of the seed SSR (without the insertion of a random nucleotide in the middle of the SSR; for example, a dinucleotide SSR repeated 4 times with an insertion occurring after the 7th nucleotide would have an SSR seed minimum length of 6); and (4) SSR seed minimum repeats of 3, the mismatch penalty, or the number of repeats occurring in the seed SSR. Once collected, this data was uploaded to Microsoft Excel (2019), then parsed and sorted to focus on the SSRs that were dinucleotide to hexanucleotide in repeat type (excluding mononucleotide repeats) and occurred at least 3 times.

2.2.4 Translations, Alignments, and creation of BED file

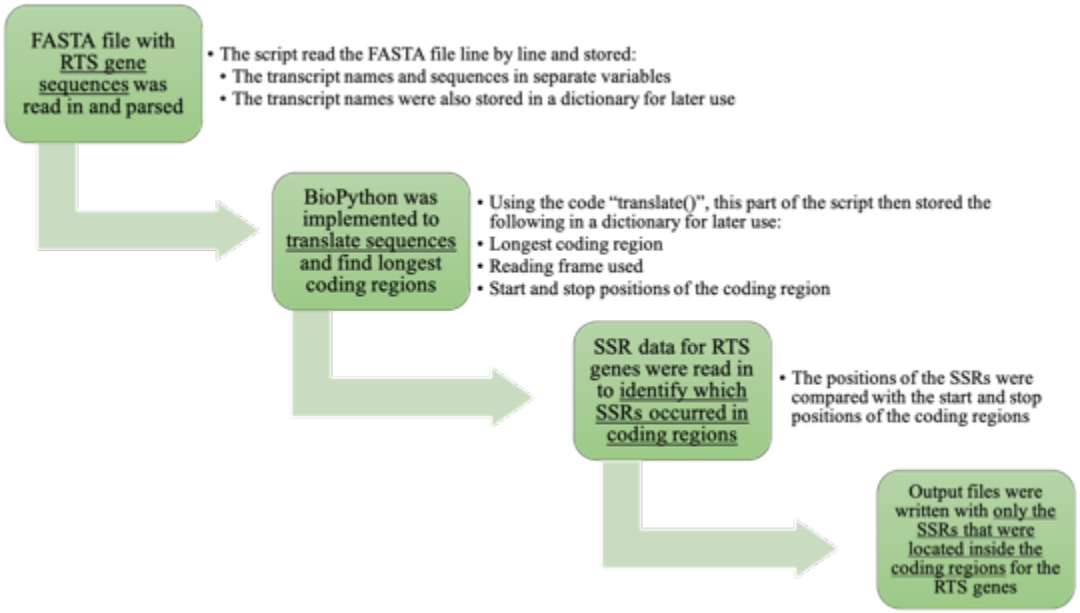


Figure 2.5 Flow chart for Python script developed to identify SSRs associated with the RTS genes and their coding regions.

Once the SSRs in the RTS transcripts were identified, another in-house Python script using Biopython version 1.78 (Cock et al., 2009) was developed to translate the different reading frames of the RTS transcripts to identify the longest coding region and store their start and stop positions, here forward referred to as the Chapter 2 Translate (C2T) Script and is highlighted in Figure 2.5. The script then parsed positions of the SSRs from the output file from SciRoKo 3.4 (Kofler et al., 2007) and compared them to the start and stop of the genes to determine if they occur inside or outside of the coding regions of the RTS transcripts. The SSRs that occurred inside and outside the coding regions were then written to separate output files. Another output file of the C2T script was a text document that gave details on the type of SSR that occurred in the transcript as well as its positions (see appendix A3 for the complete script).

Once the SSRs occurring inside and outside of the coding regions of the RTS transcripts were identified and stored in separate output files, the SSRs were aligned with the full RTS transcript using Clustal Omega (Sievers et al., 2011) to visually see where they occurred and to verify that the translate script was working correctly (Figure 2.6). The output file from the C2T script that contained details on the SSRs were then parsed for start and stop positions to create a BED file that was used with VCFtools (Danecek et al., 2011). In order to identify variability in SSR lengths, the start and stop positions in the BED file that correspond with the positions of the SSRs in the RTS transcripts were altered to extract mutations within 10 bps before the start position and 10 bps after the stop position. VCFtools (Danecek et al., 2011) was used to extract mutations within the positions of the BED file from the 8 samples in the *E. vaginatum* transcriptome. This was completed for SSRs occurring inside the RTS coding regions and SSRs occurring outside the coding regions.

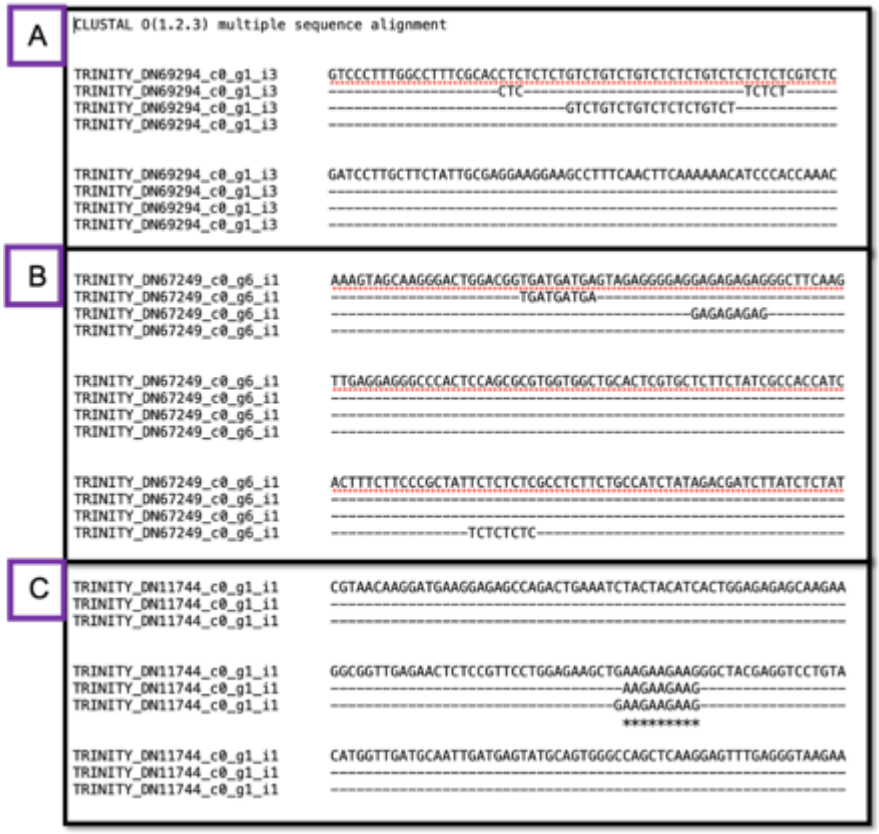


Figure 2.6 Examples of different alignments of RTS transcripts that had multiple SSRs using Clustal Omega. (A) RTS gene that had a long region of repeats, both Perfect and Imperfect. (B) RTS gene that contained multiple SSRs motifs within. (C) RTS gene where a perfect and imperfect SSR were recognized.

The output VCF files were then examined by hand to determine if the Variant Allele Frequency suggested a real SNP, or if a sequencing error occurred. The Variant Allele Frequency (VAF) is a breakdown of the different alleles present at a mutation, this is how a mutation can be classified as homozygous for the reference, heterozygous, or homozygous for the alternate. The VAF is a calculation determined by dividing the sequences the mutation is present in by the total allele depth. For example, if a VCF file shows that a mutation is present in a sample 5 times with a total depth of 9, the VAF would be calculated by dividing 5 by 9, yielding a VAF of 0.56. Generally speaking, I used a VAF from 0-0.25 is homozygous for the reference, 0.25-0.75 is heterozygous, and 0.75-1 is homozygous for the alternate, which is considered to be very

conservative. In this case, mutations that yielded VAFs between 0.25 and 1 were ideal, all other mutations (that yielded a VAF of <0.25 or were homozygous for the reference) were parsed out.

2.2.5 SNP Identification in RTS Genes

Using the FASTA file that was previously parsed with all the RTS transcript sequences and the same in-house python script that was constructed in Chapter 1, the Chapter 1 SNP Script, here forward referred to as C1S script, (to locate the Phytochrome genes; Figure 1.3) the full transcriptome VCF was parsed for SNP mutations that occur in these RTS genes. Parameters included an allele depth of 60 and a sample depth of 8, these parameters were used in an effort to narrow down the number of mutations found in these transcripts to a workable number. The RTS genes that displayed mutations within these parameters were written to an output directory. This script (C1S) was also altered from its use in Chapter 1 to create a BED file with positions of all RTS SNPs that were identified.

2.2.6 Primer Design

After the SSRs and SNPs in the RTS transcripts are identified, primer design for RTS SSRs will be conducted using SciRoKo 3.4, SciRoKo's Little Helper, and a Perl script that incorporates Primer3 (Kofler et al., 2007). All of these components are included with the SciRoKo 3.4 software (Kofler et al., 2007), primer design will follow similar parameters to Chapter 1. Sequence extractions would be conducted with SciRoKo's Little Helper, to collect extractions of up to 200bp on either side of the SSR of interest. The Perl script then uses the SciRoKo 3.4 output file with information on the identified SSRs and sequence extractions to create an output file with designed primers using Primer3. For RTS SNPs, primer design will be conducted in Geneious 10.0.9 (Kearse et al., 2012) to target regions of 200bps that include the SNPs of interest and will be dependent on the number of SNPs at a locus as well as the distance

apart. For both SSRs and SNPs, primer design will use the following parameters: Primer Size; 18bp to 24bp, Temperature melting point; 52°C to 58°C, GC content; 40% to 60%, and recognition of hairpins and primer dimers will be applied.

2.3 RESULTS

2.3.1 SSR Detection in RTS Genes

In efforts to only focus on transcripts with expression, the Data Matrix file from the full *E. vaginatum* transcriptome (Mohl et al., 2020) was used to parse transcripts that contained the RTS identifier. If the full WEGO file from the *E. vaginatum* transcriptome was used, the script would have yielded approximately 200 RTS transcripts, regardless of expression. There were 47 transcripts detected in the transcriptome that showed expression and were identified as RTS genes using the MS1 Python Script. The MS2 Python script was used to create a separate FASTA file with transcript names and sequences of the 47 RTS transcripts. The program SciRoKo 3.4 (Kofler et al., 2007) found a total of 44 Perfect SSRs and 104 Imperfect SSRs associated within these 47 RTS transcripts. The C2T script identified that of these SSRs, there was 1 hexanucleotide, 1 tetranucleotide, 40 trinucleotide, and 14 dinucleotide SSRs inside the coding regions of the RTS transcripts. The C2T script also identified that there were 4 pentanucleotide, 4 tetranucleotide, 25 trinucleotide, and 15 dinucleotide SSRs outside the coding region. After the RTS genes were translated and aligned with Clustal Omega (Sievers et al., 2011), there were 16 RTS transcripts that contained Perfect SSRs inside the coding region and the same 16 transcripts and an additional 13 transcripts contained Imperfect SSRs inside the coding region. There were 13 transcripts with Perfect SSRs and 26 transcripts with Imperfect SSRs that occurred outside of the RTS coding regions. All 13 transcripts that contained Perfect SSR motifs also had variability in an Imperfect SSR motif, similarly to the SSRs that occurred in

the coding regions. After creating alignments of the SSRs in the coding region and the corresponding full RTS transcript using Clustal Omega (Sievers et al., 2011), there were multiple SSR regions that had both Perfect and Imperfect motifs (Figure 2.6; Table 2.2). Altogether, there were 56 SSRs detected inside the coding region of the RTS transcripts and 48 SSRs detected outside the coding region (Table 2.2).

2.3.2 Variable SSRs in Transcriptome Samples

Using VCFtools (Danecek et al., 2011) and the constructed BED file, there were 7 SSRs that provided length variation among the transcripts from Mohl et al (2020) that occurred within the coding regions, and 1 outside the coding region that were parsed and added to new VCF files. However, once the Variant Allele Frequency was assessed, it was determined that these mutations were likely sequencing errors rather than true mutations occurring in the SSRs inside the coding regions as well as outside of the RTS transcripts.

2.3.3 SNP Detection in RTS Genes

Previously, arguments were implemented in the C1S script (Figure 1.3, see appendix A2 for full script) for ease of use in the future. The C1S Script (Figure 1.3) was written to take a subset of genes in the form of a FASTA file and transform sequence data with SNPs in the form of a VCF file. Parameters pertaining to sample depth and allele depth were then implemented to target only the genes with mutations containing a sample depth of 8 and allele depth of 60 (Figure 1.3). This script was altered to create a BED file of the positions of the SNPs in the RTS transcripts, this was done to keep track of how many SNPs met the given parameters for each transcript, and their exact positions can be used to identify which ecotypes the mutations occurred in as well as whether or not they occur in the coding regions of the RTS transcripts. The modified C1S script was used to detect SNPs occurring in the RTS genes with at least an allele

depth of 60 and a sample depth of 8. From the newly created VCF file, there were 28 RTS transcripts detected with 170 SNPs that correspond with the parameters mentioned above (Table 2.2).

2.3.4 Primer Design

Due to the limitations of SciRoKo 3.4 in re-uploading and reading SSR output files, I was unable to use this program to design primers for the identified RTS SSRs. The original SciRoKo 3.4 output file needed to be parsed for the RTS SSRs and re-uploaded to the program in order for SciRoKo's Little Helpers to perform the extractions that were then to be used with the Perl script and Primer3 to design the primers. These RTS SSRs included trinucleotide repeats that occur inside the coding regions and all other forms of SSRs that occur outside of the coding regions, these were identified using the C2T script. Due to time limitations, the SNP primers weren't designed. The C2T script is being developed for SNPs that occur in the coding regions. Once identified, primer pairs can be designed for SNP markers that show variability in the coding regions of RTS transcripts.

Table 2.2 Displays RTS transcripts with SSRs or SNPs among ecotypes with gene description and the types of genetic marker identifiers they contain. SSRs are categorized by whether or not they occur in the coding regions. The number of SNPs within each RTS transcript are noted in the last column.

Transcript	Gene Description	SSRs		SNPs
		Inside	Outside	
DN2248_c0_g1_i1	retrotransposon protein, putative, unclassified	1	3	
DN11744_c0_g1_i1	heat shock cognate protein 80	2		
DN60087_c0_g2_i1	heat shock 90	6		
DN62293_c0_g2_i1	predicted protein	5		1
DN65296_c0_g1_i1	universal stress family expressed	1	1	
DN66896_c0_g1_i5	histone deacetylase HDT1-like	1		3
DN66904_c0_g1_i6	ubiquitin receptor RAD23c-like	2		15
DN67150_c0_g1_i1	calmodulin-binding 60 A isoform X1		3	2
DN67150_c0_g1_i7	calmodulin-binding 60 A-like	1	1	2
DN67249_c0_g6_i1	universal stress A isoform X1	3	3	
DN68049_c0_g1_i5	universal stress A	2	2	2
DN68720_c0_g2_i1	ASR2	1		
DN69283_c1_g1_i1	Adenine nucleotide alpha hydrolase-like superfamily		2	
DN69294_c0_g1_i3	predicted protein	3	2	1
DN70026_c0_g1_i2	activator of 90 kDa heat shock ATPase homolog		2	11
DN70789_c0_g2_i1	heat shock cognate 80		2	4
DN70789_c0_g2_i5	heat shock 83	4	1	13
DN71373_c0_g1_i7	calmodulin-binding 60 B	1	5	4
DN72618_c0_g6_i1	U-box domain-containing 33-like isoform X1		1	
DN72644_c2_g4_i5	xanthoxin dehydrogenase-like		2	
DN73671_c0_g2_i7	predicted protein	1		
DN73755_c0_g2_i1	Pre-mRNA-processing factor 19 homolog 2		1	
DN74061_c0_g1_i1	ethylene-responsive transcription factor 5-like		1	17
DN74431_c0_g7_i1	mediator of RNA polymerase II transcription subunit 32-like			3
DN75232_c0_g9_i4	XP_008778664.1 enolase-like	2	1	5
DN75407_c0_g1_i4	probable zinc metallopeptidase EGY3, chloroplastic	3	2	7
DN75807_c2_g1_i1	plant UBX domain-containing 2-like			3
DN75807_c2_g1_i2	XP_020083507.1plant UBX domain-containing protein 2			6
DN75807_c2_g1_i3	XP_020083507.1plant UBX domain-containing protein 2			9
DN76530_c0_g1_i3	retrotransposon unclassified	2	5	4
DN76906_c1_g2_i1	fumarate hydratase 1	2	1	
DN77223_c0_g1_i2	Universal stress protein A-like protein	1		1
DN77223_c0_g1_i3	Universal stress protein A-like protein	1		6
DN77621_c0_g1_i5	bromodomain-containing protein, putative	1		3
DN78174_c1_g6_i1	water-stress inducible	1	1	4
DN78174_c1_g8_i2	abscisic stress ripening	1	1	3
DN78253_c2_g14_i10	heat shock 70 kDa 17	2	1	
DN78557_c0_g15_i1	ycf3-interacting protein 1, chloroplastic			3
DN78557_c0_g15_i2	protein CHLOROPLAST ENHANCING STRESS TOLERANCE, chloroplastic			4
DN78639_c0_g16_i1	heat shock protein 90-6, mitochondrial	2	2	30
DN79210_c3_g3_i2	probable serine/threonine-protein kinase GCN2-like	1	1	
DN79301_c0_g4_i3	single-stranded DNA-binding mitochondrial	1	1	4
DN130620_c0_g1_i1	endoplasmin	2		

2.4 DISCUSSION

Given the evidence of phenotypic variation in *E. vaginatum* (Parker et al., 2017) along the latitudinal gradient in northern Alaska, a better understanding of *E. vaginatum's* molecular response to the changing climate in the Arctic is needed. The purpose of this work was to identify genetic markers in the RTS genes of the *E. vaginatum* transcriptome that are likely to be variable among ecotypes, then to develop tools that can be used for future selection studies directed toward understanding the molecular response of *E. vaginatum* under environmental pressures. A major goal of this study was to design bioinformatic tools to utilize in this and future work that can be easily altered to incorporate different input data files and parameters developed for specific tasks related to the transcriptome. Here, bioinformatic tools were designed for a subset of RTS transcripts but can be applied to the full *E. vaginatum* transcriptome or other large data sets of interest.

2.4.1 RTS Gene Identification

The scripts developed for this research were highly effective for identifying and extracting the RTS gene transcripts for this project. 47 RTS transcripts were identified in the *E. vaginatum* transcriptome, of which, 44 RTS transcripts, or 36 RTS genes contained at least one of the 274 SSRs or SNPs identified (Table 2.2). The program VCFtools (Danecek et al., 2011) and the C2T script did not identify SSRs located inside the coding regions of the RTS genes that varied among the 8 samples in the transcriptome. The C1S script identified that there were 28 RTS transcripts with SNPs that were variable among the samples in the transcriptome. The scripts (C1S and C2T) were designed to be easily modified to search for other regions of interest in the transcriptome such as different gene families or genes with different GO IDs. This was done by implementing arguments in the scripts that can be called from the command line, these

different arguments correspond with critical components of the script such as input files, parameters, output files, and output directories.

Some of the RTS transcripts identified that contained SSRs or SNPs could be associated to transcripts that showed differential expression associated with adaptation found by Mohl et al. (2020). For example, one RTS gene, DN70789, is a heat shock protein (HSP) that, showed expression level variation in all eight samples in the transcriptome on both the ambient temperature day (13.8°C) and the extreme heat day (26.6°C) (Mohl et al., 2020). This gene had two isoforms and contained SSRs both inside and outside of the coding region for which markers could be designed. There are also SNPs identified among the DN70789 isoforms (Table 2.2). These findings make this transcript a prime candidate to investigate selection among the latitudinal gradient in Alaska. There were other HSPs with associated SSRs and SNPs (e.g. DN78253, DN11744, DN60087, and DN78639) that did not show variation in expression levels in response to heat stress (Mohl et al., 2020), but may still be useful for examining other stress responses.

2.4.3 SSRs and SNPs

SSRs with multiple alleles were found to be associated with 22 RTS transcripts. Of these transcripts, 19 had more than one SSR region for which markers could be developed. Due to the higher likelihood of variability found when using dinucleotide SSR markers they are frequently targeted for designing and selecting primer pairs (Chakraborty et al., 1997; Levinson & Gutman, 1987; Merritt et al., 2015), despite their locations likely being outside of the coding regions. There were almost the same number of dinucleotide repeats located inside the coding regions versus outside for the RTS transcripts. When the dinucleotide repeats that were found inside the coding region were examined for size variability among ecotypes, no variation occurred among

the samples in the transcriptome. Due to variability in the length of dinucleotide repeats shifting the reading frame, no variation was expected. However, variation among ecotypes was also absent for the trinucleotide and hexanucleotide repeats inside the coding regions, that would not shift the reading frame but could still alter amino acids. While SSRs inside the coding regions could directly identify selection directed at RTS genes, SSRs closely associated outside the coding regions of the RTS transcripts is more likely and targeting length variability among these for multiple ecotypes will be a future priority.

For future work, the C2T script (Figure 2.5) needs to be enhanced to identify genetic markers at a more proximal distance to the coding regions of genes of interest, starting at 10 nucleotides and expanding further if needed, rather than just identifying if a genetic marker is inside or outside of the coding region. Another component that would be valuable is implementing an argument that can be called from the command line in order to easily alter the number of base pairs on either side of the coding region identified. Interest lies in the region just outside of coding regions due to its close linkage to the gene. This region is associated ribosomal recruitment in the 5' cap, where the ribosome binds and translation is initiated (Hellen & Sarnow, 2001) and is more prone to variability as it lacks the structural constraints of the coding region.

The C1S script was created to identify transcripts containing SNPs that met specified parameters for allele and sample depth and store the transformed sequence data in an output directory (Figure 1.3). However, due to the need to identify how many SNPs met the parameters of the script and occurred in each transcript, the C1S script was altered to store the positions of the SNPs in the RTS genes by implementing a Boolean type, and then creating a BED file with these positions. The program VCFtools (Danecek et al., 2011) used the created BED file and the

full *E. vaginatum* VCF file to parse the SNPs occurring in the RTS transcripts. Using this modified script 170 individual SNPs were identified in 28 transcripts. Modification of the script will be needed to identify if there are ecotype specific alleles among the samples used in Mohl et al. (2020) or allelic bias for northern vs southern ecotype. The C1S script will be modified by implementing a loop that sorts the mutations by sample, which ecotype they occurred in and on which temperature day (either of ambient or extreme heat).

There is also a need to further enhance the C2T script to read in the positions of the SNPs to determine if they occur inside or outside of the RTS transcript coding regions. This will be done by altering the part of the script that reads the positions of the SSR BED files. There is currently a parameter for the type of SSR (dinucleotide, trinucleotide, tetranucleotide…), which will be removed in order for the SNP BED file to be read in and processed. Ideally, the C1S and C2T scripts will be combined to form one script that transforms RTS sequence data and identifies prominent SNPs, creates a BED file with the positions of these SNPs, translates the RTS sequences to find the longest coding region, and stores SNPs that occur in the coding regions of the RTS transcript in an output directory.

Although, not feasible under the time constraints of this project, the identified SSR and SNP markers can be examined across the populations sampled for Mohl et al. (2020) to identify if there is allelic difference that correlates with northern and southern ecotypes. This can be done using a series of contingency Fisher's Exact Tests (FET), Bonferroni correction tests, and R statistical package (R Core Development Team, 2019). If there is a significant relationship between the presence of an allele with an ecotype these markers will be candidate markers for studying RTS genes involved in adaptation with additional population level sampling.

2.4.5 Future Research

Due to limitations of RNA-seq and the quality of the transcriptomic sequence data available for *E. vaginatum,* the parsed RTS genes were referred to as transcripts, this is due to some genes having multiple isoforms (example: DN77223_c0_g1_i2 and DN77223_c0_g1_i3; Table 2.2) that vary by length and coverage in the transcriptome sequence data. With the limited number of samples in the *E. vaginatum* transcriptome, and the possibility of multiple gene copies present in the full genome, it is not currently possible to determine if the isoforms are duplicate genes. Complete genome sequence data of *E. vaginatum* would potentially lead to more clarity on whether these isoforms are functionally different, and the associate allelic diversity would be informative for each.

References

Ashburner, M., Ball, C., Blake, J., Botstein, D., & Butler, H. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(may), 25–29. https://doi.org/10.1038/75556

Bagnall, D. J., King, R. W., Whitelam, G. C., Boylan, M. T., Wagner, D., & Quail, P. H. (1995). Flowering responses to altered expression of phytochrome in mutants and transgenic lines of Arabidopsis thaliana (L.) Heynh. *Plant Physiology*, *108*(4), 1495–1503. https://doi.org/10.1104/pp.108.4.1495

Bennington, C. C., Fetcher, N., Vavrek, M. C., Shaver, G. R., Cummings, K. J., & Mcgraw, J. B. (2012). Home site advantage in two long-lived arctic plant species: Results from two 30-year reciprocal transplant studies. *Journal of Ecology*, *100*(4), 841–851. https://doi.org/10.1111/j.1365-2745.2012.01984.x

Borner, A. P., Kielland, K., & Walker, M. D. (2013). Effects of Simulated Climate Change on Plant Phenology and Nitrogen Mineralization in Alaskan Arctic Tundra. *Arctic, Antarctic, and Alpine Research*, *40*(1), 27–38. https://doi.org/10.1657/1523-0430(06-099)

Casal, J. J., Sanchez, R. A., & Yanovsky, M. J. (1997). The function of phytochrome A. *Plant, Cell and Environment*, *20*(6), 813–819. https://doi.org/10.1046/j.1365-3040.1997.d01-113.x

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J., & Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(3), 1041–1046. https://doi.org/10.1073/pnas.94.3.1041

Chapin, F. S., Shaver, G. R., Giblin, A. E., Nadelhoffer, K. J., & Laundre, J. A. (1995). Responses of Arctic tundra to experimental and observed changes in climate. *Ecology*, *76*(3), 694–711. https://doi.org/10.2307/1939337

Chen, A., Li, C., Hu, W., Lau, M. Y., Lin, H., Rockwell, N. C., … Dubcovsky, J. (2014). PHYTOCHROME C plays a major role in the acceleration of wheat flowering under long-day photoperiod. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), 10037–10044. https://doi.org/10.1073/pnas.1409795111

Cleland, E. E., Chuine, I., Menzel, A., Mooney, H. A., & Schwartz, M. D. (2007). Shifting plant phenology in response to global change. *Trends in Ecology and Evolution*, *22*(7), 357–365. https://doi.org/10.1016/j.tree.2007.04.003

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Curasi, S. R., Parker, T. C., Rocha, A. V., Moody, M. L., Tang, J., & Fetcher, N. (2019). Differential responses of ecotypes to climate in a ubiquitous Arctic sedge: implications for future ecosystem C cycling. *New Phytologist*, *223*(1), 180–192. https://doi.org/10.1111/nph.15790

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

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). JModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*, *9*(8), 772. https://doi.org/10.1038/nmeth.2109

Ding, J., & Nilsson, O. (2016). Molecular regulation of phenology in trees-because the seasons they are a-changin'. *Current Opinion in Plant Biology*, *29*(Figure 1), 73–79.

https://doi.org/10.1016/j.pbi.2015.11.007

Elmer, K. R., & Meyer, A. (2011). Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Erschienen in: Trends in Ecology & Evolution*, *26*, 298–306. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169534711000528

Emanuelli, F., Lorenzi, S., Grzeskowiak, L., Catalano, V., Stefanini, M., Troggio, M., … Grando, M. S. (2013). Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biology*, *13*(1), 1–17. https://doi.org/10.1186/1471-2229-13-39

Fetcher, N., & Shaver, G. (1982). *Growth and tillering patterns within tussocks of Eriophorum vaginatum*. *5*(2).

Gaut, B. S. (2001). Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses [3]. *Genome Research*, *11*(1), 55–66. https://doi.org/10.1101/gr.160601

Halliday, K. J., & Davis, S. J. (2016). Light-sensing phytochromes feel the heat. *Science*, *354*(6314), 832–833. https://doi.org/10.1126/science.aaj1918

Han, Z., Ma, X., Wei, M., Zhao, T., Zhan, R., & Chen, W. (2018). SSR marker development and intraspecific genetic divergence exploration of Chrysanthemum indicum based on transcriptome analysis. *BMC Genomics*, *19*(1), 291. https://doi.org/10.1186/s12864-018-4702-1

Hellen, C. U. T., & Sarnow, P. (2001). Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes and Development*, *15*(13), 1593–1612. https://doi.org/10.1101/gad.891101

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., & Hood, L. (1997). Gene

families: The taxonomy of protein paralogs and chimeras. *Science*, *278*(5338), 609–614. https://doi.org/10.1126/science.278.5338.609

Hill, C. B., & Li, C. (2016). Genetic architecture of flowering phenology in cereals and opportunities for crop improvement. *Frontiers in Plant Science*, *7*(DECEMBER2016), 1–23. https://doi.org/10.3389/fpls.2016.01906

Hobbie, J., & Kling, G. (Eds.). (2014). *Alaska's Changing Arctic Ecological Consequences for Tundra, Streams, and Lakes*. Oxford.

Hodel, R. G. J., Gitzendanner, M. A., Germain-Aubrey, C. C., Liu, X., Crowl, A. A., Sun, M., … Soltis, P. S. (2016). A New Resource for the Development of SSR Markers: Millions of Loci from a Thousand Plant Transcriptomes. *Applications in Plant Sciences*, *4*(6), 1600024. https://doi.org/10.3732/apps.1600024

Holmes, M. G., & Smith, H. (1977). the Function of Phytochrome in the Natural Environment—Ii. the Influence of Vegetation Canopies on the Spectral Energy Distribution of Natural Daylight. *Photochemistry and Photobiology*, *25*(6), 539–545. https://doi.org/10.1111/j.1751-1097.1977.tb09125.x

Honig, J. A., Zelzion, E., Wagner, N. E., Kubik, C., Averello, V., Vaiciunas, J., … Meyer, W. A. (2017). Microsatellite Identification in Perennial Ryegrass using Next-Generation Sequencing. *Crop Science*, *57*, 331–340. https://doi.org/10.2135/cropsci2016.07.0608

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Hyles, J., Bloomfield, M. T., Hunt, J. R., Trethowan, R. M., & Trevaskis, B. (2020). Phenology and related traits for wheat adaptation. *Heredity*, *125*(6), 417–430. https://doi.org/10.1038/s41437-020-0320-1

Jain, S. M., Darshan, S. B., & B. S. Ahloowalia, E. (2010). *Molecular Techniques in Crop Improvement*. New York, NY: Springer US.

Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., & Dhawan, A. K. (2011). Microsatellite markers: An overview of the recent progress in plants. *Euphytica*, *177*(3), 309–334. https://doi.org/10.1007/s10681-010-0286-9

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kofler, R., Schlötterer, C., & Lelley, T. (2007). SciRoKo: A new tool for whole genome microsatellite search and investigation. *Bioinformatics*, *23*(13), 1683–1685. https://doi.org/10.1093/bioinformatics/btm157

König, K., Peifer, M., Fassunke, J., Ihle, M. A., Künstlinger, H., Heydt, C., … Heukamp, L. C. (2015). Implementation of amplicon parallel sequencing leads to improvement of diagnosis and therapy of lung cancer patients. *Journal of Thoracic Oncology*, *10*(7), 1049–1057. https://doi.org/10.1097/JTO.0000000000000570

Krinner, G., Germany, F., Shongwe, M., Africa, S., France, S. B., Uk, B. B. B. B., … Lucas, C. (2013). Long-term climate change: Projections, commitments and irreversibility. In P. M. M. Thomas F. Stocker, Dahe Qin, Gian-Kasper Plattner, Melinda M.B. Tignor, Simon K. Allen, Judith Boschung, Alexander Nauels, Yu Xia, Vincent Bex (Ed.), *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Vol. 9781107057, pp. 1029–1136). https://doi.org/10.1017/CBO9781107415324.024

Kudoh, H. (2016). Molecular phenology in plants: In natura systems biology for the comprehensive understanding of seasonal responses under natural environments. *New Phytologist*, *210*(2), 399–412. https://doi.org/10.1111/nph.13733

Labbé, J., Murat, C., Morin, E., Le Tacon, F., & Martin, F. (2011). Survey and analysis of simple sequence repeats in the Laccaria bicolor genome, with development of microsatellite markers. *Current Genetics*, *57*(2), 75–88. https://doi.org/10.1007/s00294-010-0328-9

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, *4*(3), 203–221. https://doi.org/10.1093/oxfordjournals.molbev.a040442

Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A., & Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, *11*, 2453–2465.

Lin, C. (2000). Photoreceptors and Regulation of Flowering Time. *Plant Physiology*, *123*(1), 39–50.

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., & Abe, J. (2008). Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics*, *180*(2), 995–1007. https://doi.org/10.1534/genetics.108.092742

Mattila, T. M., Aalto, E. A., Toivainen, T., Niittyvuopio, A., Piltonen, S., Kuittinen, H., & Savolainen, O. (2016). Selection for population-specific adaptation shaped patterns of variation in the photoperiod pathway genes in Arabidopsis lyrata during post-glacial colonization. *Molecular Ecology*, *25*(2), 581–597. https://doi.org/10.1111/mec.13489

Mcgraw, J. B., Turner, J. B., Souther, S., Bennington, C. C., Vavrek, M. C., Shaver, G. R., & Fetcher, N. (2015). Northward displacement of optimal climate conditions for ecotypes of

Eriophorum vaginatum L. across a latitudinal gradient in Alaska. *Global Change Biology*, *21*(10), 3827–3835. https://doi.org/10.1111/gcb.12991

Merritt, B. J., Culley, T. M., Avanesyan, A., Stokes, R., & Brzyski, J. (2015). An Empirical Review: Characteristics of Plant Microsatellite Markers that Confer Higher Levels of Genetic Variation. *Applications in Plant Sciences*, *3*(8), 1500025. https://doi.org/10.3732/apps.1500025

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop, GCE 2010*. https://doi.org/10.1109/GCE.2010.5676129

Mockler, T. C., Guo, H., Yang, H., Duong, H., & Lin, C. (1999). Antagonistic actions of Arabidopsis cryptochromes and phytochrome B in the regulation of floral induction. *Development*, *126*(10), 2073–2082.

Mohl, J. E., Fetcher, N., Stunz, E., Tang, J., & Moody, M. L. (2020). Comparative transcriptomics of an arctic foundation species, tussock cottongrass (Eriophorum vaginatum), during an extreme heat event. *Scientific Reports*, *10*(1), 1–14. https://doi.org/10.1038/s41598-020-65693-8

Morgante, M., & Olivieri, A. M. (1993a). PCR-amplified microsatellites as markers in plant genetics. *The Plant Journal*, *3*(1), 175–182.

Morgante, M., & Olivieri, A. M. (1993b). PCR-amplified microsatellites as markers in plant genetics. *Plant Journal*, *3*(1), 175–182. https://doi.org/10.1111/j.1365-313x.1993.tb00020.x

Mudunuri, S. B., & Nagarajaram, H. A. (2007). IMEx: Imperfect microsatellite extractor. *Bioinformatics*, *23*(10), 1181–1187. https://doi.org/10.1093/bioinformatics/btm097

Oberbauer, S. F., Tweedie, C. E., Welker, J. M., Fahnestock, J. T., Henry, G. H. R., Webber, P.

J., … Starr, G. (2007). Tundra CO2 fluxes in response to experimental warming across latitudinal and moisture gradients. *Ecological Monographs*, *77*(2), 221–238. https://doi.org/10.1890/06-0649

Parker, T. C., Tang, J., Clark, M. B., Moody, M. M., & Fetcher, N. (2017). Ecotypic differences in the phenology of the tundra species Eriophorum vaginatum reflect sites of origin. *Ecology and Evolution*, *7*(22), 9775–9786. https://doi.org/10.1002/ece3.3445

Parker, T. C., Unger, S., Moody, M. L., Tang, J., & Fetcher, N. (2021). *The North Remembers: Phenology of northern ecotypes of Eriophorum vaginatum responds less to simulated climate change in the tundra*.

Pavey, S. A., Bernatchez, L., Aubin-Horth, N., & Landry, C. R. (2012). What is needed for next-generation ecological and evolutionary genomics? *Trends in Ecology and Evolution*, *27*(12), 673–678. https://doi.org/10.1016/j.tree.2012.07.014

Pramod, S., Perkins, A., & Welch, M. (2014). Patterns of microsatellite evolution inferred from the Helianthus annuus (Asteraceae) transcriptome. *Journal of Genetics*, *93*(2), 431–442. https://doi.org/10.1007/s12041-014-0402-z

Provan, J., Powell, W., & Hollingsworth, P. M. (2001). Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends in Ecology and Evolution*, *16*(3), 142–147. https://doi.org/10.1016/S0169-5347(00)02097-8

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., & Fischer, M. C. (2013). Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0080422

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-

Onsins, S. E., & Sanchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, *34*(12), 3299–3302. https://doi.org/10.1093/molbev/msx248

Schippers, J. H. M. (2015). Transcriptional networks in leaf senescence. *Current Opinion in Plant Biology*, *27*(Figure 1), 77–83. https://doi.org/10.1016/j.pbi.2015.06.018

Schmitt, J., Dudley, S. A., & Pigliucci, M. (1999). Manipulative approaches to testing adaptive plasticity: Phytochrome-mediated shade-avoidance responses in plants. *American Naturalist*, *154*(SUPPL.). https://doi.org/10.1086/303282

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18. https://doi.org/10.1038/nmeth1156

Scott, K. D., Eggler, P., Seaton, G., Rossetto, M., Ablett, E. M., Lee, L. S., & Henry, R. J. (2000). Analysis of SSRs derived from grape ESTs. *Theoretical and Applied Genetics*, *100*(5), 723–726. https://doi.org/10.1007/s001220051344

Senetar, M. A., & McCann, R. O. (2005). Gene duplication and functional divergence during evolution of the cytoskeletal linker protein talin. *Gene*, *362*(1–2), 141–152. https://doi.org/10.1016/j.gene.2005.08.012

Sheehan, M. J., Farmer, P. R., & Brutnell, T. P. (2004). Structure and expression of maize phytochrome family homeologs. *Genetics*, *167*(3), 1395–1405. https://doi.org/10.1534/genetics.103.026096

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal

Omega. *Molecular Systems Biology*, *7*(539). https://doi.org/10.1038/msb.2011.75

Souther, S., Fetcher, N., Fowler, Z., Shaver, G. R., & McGraw, J. B. (2014). Ecotypic differentiation in photosynthesis and growth of Eriophorum vaginatum along a latitudinal gradient in the Arctic tundra. *Botany*, *92*(8), 551–561. https://doi.org/10.1139/cjb-2013-0320

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M. M. B., Allen, S. K., Boschung, J., … Midgley, P. M. (2013). Climate change 2013 the physical science basis: Working Group I contribution to the fifth assessment report of the intergovernmental panel on climate change. *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, *9781107057*, 1–1535. https://doi.org/10.1017/CBO9781107415324

Team, R. D. C. (2019). *R: A language and environment for statistical computing*. (Vol. 3). Retrieved from https://www.r-project.org/

Temnykh, S., Park, W., Ayres, N., Cartinhour, S., Hauck, N., Lipovich, L., … McCouch, S. R. (2000). *Mapping and genome organization of microsatellite sequences in rice (Oryza sativa L.)*. 697–712.

van Inghelandt, D., Melchinger, A. E., Lebreton, C., & Stich, B. (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theoretical and Applied Genetics*, *120*(7), 1289–1299. https://doi.org/10.1007/s00122-009-1256-2

Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genic microsatellite markers in plants: Features and applications. *Trends in Biotechnology*, *23*(1), 48–55. https://doi.org/10.1016/j.tibtech.2004.11.005

Varshney, R. K., Mahendar, T., Aggarwal, R. K., & Börner, A. (2007). Genic molecular markers in plants: Development and applications. *Genomics-Assisted Crop Improvement*, *1*, 13–29. https://doi.org/10.1007/978-1-4020-6295-7_2

Wang, Z., Weber, J. L., Zhong, G., & Tanksley, S. D. (1994). Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics*, *88*(1), 1–6. https://doi.org/10.1007/BF00222386

Waples, R. S., & Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, *15*(6), 1419–1439. https://doi.org/10.1111/j.1365-294X.2006.02890.x

Wessinger, C. A., Kelly, J. K., Jiang, P., Rausher, M. D., & Hileman, L. C. (2018). SNP-skimming: A fast approach to map loci generating quantitative variation in natural populations. *Molecular Ecology Resources*, *18*(6), 1402–1414. https://doi.org/10.1111/1755-0998.12930

Wright, S. I., & Charlesworth, B. (2004). The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics*, *168*(2), 1071–1076. https://doi.org/10.1534/genetics.104.026500

Zane, L., Bargelloni, L., & Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*, *11*(1), 1–16. https://doi.org/10.1046/j.0962-1083.2001.01418.x

Zhang, D. X., & Hewitt, G. M. (2003). Nuclear DNA analyses in genetic studies of populations: Practice, problems and prospects. *Molecular Ecology*, *12*(3), 563–584. https://doi.org/10.1046/j.1365-294X.2003.01773.x

Zhao, X., Liu, H., Wei, X., Wu, L., Cheng, F., Ku, L., … Chen, Y. (2014). Promoter region characterization of ZmPhyB2 associated with the photoperiod-dependent floral transition in maize (Zea mays L.). *Molecular Breeding*, *34*(3), 1413–1422.

https://doi.org/10.1007/s11032-014-0125-0

Appendix

A. Python Scripts

A1. Chapter 1 Preliminary Script

```
import re,random, argparse

#looking for mutations in general across all Phytochrome genes

#Arguments that can be altered in the terminal
parser = argparse.ArgumentParser()
parser.add_argument('-i', '--VCF_input_file', type=str)
parser.add_argument('-o', '--parsed_output_file', type=str)
parser.add_argument('-a', '--allele_depth', type=int)
parser.add_argument('-m', '--min_allele_freq', type=float)

args = parser.parse_args()

#VCF input file
if args.VCF_input_file:
    vcfIn = args.VCF_input_file
else:
    print('Error: Need VCF file')

#parsed output file (showing there are mutations in samples)
if args.parsed_output_file:
    fileOut = args.parsed_output_file
else:
    print('Error: Output file already exists')

#allele depth (only care about transcripts with _% for alternate)
if args.allele_depth:
    Depth = args.allele_depth
else:
    print('Error: No mutations at that allele depth')

#minimum allele frequency
if args.min_allele_freq:
    Freq = args.min_allele_freq
else:
    print('Error: No mutations at that allele frequency')

#Begin script
#PART 1 reading in and parsing FASTA file
f = open(vcfIn,'r')
```

```
lines = f.readlines()
f.close()
lines
temp = ''.join(lines).split('#')[-1]
geno = temp.strip().split('\n')

need = [] #list of sample lists
perc = [] #alternate numbers

#making a list of lists for samples
for g in geno[1:]:
        ge = g.strip().split('\t')
        need.append(ge)



#PART 2 creating output file with genes containing mutations
#Writing samples of interest (>= _allele depth and >_% for alt min allele freq)
f = open(fileOut,"w+")
f.write(geno[0] + '\n')
focus = []



#PART 3 calculating allele depth and minimum allele frequencey
#focus on 7th element where mutation quality is located
for ge in need:
        if ge[4] != '<*>':
          if ge[7].split(';')[0] == 'INDEL':
                  temp = ge[7].split(';')[4].strip('AD=').split(',')
                  if int(temp[0]) + int(temp[1]) >= Depth and float(temp[1]) / (int(temp[0]) +
int(temp[1])) > Freq:
                          focus.append([temp[0],temp[1]])
                          f.write('\t'.join(ge) + '\n')
          else:
                  temp = ge[7].strip().split(';')[1].strip('AD=').split(',')
                  perc.append(temp)
                  if int(temp[0]) + int(temp[1]) >= Depth and float(temp[1]) / (int(temp[0]) +
int(temp[1])) > Freq:
                          focus.append([temp[0],temp[1]])
                          f.write('\t'.join(ge) + '\n')
f.close()
```

A2. Chapter 1 SNP Script (C1S)

```
import re,random, os, sys, argparse

#Arguments that can be altered in the terminal
parser = argparse.ArgumentParser()
parser.add_argument('-f', '--fasta_file_name', type=str) #look up add.argument = required
parser.add_argument('-v', '--vcf_file_name', type=str) #look up add.argument = required. Need to
have in order to run.
parser.add_argument('-ad', '--allele_depth', type=int)
parser.add_argument('-sd', '--sample_depth', type=int)
parser.add_argument('-o', '--output_directory', type=str) #send files to an output directory

args = parser.parse_args()

#fasta file
if args.fasta_file_name:
        fastaIn = args.fasta_file_name
else:
        print('Error: Need file containing fasta sequences')

#vcf file
if args.vcf_file_name:
        vcfIn = args.vcf_file_name
else:
        print('Error: Need file containing variant call format')

#allele depth
if args.allele_depth:
        allele_depth = args.allele_depth
else:
        allele_depth = 60

#sample depth
if args.sample_depth:
        sample_depth = args.sample_depth
else:
        sample_depth = 8

#output file make directory
if args.output_directory:
    Outdir = args.output_directory
    if os.path.exists(Outdir):
        print('output folder already exists, choose new folder')
        quit()
    else:
```

```
        os.mkdir(Outdir)
else:
        print('Error: Output directory name not given')



#PART 1
#Bring in sequence dictionary (find_seq.py)

f = open(fastaIn, 'r')
lines = f.readlines()
f.close()

#cleaning up sequences
seq = ''.join(lines)[4:].split('>')

search = {}

#storing headers in dictionary
for s in seq:
        temp = s.split('\n')
        head = temp[0]
        seqs = temp[1]
        search[head] = seqs



#PART 2
#Reading in VCF and determining if allele and sample depth qualify per parameters
positions = []

g = ''
#Reading in VCF
#For loop to go through sites on protein
with open(vcfIn, 'r') as fp:
    l = fp.readline()
    while l[0:2] == '##':
        l = fp.readline()
    #headers of individual sequence locations
    header2 = l.split('\t')[9:]

    head =[]
    all = []
    #creating headers for transformed sequence data
    #'a' for reference and 'b' for alternate mutations
    for h in header2:
```

```
        h = h.split('.')[0]
        head.append('>' + h + 'a')
        head.append('>' + h + 'b')

#Begin code for AD and SD parameters
    l = fp.readline()
    while l:
        sl = l.strip().split('\t')

        #determine AD
        if sl[0] in search.keys():
            AD1 = sl[7].split(';')
            AD2 = AD1[1].split('=')
            ADfinal = AD2[1].split(',')
            ADtotal = 0
            for x in range(0,len(ADfinal)):
                ADtotal += int(ADfinal[x])
            if ADtotal >= allele_depth and sl[0] in search.keys():

                #if the allele depth qualifies, then move to sample depth
                #determine SD
                SD1 = sl[9:]
                SD = 0
                for s in SD1:
                    if re.search('[1-9]',s):
                        SD += 1
                if SD >= sample_depth:
                #if both allele depth and sample depth qualify, transcript name will be printed in the
terminal
                    print(sl[0])
                    if sl[0] != g:
                        #Wrap up old
                        if g != '':

                            #writing individual output files with transformed data
                            f = open("%s/TRINITY_%s_individ_sequences.txt"%(Outdir,g),"w+")
                            for a in range (0,16):
                                f.write(head[a] + '\n')
                                f.write(''.join(all[a]) + '\n')
                            f.close()
                        #Start new
                        g=sl[0]
                        all = []
                        print(search[g])
                        for r in range(0,16):
                            all.append(list(search[g]))
```

```
            print(all.append(list(search[g])))



#PART 3
#Begin code for data transformation, make 16 copies of each sequence, 2 for each sample
(reference and alternate)
            #Get info for all samples
            ref = l.split('\t')[3]
            alt = l.split('\t')[4].split(',')[0]

            #For loop to go through samples
            i=0
            pos = int(l.split('\t')[1])-1

            #Begin boolean type for SNP BED file
            Boo = False
            for t in l.split('\t')[9:]:
                st = t.split(':')[-1].split(',')
                #assigning variables for reference and alternate nucleotides
                trc = int(st[0])
                tac = int(st[1])

                #for heterozygotes, randomly assigned to use ref or alt
                if trc != 0 and tac != 0:
                    if float(trc) / (trc+tac) > 0.4 and float(trc) /(trc+tac) < 0.7: #hetero
                        Boo = True
                        r = random.random()
                        if r >= 0.5:
                            all[i][pos] = alt #alt
                        else:
                            all[i+1][pos] = alt #ref

                    #homozygous for the alternate
                    elif float(tac) / trc > 0.2: #homo alt
                        all[i][pos] = alt
                        all[i+1][pos] = alt
                        Boo = True
                elif trc == 0 and tac != 0: #homo alt
                    all[i][pos] = alt
                    all[i+1][pos] = alt
                    Boo = True
                i+=2

            #Ending boolean loop and storing in "positions" variable
            if Boo == True:
```

```
                positions.append(sl[0]+'\t'+l.split('\t')[1])
        l = fp.readline()



#PART 4
#Writing positions of SNPs to BED file
#BED file contains positions of SNPs in transcript and how many SNPs each transcript contains that meet the parameters
#Boolean type parameters, True = hetero and homo for the alternate, no homo for the reference
v = open("SNP_positions.txt", "w+")
v.write('\n'.join(positions))
v.close()



#PART 5
#Write to Output directory (with individual fasta files)
if g != '':
        f = open("%s/TRINITY_%s_individ_sequences.txt"%(Outdir,g),"w+")
        for a in range (0,16):
          f.write(head[a] + '\n')
          f.write(''.join(all[a]) + '\n')
          f.close()
```

A3. Chapter 2 Translate Script (C2T)

```
from Bio import Seq
from Bio.Seq import Seq
import re, sys, argparse, os

#Arguments that can be altered in the terminal
parser = argparse.ArgumentParser()
parser.add_argument('-f', '--fasta_file_name', type=str)
parser.add_argument('-s', '--SciRoKo_output_file_name', type=str) #td file type
parser.add_argument('-l', '--SSR_minimum_repeat', type=int) #2=dinucleotide
parser.add_argument('-out_file', '--output_file_name', type=str) #Output file contains transcript name, motif, and positions only where SSRs occur in the coding regions
parser.add_argument('-out_fasta', '--output_fasta_file_names', type=str) #fasta files contain full gene sequence with SSR sequences that occur in the coding regions

args = parser.parse_args()

#FASTA file input
if args.fasta_file_name:
   fastaIn = args.fasta_file_name
```

```
else:
    print('Error: Need file containing fasta sequences')

#SciRoKo output file
if args.SciRoKo_output_file_name:
    SciRoKoIn = args.SciRoKo_output_file_name
else:
    print('Error: Need file containing SSR locations')

#SSR minimum repeat type
if args.SSR_minimum_repeat:
    Minrep = args.SSR_minimum_repeat
else:
    SSR_minimum_repeat = 2

#Output file with transcript name, SSR, start, stop
if args.output_file_name:
    FileOut = args.output_file_name
else:
    print('Error: Output file name not given, or already exists')

#FASTA output files, to output directory
if args.output_fasta_file_names:
    fastaOut = args.output_fasta_file_names
    if os.path.exists(fastaOut):
        print('output folder already exists, choose new folder')
        quit()
    else:
        os.mkdir(fastaOut)
else:
    print('Error: Output directory name not given')

ssr_min = Minrep



#Begin script
#PART 1: reading in fasta, translating, and storing info

f = open(fastaIn,'r')
lines = f.readlines()
f.close()

#Dictionary with transcript and coding region information
stored = {}
#transcript:[coding, str(frame), str(start), str(stop)]
```

```
transcript = []
seqs = []
#to store transcript name (start, stop, increments)
for l in range(0,len(lines),2):
    transcript.append(lines[l].strip())
    #defining 'key', transcript-1 gives first element in transcript
    key = transcript[-1].strip('>')
    #to store sequences (transcript+1 = sequences)
    s = lines[l+1].strip()
    seqs.append(s)
    #begin translation code
    translated = []
    for x in range(0,len(lines)):
        coding_dna = Seq(s[x:])
        translated.append(coding_dna.translate())
    coding = ''
    start = 0
    frame = 0
    stop = 0
    for t in range(0,len(lines)):
        c = translated[t].split('*')
        a = 0
        for r in c:
            if len(r) > len(coding):
                #if r is greater than coding, update coding with new r
                coding = r
                #if r is greater than coding, store t(translated) in frame
                frame = t
                #if r is greater than coding, multiply amino acids by 3 and add one for the stop codon
that was split on, need nucleotide position
                start = a*3 + t
                #if r is greater than coding, use start and the length of coding x3, need nucleotide
position
                stop = start + len(coding)*3
                #for dictionary (stored)
            a += len(r)+1
    #for dictionary (stored)
    stored[key] = [coding, int(frame), int(start), int(stop)]



#PART 2: reading in SSR data (transcript, start, stop positions) to tell if it falls in coding region

#from SciRoKo td output file
#this is where script can be altered to read regular BED file
```

```
m = open(SciRoKoIn,'r')
lines2 = m.readlines()
m.close()

info = ''.join(lines2).split('\r\n')

ssr1 = []

ssr_in_coding = []
ssr_out_of_coding = []

for i in info[1:]:
    sl = i.split('\t')
    print('test',sl)
    if len(sl) > 5 and len(sl[1]) >= ssr_min:
        #storing information from SciRoKo td output file (transcript, motif, start, stop)
        ssr1.append([sl[0],sl[1],int(sl[3]),int(sl[4])])

for s in ssr1:
    if s[0] in stored.keys():
        if s[2] > stored[s[0]][2] and s[3] < stored[s[0]][3]:
            ssr_in_coding.append(s)
        else:
            ssr_out_of_coding.append(s)



#PART 3: creating file output

forfasta = {}

#Writing file with info on just SSRs in coding region
c = open(FileOut, "w+")

c.write('Transcript'+'\t'+'SSR'+'\t'+'Start'+'\t'+'Stop'+'\n')
for r in ssr_in_coding:
    c.write('\t'.join(map(str,r))+'\n')



#PART 4: FASTA file outputs (to output directory)

#These fasta files can be used with Clustal Omega to visually see where SSRs occur
#Writing individual fasta files for transcripts (cycle through the forfasta key)

    if r[0] not in forfasta.keys():
```

```
        print(r[0])
        forfasta[r[0]] = ['>'+r[0],seqs[transcript.index('>'+r[0])]]

    forfasta[r[0]].append('>'+r[0])
    forfasta[r[0]].append(seqs[transcript.index('>'+r[0])][r[2]-1:r[3]])
c.close()

for k in forfasta.keys():
    y = open("%s/RTS_perfect_SSR_repeat_inside_coding_%s.txt"%(fastaOut,k), "w+")
    y.write('\n'.join(forfasta[k]))
    y.close
```

Vita

Carmen Paige Webster attained her Bachelor of Arts in Biological Sciences at the University of Texas at El Paso in December of 2018. While completing her undergrad, she conducted research for Dr. Kyung-An Han where she investigated genes that were likely to be important for learning and memory with age in *Drosophila melanogaster*. Carmen began her Master of Science in Biological Sciences degree in June 2019, where her thesis research was conducted under the supervision of Dr. Michael Moody in his Plant Evolution lab, and her project was funded by the Fall 2019 cycle of the Dodson Research Grant. While a graduate student, Carmen worked as a Teaching Assistant under Dr. Horacio Gonzalez and Dr. Marie Smith, teaching general and organismal Biology labs. Upon graduation, Carmen will further her education with a second Master's degree in Biological Data Science at Arizona State University in Phoenix, AZ under the supervision of Dr. María José Sanín, where she will further develop her skillset with computer programing and statistical analyses concerning biological data.

Contact Information: cwebster@miners.utep.edu