

# UVM ScholarWorks

## Inferring the demographic history of red spruce (*Picea rubens*) from chloroplast genome sequences

Item Type	undergraduate thesis
Authors	Bardsley, Katherine Anne
Download date	2026-05-19 09:20:07
Item License	<a href="http://creativecommons.org/licenses/by-nc-nd/3.0/">http://creativecommons.org/licenses/by-nc-nd/3.0/</a>
Link to Item	<a href="https://hdl.handle.net/20.500.14849/5306">https://hdl.handle.net/20.500.14849/5306</a>

Inferring the demographic history of red spruce (*Picea rubens*)  
from chloroplast genome sequences

Katherine Bardsley

University of Vermont, Plant Biology Department

Honors Thesis

Spring 2021

Advisor: Dr. Stephen Keller

**Abstract:**

Understanding the demographic history of a species is pivotal for assessing current population structure and predicting future viability. Here, I use the chloroplast genome to gain a better understanding of red spruce demographic dynamics since the last glacial maximum. This information can be applied to assessments of the future viability of the species, especially in the context of ongoing anthropogenic climate change. Data from whole-exome sequence capture of 340 individuals of red spruce sampled from across the current range were utilized in the phylogeographic analyses that form the basis of this work. The chloroplast genome coverage was greatly improved upon via imputation and the sequence data used to predict changes in effective population size, range-wide genetic structure, and hybridization. In addition to optimizing a chloroplast isolation protocol that will enable greater imputation accuracy in future work, I found a surprising trend in increasing chloroplast effective population size over the last 100,000 years. Furthermore, I showed the absence of an isolation by distance effect in the chloroplast genome, a strong deviation from conclusions drawn using the nuclear genome, supporting the idea that long distance pollen dispersal could potentially connect otherwise isolated populations across the highly fragmented range. Tied through it all, there was a clear signature of hybridization between red and black spruce in the northern part of the current species range. Through these results, we confirm that the chloroplast genome provides valuable insight into the demographic history of red spruce and can be used to inform assessments of future population viability.

## **Introduction:**

Climate change is taking a profound toll on the world's ecosystems, causing significant changes in the distribution and abundances of species (Thomas et al., 2004). This is especially true of endemic or regionally restricted species like red spruce (*Picea rubens*), which has experienced a significant decline in population size and has one of the most fragmented distributions of any eastern North American tree (Capblancq et al., 2020; Hamburg & Cogbill, 1988; Mosseler et al., 2000). Demographic events throughout its history, such as founder effects and range contractions, have likely shaped contemporary populations of red spruce by influencing population size, diversity, and geographic distribution. These events may help explain its current distributions of genetic diversity and observed low levels of heterozygosity and polymorphism (Capblancq et al., 2020; De Hayes & Hawley, 1992; Dumais & Prévost, 2007), important indicators of future population viability (Breed et al., 2011; Capblancq, Munson, Butnor, & Keller, 2021; Ellegren & Galtier, 2016; Elmqvist et al., 2003; Thomas et al., 2014).

One hypothesis is that reductions in red spruce population size and genetic diversity were triggered by the warming climate at the end of the last ice age (ca. 21,000 years ago), which dramatically re-structured the distributions of many species (Davis & Shaw, 2001; Hewitt, 1996). At the height of the last glacial maximum (LGM), an ice sheet covered much of northeastern North America. Surviving plants were located in refugia, areas where suitable conditions could be found beyond the ice. These refugia provided a potential opportunity for hybridization between species that were forced to share the same limited geographic space. Once temperatures warmed and the ice retreated, species were able to expand their ranges into areas that were previously inhospitable (Hewitt, 1996). As populations ventured into new territories,

there was a co-occurrence of species whose ranges had not previously overlapped, providing another potential opportunity for hybridization (Davis & Shaw, 2001).

While the warming climate enabled range expansion and provided opportunities for hybridization, it also made it more challenging for individuals to survive in the southern, warmer parts of the range (Hewitt, 1996). As populations spread out from refugia, bottlenecks led to losses of alleles and reductions in genetic diversity in the northern leading edge. Due to the climatic pressure on red spruce populations in the trailing edge and the occurrence of bottlenecks in the leading edge, we predict that the expansion from glacial refugia brought about by climatic warming triggered a decline in both the population size and genetic diversity of the species. By gaining insight into this reduction in past red spruce diversity, it helps us to understand the impacts of a changing climate on the species and may inform assessments of future responses and long-term population stability

The advent of DNA sequencing technology enabled the discovery that genetic variation present within a species can reveal details about its history of population size and movement, especially in the context of how species have responded to past ice ages (Cavalli-Sforza, 1997; Davis & Shaw, 2001; Hewitt, 1996). This field is known as “phylogeography,” an interdisciplinary fusion of molecular phylogenetics and biogeography (Avice et al., 2016). Phylogeographic studies use DNA sequences from many individuals to determine the genealogy of the genetic variants (alleles) as well as their geographic distribution. The resulting patterns of relatedness can be analyzed by applying statistical methods to infer the demographic history of an individual, showing how gene variants from a population diversified from their common ancestor (Fu & Li, 1999; Rosenberg & Nordborg, 2002; Wakeley, 2009).

The application of phylogeography to model historic changes in population size and migration requires generating reliable allele genealogies from highly variable regions of DNA, making the chloroplast genome of plants a great target for such research. While chloroplast genomes contain many highly conserved genes that are vital for growth and development across land plants, there are also rapidly evolving variable regions present that accumulate sequence polymorphism at the individual level, usually contained in non-coding sequences between genes. This assortment of genetic material makes the chloroplast genome extremely valuable for comparative evolutionary studies at a variety of spatial and temporal scales (Clegg, Gaut, Learn, & Morton, 1994; Vieira et al., 2014).

The chloroplast genome is also a good target for phylogeographic analysis because it is small in size, haploid, and generally non-recombinant -- factors that set it apart from the nuclear genome and make its analysis and interpretation more straight-forward (Du et al., 2015). The chloroplast genome of conifers is particularly unique, as unlike in angiosperms, it is paternally inherited and dispersed through pollen flow (Sutton et al., 1991). Its mode of inheritance makes the chloroplast genome an informative marker to investigate the effects of long-distance gene flow through pollen across the species range (Du, Petit, & Liu, 2009). This is especially relevant in the case of red spruce, where the species distribution is extremely fragmented and there is the potential for isolated patches to be cut off from seed-mediated gene flow, yet still be connected by wind-dispersed pollen.

The chloroplast genome is an immensely informative resource. However, targeting the chloroplast of a conifer such as red spruce is not without its challenges. While older studies relied on sequencing a small sample of chloroplast marker loci, the advent of next generation sequencing makes it feasible to sequence whole chloroplast genomes, which results in much

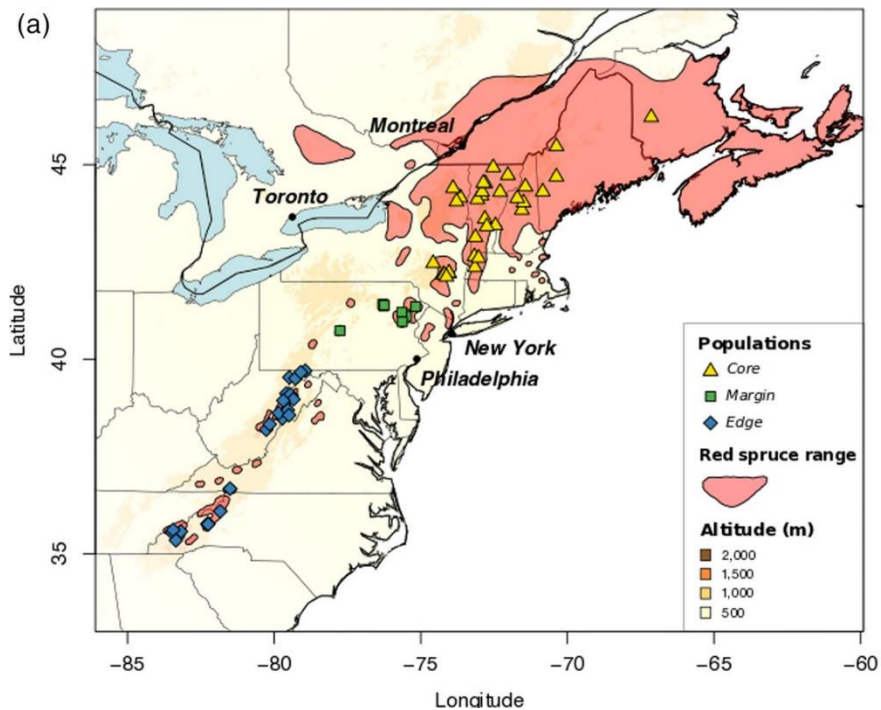
higher resolution genetic data for coalescent-based analysis. In most plant species, it is possible to sequence the nuclear and chloroplast genomes simultaneously; however, in the conifers (including spruce), the very large size of the nuclear genome necessitates biochemical isolation of whole intact chloroplasts first, prior to sequencing. Isolation techniques that employ a combination of high salt buffers and Percoll density gradients have been successfully used for chloroplast isolation in several species, often necessitating a large volume of sample tissue (Du et al., 2015; Sakaguchi et al., 2017; Vieira et al., 2014).

To address this obstacle to using the red spruce chloroplast genome in phylogeographic studies, the first objective of my thesis work is to develop laboratory methods that allow for the efficient enrichment of chloroplast DNA away from nuclear DNA, which is necessary prior to generating sequence libraries for Next Generation Sequencing. The next three objectives aim to inform ongoing conservation efforts using chloroplast genome sequence data derived from the whole-exome sequencing effort described in Capblancq et al. 2020. These goals are as follows: (a) to assess evidence for changes in red spruce population size since the last glacial maximum by applying Bayesian Coalescent analyses; (b) to investigate potential introgression between red spruce and black spruce (*Picea mariana*) in different regions using haplotype networks and principal component analysis; and (c) to explore the effects of long-distance pollen flow on range-wide genetic structure and the potential implications for spatially isolated populations by analyzing measures of genetic variation. From the unique perspective of the haploid, non-recombinant, and paternally-inherited chloroplast genome, I hope to address a major question in how climate change impacts population size and genetic diversity of species, painting a better picture of red spruce population dynamics.

## Materials and Methods:

### *Sampling Design*

The sampling for this study comes from DNA sequencing of 340 red spruce individuals that were sampled from 65 distinct populations. These populations span the current range in eastern North America, from New England and southeastern Canada south to North Carolina and Tennessee (Figure 1). This broad distribution of red spruce has been shown to consist of three distinct genetic groups with divergent ancestry and exhibiting clear differences in genetic diversity and population structure (Capblancq et al., 2020). These three groups will be referred to as the Core (the northernmost part of the range), Margin (fragmented section in the Pennsylvania region), and Edge (highly fragmented trailing edge region in the southern Appalachians) (Figure 1) (Capblancq et al., 2020).



**Figure 1** Geographic distribution of red spruce in eastern North America, including the sampled populations in the Core (yellow), Margin (green), and Edge (blue) regions. Figure adapted from Capblancq et al., 2020.

### *Chloroplast Isolation*

A small subset of seedlings grown from the original sampled mother trees were the focus of targeted chloroplast DNA extraction. Chloroplast DNA was isolated from 8 total seedlings and included individuals to represent each of the three geographically and genetically distinct groups. In order to extract intact chloroplasts from each of the 8 saplings, a high salt plus saline Percoll gradient method was used (Du et al., 2015; Sakaguchi et al., 2017; Vieira et al., 2014). Prior to the extraction, the needle tissue was washed and incubated to reduce microbial contamination and starch and resin content (Vieira et al., 2014). Needles were homogenized in a high salt isolation buffer and the homogenate filtered using Miracloth. Multiple centrifugation steps were performed to first pellet the nucleus and cell wall debris, then spin the remaining supernatant to pellet the chloroplast. A Percoll density gradient was implemented to isolate the chloroplasts from remaining starch and debris (Vieira et al., 2014).

Successful chloroplast isolation prior to DNA extraction was essential in order to limit the amount of nuclear material present. This is an especially crucial step for red spruce as it has a very large nuclear genome size (ca. 22 Gbp), with a chloroplast to nuclear genome ratio of  $5.6 \times 10^{-6}$  (Lin et al., 2019). In order to obtain enough coverage of the chloroplast genome during sequencing, it is vital that the proportion of chloroplast DNA is enriched so fewer reads are used (e.g., wasted) on nuclear DNA. Once intact chloroplasts were isolated, I extracted the DNA from the enriched chloroplast samples using a single-tube DNA extraction protocol optimized in the Keller lab.

### *qPCR Test of Enrichment*

A qPCR experiment was developed to test the enrichment (i.e. ratio of chloroplast to nuclear genetic material present compared to normal levels). Genomic DNA was extracted from needle tissue from the same individuals targeted for chloroplast isolation using the same single-tube DNA extraction protocol. These samples acted as controls, providing a way to measure the amount of chloroplast DNA present relative to nuclear DNA under normal extraction conditions and without extra chloroplast isolation steps. The qPCR reaction used two sets of primers, one that targeted the nuclear genome (“nucPrimers”) and one that targeted the chloroplast genome (“cpPrimers”). The nuclear primers amplified a 172 bp segment of YLS8, a robust, single-copy, nuclear gene (Rutledge et al., 2013). These primers (YLS8\_F4 and YLS8\_R1) were designed and optimized by Rutledge et al. 2013. The chloroplast primers amplified a 220 bp segment of the trnT-trnL intergenic spacer. These primers (aDNA\_15 and aDNA\_16) were designed and optimized by Ethan Thibault for work with red spruce ancient DNA samples and demonstrated robust amplification.

qPCR reactions were performed for each plant using Bio-Rad iTaq Universal SYBR Green Supermix. Chloroplast-enriched samples and their corresponding genomic DNA controls from each plant were tested in parallel with each set of primers, one set targeting the nuclear DNA and the other the chloroplast. During each reaction, the  $C_T$  values were recorded, indicating the cycle number at which enough amplified product had accumulated to produce a detectable fluorescent signal. These values provide a relative measure of the amount of starting template material and were the primary input for enrichment calculations. The  $2^{-\Delta\Delta C_t}$  method, alternatively known as the Livak method, was used for relative quantification (Bio-Rad Laboratories, 2006). To perform these calculations, the  $\Delta C_T$  values were first calculated for each sample by

subtracting the  $C_T$  for the nuclear locus from that of the  $C_T$  from the chloroplast locus (Eq. 1, 2). Next,  $\Delta C_T$  for the control genomic sample (“gDNA”) was subtracted from the  $\Delta C_T$  for the chloroplast-enriched sample (“cpDNA”) for the same plant (Eq. 3), and the difference exponentiated to determine the fold increase in chloroplast enrichment (Eq. 4).

$$\Delta C_{T(\text{cpDNA})} = C_{T(\text{cpPrimers, cpDNA})} - C_{T(\text{nucPrimers, cpDNA})} \quad (1)$$

$$\Delta C_{T(\text{gDNA})} = C_{T(\text{cpPrimers, gDNA})} - C_{T(\text{nucPrimers, gDNA})} \quad (2)$$

$$\Delta \Delta C_T = \Delta C_{T(\text{cpDNA})} - \Delta C_{T(\text{gDNA})} \quad (3)$$

$$2^{-\Delta \Delta C_T} = \text{Normalized expression ratio} \quad (4)$$

The fold increase in chloroplast enrichment was then used to determine which samples were adequately enriched for sequencing. The necessary threshold frequency was determined by performing the following calculations. The number of chloroplast copies per cell before enrichment was estimated from Eq. 2 and then exponentiated (in the same way  $\Delta \Delta C_T$  is exponentiated in Eq. 4), and averaged the values across all the samples. This result, along with the nuclear and chloroplast genome sizes in base pairs ( $2.2 \times 10^{10}$  bp and 124,000 bp respectively), and the knowledge that there are only two copies of the nuclear genome per cell, made it possible to predict the proportion of DNA base pairs per cell contributed by the chloroplast genome. With this information, I predicted the expected coverage per sample that would result from a shot-gun library sequencing run (Eq. 5, 6) and used this to decide how many samples could be sent for sequencing in a single multiplexed run and determined a threshold enrichment level required for sequencing to an adequate depth of coverage per site.

$$\begin{aligned}
& \text{Expected bp of chloroplast reads per Miseq run} && (5) \\
& = (\text{proportion of bp per cell contributed by chloroplast genome}) \\
& \times (\text{bp Miseq output})
\end{aligned}$$

$$\begin{aligned}
& \text{Expected coverage (read depth) for each site for a given sample} && (6) \\
& = \frac{\text{expected bp of chloroplast reads/Miseq run}}{\text{chloroplast genome length}} \times \frac{1}{\# \text{ samples in single run}}
\end{aligned}$$

### *Whole-exome Sequence Capture*

Whole-exome sequence capture was performed on the full set of 340 red spruce individuals, as described in Capblancq et al. 2020. Briefly, 80,000 120 bp probes were designed based on an extensive catalog of expressed genes across multiple tissue types, developmental stages, and environments in white spruce, a closely related spruce species. These probes were hybridized to genomic DNA of red spruce, and the resulting captured fragments barcoded and sequenced using an Illumina HiSeq X. After trimming and quality filtering, this resulted in an average of 2.56 million 2 x 150 bp reads per individual, which were mapped to the *Picea* reference genome using BWA (Burrows-Wheeler Aligner) (Capblancq et al., 2020).

### *Imputation*

The reads generated through previous whole-exome sequence capture efforts (Capblancq et al., 2020) were subset to only include segments that could be aligned to the published Norway Spruce (*Picea abies*) chloroplast reference (Nystedt et al., 2013), resulting in partial coverage of this genome. It was then possible to use imputation to produce more complete genome coverage using a panel of haplotype references, generated by calling variants from four *Picea* reference genomes from NCBI, two from red spruce (*Picea rubens*) (accession numbers LT727875 and

LT727876) and two from black spruce (*Picea mariana*) (accession numbers LT727842 and LT727861) (Lo et al., 2020).

Imputation is the statistical inference of missing bases based upon known haplotypes in a population (Yun, Willer, Sanna, & Abecasis, 2009). Since the chloroplast genome is haploid and non-recombining, haplotypes extend across its entire length. This makes imputation more effective and straightforward using the chloroplast genome compared to the recombining nuclear genome. Imputation was performed using IMPUTE2 (Howie, Donnelly, & Marchini, 2009; Kraja et al., 2019; Whalen, Gorjanc, Ros-Freixedes, & Hickey, 2018). Imputation accuracy increases with number of reference haplotypes available and is limited by the amount of missing data within reference genomes. To address this issue and generate a larger pool of chloroplast haplotype references for red spruce, one must isolate intact chloroplasts away from the nucleus before sequencing can occur and these references can be incorporated into imputation.

#### *Assessing Hybridization with Black Spruce (Picea mariana)*

Haplotype networks were created using the R package pegas (Population and Evolutionary Genetics Analysis System) (Paradis, 2010, 2018). These networks included the 340 red spruce samples as well as four NCBI chloroplast genome references, two from red spruce (*Picea rubens*) (accession numbers LT727875 and LT727876) and two from black spruce (*Picea mariana*) (accession numbers LT727842 and LT727861). Networks were created with various levels of filtering of missing data using the maxmiss parameter in VCFtools. The three investigated networks allowed for up to 0%, 10%, and 20% of the data at a given site to be missing, respectively. For clarity, this parameter will be referred to as the maximum missingness level hereafter.

Within these haplotype networks it is expected that the haplotypes associated with the red and black spruce references would be genetically divergent, falling into two separate and distantly related clusters by species. In the case of no hybridization, the haplotypes associated with all 340 samples would cluster around the red spruce references and be more distantly related to the black spruce reference. However, in the case of hybridization, you might expect to see a subsample of red spruce haplotypes that cluster more closely with the black spruce references, sharing more similarities with these references than those from red spruce. By identifying the sample regions of origin within the haplotype networks, observing which sample haplotypes cluster near each subset of references can be an informative way to assess the nature of potential hybridization.

To assess the genetic structure from another perspective, a principal component analysis (PCA) was performed using the R package SNPRelate (Zheng et al., 2012). The vcf file used did not undergo any prior filtering of missing sites, but SNPRelate performed its own automatic imputation process by replacing missing data with the average genotype value at that locus before analysis. The same four NCBI references were included in both the haplotype networks and the PCA, with two from red spruce and two from black spruce. A total of 467 SNPs (single nucleotide polymorphisms) were included in the analysis.

### *Estimating Changes in Effective Population Size ( $N_e$ )*

The BEAST2 software package and its complement BEAUti were used to perform Bayesian evolutionary analyses that provide insight into the past population dynamics of red spruce (Bouckaert et al., 2019; Suchard et al., 2018). Bayesian Evolutionary Analysis by Sampling Trees (BEAST) uses molecular data and samples from the posterior distribution of parameters and trees using the Markov chain Monte Carlo (MCMC) algorithm. Specifically, I

used the Coalescent Bayesian Skyline plot method in the BDSKY package to predict changes in effective population size ( $N_e$ ) through time (A. J. Drummond, Rambaut, Shapiro, & Pybus, 2005).

To determine which of sequence evolution best fit our alignment of chloroplast sequences, we used jModelTest (Darriba, Taboada, Doallo, & Posada, 2012; Guindon & Gascuel, 2003). The Bayesian Information Criteria (BIC) generated suggested that the model with the highest likelihood of fitting the data that was also supported by the BEAST package was in the GTR class. We used estimates from jModelTest for the shape parameter of the gamma distribution of rate variation among sites, the proportion of invariant sites, and the nucleotide substitution rates for each possible substitution type to inform the GTR + invariant sites + gamma model in BEAST2. The clock rate was determined based on the work of Lockwood et al., who used fossil dating to calibrate the molecular clock for two chloroplast genes in *Picea* (*rbcL* and *matK*) in units of substitutions per site per million years (Lockwood et al., 2013). The average nucleotide substitution rate of the two genes was  $1.21 \times 10^{-10}$  substitutions per site per year, which was used as input for BEAST2.

The BDSKY model was run with a MCMC chain length of 90,000,000 iterations, logged every 5,000, to provide 18,000 sampling iterations to evaluate the parameters, split between two separate runs (with chain lengths of 60,000,000 and 30,000,000 respectively). The runs were combined using LogCombiner (Alexei J. Drummond & Rambaut, 2007). Tracer v1.7.1 was used to summarize and visually inspect the posterior parameter estimates produced by BEAST2 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018).

To investigate the impact of potential hybridization of black spruce (*Picea mariana*) with red spruce on our demographic inferences of population size changes in the coalescent model, a

second BEAST model was generated, this time by excluding samples from individuals whose haplotypes clustered more closely with black spruce references than red spruce references based on the haplotype network. The same filtering was used on the input files for the haplotype network and the BEAST run to determine the subset of individuals to include. This filtering excluded sites that had missing data for more than 20% of individuals using VCFtools. Of the original 340 individuals, 323 were included in the population subset used for the second BEAST model based on their proximity to red spruce references in a haplotype network similar to the one shown in Figure 5. The same GTR + I + G model and parameters were used for the original model and the subset. The subset was run with a total chain length of 155,000,000 samples, logged every 5,000, to provide 31,000 sampling iterations, split between four separate runs (with chain lengths of 30 million, 30 million, 45 million, and 50 million respectively).

### *Population Genetic Structure*

Population structure was assessed using the *vcfR* package in R (Knaus & Grünwald, 2017). The samples were divided into populations based on gridded landscape pixels of 50 km x 50 km following Capblancq et al. (in review), which allowed the groupings of nearby individuals together into populations based on their geographic proximity in order to ensure each population included in the analyses had at least 10 individuals. Overall  $G_{ST}$ , pairwise  $G_{ST}$ , and population heterozygosity were all estimated using *vcfR*. The *vcfR* `genetic_diff()` function calculates Nei's  $G_{ST}$  which for biallelic data like that used here, is equivalent to  $F_{ST}$ . This makes it possible to compare  $G_{ST}$  measures with  $F_{ST}$  values estimated by Capblancq et al. (in review) for the nuclear genome.

To investigate the effect of geographic distance on genetic differentiation, the distance between each pixel was calculated. The center of each pixel was determined by finding the

average latitude and longitude for all the samples within a pixel. The geographic distances between pixel centroids were found using the raster package in R. These analyses were all run on both the full data set as well as the subset that clustered more closely to red spruce in the haplotype network (maximum missingness = 20%). A Mantel test was performed using the R package ape to statistically test the relationship between genetic differentiation and geographic distance (Paradis & Schliep, 2019).

Populations that experience significant changes in population size, either growing or shrinking, also harbor signals in the distribution of the frequency of mutations. When a population is growing rapidly, there are many new mutations entering that will be initially rare. This leads to an over-abundance of rare polymorphisms relative to a population at demographic equilibrium. Conversely, when the population size is rapidly shrinking or has gone through a demographic bottleneck, many of the rare alleles tend to be lost and the number of rare polymorphisms is lower than it would be at demographic equilibrium. These trends can be captured in both site-frequency spectra of mutation frequency and Tajima's D statistic.

Site-frequency spectra capture the frequency of rare alleles within a population. A skew towards rare alleles would indicate a growing population. If fewer rare alleles are present than expected under equilibrium, the data suggest a shrinking population or demographic bottleneck. Serving as a complement to the site-frequency spectra, Tajima's D compares the average number of pairwise differences with the number of segregating sites. In a demographically neutral population, these values would be equivalent. However, when the population is growing, there are more segregating sites and a lot of rare alleles. The presence of many new rare mutations is a signature of recent growth and reflected in a negative Tajima's D.

To provide an additional perspective on the question of demographic change in the history of red spruce, the site-frequency spectrum and Tajima's D statistic were calculated. Minor allele frequency was calculated using vcfR for all of the SNP (single nucleotide polymorphism) sites. Tajima's D was estimated for the full population as well as for Core, Margin, and Edge subsets using the R package pegas (Paradis, 2010).

## **Results:**

### *Chloroplast Enrichment*

The mean  $\Delta\Delta C_T$  value from qPCR tests of the eight most promising chloroplast isolations was -2.59, leading to a mean fold-enrichment of chloroplast DNA over nuclear DNA by 6.54X (Table 1). This suggests that on average, the mass of chloroplast DNA relative to nuclear DNA present per cell increased by a factor of 6.5 when the chloroplast isolation protocol was performed. Based on standardized  $\Delta C_T$  values for the control genomic DNA samples, we predicted that the chloroplast genome makes up about 0.005 of the base pairs in a given red spruce needle cell. With the 500,000,000 bp of reads expected as output from a single Illumina MiSeq run, this suggests coverage of about 15X per site per individual, which is more than adequate for our good estimations of the chloroplast genome sequence at each site and the calling of polymorphism.

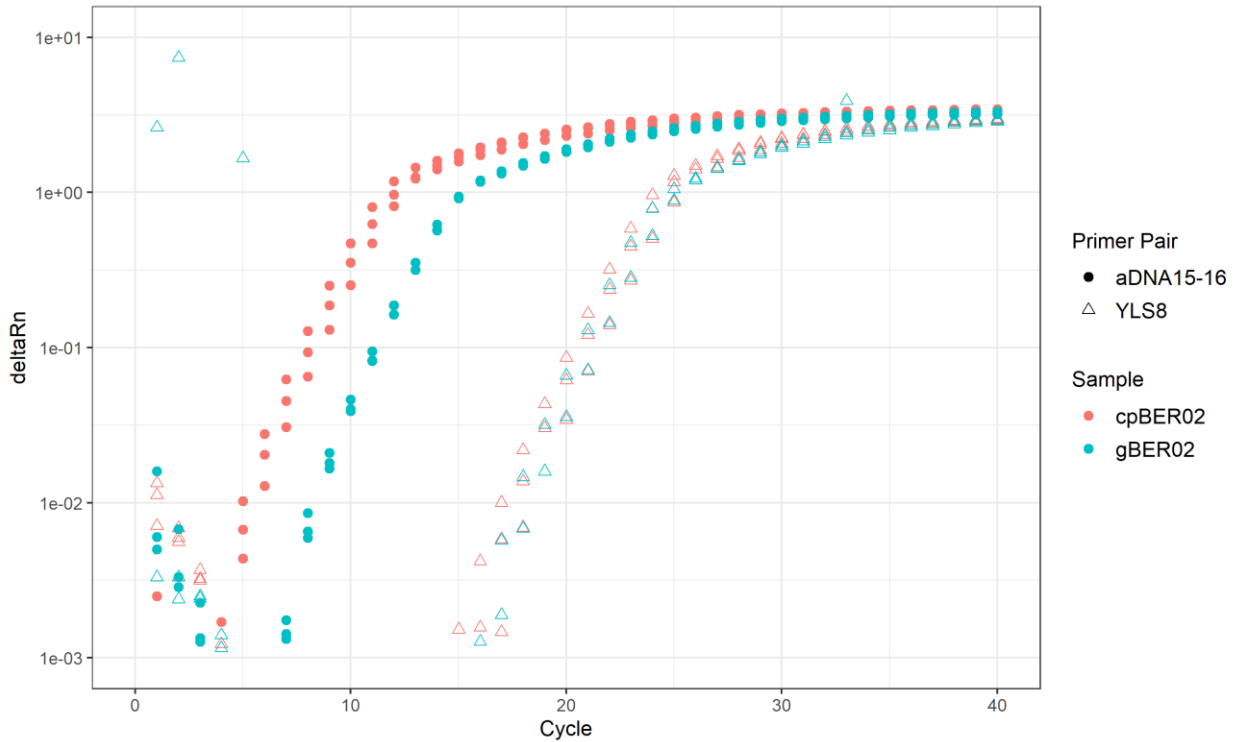
**Table 1.**  $2^{-\Delta\Delta C_T}$  calculations for each individual.

qPCR cohort refers to which samples were processed in parallel. cpDNA refers to chloroplast-enriched samples generated through chloroplast isolation protocol and gDNA refers to samples produced using a standard single-tube genomic DNA extraction as a control.

Sample Name	Region	qPCR Cohort	$\Delta C_T$ for cpDNA	$\Delta C_T$ for gDNA	$\Delta\Delta C_T$	Fold Increase
MMF_02	Core	A	-13.017	-10.171	-2.846	7.192
MMF_32	Core	A	-13.841	-11.630	-2.211	4.630
ALB_05	Core	A	-13.393	-9.597	-3.796	13.891
WHI_05	Core	A	-13.612	-10.817	-2.795	6.941
BER_02	Core	A	-12.936	-10.211	-2.726	6.614
MMF_37	Core	B	-12.085	-10.016	-2.068	4.194
CRR_02	Margin	C	-7.933	-5.684	-2.249	4.755
KOS_03	Edge	C	-7.573	-5.534	-2.038	4.108

The enrichment is also illustrated by the amplification plots generated during the qPCR experiments (Figure 2). These plots show the fluorescent signal (y-axis), which is proportional to the amount of amplified product, as a function of the cycle number (x-axis). In the beginning of the reaction, even though amplification is occurring, the fluorescent signal remains at background levels. The cycle at which the product has accumulated enough to yield a detectable fluorescent signal is known as the threshold cycle ( $C_T$ ). Figure 2 shows the amplification plot for the samples associated the BER\_02 individual. The curves for the nuclear primer pair (hollow triangles) are nearly superimposed for both the genomic DNA and chloroplast-enriched DNA samples (pink and blue curves respectively). Both curves have mean  $C_T$  values around 19. However, the amplification curves for the chloroplast primer pair (solid circles) differs substantially between these two samples (pink and blue curves), with the chloroplast-enriched DNA sample achieving greater amplification prior to the genomic DNA sample. The mean  $C_T$  values for the chloroplast-enriched sample and the genomic DNA sample were 5.9 and 8.9

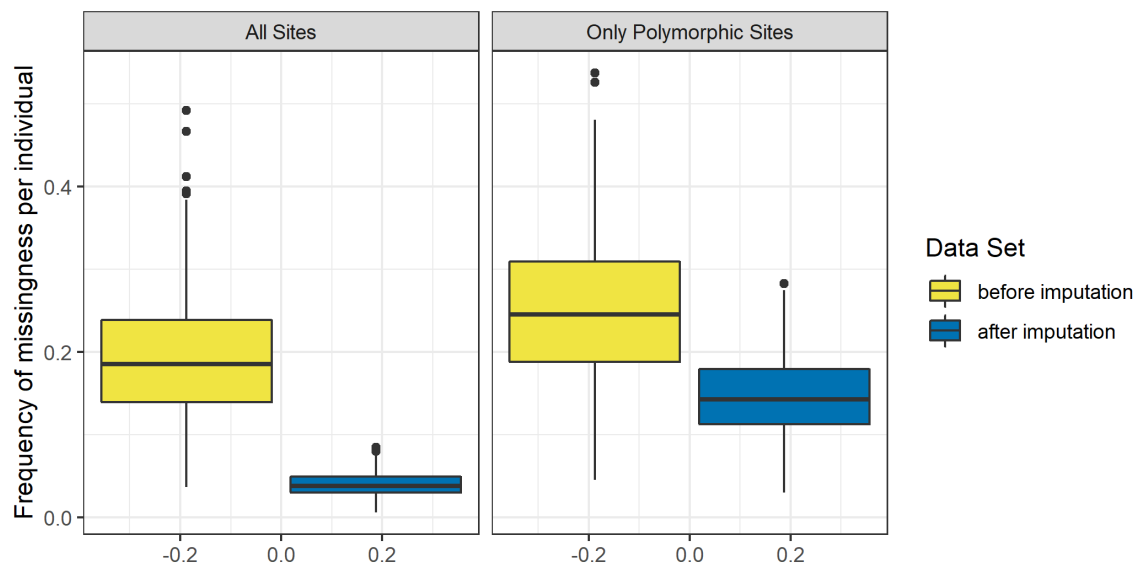
respectively. Since the rate of amplification is directly related to the amount of template available, faster amplification and lower  $C_T$  values are indicative of an increase in the proportion of chloroplast material present.



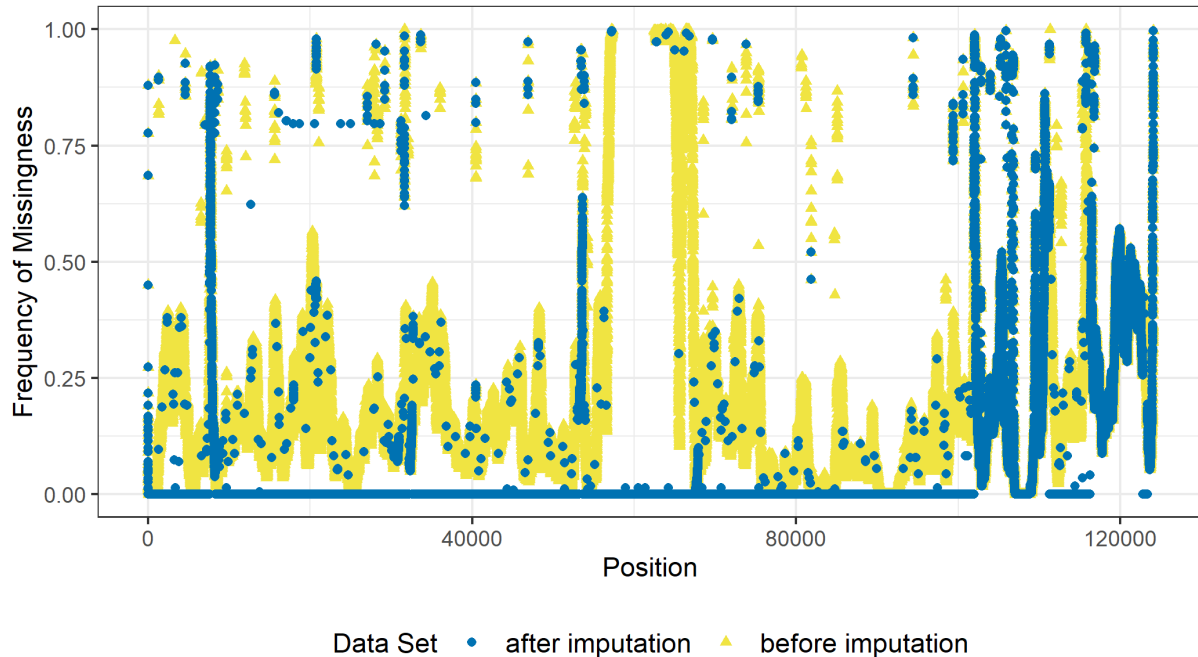
**Figure 2.** Amplification plot for sample BER\_02 that illustrates the change in  $C_T$  values for the different samples and primer pairs. Amplification of the chloroplast-enriched sample (cpBER02) is shown in pink and the genomic DNA control (gBER02) is shown in blue. Amplification using the chloroplast primer pair, aDNA15-16, is indicated by the solid circles and amplification using the nuclear primer pair, YLS8, is indicated by the hollow triangles. There are three replicates for each primer pair-sample combination.

## Imputation

The per-site missingness was assessed with VCFtools both pre- and post-imputation to determine the imputation success (Danecek et al., 2011). Before imputation, the exome-capture sequence data had an average missingness of 19% per site across the chloroplast genome. After running IMPUTE2, the average per-site missingness decreased to 4% (Figures 3 and 4). The median missingness dropped from 14% to 0%. Since the polymorphic sites are the ones more meaningful for analyses, I also subset the data to look at only the sites labeled as polymorphic by VCFtools. For these polymorphic sites, the average per-site missingness pre-imputation was 25%. Post-imputation, this number decreased to 14% (Figure 3). The median missingness dropped from 19% to 0% for these polymorphic sites.



**Figure 3.** Frequency of missingness per individual before imputation (yellow) and after imputation (blue) for the full data set with all sites in the panel on the left and only the polymorphic sites in the panel on the right.



**Figure 4.** Graph of the frequency of missing data by position across the chloroplast genome averaged across the 340 individuals represented in the exome-capture data (Capblancq et al., 2020a) before imputation (yellow) and after imputation (blue).

Another key part of assessing the imputation success was determining the accuracy of the called genotypes. IMPUTE2 performs an internal cross-validation during each run, masking the genotypes of one variant at a time and imputing the masked genotypes based on the remaining samples. The imputed genotypes are compared with the original genotypes to calculate concordance statistics. Here IMPUTE2 repeated this process with 74,112 genotypes that were called with high confidence. The results, which are broken down by genotype probability interval, are shown in Table 2 on a cumulative scale. All of the genotypes that were called had a probability greater than 0.5 and the overall concordance was 95.9%. As the genotype probability interval is more selective, the percent of total called genotypes decreases slightly and the reported percent concordance increases slightly. Overall, there was high accuracy across the called genotypes.

**Table 2.** IMPUTE2 Imputation Concordance Table.

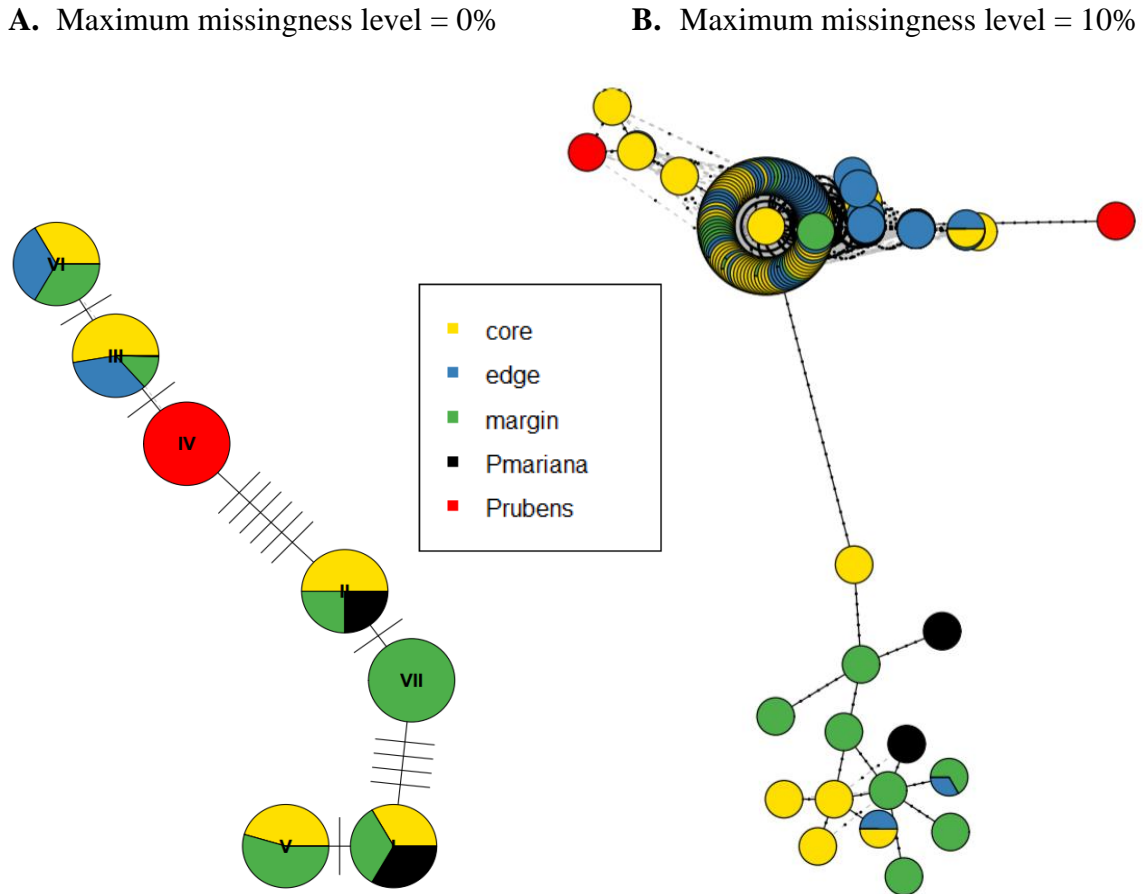
Genotype Probability Interval	Percent of Total Genotypes Called	Percent Concordance
$\geq 0.5$	100.0 %	95.9 %
$\geq 0.6$	99.6 %	96.2 %
$\geq 0.7$	99.3 %	96.2 %
$\geq 0.8$	98.4 %	96.5 %
$\geq 0.9$	96.6 %	97.6 %

*Hybridization with Black Spruce (Picea mariana)*

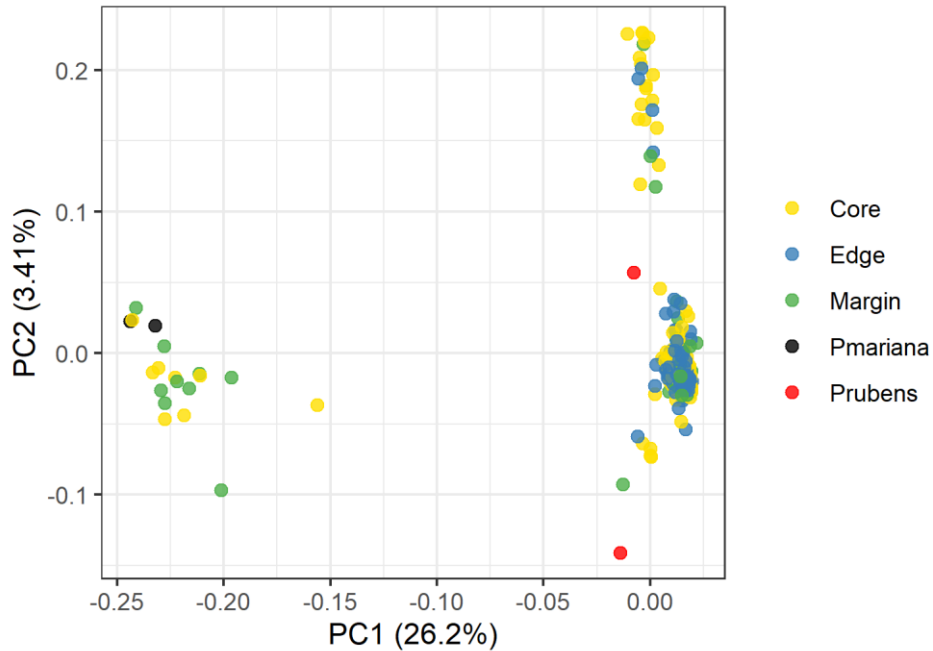
Haplotype networks were generated using pegas at multiple levels of maxmiss (i.e., maximum missingness level), a VCFtools parameter that dictates the filtering of missing data. At a maximum missingness level of 0% (allowing only for sites where zero individuals were missing data in that location), there were 7 haplotypes generated based on 15 polymorphic sites (Figure 5a). The most common haplotype was III, with 321 individuals classified under this haplotype. This haplotype is the closest in proximity to the two red spruce references (Figure 5a). The other six haplotypes contained samples from 11 or fewer individuals. At a maximum missingness level of 10% , there were 282 haplotypes based on 139 polymorphic sites (Figure 5b). These haplotypes were not labeled due to the increased complexity of the network, but the trend of the majority of individuals clustering near the red spruce references (top of Figure 5b) holds true for the less stringent level of maximum missingness filtering.

Similar trends can be seen in results from the principal component analysis (PCA). The analysis was completed using all 344 samples, including 467 SNPs. The first principal component accounted for 26.16% of the variation, the second principal component accounted for 3.41% of the variation, and each following principal component accounted for less than 2% of the variation. Plotting the first two principal components shows a clear division into two clusters

along PC1, with the red spruce and black spruce references falling into separate clusters (Figure 6). In both the haplotype networks and the PCA, we do not see region-specific clusters forming.



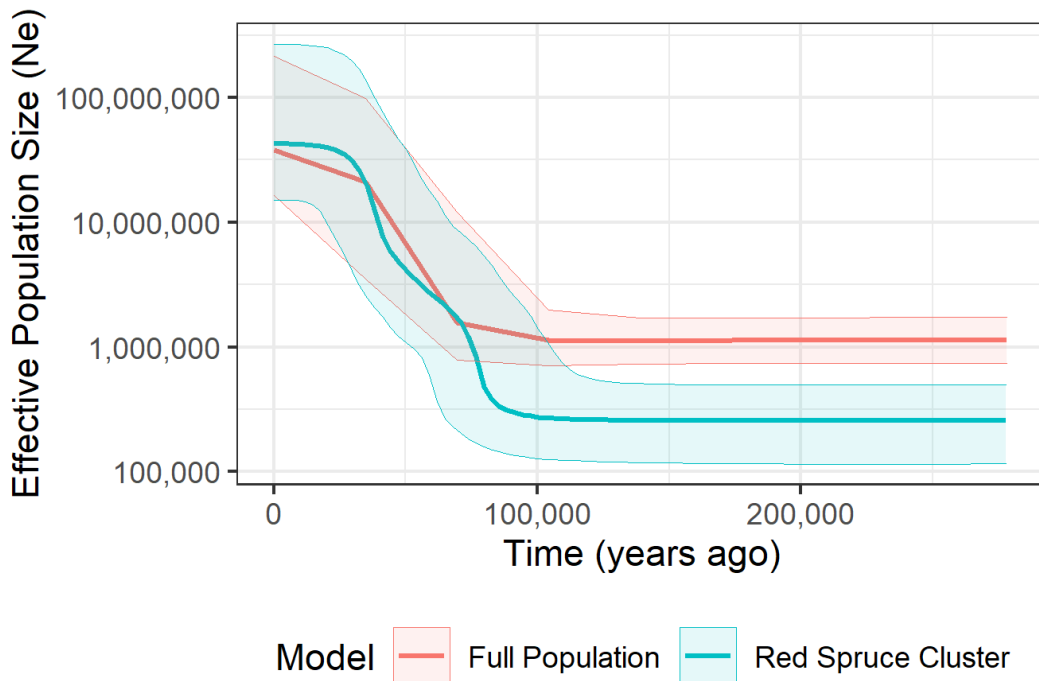
**Figure 5.** Haplotype networks generated using the R package *pegas* with 340 samples of red spruce plus 4 NCBI references, two from red spruce (*Picea rubens*) (shown in red) and two from black spruce (*Picea mariana*) (shown in black). The 340 samples are color-coded based upon their region of origin (Figure 1), with the Core in yellow, the Margin in in green, and the Edge in blue. The tick marks (a) or points along the connectors (b) indicate number of single nucleotide changes between the two haplotypes in question. Two different levels of maximum missingness filtering were used: 0% (a) and 10% (b).



**Figure 6.** Genetic structure across the red spruce chloroplast genome shown by the first two principal components (accounting for 26.2% and 3.4% of the variation respectively) of a genetic principal component analysis (PCA) of single nucleotide polymorphism (SNPs).

### *Estimating Changes in Effective Population Size*

The effective sample sizes (ESS) for all of the statistics calculated by BEAST2 for the full data set of 340 individuals were above the threshold of 100, with a minimum value of 122 (corresponding with likelihood and treeLikelihood) and a maximum value of 8555 (corresponding with TreeHeight). The median ESS value was 243. The Bayesian Skyline Reconstruction Analysis predicted a current effective population size ( $N_e$ ) of about 38 million (Figure 7). The model illustrates an increase in chloroplast effective population size over the last 100,000 years with a consistent  $N_e$  around 1.1 million before that point.



**Figure 7.** Bayesian Skyline Reconstruction analysis based on models from BEAST2 for the full data set of 340 individuals of red spruce (pink) and for the subset of 323 individuals that clustered near red spruce in the haplotype networks (blue), filtered to exclude sites where more than 20% of the data was missing.

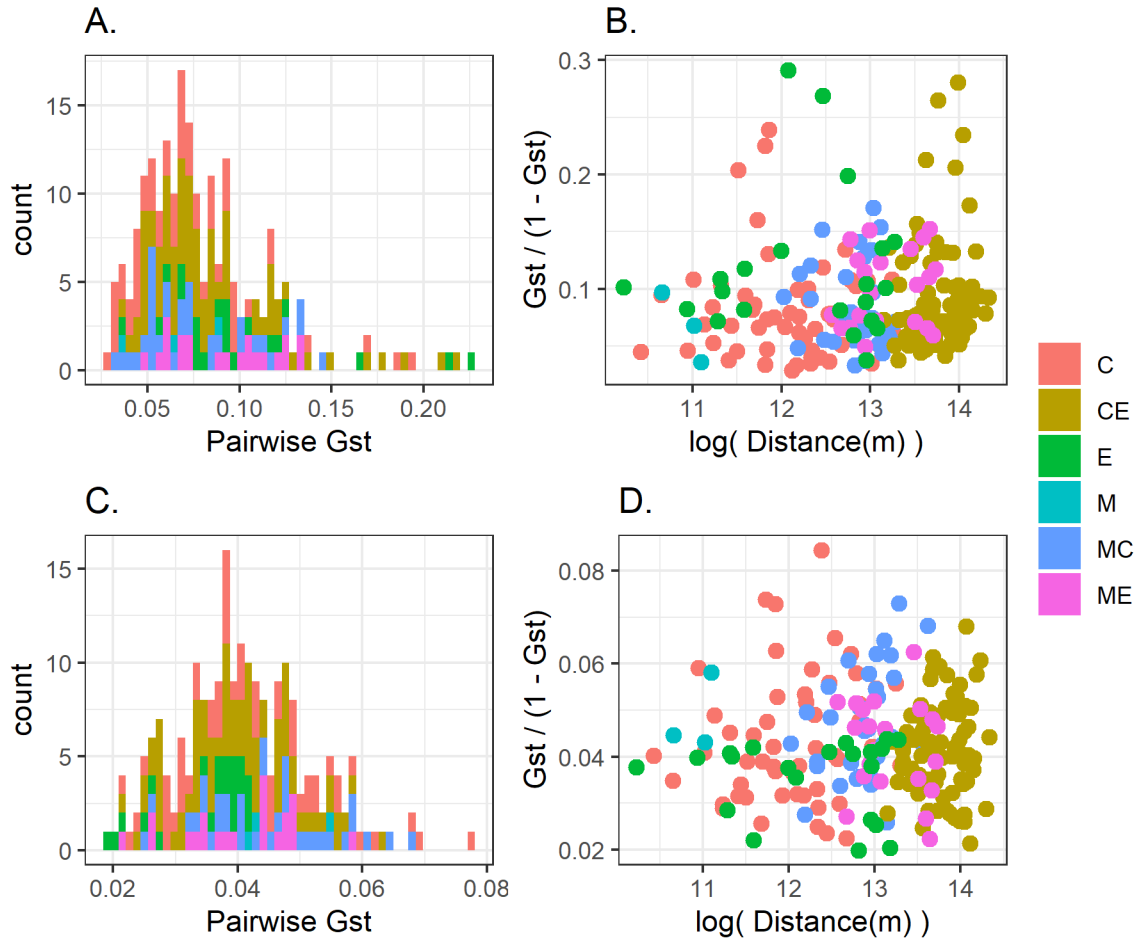
To assess the impact of hybridization with black spruce as demonstrated by haplotype networks, the analyses were repeated with only the subset of individuals clustering near the red spruce NCBI references and away from the black spruce NCBI references. The effective sample sizes (ESS) for all but one of the statistics calculated by BEAST2 for the subset of 323 individuals were above the threshold of 100, with a minimum value of 56 (corresponding with bGroupSizes.1) and a maximum value of 6153 (corresponding with gammaShape). The median ESS value was 175. The Bayesian Skyline Reconstruction analysis was consistent with the results from the full data set in its prediction of a current effective population size (Ne) of 43 million (Figure 7). The model illustrates an increase in chloroplast effective population size over

the last 100,000 years with a consistent  $N_e$  around 260,000 before that point, a level far lower than that seen in the full model.

### *Population Genetic Structure*

The overall  $G_{ST}$  for the full population was 0.115. The average pairwise  $G_{ST}$  was 0.0835 and values did not appear to vary significantly depending on the regions represented in the pair (Figure 8a). There was not a significant relationship between standardized  $G_{ST}$  and geographic distance between the pixels under consideration (two-sided Mantel test  $p$ -value = 0.489) (Figure 8b).

To investigate the impact of individuals that had greater black spruce ancestry in the haplotype networks, these analyses were repeated with only the subset of individuals that clustered near the red spruce references in the haplotype network (maximum missingness level = 20%). The overall  $G_{ST}$  for the subset population was 0.106 and the average pairwise  $G_{ST}$  was 0.0409. Again, there was not a significant relationship between standardized  $G_{ST}$  and geographic distance (two-sided Mantel test  $p$ -value = 0.961) (Figure 8d). The most notable difference between the analyses with and without these individuals is the range of  $G_{ST}$  values seen. Before subsetting the individuals included, the pairwise  $G_{ST}$  values ranged from 0.028 to 0.23 (Figure 8a). Post subsetting, they ranged from 0.020 to 0.078 (Figure 8c). For comparison, the nuclear estimates of  $F_{ST}$  from Capblancq et al. (in review) ranged from 0.00 to 0.04 and showed a significant positive relationship between  $F_{ST}$  and geographic distance for the nuclear genome.



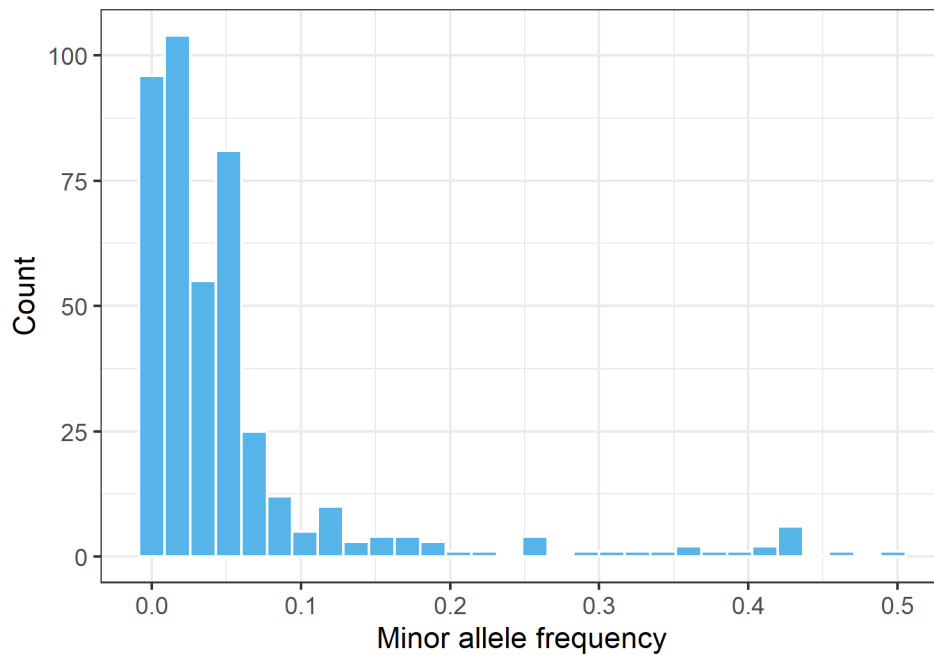
**Figure 8.** Pairwise  $G_{ST}$  distributions (a, c) and  $G_{ST}$  as a function of geographic distance (b,d) for both the full population (a, b) and the subset of the population less influenced by black spruce haplotypes (c, d).  $G_{ST}/(1-G_{ST})$  is used to standardize  $G_{ST}$ , and is graphed as a function of the distance in meters, which has been log transformed (b, d). The colors are used to indicate the regions of origin of the two samples included in the pairwise comparison (C = Core, M = Margin, E = Edge, and combinations of the three indicate one of the two samples is from each region).

Looking more closely at the frequency of rare alleles, Tajima's D was estimated for various subsets of the population (Table 3). Across the full range and in each of the regions individually, Tajima's D was below -3.0 and significantly different from 0 at the 0.01 level. This

same pattern can be seen in the plot of minor allele frequency, where there is a high frequency of rare alleles and fewer minor alleles that are common in the population.

**Table 3.** Tajima's D for the full data set and the 340 individuals subset by region

Region	Tajima's D	P-value
Full Range	-3.037	0.00239
Core	-3.223	0.00127
Margin	-3.546	0.00039
Edge	-3.348	0.00081



**Figure 9.** Frequency histogram of single nucleotide polymorphisms (SNPs) across the chloroplast genome. For each variable site, the number of individuals with the minor allele were recorded and represented as a proportion of the overall population. This is shown as the minor allele frequency.

## **Discussion:**

This research aimed to investigate how past demographic events and historic changes in climate have shaped the genetic structure, diversity, and population size of red spruce across eastern North America. The goal was to use this context to assess the long-term stability of contemporary red spruce populations and inform assessments of future responses to changes in environment. Based upon the observed trends of climate warming, expansion from glacial refugia, and an overall decline red spruce populations since the last glacial maximum, we expected to observe evidence for decline in effective population size reflected in BEAST2 Bayesian Skyline analyses. Additionally, we anticipated signs of population structure and some degree of isolation by distance reflected in range-wide estimates of genetic differentiation, though to a lesser to degree than seen in the nuclear genome based on their different forms of inheritance. While many of the following findings lend support to components of these hypotheses, one of the most divergent and noteworthy results of this study is the predicted increase in effective population size based on BEAST2 analyses.

### *Changes in Effective Population Size and Frequency of Rare Alleles*

The BEAST2 Bayesian Skyline analysis models show increasing chloroplast effective population size since about 100,000 years ago with an increase in effective population size from about 1,000,000 (or 260,000 based on the subset model) to 40,000,000 million during that period (Figure 7). This increase in  $N_e$  is intriguing as it has been shown that populations of red spruce have been declining since at least 1800 and likely much earlier, not flourishing as this analysis seems to suggest (Hamburg & Cogbill, 1988). When applying the STAIRWAY PLOT method to the same exome capture data used in the present study, Capblancq et al. saw a steady decrease in nuclear effective population size over the last 800,000 years with current effective populations in

each of the three regions (Core, Margin, and Edge) below 10,000 (Capblancq et al., 2020). Similar trends have been seen in studies that focused on the chloroplast, with Jaramillo-Correa et al. showing a decline in red spruce effective population size using chloroplast SSRs (simple sequence repeats) (Jaramillo-Correa, Gérardi, Beaulieu, Ledig, & Bousquet, 2015).

Not only is the change in chloroplast effective population size as a function of time in the opposite direction of what might be expected, the scale is also intriguing (Figure 7). The present day predicted chloroplast effective population size is more than 1,000 times larger than the prediction generated based on the nuclear genome (Capblancq et al., 2020). Under neutral conditions, we would expect the red spruce chloroplast genome to have half the effective population size of the nuclear genome because the chloroplast genome is haploid and red spruce is a monoecious species, meaning each plant has both male and female reproductive parts. However, the difference in  $N_e$  between chloroplast and nuclear predictions shown here does not follow these assumptions. This suggests the population is not in a neutral, stable state and there are other demographic forces at play.

When comparing the Bayesian Skyline analysis with the full population to the one performed on the subset red spruce cluster, the most noticeable difference is in the estimated effective population size as of 100,000 years ago when both models predict  $N_e$  to have stabilized (Figure 7). Using the full population,  $N_e$  was predicted to level off at about 1,100,000. With only the subset of 323 individuals in the red spruce cluster, the  $N_e$  was predicted to level off at about 260,000. Since a homogenous population of red spruce is less diverse than a mixed population of red and black spruce, one might expect that removing individuals with a black spruce background would lead to lower predicted levels of effective population size in the BDSKY models. This is what we see during the window from 100,000 to 300,000,000 years ago.

Based on the very high present-day  $N_e$  and growing population supported by the BEAST2 models, we would expect a large amount of low frequency mutations to be present, reflected in a negative measure of Tajima's  $D$ . In these results, the presence of a high number of low frequency mutations is confirmed by the population Tajima's  $D$  of -3.0 (Table 3) and high frequency of rare minor alleles (Figure 9). These measures suggest increasing effective population size in the red spruce chloroplast, which is in concordance with the BEAST2 analyses (Figure 7). The presence of many recent, rare alleles is also reflected in the haplotype networks (Figure 5). Particularly when more sites are included in the analyses by decreasing the stringency of the maximum missingness filtering, there are more haplotypes generated. The large number of haplotypes supports the notion that many different rare mutations are present within the population.

In the haplotype network for a maximum missingness level of 10%, there are many haplotypes that cluster in a ring in the upper part of the plot (Figure 5b). There is a single haplotype represented by at least one individual from the Core (yellow), hereafter referred to as the "central" haplotype, surrounded by many haplotypes that are equidistant from this central Core haplotype. The perimeter haplotypes are all one nucleotide different from the central haplotype. One possible explanation for this trend, and the surprisingly high frequency of rare alleles, is sequencing error. To further investigate this theory, the data were filtered by minor allele frequency using VCFtools to only keep alleles that were present in at least 3% of individuals. This filtering is based on the assumption that sequencing errors occur at random, making it unlikely that the same errors would be present across multiple individuals. By filtering out sites where the minor allele is only seen in a few individuals, it is possible to disentangle the influence of sequencing error. After filtering, 214 SNPs were retained and Tajima's  $D$  was

recalculated to be -2.99, compared to an initial value of -3.04 prior to filtering. The persistence of this highly negative Tajima's D statistic suggests that the high frequency of rare alleles is not purely an artifact of sequencing error.

The Bayesian Skyline analyses, Tajima's D statistics, and subsequent investigations all suggest that the high frequency of rare polymorphism is a reliable feature of the chloroplast genome and not an artifact. One potential explanation is the impact of selection. Since populations recovering from selective sweeps are known to generate a large frequency of rare alleles, a period of strong selection followed by a period of relaxed selection could have triggered the generation of new variants we see here. However, this hypothesis remains to be explored and would be an interesting direction for future work.

### *Population Structure*

Pairwise  $G_{ST}$  was estimated to address the question of how the genetic structure of the chloroplast genome varies across the current range. Focusing first on the entire population (Figure 8a,b), the pairwise  $G_{ST}$  is fairly independent of the pair of regions represented (Figure 8a). The overall  $G_{ST}$  for the full population was 0.115, which indicates a moderately high degree of differentiation between populations. The pairwise  $G_{ST}$  values ranged from 0.028 to 0.23 with an average of 0.084. This maximum value of 0.23 is notably high. Capblancq et al. (in press at Molecular Ecology) calculated pairwise  $F_{ST}$  using nuclear reads from the same exome capture sequence data, finding that a majority of the pairwise  $F_{ST}$  values ranged from 0 to 0.04, much smaller values than we see here for the chloroplast. These low nuclear  $F_{ST}$  values fall within the range of results seen in previous studies, suggesting a noteworthy deviation from expectations in the high chloroplast  $G_{ST}$  results (Bashalkhanov, Eckert, & Rajora, 2013; Rajora, Mosseler, & Major, 2000).

This striking deviation from expected  $F_{ST}$  values warrants further inquiry. One possible explanation for differences in  $F_{ST}$  is the difference in the way the chloroplast and nuclear genomes are inherited, the former uniparentally through pollen and the second biparentally. However, one might expect this difference in inheritance to have the opposite effect with pollen transfer allowing for a greater possibility of long-distance gene flow, thus homogenizing the different populations. The homogenization would cause the  $F_{ST}$  calculated from the chloroplast to be smaller than the  $F_{ST}$  calculated from nuclear sequences. This is the opposite of what we see here, suggesting the pollen dispersal hypothesis is not a plausible explanation for the observed differences in  $F_{ST}$ .

An important element of all of the analyses performed was monitoring the influence of putative hybridization with black spruce. This factor is relevant here as it can be used to explain the elevated  $G_{ST}$  values we see for the chloroplast data. When the  $G_{ST}$  calculations were performed again, this time only including the subset of individuals that clustered near the red spruce references in the haplotype network (maximum missingness level = 20%), there was an interesting shift in the results (Figure 8). The average pairwise  $G_{ST}$  dropped from 0.084 to 0.041 with a new minimum value of 0.020 and a maximum value of 0.078 – all values much more similar to those seen in the nuclear analyses as we would expect to see.

The inclusion of individuals with a black spruce background in these analyses had such a striking impact on the results due to the significant enough differences between the red and black spruce genomes on this small scale. The large values of  $G_{ST}$  reflect the differences between red spruce and black spruce, rather than the genetic differences between pairs of populations of pure red spruce. It is intriguing to note that there was far less of a signature of black spruce background in the analyses run using nuclear data, which is highlighted by the differences in

$G_{ST}/F_{ST}$  first described here. This can be explained by the haploid, non-recombining nature of the chloroplast genome that causes remnants of past demographic events to remain present for generations. Since the chloroplast genome is essentially a single locus, the sudden addition of new variation stands out. In the nuclear context, these same traces of past events are more likely to be hidden by generations of recombination and mutation and covered up by the presence of thousands of loci.

When looking at  $G_{ST}$  as a function of geographic distance, there is no correlation between standardized  $G_{ST}$  and distance between the pixels (populations) under consideration (Figure 8b,d) (p-value 0.489). Often there is a positive correlation between these two variables as it is logical that gene transfer is spatially limited, preventing total homogenization of populations. This positive trend is what we see in similar analyses performed on the nuclear genome where Capblancq et al. found that  $F_{ST}$  and geographic distance had a significant positive relationship through linear regression. Populations that were farther apart geographically had higher levels of genetic differentiation reflected in higher  $F_{ST}$  values.

However, in the context of the chloroplast genome it is interesting to see the lack of a relationship between  $G_{ST}$  and geographic distance. Whereas the nuclear genetic information is transmitted through both the male and female gametes and spatially restricted by both pollen flow and seed dispersal, of which seed dispersal is a more limiting factor, chloroplasts are only uniparentally inherited (Sutton et al., 1991). In conifers the chloroplast is paternally inherited through pollen. Since the spatial limitations of pollen transfer are less restrictive, it is reasonable to expect distance to have less of an impact on genetic differentiation in these populations (Davis & Shaw, 2001; Du et al., 2009). Seed dispersal is much more limiting, which explains the much

larger effect of geographic distance on red spruce genetic differentiation in the nuclear genome than the chloroplast genome.

The lack of isolation by distance suggests that expansion from glacial refugia was potentially a homogenizing force. Much of the homogeneity across chloroplast genome samples has held up to present day. This has not been the case with the nuclear genome where there has been genetic structural development post expansion due to spatially restricted gene flow. Since the chloroplast is less spatially restricted, we see less of a signature of structural differences that developed post-expansion. This also implies that gene flow via pollen dispersal has the potential to be a viable way of connecting populations that are otherwise spatially isolated, which is of particular relevance for the highly fragmented southern portion of the red spruce range. The “genetic rescue” effect provided by this pollen-facilitated high gene flow may prove especially critical under conditions of ongoing climate warming.

#### *Hybridization with Black Spruce*

While there is little evidence of genetic structure in the framework of geographic distance, structure becomes more noticeable in the context of hybridization between red and black spruce. Independent of the level of maximum missingness, all haplotype networks generated had two distinct clusters. One cluster contained the two red spruce references and haplotypes representing a majority of the 340 samples and the second contained the two black spruce references and a smaller subset of red spruce haplotypes, mostly from Core and Edge regions (Figure 5). This is further supported by the PCA results, which showed two distinct groupings specific to each of the red and black spruce references (Figure 6). The individuals within the black spruce cluster that formed based on the first principal component from the PCA were identical to those identified using the haplotype network (maximum missingness level =

20%). These were the same individuals excluded from the second round of BEAST2 analyses and  $G_{ST}$  calculations.

Taken together, the haplotype networks and PCA suggest that there has been introgression between red spruce and black spruce in parts of the species range. This activity appears to be more common in the northern part of the species range, areas represented by the Core and Margin. Since black spruce is most commonly found in Canada and the northeastern U.S., it is logical that these hybridization events would be most likely to occur in the overlap of the two species ranges, and not further south in the Edge region (De Lafontaine, Prunier, Gérardi, & Bousquet, 2015).

In the haplotype network using data that was filtered at the maximum missingness level of 10%, three Edge families were unexpectedly represented in the black spruce cluster (Figure 5b). These families were MT\_10 (North Carolina), RP\_01 (Tennessee), and HR\_02 (Tennessee). HR\_02 and RP\_01 were collected along roads within Great Smoky Mountains National Park, making it possible that they could have been planted there as part of road building and reforestation in the 1930s. The first two axes of a genetic principal component analysis completed based on the whole-exome sequence capture data shows HR\_02 clustering with families in the Core background rather than the Edge background that would be expected, providing genetic context for this explanation (Capblancq et al., 2020). MT\_10 was collected near a red spruce plantation on Higgins Bald, again making it possible that its background was not local. This provides some explanation for why we see that these families contain traces of black spruce ancestry, even though their sampling sites are outside the normal black spruce range.

The idea of introgression between red and black spruce in the Core and particularly the Margin is further supported by previous work by Capblancq et al. 2020. Using the same exome capture data that formed the basis of this work, they found high levels of genetic diversity within the Margin populations, even though these samples were highly geographically localized within the current range of red spruce (Capblancq et al., 2020). Hybridization can increase genetic variability and therefore, this could be one possible explanation for increased genetic diversity in this marginalized population of red spruce. Hybridization could have occurred between black spruce and red spruce, two species which are known to hybridize extensively (De Lafontaine et al., 2015). Future work clarifying the contribution of hybridization will be informed by a new set of whole exome-capture sequences that include black spruce samples.

### *Looking Forward*

This research uncovered intriguing trends in increasing chloroplast effective population size and a high number of rare alleles across the chloroplast genome. Additionally, we saw a lack of isolation by distance suggesting the that high gene flow could potentially provide a “genetic rescue” effect to isolated populations during climate change. Woven through these other findings is the pervasive theme of hybridization between red spruce and black spruce, which has clearly played a key role in shaping the demographic history of red spruce. All of this work was built on imputation procedures that helped to drastically reduce the quantity of missing information from 19% to 4% across the chloroplast genomes of the 340 individuals sequenced through whole-exome sequence capture (Figure 3). As imputation success is largely dependent on the size and completeness of the haplotype reference panel, the imputation procedures could no doubt be further improved upon by building a larger reference panel with greater genome coverage than could be provided by the four NCBI references. The optimized chloroplast isolation protocols

developed here are a major step in that direction and will help provide the sequence data to support future work.

## References:

- Avise, J. C., Bowen, B. W., & Ayala, F. J. (2016). In the light of evolution X: Comparative phylogeography. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 7957–7961. doi: 10.1073/pnas.1604338113
- Bashalkhanov, S., Eckert, A. J., & Rajora, O. P. (2013). Genetic signatures of natural selection in response to air pollution in red spruce (*Picea rubens*, Pinaceae). 5877–5889. doi: 10.1111/mec.12546
- Bio-Rad Laboratories. (2006). *Real-Time PCR Applications Guide* [Brochure]. Bio-Rad Laboratories, Inc.
- Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., ... Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*(4), 1–28. doi: 10.1371/journal.pcbi.1006650
- Breed, M. F., Ottewell, K. M., Gardner, M. G., & Lowe, A. J. (2011). Clarifying climate change adaptation responses for scattered trees in modified landscapes. *Journal of Applied Ecology*, *48*(3), 637–641. doi: 10.1111/j.1365-2664.2011.01969.x
- Capblancq, T., Butnor, J. R., Deyoung, S., Thibault, E., Munson, H., Nelson, D. M., ... Keller, S. R. (2020). Whole-exome sequencing reveals a long-term decline in effective population size of red spruce (*Picea rubens*). *Evolutionary Applications*, (April 2020), 1–16. doi: 10.1111/eva.12985
- Capblancq, T., Munson, H., Butnor, J. R., & Keller, S. R. (2021). Genomic drivers of early-life fitness in *Picea rubens*. *Accepted in Conservation Genetics*.
- Cavalli-Sforza, L. L. (1997). Genes, peoples, and languages. *Proceedings of the National Academy of Sciences*, *94*(July), 7719–7724.
- Clegg, M. T., Gaut, B. S., Learn, G. H., & Morton, B. R. (1994). Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(15), 6795–6801. doi: 10.1073/pnas.91.15.6795

- 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. doi: 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. doi: 10.1038/nmeth.2109
- Davis, M. B., & Shaw, R. G. (2001). Range shifts and adaptive responses to quaternary climate change. *Science*, 292(5517), 673–679. doi: 10.1126/science.292.5517.673
- De Hayes, D. H., & Hawley, G. J. (1992). Genetic implications in the decline of red spruce. *Water, Air, & Soil Pollution*, 62(3), 233–248. doi: 10.1007/BF00480258
- De Lafontaine, G., Prunier, J., Gérardi, S., & Bousquet, J. (2015). Tracking the progression of speciation: Variable patterns of introgression across the genome provide insights on the species delimitation between progenitor-derivative spruces (*Picea mariana* × *P. rubens*). *Molecular Ecology*, 24(20), 5229–5247. doi: 10.1111/mec.13377
- Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5), 1185–1192. doi: 10.1093/molbev/msi103
- Drummond, Alexei J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1), 1–8. doi: 10.1186/1471-2148-7-214
- Du, F. K., Lang, T., Lu, S., Wang, Y., Li, J., & Yin, K. (2015). An improved method for chloroplast genome sequencing in non-model forest tree species. *Tree Genetics and Genomes*, 11(6). doi: 10.1007/s11295-015-0942-2
- Du, F. K., Petit, R. J., & Liu, J. Q. (2009). More introgression with less gene flow : chloroplast vs . mitochondrial DNA in the *Picea asperata* complex in China , and comparison with other Conifers. *Molecular Ecology*, 18, 1396–1407. doi: 10.1111/j.1365-294X.2009.04107.x
- Dumais, D., & Prévost, M. (2007). Management for red spruce conservation in Québec: The importance of some physiological and ecological characteristics - A review. *Forestry Chronicle*, 83(3), 378–392. doi: 10.5558/tfc83378-3

- Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), 422–433. doi: 10.1038/nrg.2016.58
- Elmqvist, T., Folke, C., Nyström, M., Peterson, G., Bengtsson, J., Walker, B., & Norberg, J. (2003). Response diversity, ecosystem change, and resilience. *Frontiers in Ecology and the Environment*, *1*(9), 488–494.
- Fu, Y. X., & Li, W. H. (1999). Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoretical Population Biology*, *56*(1), 1–10. doi: 10.1006/tpbi.1999.1421
- Guindon, S., & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, *52*(5), 696–704. doi: 10.1080/10635150390235520
- Hamburg, S. P., & Cogbill, C. V. (1988). *Historical decline of red spruce populations and climatic warming*. *331*(1), 428–431.
- Hewitt, G. M. (1996). Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, *58*(3), 247–276. doi: 10.1006/bijl.1996.0035
- Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6). doi: 10.1371/journal.pgen.1000529
- Jaramillo-Correa, J. P., Gérardi, S., Beaulieu, J., Ledig, F. T., & Bousquet, J. (2015). Inferring and outlining past population declines with linked microsatellites: a case study in two spruce species. *Tree Genetics and Genomes*, *11*(1). doi: 10.1007/s11295-015-0835-4
- Knaus, B. J., & Grünwald, N. J. (2017). vcfR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44–53. doi: 10.1111/1755-0998.12549
- Kraja, A. T., Liu, C., Fetterman, J. L., Graff, M., Have, C. T., Gu, C., ... North, K. E. (2019). Associations of Mitochondrial and Nuclear Mitochondrial Variants and Genes with Seven Metabolic Traits. *American Journal of Human Genetics*, *104*(1), 112–138. doi:

10.1016/j.ajhg.2018.12.001

- Lin, D., Coombe, L., Jackman, S. D., Gagalova, K. K., Warren, R. L., Hammond, S. A., ... Birol, I. (2019). Complete Chloroplast Genome Sequence of a White Spruce ( *Picea glauca* , Genotype WS77111) from Eastern Canada . *Microbiology Resource Announcements*, 8(23), 13–16. doi: 10.1128/mra.00381-19
- Lo, T., Coombe, L., Lin, D., Warren, R. L., Kirk, H., Pandoh, P., ... Birol, I. (2020). Complete Chloroplast Genome Sequence of a Black Spruce ( *Picea mariana* ) from Eastern Canada . *Microbiology Resource Announcements*, 9(39), 1–3. doi: 10.1128/mra.00877-20
- Lockwood, J. D., Aleksić, J. M., Zou, J., Wang, J., Liu, J., & Renner, S. S. (2013). A new phylogeny for the genus *Picea* from plastid, mitochondrial, and nuclear sequences. *Molecular Phylogenetics and Evolution*, 69(3), 717–727. doi: 10.1016/j.ympev.2013.07.004
- Mosseler, A., Major, J. E., Simpson, J. D., Daigle, B., Lange, K., Park, Y. S., ... Rajora, O. P. (2000). Indicators of population viability in red spruce, *Picea rubens*. I. Reproductive traits and fecundity. *Canadian Journal of Botany*, 78(7), 928–940. doi: 10.1139/cjb-78-7-928
- Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., ... Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451), 579–584. doi: 10.1038/nature12211
- Paradis, E. (2010). *pegas* : an R package for population genetics with an integrated – modular approach. 26(3), 419–420. doi: 10.1093/bioinformatics/btp696
- Paradis, E. (2018). Analysis of haplotype networks: The randomized minimum spanning tree method. *Methods in Ecology and Evolution*, 9(5), 1308–1317. doi: 10.1111/2041-210X.12969
- Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3), 526–528. doi: 10.1093/bioinformatics/bty633
- Rajora, O. P., Mosseler, A., & Major, J. E. (2000). *Indicators of population viability in red spruce , Picea rubens . II . Genetic diversity , population structure , and mating behavior.* 956, 941–956.

- Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. doi: 10.1093/sysbio/syy032
- Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*(5), 380–390. doi: 10.1038/nrg795
- Rutledge, R. G., Stewart, D., Caron, S., Overton, C., Boyle, B., MacKay, J., & Klimaszewska, K. (2013). Potential link between biotic defense activation and recalcitrance to induction of somatic embryogenesis in shoot primordia from adult trees of white spruce (*Picea glauca*). *BMC Plant Biology*, *13*(1), 1–17. doi: 10.1186/1471-2229-13-116
- Sakaguchi, S., Ueno, S., Tsumura, Y., Setoguchi, H., Ito, M., Hattori, C., ... Isagi, Y. (2017). Application of a Simplified Method of Chloroplast Enrichment to Small Amounts of Tissue for Chloroplast Genome Sequencing. *Applications in Plant Sciences*, *5*(5), 1700002. doi: 10.3732/apps.1700002
- Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), 1–5. doi: 10.1093/ve/vey016
- Sutton, B. C. S., Flanagan, D. J., Gawley, J. R., Newton, C. H., Lester, D. T., & El-Kassaby, Y. A. (1991). Inheritance of chloroplast and mitochondrial DNA in *Picea* and composition of hybrids from introgression zones. *Theoretical and Applied Genetics*, *82*(2), 242–248. doi: 10.1007/BF00226220
- Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Baumont, L. J., Collingham, Y. C., ... Williams, S. E. (2004). Extinction risk from climate change. *Nature*, *427*, 145–148. doi: 10.1038/nature02121
- Thomas, E., Jalonen, R., Loo, J., Boshier, D., Gallo, L., Cavers, S., ... Bozzano, M. (2014). Genetic considerations in ecosystem restoration using native tree species. *Forest Ecology and Management*, *333*(2014), 66–75. doi: 10.1016/j.foreco.2014.07.015
- Vieira, L. D. N., Faoro, H., De Freitas Fraga, H. P., Rogalski, M., De Souza, E. M., De Oliveira Pedrosa, F., ... Guerra, M. P. (2014). An improved protocol for intact chloroplasts and

- cpDNA isolation in conifers. *PLoS ONE*, 9(1), 1–8. doi: 10.1371/journal.pone.0084792
- Wakeley, J. (2009). *Coalescent theory: an introduction*. Greenwood Village, Colo.: Roberts & Co. Publishers.
- Whalen, A., Gorjanc, G., Ros-Freixedes, R., & Hickey, J. M. (2018). Assessment of the performance of hidden Markov models for imputation in animal breeding. *Genetics Selection Evolution*, 50(1), 1–10. doi: 10.1186/s12711-018-0416-8
- Yun, L., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10, 387–406. doi: 10.1146/annurev.genom.9.081307.164242
- Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24), 3326–3328. doi: 10.1093/bioinformatics/bts606