Lingonberry (*Vaccinium vitis-idaea* L.) species origin and subspecies divergence through genome sequencing and assembly

by

Kaede Hirabayashi
Bachelor of Science (Honours), University of British Columbia Okanagan, 2021

A Thesis Submitted in Partial Fulfillment
of the Requirements for Degree of

MASTER OF SCIENCE

in the Department of Biology

Supervisory Committee

Lingonberry (*Vaccinium vitis-idaea* L.) species origin and subspecies divergence through genome sequencing and assembly

by

Kaede Hirabayashi
Bachelor of Science (Honours), University of British Columbia Okanagan, 2021

Supervisory Committee

Dr. Gregory L. Owens (Department of Biology)
Supervisor

Dr. Peter C. Constabel (Department of Biology)
Department Member

Dr. Samir Debnath (Agriculture and Agrifood Canada)
Outside Member

Abstract

Lingonberry (*Vaccinium vitis-idaea* L.) produces tiny red berries that are tart and nutty in flavour. It grows widely in the circumpolar region, including Scandinavia, northern parts of Eurasia, Alaska, and Canada. Although cultivation is currently limited, the plant has a long history of cultural use among indigenous communities. Given its potential as a food source, genomic resources for lingonberry are significantly lacking. To advance genomic knowledge, the genomes for two subspecies of lingonberry (*V. vitis-idaea* ssp. *minus* and ssp. *vitis-idaea* var. 'Red Candy') were sequenced and *de novo* assembled into contig-level assemblies. The assemblies were scaffolded using the bilberry genome (*V. myrtillus*) to generate chromosome-anchored reference genome consisting of 12 chromosomes each with total length 548.071 Mbp (contig N50 = 1.170 Mbp, BUSCO (C%) = 96.5%) for ssp. *vitis-idaea*, and 518.704 Mbp (contig N50 = 1.400 Mbp, BUSCO (C%) = 96.9%) for ssp. *minus*. RNA sequencing based gene annotation identified 27,243 genes on the ssp. *vitis-idaea* assembly, and transposable element detection methods found that 45.82% of the genome was repeats. Phylogenetic analysis confirmed that lingonberry is most closely related to bilberry and is more closely related to blueberries than cranberries. Estimates of past effective population size suggested a continuous decline over the past 1–3 MYA, possibly due to the impacts of repeated glacial cycles during Pleistocene leading to frequent population fragmentation. The genomic resource created in this study can be used to identify industry relevant genes (e.g., flavonoid genes), infer phylogeny, and call sequence-level variants (e.g., SNPs) in future research.

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

The following is a list of abbreviations and symbols employed in this Thesis.

BC	British Columbia, Canada
bp	base pair
°C	degree Celsius
CTAB	cetyltrimethylammonium bromide
D_a	Average amount of nucleotide divergence
D_{XY}	Amount of nucleotide divergence between X and Y populations
DNA	deoxyribonucleic acid
dsDNA	double stranded DNA
FPKM	fragments per kilobase of transcript per million mapped fragments
g	grams
gDNA	genomic DNA
gen.	generation
Gbp	giga base pair, 1,000,000,000 bp, 10^9 bp
GB	giga byte; unit for data, 1,000,000,000 bytes, 10^9 bytes
het	heterozygous
HMW	high-molecular-weight
k	kilo, 10^3
kbp	kilo base pair, 1,000bp, 10^3 bp
M	million, 10^6
Mbp	mega base pair, 1,000,000 bp, 10^6 bp
mg	milligrams
min	minutes
mut	mutation
MYA	million years ago
Ne	effective population size
NGS	next-generation sequencing
NL	Newfoundland and Labrador, Canada
no. / #	number of
PCR	polymerase chain reaction
pi / π	nucleotide diversity
pop.	population
RNA	ribonucleic acid
RNAseq	RNA sequencing
sect.	section, in plant taxonomy
SNPs	single nucleotide polymorphisms
ssDNA	single stranded DNA
ssp.	subspecies
subs.	substitutions
US	United States
var.	variety, in plant taxonomy
X	unit for sequencing coverage, “-fold”

Acknowledgements

I would like to first respectfully acknowledge the lək̓ʷəŋən peoples on whose traditional territory the University of Victoria stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day. Thank you for letting us work on your territory. I would like to also acknowledge the funding sources, NSERC Discovery Grant and the graduate scholarships and awards from the University of Victoria Department of Biology.

I am sincerely grateful for all the people I met during my degree. First and foremost, thank you to my supervisor, Dr. Greg Owens, for his dedication and belief in my potentials, and taking me on board with your brand-new lab in the amid of COVID back in January 2021, who had zero experience with coding and bioinformatics whatsoever. I cannot think of a supervisor who is so intelligent and patient, yet very humble and always available for feedback. Thank you for supporting me in every way I can ask for. I would like to present my genuine appreciation to my labmates, past and present: Dr. Jordan Bemmels, Dr. Sara Smith, Nathan Sykes, Justin Merondun, Koa, Alex, Adam, Cam, and Martin, for listening to me rant on occasional (maybe daily) complaints, troubleshooting together, and creating such a productive and fun work environment. I appreciate my former supervisors Dr. Lauren Erland and Dr. Susan Murch for continuously guiding me on the right path forward. I thank my committees, Dr. Peter Constabel and Dr. Samir Debnath, for being understanding and supportive of my progress. I thank my collaborators at Agriculture and Agri-Food Canada at St. Johns Research Centre for providing the materials, and peers from Koop lab, Constabel lab and Ehlting lab for troubleshooting together and sharing their facilities, chemicals, and protocols as necessary. Many thanks to Compute Canada resource for decades of computing hours, and their support team for dealing with my numerous troubleshooting emails. Lastly, I would like to say thank you to all my friends and family for making my grad school experience positive. I don't know how my grad student life would have gone if I didn't meet you two especially, Ashley and Anne-Marie, thank you both for being my mental support when I needed it!

Dedication

To all women in STEM, past mentors, and strong mothers, who proved me that becoming a scientist is an achievable career without giving up on personal life goals.

Chapter 1: Literature review

1.1 Lingonberry – *Vaccinium vitis-idaea L.* (Ericaceae)

1.1.1 Botany

Lingonberry (*Vaccinium vitis-idaea* Linnaeus) is an evergreen, berry-bearing dwarf shrub that inhabits a wide range across the Northern hemisphere (Figure 1.1; U.S. Department of Agriculture Natural Resources Conservation Service, 2021). The plant grows decumbent or ascending and can reach about 10–30 cm in height, with elliptic to egg-shaped leathery/shiny leaves with rounded ends. It produces pinkish-white bell-shaped flowers in summer and bright red-coloured fruits towards the end of summer to late autumn (Figure 1.2a). It can inhabit various types of habitats ranging from montane coniferous forests to arctic region, under sunny locations on sandy ground, rock, heath, or bogs. Acidic soil is generally preferred (pH 3.5–5), with little nutritional requirements. It reproduces vegetatively through rhizomes as a mat-forming forest understoryplant, or sexually through seeds (Tirmenstein 1991; Douglas and Meidinger 2002).

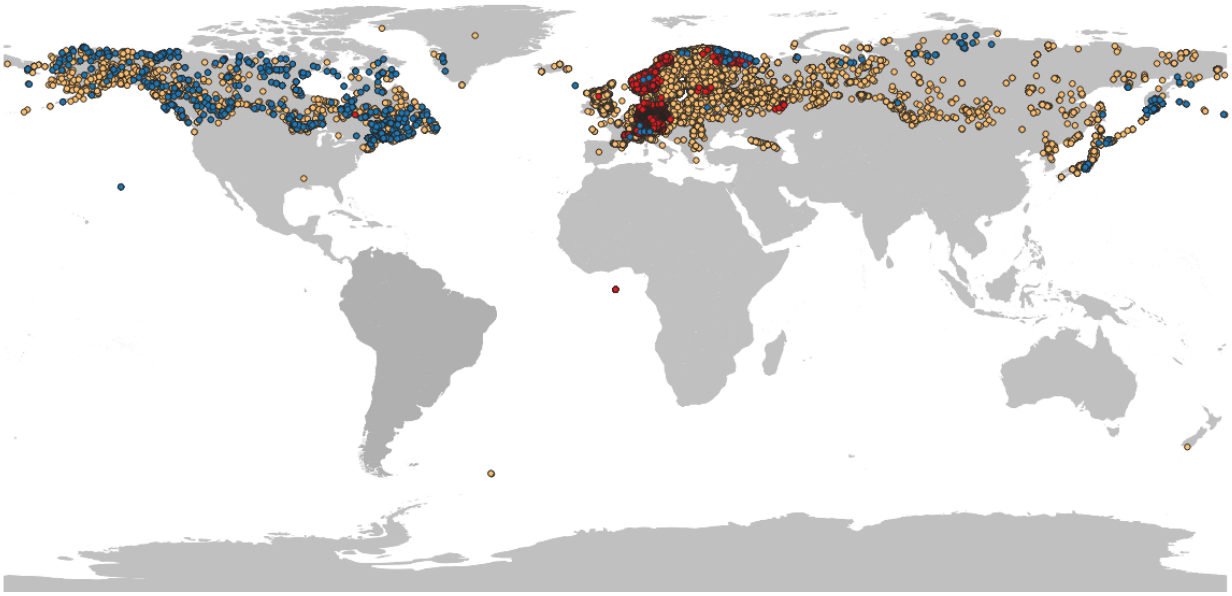

Figure 1.1: Worldwide distribution of *Vaccinium vitis-idaea* L. (www.gbif.org) Dots represent occurrence records registered as: *V. vitis-idaea* ssp. *minus* (blue), *V. vitis-idaea* ssp. *vitis-idaea* (red), *V. vitis-idaea* L. ssp. unidentified (yellow).

Floral characteristics, field observations, and hand-pollination experiments have agreed on its main mode of reproduction being outcrossing driven by insect pollination, with bumblebees being the most frequent (Jacquemart and Thompson 1996; Hjalmarsson and Ortiz 1998; Guillaume and Jacquemart 1999; Nuortila *et al.* 2002). Lingonberry is obligately outcrossing or partially self-incompatible. There is no evidence of outcrossing, but a reduced number of fruit set (Nuortila *et al.* 2002) and increased number of aborted seeds in self-fertilized samples compared to outcrossed ones suggests some mechanism of early inbreeding depression (Guillaume and Jacquemart 1999). However, in nature lingonberry predominantly propagates through rhizomes rather than seeds (Hjalmarsson and Ortiz 1998). Other pollinators like mammals or birds may spread the seeds through berries, but the seed germination success upon passing through the gut is infrequent (Nuortila *et al.* 2002). Nonetheless, occasional long-distance dispersal through the successful seeds in the berries is possible especially considering the wide range of habitat they thrive and the rapid recolonization history (Ikeda *et al.* 2015).

a **b**

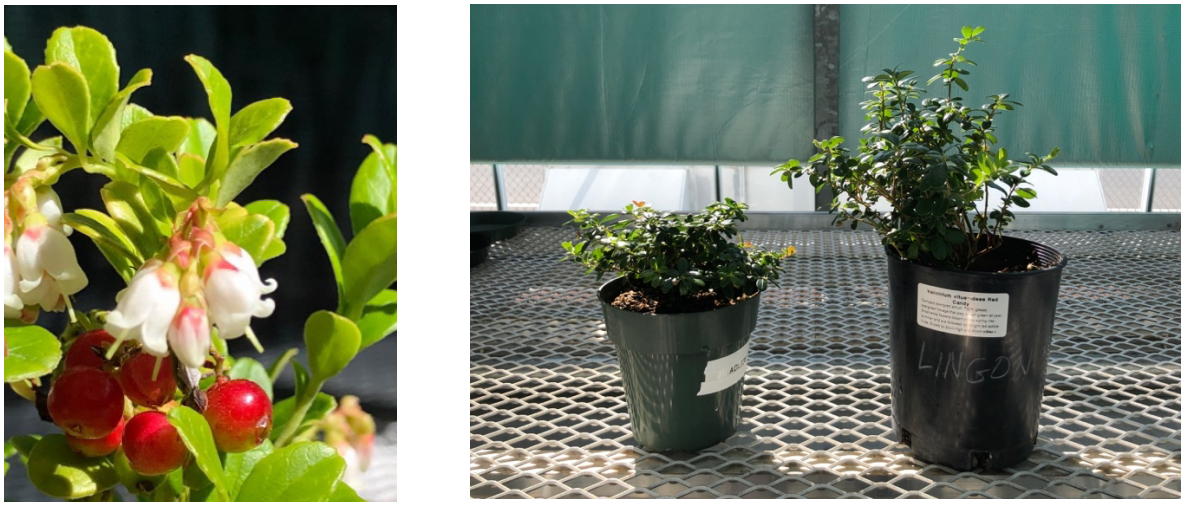

Figure 1.2: a) *Vaccinium vitis-idaea* ssp. *vitis-idaea* flowers and fruits. b) *V. vitis-idaea* ssp. *minus* (left) and ssp. *vitis-idaea* var. 'Red Candy' (right) grown in greenhouse.

The species has two recognized subspecies: *V. vitis-idaea* ssp. *vitis-idaea* and *V. vitis-idaea* ssp. *minus*, differing in morphology and geographical origin (Figure 1.2b, Table 1.1). Compared to the

European ssp. *vitis-idaea* that grows vertically upwards and produces crops twice a year, the North American ssp. *minus* is generally smaller in size, grows sideways by trailing just above the ground, and produces crops once a year (Gustavsson 2001; Penhallegon 2009; Debnath and Arigundam 2020). Individuals from ssp. *vitis-idaea* have been selected for commercial cultivation for at least several decades (Penhallegon 2006; Agriculture and Agri-food Canada 2023).

Table 1.1: Subspecies of lingonberry (*Vaccinium vitis-idaea*) and their distinct characters*.

Character:	*V. vitis-idaea* ssp. *vitis-idaea*	*V. vitis-idaea* ssp. *minus*
Variety	European	North American
Race	Large low land	Small arctic
Distribution	Europe, Asia	Iceland, Greenland, North America, northern Asia, Scandinavia
Plant height	10–30 cm	5–20 cm
Leaf	length: 2.5 cm, width: 1.0 cm	length: 1.0 cm, width: 0.5 cm
Berry	red, globular, 5-10 mm diam. acidic or slightly bitter	red, globular, 8-10 mm diam. acidic or tart
Crop per year	Two crops; First flowering in May and June, second blossoms in October producing more fruits.	One crop; Flowering beginning of June, fruiting late August to early September.
Flower	pinkish white	pinkish white
Habitat/Climate resilience	Found in lowland to mountain. Extremely frost tolerant, can tolerate -40˚C or lower, poor growth in areas with warm summers.	Found in well-drained dry habitats in arctic and alpine.

*Information collected from (Vander Kloet 1988; Douglas and Meidinger 2002; Nestby *et al.* 2019; Debnath and Arigundam 2020).

1.1.2 Ethnobotany

The use of wild-harvested lingonberry by indigenous peoples has been documented throughout the circumpolar region, including Nunavut, coastal and interior BC, and Haida Gwaii in Canada alone (Turner 1975, 1978, 2004; Moerman 2010; Mallory and Aiken 2012; Boulanger-Lapointe 2017). Wild berry-picking is a popular cultural practice among indigenous communities; the

berries are eaten raw or cooked with oil as a relish and served with meat and fish in traditional meals. Moreover, according to ethnobotanical surveys from Inuit people, Scandinavians, and Russian healers, berries and leaves are traditionally used for medicinal purposes as well, such as disinfecting bladder and kidney, lowering cholesterol level, treating hypertension, rheumatic diseases, mouth infections, sore throat, ulcers of the mouth, and scurvy (Jun *et al.* 1993; Cuerrier 2011; Belichenko *et al.* 2022).

1.1.3 Nutrition and health benefit

A growing body of research has focused on the potential medicinal benefits of lingonberry for human health such as its anti-cancer (Misikangas *et al.* 2007; Kondo *et al.* 2011; Onali *et al.* 2021; Zhu *et al.* 2022), cardioprotective (Isaak *et al.* 2017), and neuroprotective (Hossain *et al.* 2016) properties. The fruits and juice are known to contain varying amounts of nutrients including phenolic compounds with a high antioxidant capacity (e.g., anthocyanins, proanthocyanidins, flavanols, simple phenols such as arbutin, coumarins, phenolic acids), triterpenoids (C_{30}), fatty acids (alpha-linoleic acid, palmitic acid, stearic acid, oleic acid, linoleic acid, arachidic acid), and minerals (K, Fe, Cr, Cu, Zn) (Kowalska 2021). Lingonberry fruits contain 315–770 mg/100 g (dry weight) total anthocyanins composed of cyanidin-galactoside (69–90%), cyanidin-glucoside (2–10%), and cyanidin-arabinoside (6–23%); they have no peonidin-based anthocyanins which are common in cranberries (Brown *et al.* 2012; Amundsen *et al.* 2021). Stems and leaves are nutritious as well and can be used to make tea, now readily available commercially (Cuerrier 2011; Ferlemi and Lamari 2016; Raudone *et al.* 2019). The amount and composition of the above nutrients vary greatly among individuals and across wild populations (Alam *et al.* 2018; Amundsen *et al.* 2021; Vilkickyte and Raudone 2021), indicating an opportunity to explore their genetic basis, which may be useful in cultivar improvement.

1.1.4 Cultivation status – Emerging "Superfood"

While many fruit crops have been domesticated for thousands of years (Olver 1999), *Vaccinium* berries are considered New World foods with relatively recent histories of domestication (Olver 1999; Edger *et al.* 2022). In particular, lingonberry has only recently gained recognition as a potential crop species. Commercial investment for its cultivation was initiated in Sweden around the 1960s (Hjalmarsson and Ortiz 1998) and the oldest commercial field of lingonberry is in Germany and has been running since the 1970s (Jun *et al.* 1993). The selection of wild populations for cultivation began in the 1980s in Scandinavia (Gustavsson 2001), followed by the first North American research program investigating its physiological requirements established in 1987 in the US (Stang *et al.* 1993). Because cultivation strategies are underdeveloped (e.g., challenges in machine harvesting), lingonberry products sold commercially are mostly from wild-harvested fruits rather than domesticated fruits (Penhallegon 2009; Turner *et al.* 2011). Examples of wild-harvested fruit include the Finnish forest natural berry stand. There the public can freely access and pick berries during the harvest season, and about half of its harvest gets processed into products including jam and juice ("Arctic Lingonberry" 2022). In Canada, lingonberry is regionally cultivated in Newfoundland and Labrador and in Quebec. A few of the earliest facilities have been maintained by Agri-Foods Canada Research and Development Centre since 1999 (Debnath 2007b). A recently published news article suggests that lingonberry's potential value as a crop species is just getting recognized, and so far there are not enough farmers growing lingonberries to establish the market in North America (Arnason 2023). Lingonberry products sold at local grocery stores in BC are generally imported products from Europe (personal observation).

1.1.5 Genome structure

Flow-cytometry based studies found that lingonberry is predominantly diploid (2x) and has 12 chromosomes per haplotype (2n = 24), with an estimated genome size of 550 Mbp (Redpath *et al.* 2022). Naturally occurring tetraploid populations (4x, 2n = 48) or triploid individuals have been

recorded (Ahokas 1971; Lyrene *et al.* 2003; Wakui and Kudo 2021). No complete or draft nuclear genome assembly has been released for this species to date (https://www.vaccinium.org). However, its chloroplast genome (Kim *et al.* 2020) and full-length transcriptome assembly from berry tissues (Tian *et al.* 2020) are available.

1.1.6 Genomic and genetic diversity

Lingonberry (*V. vitis-idaea*) has been observed to frequently hybridize with species in sect. *Myrtillus* (*V. myrtillus*) in southeastern Europe, which is named *V. intermedium* Ruthe (Vander Kloet 1988; Bjedov *et al.* 2015), although contrasting observation of these two species growing in sympatry without hybrid formation has been made in northeastern Europe (Gailīte *et al.* 2020). Less commonly, lingonberry can cross with species in sect. *Vaccinium* (e.g. *V. uliginosum*) and sect. *Pyxothamnus* (e.g. *V. corymbodendron*) (Bjedov *et al.* 2015; Ehlenfeldt and Ballington 2018). Natural hybrids of lingonberry with the morphologically similar cranberry (*V. macrocarpon*) has not been confirmed, but they produce sterile offspring when crossed artificially (Edger *et al.* 2022).

The phylogenetic position of lingonberry within the genus *Vaccinium* is not clear. The earliest molecular study using two chloroplast genes (*mat*K gene and the nrITS region gene) supported a sister species relationship of lingonberry with cranberry (Kron, Powell, & Luteyn, 2002; Figure 1.3a), but a recent study using 30 chloroplast and 23 mitochondrial simple sequence repeats (SSR) markers suggested a different topology where cranberry is the sister group of lingonberry, farkleberry (*V. aboreum*), and the rest of the blueberries in the genus (Schlautman *et al.* 2017; Figure 1.3b). Whole chloroplast genome comparison supported the latter topology with cranberries sister to lingonberry, bog bilberry (*V. uliginosum*), and bilberry (*V. myrtillus*) (Kim *et al.* 2020; Fahrenkrog *et al.* 2022). A study comparing 507 SSR markers specifically developed for cranberry showed the sister-group status of a clade comprising lingonberry to huckleberry (*V. ovatum*), which is in turn sister to creeping blueberry (*V. classifolium*) and cranberries (*V.*

macrocarpon, V. oxycoccos) (Rodriguez-Bonilla *et al.* 2019; Figure 1.3d). In their tree topology, lingonberry is more closely related to cranberry than blueberry, though it should be noted that they did not include bilberry, so it is difficult to compare those results side by side.

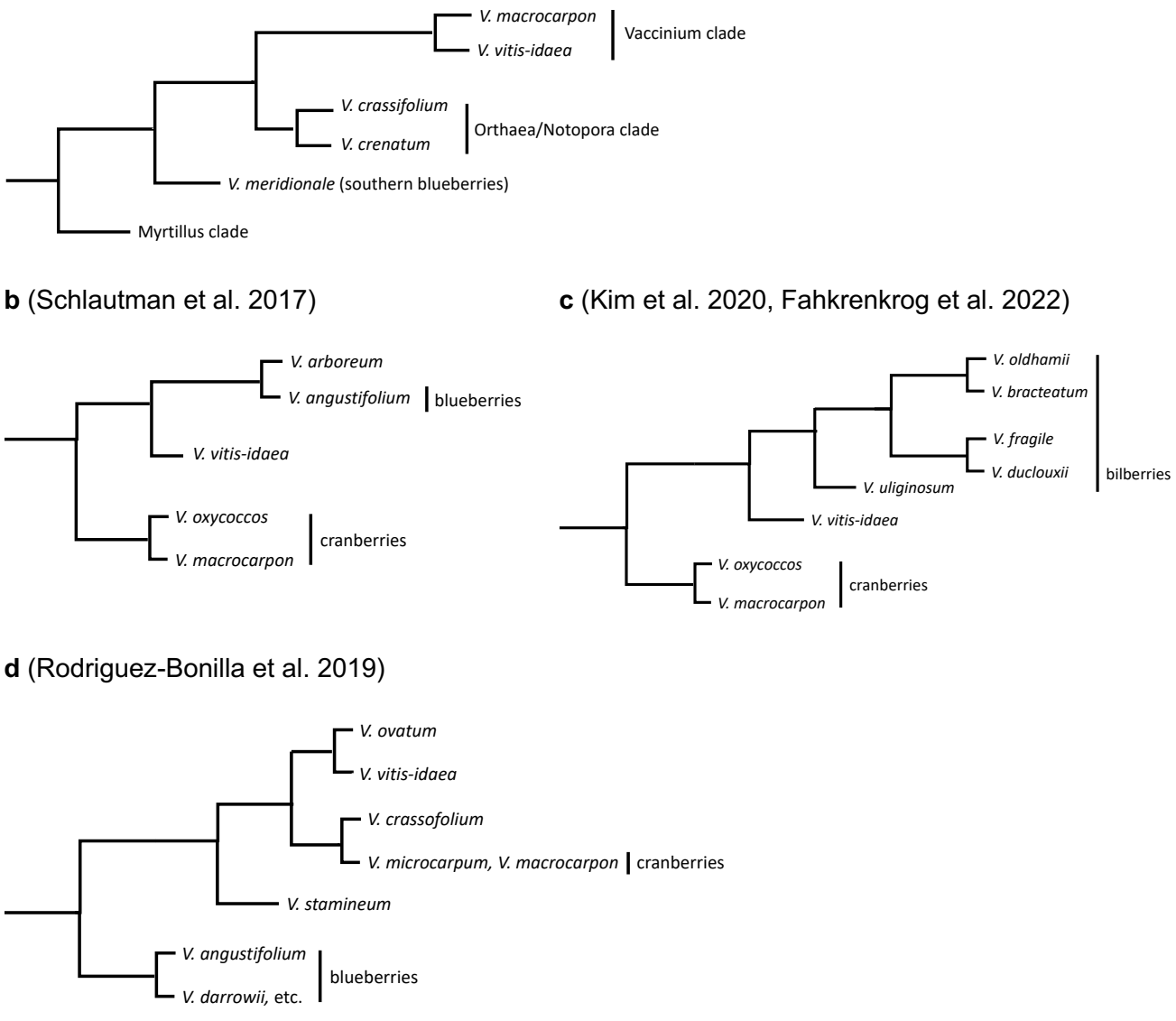

Figure 1.3: Published phylogenies of lingonberry (*Vaccinium vitis-idaea*).

The genetic diversity of wild lingonberry populations has been investigated using several neutral genetic markers, including random amplified polymorphic DNA (RAPD; Bjedov et al., 2015; Garkava-Gustavsson et al., 2005; Persson & Gustavsson, 2001), inter simple sequence repeat

(ISSR; Debnath, 2007a; Debnath & Sion, 2009), and microsatellite markers (Gailīte *et al.* 2020; Wakui and Kudo 2021). These genetic studies suggest that more genetic variation is found within-population comparisons than among-population or among-regional-populations. More recent studies have looked at SNPs and ploidy as additional genetic variation measures (Alam *et al.* 2018; Wakui and Kudo 2021) in an attempt to analyze the correlation between morphological features and genotypes. This relationship tends to be weak and no significant correlation between traits and genotypes have been observed (Persson and Gustavsson 2001; Debnath 2007a; Debnath and Sion 2009; Alam *et al.* 2018), although tetraploid populations seem to consistently occur in low-latitude areas than high-latitude areas, at least for populations in Japan (Wakui and Kudo 2021).

1.2 Genus *Vaccinium*

1.2.1 Phylogeny and species diversity

The genus *Vaccinium* encompasses over 500 species (Rosindell and Harmon 2012) most of which were described in the 20th century, including Vander Kloet's monograph on North American *Vaccinium* species (Vander Kloet 1988). The genus is classified in Ericaceae, subfamily *Vaccinioideae* (Kron *et al.* 2002a). The genus is further divided into many sections based on geography, morphology, and molecular characteristics (Vander Kloet 1988; Schlautman *et al.* 2017): sect. *Cyanococcus*, includes high-bush and low-bush blueberries; sect. *Oxycoccos*, American and wild cranberries; sect. *Vitis-idaea*, lingonberry; sect. *Vaccinium*, bog bilberry; sect. *Myrtillus*, European blueberry or bilberry and huckleberries; sect. *Pyxothamnus*, tropical blueberry and evergreen huckleberry; sect. *Batodendron*, farkleberry. Although the relationships between sections is not fully resolved (see Figure 1), *Vaccinium* as a genus is suggested as monophyletic in the recent microsatellite study (Schlautman *et al.* 2017).

Most *Vaccinium* species share the same karyotype with a base chromosome number of n = 12, and many of them have natural polyploid populations (Hancock, Lyrene, Finn, Vorsa, & Lobos,

2008; Lyrene et al., 2003; Sultana et al., 2020; Vander Kloet, 1988). This lack of major chromosomal differentiations allows frequent inter- and intra-section hybridization and polyploid formation (Edger et al., 2022; Vander Kloet, 1988), as well as continuous introgression which makes delineating the species difficult (Hancock et al., 2008; Lyrene et al., 2003). At the same time, this characteristic makes *Vaccinium* species a convenient system to breed for crop development. In plant breeding, researchers use the term gene pool to describe the diversity of genetic resources available for a crop plant. The primary gene pool is composed of the main cultivated species, the secondary gene pool is generally composed of the wild populations or close relatives of the cultivated species that readily cross, and the tertiary gene pool contains species close enough in genetic architecture considered to be relevant for exploring new traits beneficial for the cultivated species (Edger *et al.* 2022). The more genetically crossable populations/varieties are within the related species, the larger the size of each gene pool. In addition, the high species diversity within *Vaccinium* provides a potential case-study for investigating evolutionary mechanisms of rapid species radiations (Sultana *et al.* 2020; Cui *et al.* 2022).

Although the primary area of research has been breeding for better cultivar development, the genus has been occasionally investigated in evolutionary studies. For instance, some of the earlier works analyzed the distribution pattern of satellite repeats among different *Vaccinium* species to gain insights about their genome evolution (Sultana *et al.* 2020). Sultana *et al*. (2020) were able to identify some species-specific satellite repeats that likely evolved after speciation. At the whole-genome level, Cui *et al*. (2022) explored the tandemly duplicated genes in *V. darrowii* – the subtropical blueberry species – which has adapted to a different environment compared to the temperate species including *V. macrocarpon* and *V. corymbosum*. The authors used whole genome alignments to target gene families that have either expanded or contracted in this subtropical blueberry compared to *V. macrocarpon* and *V. corymbosum* and showed that genes involved in anthocyanin biosynthesis are relatively contracted in subtropical blueberry. They

suggested that this may be due to the role of anthocyanins in cold temperature protection, and the lack of their necessity in the subtropics. They also suggested that expanded gene families related to DNA repair may be an adaptation to heat tolerance (Cui *et al.* 2022). Kawash *et al*. (2022) similarly investigated genome-wide features of the commercial cranberry species, *V. macrocarpon*, but focused on the signatures of selection under crop development compared to the wild cranberry, *V. oxycoccos*. They found that the commercial cranberry genes involved in stress tolerance, such as those responsible for wax layers on leaves that could be protective against pathogens or UV stress, were under selection during crop development (Kawash *et al.* 2022). They also found very different sugar composition on anthocyanin molecules between the two cranberry species; however, the genomic basis or potential biochemical paths involved in differential anthocyanin production remains uninvestigated.

1.2.2 Use of genomic resource in berry crop development

Blueberries and cranberries are popular commodities that have been steadily increasing in supply and demand; blueberry production quantity scaled from 145 ktonne in 2000 to 497 ktonne in 2021, cranberry production scaled from 328k tonnes to 477 ktonne, producing $745M and $456M market value, respectively, in North America alone (FAO, 2022; Statistics Canada, 2022; USDA National Agricultural Statistics Service, 2021). Recent reports have highlighted how high-throughput genotyping is accelerating the breeding efforts of *Vaccinium* crops. For instance, marker-assisted selection allows more rapid recombination of traits being targeted. If a marker is known to be associated with a desirable trait, then the breeders can select for the individuals specifically carrying that desirable allele to use in the next round of crossing (Torkamaneh *et al.* 2018). In marker-assisted back-crossing, breeders take the opposite route and attempt to introduce the desirable allele into an elite cultivar through several generations of back-crossing (Varshney *et al.* 2014). Genomic selection allows breeders to predict the outcome at the pre-breeding stage by applying the knowledge of all the markers and their genetic incompatibilities to model phenotypic performances. This allows for the selection of seeds or seedlings that have the

highest breeding value or potential to grow into the most desirable crop, without needing to grow them into maturity (Varshney *et al.* 2014). Molecular breeding requires baseline knowledge of the genome, as well as a genotyped and phenotyped panel to detect associations between genotypes and phenotypes (Ferrão *et al.* 2021). In *Vaccinium*, the regularly screened traits include berry size, yield, and firmness which is especially a key factor for machine harvestability (Edger et al., 2022). Current consumer trends, on the other hand, lean towards better berry quality, such as flavour, rather than the quantity. In this regard, metabolomics-assisted flavour profiling and the development of a shared platform to compile known metabolites within *Vaccinium* berries are becoming today's industry priorities (Edger et al., 2022; Ferrão et al., 2021, 2020). However, selecting for such complex traits often requires multiple levels of infrastructure, including sensory panels, and so the implementation has been slow. Although metabolomics offers a great starting point to tackle this challenge (Colantonio *et al.* 2022), multi-omics approaches connecting the ties between genomics, transcriptomics, and metabolomics could also be considered utilising the already available genomic resource in the field (Edger et al., 2022).

1.2.3 What else do we know about *Vaccinium* genomes?

The cultivated species of blueberry is tetraploid *V. corymbosum*, whose genome assembly was first published in 2019 (Colle *et al.* 2019). Although the assembly was resolved to only a scaffold-level – many gaps and contigs unanchored to pseudomolecules – the study was successfully able to phase the four sets of homoeologous chromosomes constituting the total of 48 chromosomes. Using the high sequence coverage and physical linkage data (i.e., Hi-C), the genomic evidence for its allopolyploid origin through sequence similarity and date estimates from transposable element (TE) insertions was shown. Moreover, they identified candidate genes playing a role in fruit development, suggesting tandem duplication as a common mechanism of gene evolution in addition to whole genome duplication (WGD) in *Vaccinium*. More recently, two diploid wild progenitor species to the commercial blueberry genomes were published respectively (Cui *et al.* 2022; Mengist *et al.* 2023). Cui *et al*. (2022) largely supported the finding from Colle *et al*. (2019)

through genome alignment and synteny analysis, indicating both an ancient *Vaccinium* and a *V. corymbosum*-specific WGD event. In contrast, the study by Mengist *et al*. (2023) supported an autopolyploid origin of tetraploid blueberry, based on a lack of preferential pairing during meiosis. The observation of mostly quadrivalent pairing provided strong direct evidence that the commercial blueberry genome had undergone an autopolyploidization followed by substantial inter-chromosomal translocation and other rearrangements, which could have resulted in the highly diverged homoeologous sequences seen in previous studies that suggested allopolyploidy (Colle *et al.* 2019; Mengist *et al.* 2023).

The first chromosome-level cranberry genome was published in 2021, *V. macrocarpon* var. 'Stevens' (Diaz-Garcia *et al.* 2021), soon followed by another chromosome-level assembly published in 2022 by a different group, sequencing *V. macrocarpon* var. 'Ben Lear' (Kawash *et al.* 2022). The cranberry genome is diploid, and both studies included one other wild cranberry species assembly (*V. microcarpum* and *V. oxycoccos*, respectively). Both studies also placed cranberry in a similar phylogenetic position, indicating divergence from the blueberry clade ~10 MYA, although the timing of divergence slightly differs between the two studies. The divergence between *V. macrocarpon* and *V. microcarpum* and *V. oxycoccos* was suggested to be 4.5 MYA and 2 MYA, respectively. Note that both studies used Ks (synonymous substitution rate) as the basis of calculating divergence time, but the latter study tended to have more recent divergence estimates than the former. Several observations were made in the comparison between *V. macrocarpon* and *V. oxycoccos* in terms of genetic diversity with respect to inbreeding (Kawash *et al.* 2022). They performed resequencing of the parents and the inbred commercial lines to see if there is evidence of reduced genomic diversity, by counting unique heterozygous SNPs. The wild species was found to contain the highest SNP diversity (99% of the 250 kbp windows containing unique heterozygous SNPs), as expected, and the commercial lines were much lower with large variabilities (11-45%). Following selfing, the commercial species dropped in the count of unique heterozygous SNPs by the first generation – total heterozygous SNP counts became

about half, and by the fifth generation only a single genomic window had significant heterozygous sites.

1.3 Genome sequencing and assembly

1.3.1 Historical landmarks

The concept of DNA sequencing dates back to late 1970s when Sanger (Sanger 1975) and Maxam and Gilbert (Maxam and Gilbert 1977) developed an experimental setup to sequence a nucleotide base one by one through DNA synthesis or degradation, respectively. Both methods were highly accurate but were unable to sequence a fragment longer than 2000 bases. This is a problem when attempting to sequence an entire genome, which can be millions or billions of base pairs (bp) long. Beginning 2000s, the development of next-generation sequencing (NGS) technology allowed for a large-scale sequencing through a high-throughput analysis of nucleotide bases, generating hundreds of millions of reads per run. Early technologies feature short-read sequencing (e.g., Illumina, 454 pyrosequencing), producing DNA reads of 50-500 bp (Gavrielatos *et al.* 2021).

To sequence a whole genome that is longer than those individual reads, a reliable way to properly order, orient, and put together the sequences – the process called assembly – is necessary (Simpson and Pop 2015). Staden proposed a shotgun approach to solve this dilemma by fragmenting the genome randomly into smaller pieces, sequencing in parallel, and reconstructing the original sequence based on the overlaps (Staden 1979). This idea caught the attention of early mathematicians to create a computer program to assemble the fragments (Lander and Waterman 1988). Currently, approaches based on graphical representations which include Overlap-Layout-Consensus (OLC) graph, De Bruijin graph, and String graph predominate the assembly workflow, improving the speed and computational cost. However, as the genome size increases, *de novo* assembly of short fragments becomes more challenging and computationally expensive, being prone to errors due to repetitive elements (Gavrielatos *et al.* 2021). As such, it

is only recently that the whole-genome assembly became achievable and affordable for highly heterozygous, polyploid, or large genomes via long-read sequencing technologies.

1.3.2 Genome sequencing in the 21st century

Pipeline overview

Drawing on the growing number of literature on plant genome assemblies published to date, a novel genome assembly nowadays follows a similar pipeline: 1) DNA is extracted from the organism of interest; 2) DNA is sequenced on at least one long-read and one short-read sequencing platform; 3) sequenced reads are *de novo* assembled into contiguous overlapping reads (contigs) using computational tools; 4) contigs are corrected for errors and misalignments by aligning them back to the raw reads (polishing or error correction); 5) additional read mapping or physical mapping of chromosomes is completed to assemble contigs into larger contiguous units called scaffolds.

Quality and quantity assessment of sequenced reads

In order to discuss different sequencing platforms available, it is essential to understand what quantity and quality metrics are used and how they are reported in genome assembly literature. The quality of sequenced reads is typically represented by Phred quality score or Q score, which corresponds to the error rate of basecalling defined in a logarithmic scale (Illumina 2011):

$$Q = -10 \log_{10} P$$

where *P* = probability of basecalling error. So, when the probability of basecalling error is one in a thousand bases, $P = \frac{1}{1000} = 0.001$ and thus $Q = -10 \log_{10}(0.001) = 30$. This means Q30 equals 0.1% basecalling error rate or 99.9% basecalling accuracy.

The quantity of sequenced data produced from a sequencing platform is typically in the scale of giga base pair (Gbp, 10^9). Since the genome length is variable, how much sequencing data is adequate completely depends on the organism and the purpose of the study. Thus, the measure of quantity can also be represented by a unit relative to the genome length of the target organism called coverage denoted as "X" or "-fold" computed by:

$$C = LN/G$$

where *C* = coverage, *L* = read length, *N* = number of reads, *G* = haploid genome length (Lander and Waterman 1988). When sequencing is performed for assembly purposes, we describe the amount of data used to assemble by coverage or depth. These terms can be used interchangeably (Sims *et al.* 2014), with the idea being how many times the exact same location of the genome/transcriptome is sequenced. Coverage often describes the average number of reads that would theoretically cover the same positions in the genome using alignment to a reference genome, whereas depth is used more generally to describe the total number of reads generated without necessarily mapping to a reference sequence (Illumina 2022). For instance, we estimate raw read depth from a sequencing run by taking the number of total base pairs of the data divided by the estimated genome length.

1.3.3 Sequencing platforms

While multiple options exist for NGS platforms, Illumina (Harris *et al.* 2008) represents the most widely used high-throughput short-read sequencing platform that generates a consistently accurate sequencing data (Q ≥30) with read length of 150-300 bp (Illumina 2011). Illumina uses sequencing by synthesis, which means the enzymatic reaction by DNA polymerase that adds complementary bases to the 3' OH group is monitored and recorded using fluorescently tagged deoxyribonucleotides (dNTPs). In contrast to the previous methods that required parallelizing each reaction step for each DNA template, this technology monitors and records the signals from each template strand separately, allowing fast and simultaneous sequencing (Harris *et al.* 2008).

Although Illumina generates large amounts of data and is the dominant sequencing platforms in genomics, the state-of-the-art sequencing methodology for genome assembly is long-read sequencing (Pucker *et al.* 2022). Plant genomes are often challenging to assemble due to large size and high proportion of repetitive elements. During assembly, individual reads are linked together by overlap or similarity. For reads containing repetitive regions, a potential overlap exists for each copy of the repeat leading to uncertainty in true position, and ultimately a fragmented assembly. The solution to this challenge is having a read long enough to span an entire repetitive region. In this case, read linking is done using the single-copy regions flanking the repeat region, which can be unambiguously joined with other reads. Ultimately this produces a more continuous genome assembly. Two competing long-read sequencing platforms today are single-molecule real-time sequencing or SMRT sequencing by Pacific Bioscience (PacBio) and nanopore sequencing by Oxford Nanopore (ONT).

PacBio uses sequencing-by-synthesis similar to the traditional Sanger sequencing and Illumina technology, except the incorporated dNTPs are attached to a fluorophore through phosphate group so that it does not hinder the native enzymatic reaction of DNA polymerase, maximizing its fidelity and speed (Eid *et al.* 2009). With the zero-mode waveguide invention that helps to observe fluorescence at a nano-scale (Levene *et al.* 2003), it enabled the detection of the fluorescence from the incorporation of a single base pair, in comparison to Illumina which detects the signal of multiple incorporation together in order for it to reach the detection limit. By design, SMRT sequencing allows DNA template to be sequenced as long as the DNA polymerase stays active at the zero-mode waveguide reaction site, resulting in read length of 10s of kbp which is a significant increase from ~300 bp in short-read technologies (Eid *et al.* 2009). With the improved circular consensus sequencing mode, where the same DNA molecule is read multiple times by the enzyme, the accuracy consistently exceeds 99.9% (>Q30) (Wenger *et al.* 2019).

On the other hand, ONT uses a completely different approach to sequence a DNA molecule. It focuses on the chemical differences between nucleotide bases themselves. DNA is fed through a nanopore and the resulting changes in electric current are recorded and analyzed to identify the base pair sequence. In the conventional library preparation stage (Ligation Sequencing Kit-110 or older), adapters are ligated to the DNA, one side of which is attached to a motor enzyme and the other is looped by a hairpin. The adapter-ligated DNA molecule is then attracted and captured by the nanopore where the dsDNA is unwound, and one strand is guided to pass through the pore. When the motor enzyme reaches the adapter, ideally, the strand switches and the complementary strand starts passing the pore. When the complementary strand is successively sequenced, these paired sequences connected by the adapter sequence can be obtained, also called duplex reads. Duplex reads would provide a breakthrough in terms of read accuracy as reading the complementary strands consecutively substantially improves the basecalling accuracy (Lawrence 2022). This, in theory, is possible for all dsDNA that pass through the pore; however, the duplex reads are currently only a minor proportion of all reads.

The original version R9.4.1 of MinION sequencer (Oxford Nanopore, OX4 4DQ, UK) has one electric sensor in the nanopore reading barrel producing simplex read basecalling accuracy of ~98% (~Q17) and duplex reads are rarely observed, possibly due to the failure in ligation of hairpin loop during the library preparation and/or interference in signal caused by other secondary structures of DNA and artifacts (Oxford Nanopore, 2021). The improved R10.4 version features the dual reader barrel where the electric signals of DNA strand is recorded at two locations within the nanopore, increasing the accuracy of simplex (one strand) basecalling, specifically for homopolymer regions because a longer portion of the DNA molecule is monitored at a time instead of a single point (Oxford Nanopore, 2020). Additionally, the new library preparation (LSK-114) potentially provides a solution to duplex sequencing by encouraging the natural flow of the complementary strand passing after the template strand without using hairpins. It instead ligates adapters on both sides of the DNA molecule with a tether that secures DNA on the membrane

where the nanopore is located. This greatly increases the likelihood of the complementary strand being read successively following the template strand: up to 40% of the reads have been reported as duplex (>Q29) (Oxford Nanopore 2021).

1.3.4 Genome assemblers

In parallel to the advancements in sequencing platforms, advancements in computational technology have also been made rapidly in recent years. In contrast to the first ever assembled genome of a plant species, *Arabidopsis thaliana* (~125 Mbp), which took over ten years of assembling efforts with several world-class laboratories in collaboration (The Arabidopsis Genome Initiative, 2000), now a plant genome assembly of the comparable size can be performed solely by computer software (*de novo* assembler) within a few hours (Gavrielatos *et al.* 2021). In the next few sections, I will discuss the currently available *de novo* assemblers that perform the assembly procedures solely from sequenced reads. I will focus on the algorithms as well as the main advantages and limitations for each type of assembler.

Canu (Koren *et al.* 2017) – a fork of the conventional Celera assembler used in the Human Genome Project – represents by far the most accurate and thorough assembler with strict requirements for base call accuracy and overlap quality of the raw reads. At the same time, the software is optimized to work with the relatively noisy, high-error-rate long-reads so that it can be applied to PacBio and ONT data. The first step in Canu assembly pipeline corrects the input raw reads by detecting base-by-base overlaps using the MinHash alignment process (MHAP). This alignment algorithm uses weighted k-mer distribution to find candidates for overlaps. Multiple rounds of such error correction are done to ensure that all the reads used in the assembly are accurate and have enough coverage. Corrected reads are then trimmed to remove unsupported bases, hairpin adapters, chimeric sequences, and other anomalies present in the dataset. Then finally the software applies the greedy “best overlap graph” approach to pick the best assembly (i.e., the longest overlap of reads).

If the raw reads are accurate and abundant (~40X coverage), Canu does a great job at assembling the genome correctly. Because of rigorous error correction and trimming, misalignments and misassembly is rarer than any other tools introduced below. However, Canu requires the most computational resource as the MHAP is a memory-intense process (Chen *et al.* 2021). The fundamental approach used for its assembly algorithm is the OLC graph, which requires the computer to keep track of all the reads with overlaps and find the longest connected path without breaks. The best path is basically chosen based on consensus of all the read fragments in the dataset (Simpson and Pop 2015). When it encounters repeats, the program flags it as a suspicious overlap and makes breaks in the assembly, creating many contigs made of repeats that are then left out until the end (i.e., they are never assembled into longer contigs). Moreover, Canu error-correction may be too strict and so may not retain any valid reads or overlap candidates when there is not enough sequencing data (i.e., low coverage). Particularly important given the poor accuracy of ONT reads, Canu trims reads with low accuracy and consequently reduces the raw read length – reducing the main strength of ONT (Chen *et al.* 2021).

To accommodate such issues with ONT reads, several programs such as Miniasm (Li 2016), SmartDenovo (Liu *et al.* 2021), wtdbg2 (Ruan and Li 2020), and Flye (Kolmogorov *et al.* 2019) attempt assembly first then correction. Miniasm works by using minimap2 (Li 2018) all-vs-all read mapping, less sensitive trimming of reads so that regions covered by three good mappings are retained, and assembly using the OLC approach. Although it is fast, Miniasm has drawbacks as it skips the sequence consensus step in OLC. This creates more erroneous overlaps due to sequencing errors or repetitive regions. In addition, ambiguous overlaps are merged at the end to produce unitigs, which can also introduce misassemblies (i.e., frequent collapse of repeats and segmental duplications) (Li 2016; Sun *et al.* 2021). SmartDenovo works similarly to Miniasm, but it adds a homopolymer compressed k-mer counting function at the all-vs-all read mapping stage (Liu *et al.* 2021). This targets the inaccurate homopolymer regions in the ONT data sets and improves the assembly accuracy compared to Miniasm. Moreover, it stores the OLC graph to

clean up and rescue missed overlaps to maximize total length of valid overlaps before moving onto assembly. Wtgbg2 avoids the memory-intensive base-by-base mapping by first counting the k-mer occurrence and binning the reads into 256 bp sequences as one unit. It then constructs the fuzzy Brujin assembly graph by all-vs-all alignment between binned reads (Ruan and Li 2020). In this way, it effectively performs consensus sequence detection while retaining mismatches and gaps when needed.

Flye (Kolmogorov *et al.* 2019) is an alternative assembler developed for ONT data. Flye utilizes a strategy called disjointigs which allows it to extend the overlapping sequences with potential mismatches, errors, or gaps as contigs. This significantly reduces the computational load (i.e., it can perform faster and is less memory intensive) and increases the length of contigs from error-prone ONT reads compared to OLC or de Brujin graph-based approaches mentioned above. It then performs self-alignment of the concatenated disjointigs to identify repetitive sites and build a repeat graph. The repeat graph then resolves highly repetitive regions by looking for small differences between repeat sequences through multiple rounds of read correction. Due to the initial step not involving error correction and rigorous base-by-base alignment like MHAP, Flye fully maximizes the long read length of ONT. By taking the repeat graph approach, it can generate long contigs in a short time with high confidence which was previously impossible to do with only short-read data. However, the drawback of Flye is limited accuracy and potential misassembly. Because ONT raw reads are error-prone, the individual reads can contain numerous incorrect base pair assignments. When erroneous reads are combined as a contiguous sequence, the resulting contig could be completely a false overlap. Moreover, resolving minimal differences between repeat copies is nearly impossible with inaccurately basecalled ONT reads.

1.3.5 Post-assembly process

Polishing

As genome sequencing is never perfect, error correction or polishing of assembled draft genome is important, even when integrated into assemblers as described above. If the assembly based solely on ONT long-reads were used for variant calling, there is at least 1% error rate in predicted indels (insertions and deletions), meaning one in 100 detected indels are false positives created by sequencing error (Oxford Nanopore, 2021). A few programs are specifically designed to polish the genome after draft assembly, and most importantly this can be done with short-read data to compensate for the errors produced by long-read technologies. Pilon is a representative polishing tool that provides error correction based on paired-end Illumina reads (Walker *et al.* 2014). The ONT sequencer measures the changes in electric current, which struggles to accurately basecall when it encounters low complexity regions such as homopolymers (Delahaye and Nicolas 2021). While Illumina can use counts to determine how many C's were present in a row through fluorescence, for example, ONT needs to estimate how many C's might have passed the pore based on the constant current pattern in the given time recorded. The latter is challenging because the DNA translocation speed is not always the same and depends on the condition, and basecalling algorithms are still not perfect (Delahaye and Nicolas 2021). Additionally, ONT tends to make small indel errors, which are rare occurrences in Illumina data. Because common Illumina errors are associated with random point mutation caused by PCR (polymerase base pair mismatch) or DNA damage during library preparation (Stoler and Nekrutenko 2021), often a polisher like Pilon can significantly improve the draft genome assembly by identifying potential base errors through alignment of short-reads and evaluating pileups, then correcting the potential misassemblies caused by the indel errors (Walker *et al.* 2014).

Although the basecalling accuracy of a draft genome assembly can be significantly improved with long-read data only (such as NextPolish; Vaser *et al.* 2017, Racon; Hu *et al.* 2020), the higher accuracy provided from short-read data is beneficial when performing any kind of downstream analysis. False positives can inflate the genomic divergence between compared species, for example, and can easily overestimate their divergence time. Artificial indels or nucleotide variants

introduced due to sequencing errors can result in frame-shift mutations or incorrect gene annotations. Therefore, it is important to minimize basecalling errors as much as possible. Given the reduced cost of short-read sequencing and the significant improvement in accuracy by incorporating the data, the current standard for *de novo* genome assembly is to use the hybrid approach to polish the assembly.

Phasing and polishing haplotigs

In early genome sequencing efforts, samples would be selected that had little or no heterozygosity, for example inbred lines (e.g., Kawash et al., 2022). In this case, a single haploid reference genome was produced which selected one parental allele in heterozygous positions (Whibley *et al.* 2021). For diploid heterozygous samples, fully assembling separate parental chromosomes is highly challenging due to their similarity. For instance, with 0.5% heterozygosity, a typical value for many organisms, that means the two parental copies are ~99.5% identical. One approach for solving this challenge is using inheritance information. If both the maternal and paternal genomes are sequenced, then reads in the offspring can be partitioned into bins corresponding to maternal and paternal genomes (e.g., Trio Binning (Cheng *et al.* 2021)). Some existing assemblers using PacBio HiFi reads have a diploid-aware mode to generate a phased genome assembly by determining primary haplotype and alternate haplotype from a single heterozygous individual sample (e.g., FALCON, HiCanu, HiFiasm). These programs attempt to first produce the best assembly, aligns reads again to the best reference assembly, and then the bubbles in the assembly graph created by polymorphisms are used to identify heterozygous sites. The partially assembled reads (contigs separated by the bubbles) are then split into two separate graphs to perform assembly independently for each chromosome copy (Duitama 2023). However, this algorithm relies heavily on the raw reads basecalling and variant calling accuracy and so the accuracy is not guaranteed with more error-prone ONT reads.

Without correctly phasing the genome, parental chromosomes may or may not be collapsed together during assembly. If they are not collapsed, this can lead to false duplications, also known as haplotigs, particularly when the studied genome is highly heterozygous (Rhie *et al.* 2021). These would lead to misinterpretation regarding gene duplication, structural variation, and so on in the downstream analysis. Phasing the assembled genome or at least minimizing the dual presence of the same region of the genome in a haploid assembly is therefore very important. Tools such as purge_haplotigs (Roach *et al.* 2018) have been developed to assign whether a duplicated contig is truly a duplicate or a haplotig. This uses read depth data from the raw reads mapped to the draft genome assembly and flag contigs as primary contig (1X), suspect contig (~0.5X because it is heterozygous), or artifactual contig (too low or high coverage). After self-alignment of those suspected contigs and determining which should be a haplotig or repetition, redundant haplotigs are purged and combined with primary contigs to produce a curated haploid assembly.

Scaffolding

De novo assemblers typically generate thousands of contigs, but to produce chromosome-level assemblies that are gap-free, additional steps are taken to scaffold them. One way is using a reference genome assembly from closely related species to align, orient, and order fragmented contigs. Ragtag (Alonge *et al.* 2019) is an example of reference-based scaffolding software that does this. While this is efficient and relatively fast to do, it comes with some drawbacks. The quality and accuracy of the assembly depends on that of the reference genome assembly used. If the reference genome was not correctly assembled, then this could lead to inflated misassemblies with the input contigs. Another concern is how diverged the study organism is from that of the reference genome. A divergent reference genome may lead to a scaffolded genome that is structurally more similar to the reference used, than the true genome structure. To scaffold

the contigs more directly without using a reference genome, a few physical mapping approaches have been developed, including Hi-C, genetic marker mapping, and optical mapping.

Hi-C or the chromatin conformation capture is the most common scaffolding methods used in the modern chromosome-level genome assembly work (Whibley *et al.* 2021). This method works under the assumption that regions in the genome that are physically close together in the nucleus makes the most physical contact with each other (Belton *et al.* 2012). To capture those physical contacts, Hi-C protocol first fixes chromatin in place, then uses cross-linking enzymes to connect nearby DNA sequences. The resulting sequencing library produces read pairs that link together genomic regions that are physically nearby. After processing, the amount of read pairs linking regions is used as a measure of proximity in the genome and is used to order and orient contigs. Although this method is reliable and has been applied in numerous genome assembly projects, the cost and complexity of the library preparation procedure restricts its uses.

Quality assessment of assembled genome

Intuitively, a complete reference genome assembly should include nearly all sequences placed onto chromosome-level contigs. But this has proven very challenging to achieve, as shown by the fact that the truly gapless complete human haploid genome assembly was just recently achieved 20 years after the initial reference genome assembly (Lander *et al.* 2001; Nurk *et al.* 2022). So how do we know if a genome assembly is good enough? How do we judge which assembly is better than the other? Several metrics for assembly are typically presented in *de novo* genome assembly papers: total assembly size, number of contigs/scaffolds, N50 score of contigs/scaffolds, BUSCO score, and number of genes annotated. The total genome size can be reliably estimated either experimentally using flow cytometry or computationally using NGS coupled with *k*mer analysis (Redpath *et al.* 2022), and an assembly larger or smaller than the expected size can indicate uncollapsed haplotigs or overly-collapsed repeats respectively. Flow cytometry estimates genome size using a DNA specific dye and measures the amount of

fluorescence emitted from isolated nuclei (Hare and Johnston 2011). *K*mer spectrum, or the distribution of DNA length *k*, can be acquired from NGS data and used to estimate the total genome length as well (Chor *et al.* 2009; Simpson 2014). The number of contigs indicates how fragmented the resulting assembly is; fewer contigs mean a more contiguous assembly. Ultimately, the goal of genome assembly is to have the scaffold number matching the known chromosome number. The N50 score indicates that 50% of the total assembly length is contained in the contig or scaffold of at least this length. The N90 score is similar but represents 90% of the total assembly length. The last common way to assess genome completeness is to look for genes that should be present. Benchmarking Universal Single-Copy Orthologues (BUSCO) genes are highly conserved genes found in nearly all organisms. The BUSCO score describes whether these genes are present in the genome (Simão *et al.* 2015) – the details are described below.

1.3.6 Genome annotation

Once a genome is assembled, the next step is to understand what is in the genome. Annotation allows much broader applications in the downstream, including studies to infer phylogeny or gene orthology or address ecological questions such as selection and adaptation. Genome annotation identifies where genomic features are in the genome, for example transposable elements (TEs) and genes. It also provides information about those elements, such as gene identifiers and their protein coding products. Transposable elements are identified by unique structural components such as the known transposase sequences or tandem repeats (Bourque *et al.* 2018), which can be identified by screening a database of known TEs (e.g., Ou and Jiang 2018; Flynn *et al.* 2020).

There are two main approaches to identifying genes: 1) intrinsic or the *ab initio* predictions and 2) extrinsic or evidence-based approaches with proteins and/or transcripts. The intrinsic approach is a method that predicts gene structures solely from the nucleotide sequence (Ejigu and Jung 2020). Example programs include GlimmerHMM and Augustus (Stanke *et al.* 2006). The extrinsic approach, sometimes called the similarity-based approach, relies on information outside of the

genome sequence. For example, it can involve aligning sequenced mRNA to the genome, and guiding gene identification based on this (Ejigu and Jung 2020). While this works for many genes, it can struggle to identify all splice isoforms for a gene or identify genes with sporadic or low expression levels that may not be captured in RNA sequencing. Intrinsic annotation methods do not suffer from this issue, but rely on accurate gene models, which can vary between species. For species evolutionarily distant from well-studied species, these gene models may be less accurate. Some popular modern approaches often use a hybrid annotation pipeline, such as BRAKER and MAKER pipelines, where they start with an *ab initio* prediction using trained sets of gene models, then correct the prediction based on external evidence (Cantarel *et al.* 2008; Brůna *et al.* 2021). A recent comparative study recommends the use of both intrinsic and extrinsic methods (Vuruputoor *et al.* 2022).

Lastly, BUSCO scores are useful to assess the completeness of genes annotated. The score counts the presence of complete, duplicated, fragmented, and missing single-copy orthologs that are found in >90% of the species within a taxonomic lineage (Waterhouse *et al.* 2013; Simão *et al.* 2015). Complete BUSCO indicates that the gene was found in the query sequence at an above-expected alignment score and length, and the gene can be either single-copy or duplicated. Fragmented BUSCO, on the other hand, indicates the partial presence of gene that does not quite reach the alignment minimum, which could be due to incompleteness or significant base pair mismatch. Any orthologs that do not count towards the above categories are considered missing orthologs that should be present if the query organism is in the chosen lineage. Consequently, a large amount of missing BUSCO genes indicates the incompleteness of the query genome assembly or gene annotation. The current version of BUSCO offers a eudicot dataset (eudicots_odb10) composed of 2,326 conserved genes found across 31 species (Manni *et al.* 2021).

1.3.7 Summary – what is the best practice?

Several benchmarking studies comparing different assembler and polisher performances have been published (such as Sun et al., 2021; J. Wang et al., 2021), but with the fast pace of new tools being developed every year, the best assembler and the associated combination of polishing practices is up to the user's judgement. In addition, since different biological systems have unique genomic properties (e.g., size, ploidy, heterogeneity, GC content, repeat content), such benchmarking studies often conducted with bacterial or human genomes might be inapplicable for plant genomes.

1.4 Why lingonberry genome assembly matters

This literature review has highlighted some of the knowledge gaps that exist in lingonberry and *Vaccinium* research in general, including its phylogenetic relationships with other species in the genus and the under-investigated genomic diversity and subspecies divergence. Considering the depth of morphological, phenological, and biochemical analyses underway, there is comparatively little knowledge on lingonberry genomics. Many untouched questions regard the links between these observable phenotypes and genomic signatures, which could take advantage of the genomic resource I describe in the next chapter. For instance, why is the anthocyanin content so variable depending on the individuals? Why are there such significant morphological differences between the North American and European lingonberries? How did they become separate subspecies and why are they (not) mixing? The reference genome assemblies for lingonberry will, therefore, not only provide useful basis of molecular breeding (e.g., development of lingonberry-specific genetic markers for cultivar selection and improvement) but also add lingonberry as a study system to research edible wild berry diversity in North America.

Chapter 2: Unveiling the evolutionary history of lingonberry through genome sequencing and assembly of European and North American subspecies

2.1 Scope of the project

Vaccinium vitis-idaea L., commonly known as lingonberry or cowberry, is an evergreen dwarf shrub that has cultural, economic, and ecological importance. The bright-red coloured berries have been consumed among Indigenous communities in northern North America and Scandinavia as a relish and served with meat or fish in traditional meals (Turner 1975; Moerman 2010; Vaara *et al.* 2013). Berry picking has been a cherished cultural practice and nowadays people commonly preserve berries as jams and process them into pastries like tart or cookies, which are becoming more readily available commercially ("Arctic Lingonberry" 2022). A growing body of research suggests that lingonberry fruits have medicinal benefits to human health such as anticancer, cardioprotective, and neuroprotective properties (reviewed in Kowalska, 2021). Other studies have also described lingonberry leaves to have medicinal value for potential natural products discovery (Cuerrier 2011; Ferlemi and Lamari 2016; Raudone *et al.* 2019). Despite a long history of utilization as a culturally important food source and its recognized health benefits, the domestication of lingonberry is at its infancy in North America.

As an evergreen boreal forest understory species, lingonberry mainly propagates vegetatively by forming a mat-like clonal communities through rhizomes (Hjalmarsson and Ortiz 1998), or sexually through seeds which is primarily driven by insect pollination (Jacquemart and Thompson 1996). The species has two recognized subspecies (ssp.) based on their geographical origin: *V. vitis-idaea* ssp. *minus* and ssp. *vitis-idaea,* and the species is widely distributed in the circumpolar region (Figure 1.1). The European subspecies, ssp*. vitis-idaea*, currently has active breeding programs with more than a dozen cultivars available for commercial production, with improved yield and berry size (Penhallegon 2009). The North American ssp. *minus*, on the other hand, is considered a wild plant and little breeding efforts has taken place. The two subspecies are

distinguishable based on several morphological differences (Table 1.1) as well as genetic differences (Garkava-Gustavsson *et al.* 2005; Debnath 2007a). The extent of genomic differences between the two subspecies has not been studied before, and it is somewhat unclear whether they occur sympatrically in the overlapping ranges.

Long-read sequencing technology has fueled exponential growth in the assembly of plant genomes (https://www.plabipd.de/timeline1_view.html); there are at least 1,205 unique flowering plant species genomes assembled at higher than scaffold-level (NCBI search terms: "Magnoliopsida (flowering plants)" "scaffold+", by May 24th, 2023) and this number is likely underestimated because some genomes may not be deposited in the NCBI archive. The use of long-reads has been particularly relevant for plant genomes due to their high repeat proportion and propensity for polyploidy. Within *Vaccinium*, high quality genomes have been assembled for nine species (Colle *et al.* 2019; Diaz-Garcia *et al.* 2021; Wu *et al.* 2021; Yu *et al.* 2021; Cui *et al.* 2022; Kawash *et al.* 2022; Yang *et al.* 2022; Mengist *et al.* 2023), and there is an ongoing pangenome project for cultivated blueberry and cranberry involving 32 cultivars (Edger 2023). In contrast, genomics research on lingonberry is at its infancy; only a handful of genetic, chloroplast or mitochondrial genomic studies have been conducted (Garkava-Gustavsson *et al.* 2005; Debnath 2007a; Gailīte *et al.* 2020; Kim *et al.* 2020; Tian *et al.* 2020). My goal is to provide useful genomic resource to the lingonberry community, through genome assembly of the two distinct subspecies: *Vaccinium vitis-idaea* ssp. *vitis-idaea* and ssp. *minus*. To achieve this goal, my specific objectives are:

1) To assemble the genomes of two lingonberry subspecies to chromosome level and define the genomic differentiation between the two subspecies.
2) To explore the demographic history of the two subspecies and infer their divergence time.

3) To annotate the genome with genes using RNA sequencing and perform functional annotation using orthologous genes in model systems, then further quantify expression levels of important enzymes in anthocyanin production.
4) To compare the lingonberry genome to other related *Vaccinium* species (e.g., *V. macrocarpon, V. darrowii, V. corymbosum, V. myrtillus*) and clarify its phylogenetic position.

The resources created from my project are being made publicly available, in the hope of furthering our understanding of lingonberry evolution and aiding the future breeding efforts by accelerating the molecular screening of lingonberry cultivars.

2.2 Materials and methods

2.2.1 Plant material

The clones of a commercial lingonberry plant (*Vaccinium vitis-idaea* L. ssp. *vitis-idaea* var. 'Red Candy') were obtained from Lochside nursery (Victoria, BC) in September 2021 and July 2022 and kept in the greenhouse, designated as LC1 and LC2, respectively. The original plants were claimed to be collected from a wild-grown stand (location unknown). The wild lingonberry clone (*V. vitis-idaea* L. ssp. *minus*) designated as LW1, originally collected from Baie-Trinite, Quebec, Canada (Latitude: 49° 25'N; Longitude: 67° 18'W; Debnath, 2007a) was obtained from collaborators at Agriculture and Agri-Food Canada St. John's Research and Development Centre, NL, and kept in the greenhouse. The three accessions were vouchered and deposited at University of Victoria herbarium collection: LC1 = UVIC 48749, LC2 = UVIC 48750, LW1 = UVIC 48751, respectively.

2.2.2 High-molecular-weight DNA extraction

Young and mature shoots were excised from each subspecies (LC1, LW1). The leaves (1-2 g dry weight) were collected from shoots and wiped with 70% ethanol prior to extractions. The sterilized leaves were flash-frozen in liquid nitrogen and ground into fine powder using mortar and pestle (~5 min). High-molecular-weight (HMW) DNA was extracted using Nucleobond® HMW DNA extraction kit (Takara Bio, San Jose, CA) following the manufacturer's protocol, with double the amount of starting material and the buffers accordingly. The DNA was then size-selected using Short Read Eliminator – size XS or normal kit (Circulomics, PacBio, Menlo Park, CA) to remove fragments smaller than 10 kbp or 25 kbp, respectively. The extracted DNA was assessed for quality using Qubit (Invitrogen, Qubit® 2.0 Fluorometer) and Nanodrop (Thermo Scientific, NanoDrop® Spectrophotometer ND-1000) and stored at 4 ℃ until sequencing.

2.2.3 RNA extraction

Total RNA was extracted for the commercial lingonberry clones, LC1 or LC2, in the greenhouse from five tissue types: young expanding leaf (LC1), flower (LC2), unripe berry (greenish white; LC2), ripe berry (red; LC2), and rootstalk (underground rhizome; LC2). Note that the rootstalk was technically an underground shoot, but it does not have green leaves. The root tissue could not be sampled due to soil contaminations and difficulty in extracting enough root mass without killing the plant. For leaf and flower samples, a modified CTAB protocol was used to isolate RNA (Muoki *et al.* 2012; Yoshida *et al.* 2015). For rhizome, Spectrum™ Plant Total RNA Kit (Sigma) was used to extract RNA according to the manufacturer's protocol. For berries (LC2), a modified CTAB protocol optimized for bilberry was used to isolate and purify RNA (Jaakola *et al.* 2001). Due to low recovery of pure RNA, the unripe and ripe berries were combined to make up one berry sample in my study, resulting in a total of four RNA samples prepared for sequencing.

2.2.4 Sequencing

Long-read sequencing libraries were prepared with the Ligation Sequencing Kit (SQK-LSK110 or SQK-LSK114, ONT) which were then sequenced on MinION Flow Cell R9 (FLO-MIN106D) or R10.4.1 (FLO-MIN114), respectively, following the manufacturer's protocols. For LC1, a total of three runs were performed on a single R9 flow cell, each run followed by a washing step (Flow Cell Wash Kit; EXP-WSH004). Additionally, three runs were performed on a single R10.4.1 flow cell with 'accurate (250 bp per second)' sequencing mode. For LW1, a total of ten runs were performed on three R10.4.1 flow-cells with 'accurate (250 bp per second)' sequencing mode. All the raw output FAST5 reads were then basecalled using Guppy basecalling software v6.1.2+e0556ff (https://nanoporetech.com/) and minimap2 v2.22-r1101 (Li, 2018) using the super accurate or 'sup' model (-c dna_r9.4.1_450bps_sup.cfg). For reads generated with R10.4.1 flow cells, the reads were further duplex-basecalled according to the Guppy Duplex-basecalling pipeline v6.3.8+d9e0f64 (https://nanoporetech.com/). In brief, raw FAST5 files were basecalled

using the ‘fast’ model (dna_r10.4_e8.1_fast.cfg), and the duplex candidates were listed as read-pair candidates. Those reads were then duplex-basecalled by Guppy-duplex. The remaining reads were identified on the simplex reads already basecalled with the ‘sup’ model (dna_r10.4_e8.1_sup.cfg) using a custom perl script, and finally the duplex basecalled reads were combined with the duplex-filtered simplex reads. The generated FASTQ files were concatenated as a single raw-reads output for the downstream procedures. Note that the raw basecalled reads were filtered by the mean >Q10 prior to concatenating.

For short-read sequencing, the extracted DNA and RNA samples were sent to the Michael Smith Genome Sciences Centre at UBC for sequencing. The DNA library was prepared as a PCR-free genome and was sequenced on Illumina NovaSeq paired-end mode, targeting 75M individual reads per sample. The RNA library was prepared using the PolyA+ mRNA Library Construction service provided and sequenced on Illumina NovaSeq paired-end mode, targeting 50M reads per sample (LC1 leaf, LC2 flower/berry/rootstalk). The raw output FASTQ files were visually quality checked with fastqc v0.11.9 (Andrews 2019).

2.2.5 Assembly and polishing

Several different assemblers and polishing methods were implemented to achieve the best assembly result. The assemblers Flye v2.9-b1778 (Kolmogorov *et al.* 2019), Miniasm v0.3.r179 (Li 2016), wtdbg2 v2.5 (Ruan and Li 2020), SmartDenovo v1.4.0 (Liu *et al.* 2021), and Canu v2.2 (Koren *et al.* 2017) were tested. I have done the assembly with the above five programs and I picked the best assembly method based on the one with the least number of total contigs and the longest N50 value. The final assembly pipeline for each subspecies is as follows. For the commercial lingonberry or LC1 assembly, initial draft genome was assembled with SmartDenovo v1.4.0 (Liu *et al.* 2021), polished with the ONT reads three times using NextPolish v1.4.0 (Hu *et al.* 2020) and with Illumina reads three times using Pilon v1.24 (Walker *et al.* 2014). In brief, the raw FASTQ paired-end reads were first filtered and trimmed using Trimmomatic v0.39 (Bolger *et*

al. 2014)(the parameters used were ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36). The successfully paired reads were aligned to the respective long-read polished draft genome using BWA mem v0.7.17 (Li, 2013), then sorted and indexed with samtools v1.10 (Danecek *et al.* 2021) prior to polishing with Pilon for a total of three rounds with default parameters. Last, haplotigs and other redundant contigs were removed using purge_haplotigs v1.1.2 (parameters -l 5 -m 42 -h 95 -j 70 -s 70)(Roach *et al.* 2018). For the wild lingonberry or LW1 assembly, raw ONT reads were corrected and trimmed with Canu and then assembled by SmartDenovo. The draft assembly was similarly polished with ONT reads using NextPolish three times, with Illumina reads three times using Pilon (same parameters as LC1), and haplotigs were removed using purge_haplotigs (paramters -l 5 -m 40 -h 95 -j 70 -s 70). Note that each polishing step was done up to three rounds, or until before the BUSCO score started to decline. The *de novo* assembled genome was then scaffolded to chromosomes based on mapping contigs to the bilberry genome (*V. myrtillus*; C. Wu et al., 2021), using Ragtag v2.1.0 (Alonge *et al.* 2019). I did not enable the 'correction' mode on Ragtag, meaning it was not looking for potential misassemblies in the *de novo* assembled contigs because "misassemblies" may represent genome structure differences between bilberry and lingonberry. The final genome assembly was assessed for contiguity (N50, N90 values), per-base accuracy (QV score or consensus accuracy, error rate) and completeness (BUSCO %) using BBMap v38.86 (Bushnell 2014), Merqury meryl v1.4 (Rhie *et al.* 2020) and BUSCO v5.1.2 with parameters: --lineage_dataset eudicots_odb10, --mode genome (Simão *et al.* 2015; Manni *et al.* 2021), respectively.

2.2.6 Gene and TE annotation

Adapter trimming of Illumina RNA reads was performed by Trimmomatic v0.39 (Bolger *et al.* 2014) (parameters used are ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36). The quality of RNA reads was visually checked with fastqc, making sure that there was no sequence bias or decline in read quality throughout. The reads

were then aligned to the scaffolded genome including all contigs using Hisat2 v2.2.1 with default parameters (Kim *et al.* 2019). Additionally, published transcriptome data from *V. vitis-idaea* var. 'Sunna' (green, white, red berries) was collected (Tian *et al.* 2020) and aligned to the LC1 genome. Following alignment, transcript assembly was performed using StringTie v2.1.5 with default parameters (Pertea *et al.* 2015), and the transcripts were stored as a structural definition file gtf. Upon conversion of gtf to the appropriate gff3 format, gene features (i.e., untranslated regions (UTRs), exons, introns, genes, mRNAs) were predicted on the assembled transcripts using TransDecoder v5.5.0 (Haas 2023). The longest open reading frame (ORF) prediction (command: TransDecoder.LongOrfs) was run with -S option to ensure the orientation of the paired Illumina reads. A Blastp reference library was prepared with *Arabidopsis* and *Vaccinium* known proteins from the UniProt database, to retain homologous hits on ORFs even if they do not exceed the coding likelihood scores used to filter ORF candidates in the preceding steps. I used *Arabidopsis* and *Vaccinium* protein databases because *Arabidopsis* is the most well annotated flowering plant with gene models available in eudicots, and the *Vaccinium* database was the closest published protein gene models to lingonberry, in the hope of discovering berry-specific genes. Finally using this information, genes were predicted (command: TransDecoder.Predict) with the parameter --retain_blastp_hits. The final output was produced in annotation format gff3, relative to the LC1 scaffolded genome assembly, and used in downstream analysis. In cases where there were isoforms (genes of same genomic position, slightly different splicing pattern) or overlapping genes (splicing variants or conflicting candidate gene models), the longest gene hit was chosen as the best candidate sequence.

Transposable element (TE) annotation was done following the Extensive *de novo* TE annotator pipeline v2.0.0 (EDTA; Ou *et al.* 2019). In brief, candidate TEs were identified using LTR-Finder (Xu and Wang 2007; Ou and Jiang 2019), LTRharvest (Ellinghaus *et al.* 2020), LTR_retriever (Ou & Jiang, 2018), TIR-Learner (Su *et al.* 2019), generic repeat finder (Shi and Liang 2019), and HelitronScanner (Xiong *et al.* 2014), followed by RepeatModeler (Flynn *et al.* 2020) to find any

missed TEs based on structural-based methods. Finally, the combined repeat libraries were filtered so that coding sequences (CDS) from my transcript-based gene annotation did not get masked by repetitive regions. Additional filters to effectively remove false positives were also provided at each step of combining multiple independent programs, according to the EDTA pipeline (Ou *et al.* 2019). Centromeres regions of the bilberry genome (*V. myrtillus*) were transferred to my lingonberry genomes using syntenic positions (Wu *et al.* 2021) to approximately map the centromere location.

2.2.7 Flavonoid biosynthesis gene expression in different tissues

Flavonoids are important berry components for both flavour and health effects. To better understand flavonoid synthesis in lingonberry, enzymes in the flavonoid biosynthesis pathway in lingonberry genome were identified and then quantified using RNAseq data. Because genes that code for enzymes in anthocyanin production could be of industry and evolutionary interest, I focused my analysis on 20 enzyme-coding genes directly involved in the flavonoid biosynthesis pathway in blueberry (Colle *et al.* 2019). I first identified gene orthology between lingonberry and other *Vaccinium* species using OrthoFinder (Emms & Kelly, 2019). When running OrthoFinder, four species were included: *V. macrocarpon* var. 'Stevens', *V. corymbosum* var. 'Draper', *V. vitis-idaea* var. 'Red Candy', and *R. williamsianum* as an outgroup. Note that the tetraploid 'Draper' protein sequences were kept as a full set preserving all four haplotypes to find a potential match in lingonberry for all 20 enzymes in the pathway. Orthofinder places genes into orthogroups representing orthology. Any lingonberry gene found in the same orthogroup as a blueberry flavonoid biosynthesis gene was classified as a putative lingonberry flavonoid biosynthesis gene.

Using the LC1 assembly and gene annotation file produced above as a reference, expression levels of the annotated transcripts/genes were estimated by Hisat2 using the -A, -G and -e option (Kim *et al.* 2019). The abundance estimate was reported in the units of FPKM, corresponding to fragments per kilobase of transcript per million mapped fragments (Zhao *et al.* 2021). FPKM is

the within-sample normalized value useful for expression levels in a single sample. To estimate the gene/transcript abundance per sample type, the RNAseq data from LC1 leaf, LC2 flower, rootstalk, berry, and the published green/white/red berries (Tian *et al.* 2020) were mapped to the LC1 reference genome.

2.2.8 Genomic divergence between subspecies

To calculate pairwise nucleotide divergence between the two lingonberry subspecies genomes, the 12 scaffolded chromosomes were aligned using minimap2 v 2.24-r1122 (Li, 2018, 2021) with the LW1 scaffolded genome as a reference and the LC1 scaffolded genome as a query (default parameters: -ax ams5 --cs=long). Following data format conversions (paftools.js sam2paf | view -f maf), the alignment file was filtered to remove duplicate alignments and the pairwise divergence was calculated per 10 kbp windows using maffilter v1.3.1 (Dutheil *et al.* 2014; parameters: Subset(remove_duplicates=yes, keep=no), MinBlockLength(min_length=1000), WindowSplit (preferred_size=10000, align=ragged_left), SequenceStatistics (Pairwise Divergence)). The program computes the number of base pair mismatches based on the alignment file and reports this value as the divergence in % mismatch in the specified window size. Additionally, to explore the presence of structural variations and basic sequence variations, Synteny and Rearrangement Identifier v1.5 (SyRI; Goel *et al.* 2019) was used on the aligned chromosomes, with default parameters.

2.2.9 Demographic history estimate

To investigate the past population history and the common ancestry between the two subspecies of lingonberry, I utilized the multiple sequentially Markovian coalescent model (MSMC2; Schiffels and Wang 2020) and the pairwise sequentially Markovian coalescent model (PSMC; Li and Durbin 2011). MSMC2 requires that the analyzed populations are mapped to the same reference genome. For the purpose of comparing the two methods in parallel, I chose to use LW1 as a reference genome for both subspecies because of better contiguity and base pair accuracy than

LC1. To first calculate the effective population size of each subspecies, the paired Illumina reads were mapped to the genome using BWA mem v0.7.17 (Li, 2013). PCR and optical duplicates were then removed using GATK Picard v2.23.2 'MarkDuplicates' function (Van der Auwera and O'Connor 2020). The resulting bam file and the genome were used as an input to identify both heterozygous variants and mask files separately for each chromosome per subspecies following bamCaller.py in MSMC2 v2.1.3 (Schiffels and Wang 2020). In brief, SNPs were first called using bcftools v1.16 (Danecek *et al.* 2021) with the command 'mpileup' and 'call' with the parameters: -q 20 -Q 20 -C 50 and -c -V indels, respectively. The results were then filtered and organized based on read coverage (mean coverage set to 38 for LW1, 37 for LC1; filtering applied is the minimum of x1/2 mean coverage to the maximum of x2 mean coverage). An additional mappability mask was generated to avoid calling variants from significantly repetitive regions using GenMap v1.3.0 (Pockrandt *et al.* 2020) with the parameter: -K 30 -E 2. For the PSMC inputs, SNPs were similarly called using bcftools 'mpileup' and 'call' with the same parameters as above, and the results were filtered with the minimum coverage of x1/3 and maximum of x2 mean coverage, as recommended (Li & Durbin, 2011). No repeats mappability mask was considered in the PSMC analysis. When running the models, a generation time of 5-10 years was chosen based on a prior experiment observing minimum of 8 years required to consider a seedling fully reproductive (Hjalmarsson and Ortiz 1998) and considering the natural age of first flowering (Ritchie 1955). However, given the potential for reproduction after first maturity, I recognize that this may underestimate the average reproductive age of the natural population. A mutation rate of 3×10^{9} substitutions per generation from *Arabidopsis thaliana* was used (Exposito-Alonso *et al.* 2018).

2.2.10 Genome-wide heterozygosity percentage

To discover any potential signatures of inbreeding, the pattern of heterozygous site distribution from the SNP data prepared from above was investigated. In brief, heterozygous sites were identified using vcftools –freq2 command (Danecek *et al.* 2011). The percent heterozygosity was then calculated by:

$$\%\ het = \frac{\#\ heterozygous\ calls}{\#\ callable\ sites} \times 100\%$$

per 100 kbp window across the chromosomes. Those sites that did not have minimum coverage or were too repetitive based on the mappability masks used in MSMC2 were excluded when calculating the percent heterozygosity. In order to look for statistical evidence of inbreeding, vcftools –het command was used (Danecek *et al.* 2011). Additionally, run of homozygosity (ROH) was computed using ROHan with parameters: t -25 --tstv 1.71 --rohmu 2e-5 for ssp. *minus* and -t 25 --tstv 1.69 --rohmu 2e-5 for ssp. *vitis-idaea* (Renaud *et al.* 2019). The –tstv or transition/transversion ratio was species specific, and was determined by bcftools stats (Danecek *et al.* 2021) and taking the average ratio among 12 chromosomes for each subspecies. The rest of the parameters were kept as default.

2.2.11 Phylogenetic tree construction

Phylogenetic trees were constructed using three different approaches. The first approach follows the default pipelines provided using OrthoFinder v2.5.4 (Emms & Kelly, 2019). In brief, a total of 10 species protein sequences in amino acid fasta format were collected from published studies: eight *Vaccinium* species: *Vaccinium vitis-idaea* from this study, *V. corymbosum* var. 'Draper' v1.0 first 12 chromosomes (Colle *et al.* 2019), *V. macrocarpon* var. 'Stevens' v1.0, *V. microcarpum* v1 (Diaz-Garcia *et al.* 2021), *V. oxycoccos* NJ96-20 v1 (Kawash *et al.* 2022), *V. myrtillus* NK2018_v1 (Wu *et al.* 2021), *V. darrowii* v1.2 (Cui *et al.* 2022), and *V. caesariense* W85-20 P0 v2 (Mengist *et al.* 2023). Kiwi fruit (*Actinidia chinensis* v3.0) genome and Azalea (*Rhododendron williamsianum*) genome were used as outgroups (Tang *et al.* 2019; Soza *et al.* 2019). The species tree was constructed based on the individual gene trees inferred from the orthologous gene groups as per OrthoFinder pipeline (Emms & Kelly, 2018; Emms & Kelly, 2017).

To confirm whether the species tree with OrthoFinder was correctly inferred, species tree based on only single-copy genes was built accordingly. The protein sequences from all the 249 single-

copy orthologs identified by OrthoFinder were taken and aligned by MAFFT v7.310 (Katoh and Standley 2013) for each gene separately. Then the aligned files were used to generate species tree in IQ-TREE v2.0.7 (Minh *et al.* 2020b) with default parameters. Gene concordance factors were also calculated based on the proportion of the single-copy gene trees supporting the resulting species tree topology (Minh *et al.* 2020a).

For further validation using conserved genes only, single-copy BUSCO genes were extracted and aligned to infer the species tree. To do this, BUSCO analysis was first performed on the genome assembly itself for each 10 species included, rather than the protein sequences used in the previous two approaches, with --lineage_dataset eudicots_odb10, --mode genome (Simão *et al.* 2015; Manni *et al.* 2021). Then the identified single-copy genes were aligned using MAFFT v7.310 and the individual gene trees were inferred with IQ-TREE v1.5.5 (Nguyen *et al.* 2015; parameters: -s and -nt 1). Outlier long branches were trimmed by TreeShrink v1.3.9 (Mai and Mirarab 2018) with default parameters. Finally, the species tree was constructed using the trimmed gene tress in Astral III v5.7.8 (Zhang *et al.* 2018). For this analysis, both lingonberry subspecies (*V. vitis-idaea* ssp. *vitis-idaea* and ssp. *minus*) could be included because the BUSCO analysis was performed on the assembly not necessarily the annotated protein sequences. Consequently, a total of 11 species were included in the final tree.

For visualization and data interpretation, all the species trees were exported in Newick format, and then viewed in FigTree. Trees were rooted manually to *Actinidia chinensis* based on the known oldest divergence time with *Vaccinium* genus (Kumar *et al.* 2017).

2.3 Results

2.3.1 Sequencing and assembly statistics

A summary of sequencing data generated from this study is found in Appendix A. Collectively, 35.3 Gbp (~50.0X) of clean long-read data was generated from MinION reads (read N50 = 20.56 kbp), and additional 12.42 Gbp (~37X) of short-read data was generated from Illumina for the commercial subspecies, LC1. The draft assembly solely from long-read data by SmartDenovo and three rounds of correction resulted in 616.251 Mbp assembly consisting of 1,358 contigs with N50 = 1.039 Mbp and per-base accuracy of 99.903%. With additional three rounds of correction using short-reads, the draft genome had 614.857 Mbp in total length and improved per-base accuracy of 99.959%. The corrected assembly was then analyzed for coverage and processed to remove potential duplicates due to heterozygosity or other organismal DNA contaminations with purge_haplotigs, resulting in 548.004 Mbp haploid representative genome assembly with BUSCO (Complete) score of 96.6% and contig N50 = 1.170 Mbp. Note that the amount of duplicate BUSCO hits was reduced by 1% (9.1 to 8.1%) in the last process. Similarly, 28.6 Gbp (~46.9X) of long-read data (read N50 = 23.16 kbp) and 10.9 Gbp (~35X) of short-read data were generated for the wild subspecies, LW1. The draft assembly after polishing had 545.497 Mbp of total assembly length (contig N50 = 1.309 Mbp, BUSCO (Complete) = 96.9%, per-base accuracy = 99.975%). After cleaning the assembly with purge_haplotigs, the size of the assembly was reduced to 518.642 Mbp with improved contig N50 (1.400 Mbp) and similar BUSCO (Complete) = 96.8% with slight decrease in duplicate hits.

Initially, scaffolding to the cranberry (*V. macrocarpon* var. ‘Stevens’) chromosome-level assembly was attempted because of its similar morphological features and the uncertainty of lingonberry’s phylogenetic position. However, scaffolding to the bilberry (*V. myrtillus*) genome generated a significantly more contiguous assembly for both LC1 and LW1, resulting in the total of 757 and 696 contigs, scaffold N50 of 43.867 Mbp and 42.799 Mbp, and 98.0% and 98.5% of the contigs

anchored to chromosomes. Based on the better contiguity and the fact that bilberry was more closely related to lingonberry in our subsequent phylogenetic analyses, the bilberry genome-based assembly was considered to better represent the structure of the lingonberry genome, and thus was used as a reference genome in the downstream analysis. These final assembly statistics are found in the table below and the detailed step-by-step stats are found in the supplementary file (Table 2.1, Supplementary Tables).

Table 2.1: Genome assembly statistics. *De novo* assembly was done by SmartDenovo (LC1) or Canu + SmartDenovo (LW1). Haploid only assembly (for a diploid genome) means heterozygous alleles are represented as a mixed haplotype from either of the homologous copy, but not both. The allelic sequences with less confidence were purged during assembly correction based on sequence coverage (Roach *et al.* 2018). The *de novo* assembled draft haploid genome was then scaffolded by mapping to bilberry (*Vaccinium myrtillus*) reference genome (Wu *et al.* 2021).

***V. vitis-idaea* ssp. *vitis-idaea* (LC1)**	*De novo* assembly	Haploid only	Scaffold assembly
Total length (Mbp)	614.857	548.004	548.071
Contig N50 (Mbp)	1.028	1.170	1.170
Scaffold N50 (Mbp)	-	-	43.867
#fragments/contigs	1358	757	757
#scaffolds	-	-	92
BUSCO (C%)	96.8	96.6	96.5
BUSCO (S%)	84.3	87.5	88.4
BUSCO (D%)	12.5	9.1	8.1
QV score	33.8254	-	-
Accuracy (1-error rate)	99.959%	-	-
Genome anchored to chr (%)	-	-	98.0
#genes annotated	-	-	27,243
Coding gene content (%)	-	-	7.59
TE content (%)	-	-	45.82
***V. vitis-idaea* ssp. *minus* (LW1)**	*De novo* assembly	Haploid only	Scaffold assembly
Total length (Mbp)	545.497	518.642	518.704
Contig N50 (Mbp)	1.309	1.400	1.400
Scaffold N50 (Mbp)	-	-	42.799
#fragments/contigs	1030	696	696
#scaffolds	-	-	76
BUSCO (C%)	96.9	96.8	96.9
BUSCO (S%)	89	89.7	90.5
BUSCO (D%)	7.9	7.1	6.4
QV score	35.9577	-	-

Accuracy (1-error rate)	99.975%	-	-
Genome anchored to chr (%)	-	-	98.5
#genes annotated	-	-	NA
Gene content (%)	-	-	NA
TE content (%)	-	-	NA

RNA sequence data was produced from leaf sample (~7.8 Gbp), rhizome (~6.9 Gbp), flower (~11.4 Gbp), and berry (~11.7 Gbp) samples in the commercial subspecies. In addition, lingonberry fruit transcripts data from published work was added to our analysis (~22.5 Gbp; Tian et al., 2020). The overall alignment rate for the combined transcripts was 96.16%. With the alignment of RNA reads to the LC1 reference genome, a total of 27,243 genes were annotated (BUSCO (C): 91.4%), 25,796 of which were mapped onto chromosomes, covering 39.63% of the genome. Excluding non-coding sequences (introns, untranslated regions, etc.), the coding sequence content was 7.59% across the genome, with an average length of 237 bp. Transposable elements were also discovered using multiple independent programs, excluding regions with coding sequences already annotated, leading to cover about 45.82% of the genome overall (Table 2.1).

I found that TE density was fairly even across the genome but there tended to be less genes around the centre of the chromosomes and more on the chromosome arms (Figure 2.1, Table 2.2). When plotting the TE distributions by different types (Figure 2.2), some differences in density across the chromosome were observed. For example, long-terminal-repeat retrotransposons (LTRs) seem to mirror the distribution of the gene density, having higher density around the centre of the chromosomes, whereas the hAT-TIR or a type of terminal inverted repeat has a contrasting distribution that follows the gene density distribution quite closely (Figure 2.2).

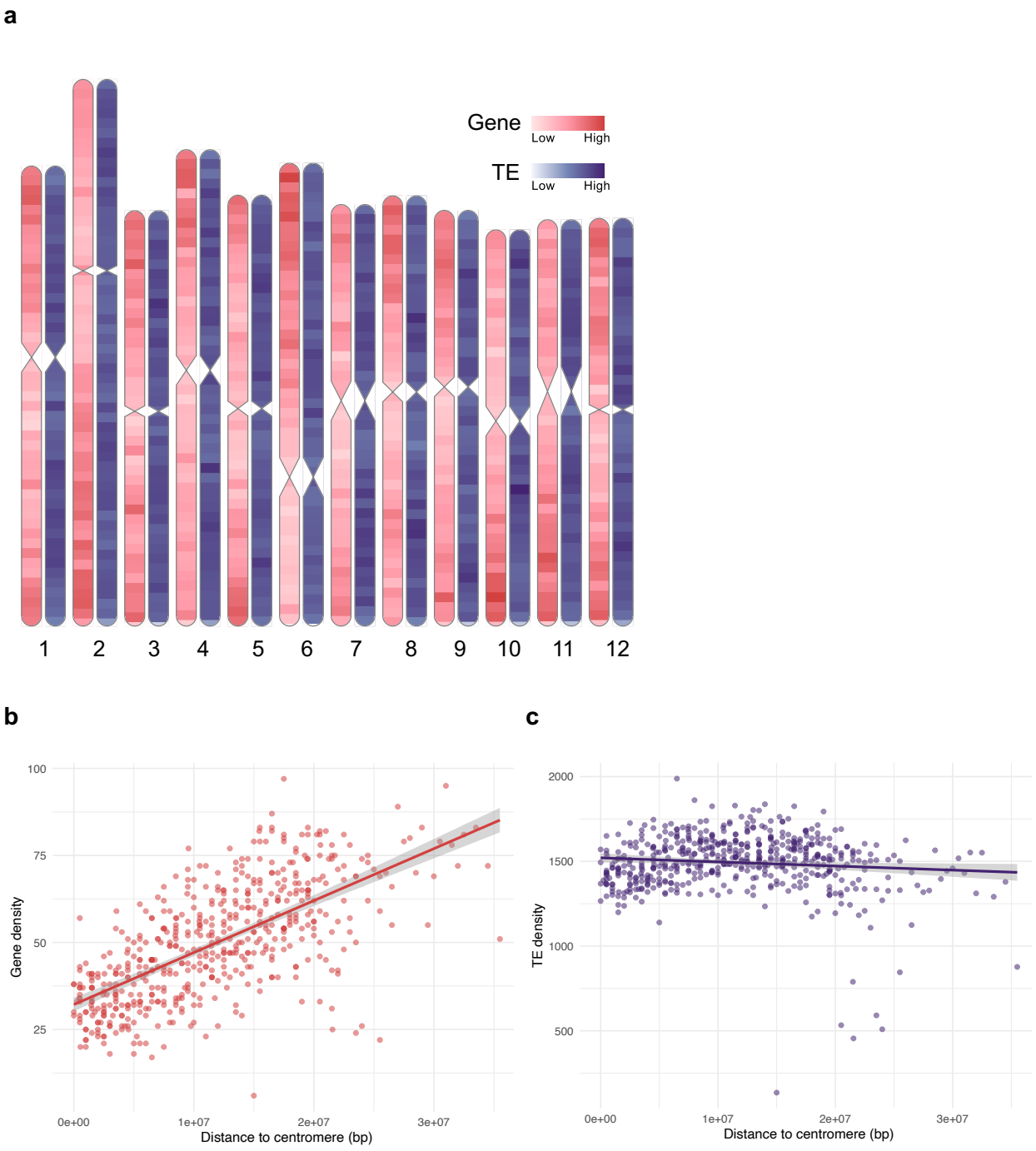

Figure 2.1: a) Gene and TE distributions in lingonberry genome (*Vaccinium vitis-idaea* ssp. *vitis-idaea*, var. 'Red Candy'). b) Gene and c) TE densities by distance from centromeres. Centromere positions are approximately mapped from bilberry genome as a range, and distance was calculated to its middle value (Wu *et al.* 2021). Red shades indicate the gene density and purple shades indicate the transposable element (TE) density. Genes were filtered to represent only the longest gene in case of isoforms and splicing variants present. All densities are presented as the number of feature counts per 1 Mbp window.

Table 2.2: Chromosome lengths and putative centromere positions on lingonberry genome. Centromere positions are putatively assigned based on the bilberry genome (Wu *et al.* 2021).

Chromosome on LC1	Length (bp)	Centromere start	Centromere end
Chr01	46899524	18000000	21000000
Chr02	55732239	19000000	20000000
Chr03	42352093	19900000	21000000
Chr04	48573227	21000000	24000000
Chr05	43929065	21000000	22500000
Chr06	47187867	30000000	34000000
Chr07	42979545	18000000	22000000
Chr08	43866791	19000000	21000000
Chr09	42388640	17000000	19000000
Chr10	40397952	18000000	21000000
Chr11	41428829	15000000	20000000
Chr12	41566430	19000000	20000000

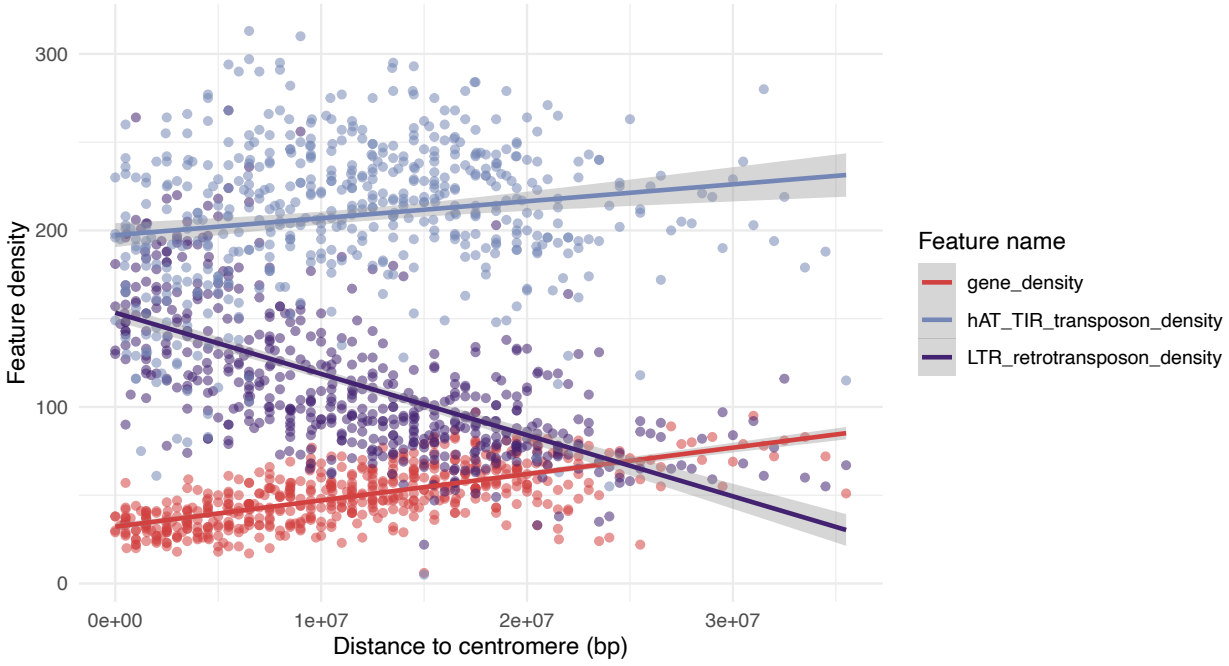

Figure 2.2: Scatter plots of long-terminal-repeat (LTR) and hAT terminal inverted repeat (hAT TIR) densities in comparison to gene density against the distance from the centre of chromosomes. Red points indicate gene density, light blue indicates hAT-TIR transposon density, and purple indicates LTR retrotransposon density. All densities are presented as the number of feature counts per 1 Mbp window.

2.3.2 The flavonoid biosynthesis pathway in lingonberry

Gene expression levels for lingonberry genes orthologous to blueberry flavonoid biosynthesis genes (Colle *et al.* 2019) were quantified in berries, flower, rhizome, and leaf tissue samples (Figure 2.3). Those enzymes include phenylalanine ammonia-lyase (PAL), 4-hydroxycinnamoyl-CoA ligase (4CL), cinnamate 4-hydroxylase (C4H), hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT), 4-coumaric acid 3'-hydroxylase (C3H), hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase (HQT), chalcone synthase (CHS), chalcone-flavanone isomerase (CHI), flavanone 3-beta-hydroxylase (FHT), flavonol synthase (FLS), flavonoid 3'hydroxylase (F3'H), flavonoid 3',5'-hydroxylase (F3'5'H), dihydroflavonol reductase (DFR), leucoanthocyanidin reductase (LAR), anthocyanidin synthase (ANS), anthocyanidin reductase (ANR), UDP glucose:flavonoid 3-O-glucosyl transferase (UFGT), anthocyanin O-methyltransferase (OMT), transparent testa 12 (TT12), and transparent testa 19 (TT19). The proposed enzymatic pathway is shown in Figure 2.3.

For all 20 genes identified in *V. corymbosum* var. 'Draper' that were described to be associated with flavonoid biosynthesis pathway I found at least one ortholog in lingonberry. These orthologs were identified based on the orthogroup classification using OrthoFinder (Emms and Kelly 2019). Many genes seem to exist in more than one copy; however, C3H, HQT, CHI, F3'H, DFR, ANS, and ANR have only one copy each. Although many of the flavonoid related genes were expected to be highly expressed in berries compared to other tissue types, rootstalk and leaf expressed C4H, HCT, HQT, CHI, FHT, F3'H, LAR, and ANS at much higher levels than the berry samples. Some genes with multiple copies seem to possess specific tissue localizations. For example, the CHS and 4CL genes were found to exist in three copies, where one is highly expressed in berries (increasing as it ripens) while one is almost completely not expressed and the other is highly expressed in rhizome and leaf. Others like PAL, HCT, F3'5'H, and OMT seem to be expressed only one of the gene copies at high level and the rest are very minimal. Notably, genes unique to different tissue types include: TT19 is very highly expressed (873 FPKM) in only the red berry;

C3H is almost close to zero except in flower (5 FPKM); FLS has one of the copies expressed highly in rhizome (58 FPKM) and the other copy highly expressed in flower (66 FPKM); HCT is only highly expressed in rootstalk (136 FPKM).

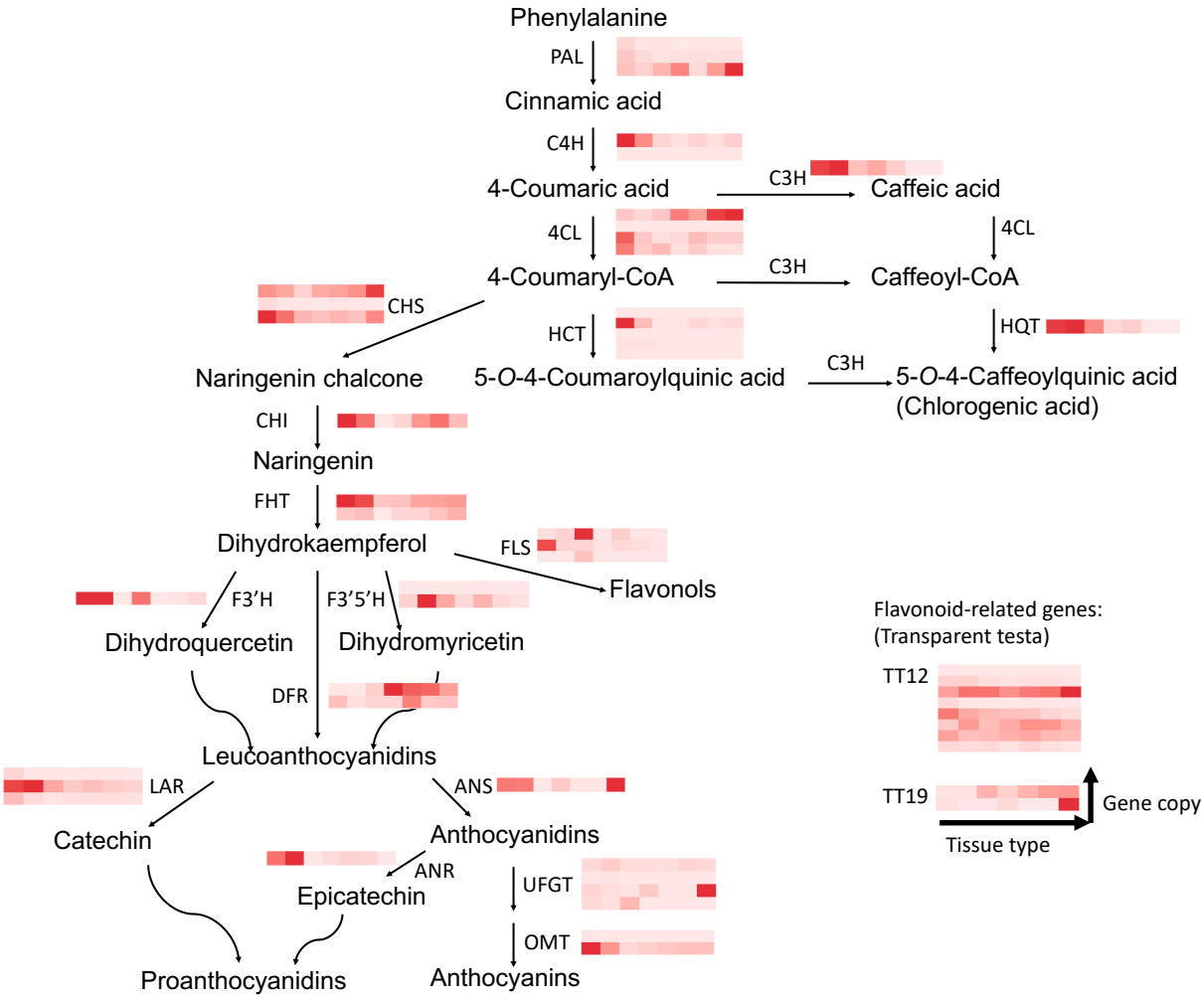

Figure 2.3: Heatmap of gene abundance related to flavonoid biosynthesis. Columns represent sample type and rows represent gene IDs on lingonberry genome. Samples (from left to right) were taken from *Vaccinium vitis-idaea* var. 'Red Candy' root, leaf, flower, and berry, as well as published transcripts data from var. 'Sunna' at different ripening stages; green berry, white berry, and red berry (Tian, et al. 2020). Abundance is measured by fragments per kilobase of transcript per million mapped fragments (FPKM). Note that the red colour gradient is normalized within each heatmap, so quantitative comparison cannot be made across heatmaps. The enzyme pathway is based on Colle *et al.* (2019).

To visually compare how much expression level differs between samples and across enzymes within the pathway, the log scaled FPKM was plotted in a heatmap (Figure 2.4). A few genes such as PAL, CHS, and ANS were expressed at much higher levels (1000+ folds) than others.

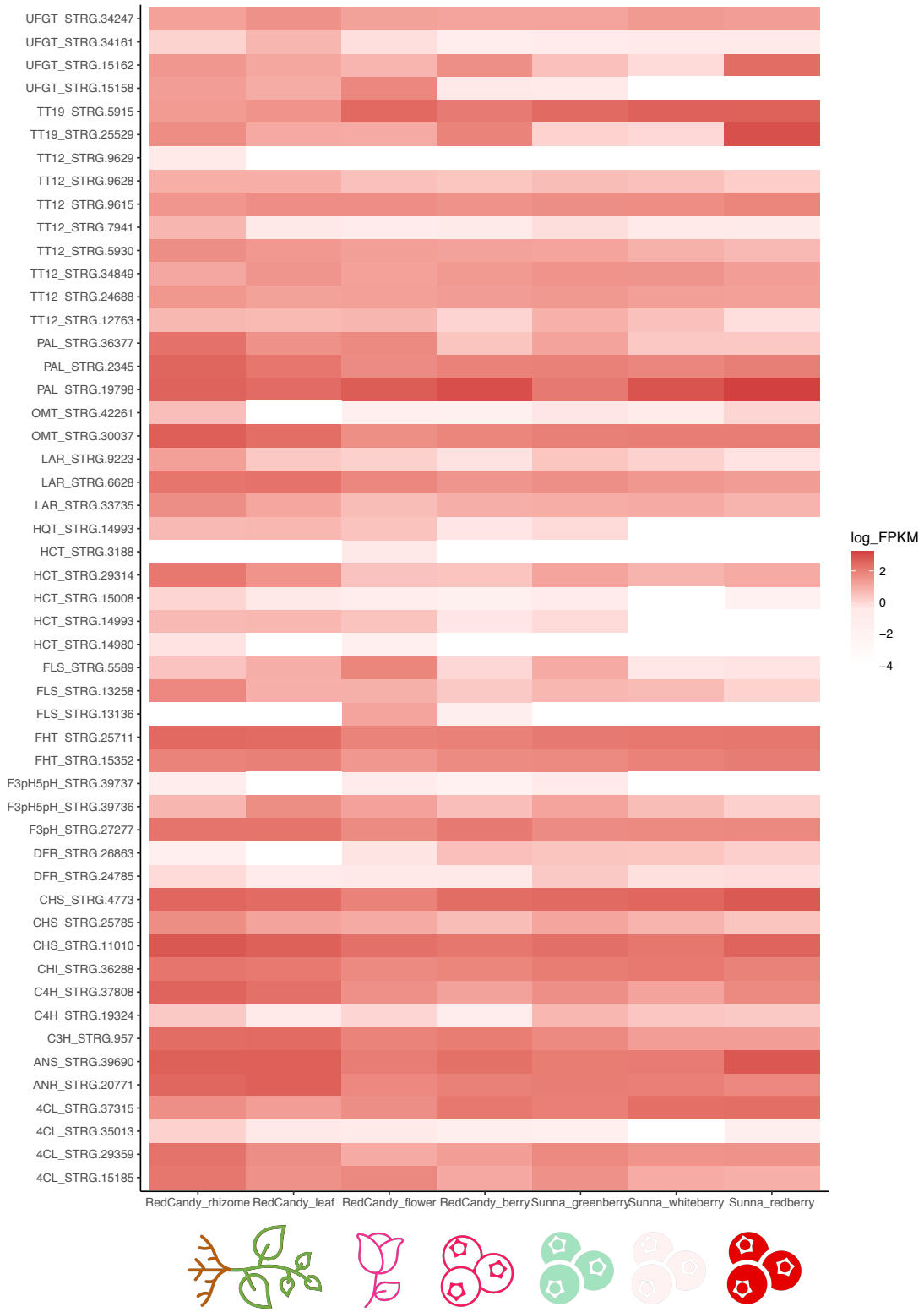

Figure 2.4: Heatmap of flavonoid biosynthesis related gene abundance in lingonberry. Gene abundance was measured in the unit of FPKM, and the values are log scaled for visualization purpose, where red indicates the most expressed and green indicates the least. Rows represent different copies of each orthologous gene in lingonberry (enzyme name_STRG-id), and columns are sample types.

2.3.3 Divergence between two subspecies genomes

To infer quantitative genomic divergence between the two lingonberry subspecies, the chromosomes from LC1 and LW1 assemblies were aligned and analyzed for sequence divergence and structural variations. The average pairwise divergence between the two subspecies genomes was 0.86% and the sequence divergence was uniformly scattered across 12 chromosomes (Figure 2.5). The most divergent window had a percent mismatch of 5.63% on chromosome 2, but no other windows had higher than 5% divergence (Figure 2.6).

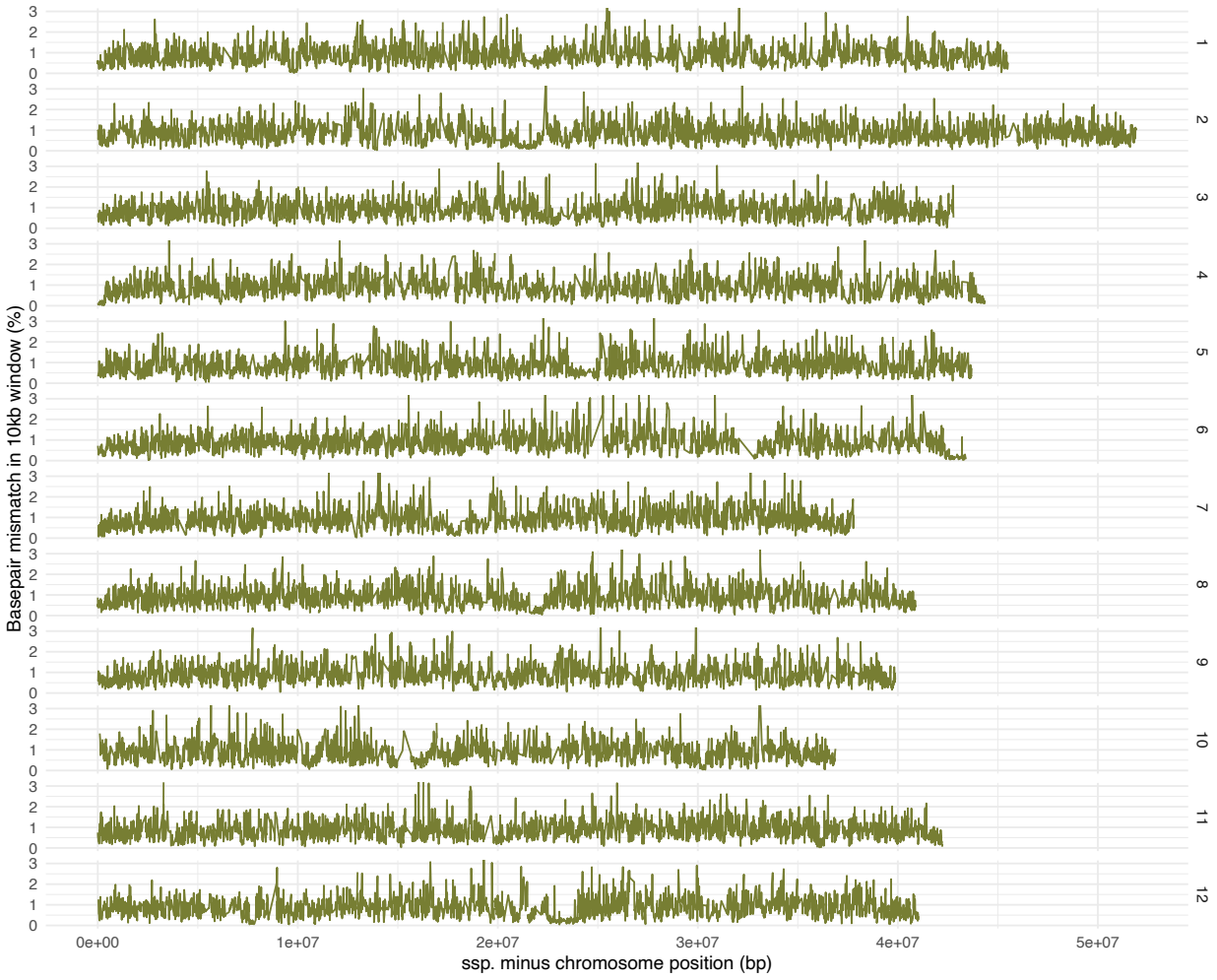

Figure 2.5: Pairwise divergence between *Vaccinium vitis-idaea* ssp. *minus* (LW1) and ssp. *vitis-idaea* (LC1). The percentage of base pair mismatch was determined based on the aligned sequences in 10 kbp windows. Any alignments <1000 bp was filtered out before plotting. The reference genome was set to ssp. *minus*. Rows are sorted by chromosomes as labelled on the right.

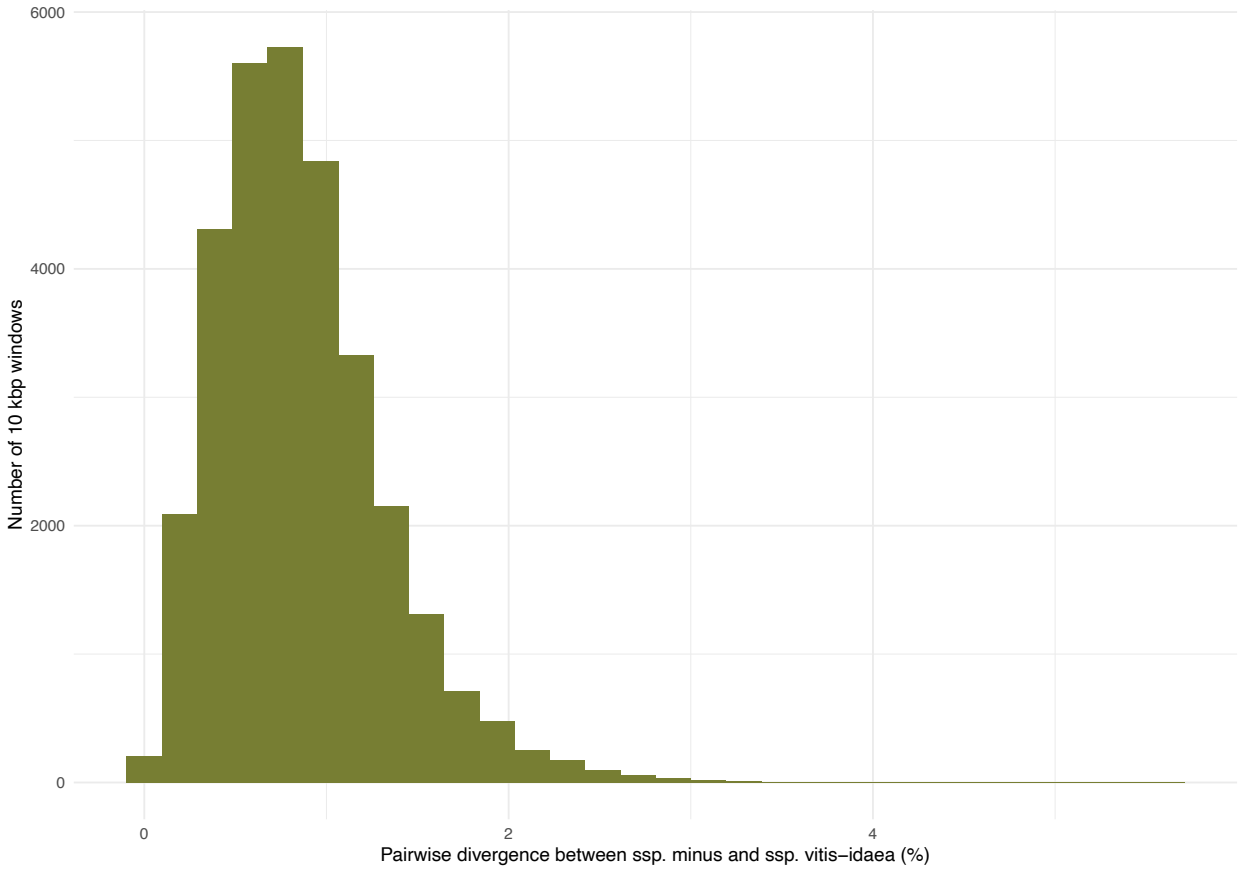

Figure 2.6: Histogram of pairwise divergence (%) between lingonberry subspecies. Chromosomes from *Vaccinium vitis-idaea* ssp. *minus* genome (LW1) and *V. vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy' (LC1) genome were aligned and windowed into 10 kbp to compute pairwise sequence divergence in % of bp mismatch. Any alignments <1000 bp was filtered out before plotting.

For structural and sequence-level variations other than the base pair mismatch, SyRI was able to detect 55.3% and 51.1% of genome in synteny while 12.6% and 10.8% were highly diverged sequences on LW1 and LC1 assembly, respectively (Table 2.3, 2.4). Numerous small translocations, inversions, and duplication were seen; in particular, chromosome 9 was enriched with translocations and chromosome 5 with many different variants (Figure 2.7).

Table 2.3: Detected structural variations between lingonberry subspecies. Reference genome was set as *Vaccinium vitis-idaea* ssp. *minus* and the query genome was *V. vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy'. Variations were detected by SyRI on the 12 chromosomes only and unplaced contigs were not considered into analysis.

Structural variation type	Count	Length LW1 (bp)	Genomic proportion LW1	Length LC1 (bp)	Genomic proportion LC1
Syntenic regions	10836	282579977	55.3%	274823289	51.1%
Inversions	217	9184582	1.8%	8521831	1.6%
Translocations	8171	63007344	12.3%	63820410	11.9%
Duplications (LW1)	1654	14209243	2.8%	NA	NA
Duplications (LC1)	10279	NA	NA	33875804	6.3%
Not aligned (LW1)	16370	165867888	32.5%	NA	NA
Not aligned (LC1)	25429	NA	NA	158473390	29.5%

Table 2.4: Detected sequence level variations between lingonberry subspecies. Reference genome was set as *Vaccinium vitis-idaea* ssp. *minus* and the query genome was *V. vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy'. Variations were detected by SyRI on the 12 chromosomes only and unplaced contigs were not considered into analysis.

Sequence variation type	Count	Length LW1 (bp)	Genomic proportion LW1	Length LC1 (bp)	Genomic proportion LC1
SNPs	2756611	2756611	0.5%	2756611	0.5%
Insertions	282583	NA	NA	6284926	1.2%
Deletions	279127	6983472	1.4%	NA	NA
Copy gains	447	NA	NA	2175958	0.4%
Copy loses	710	2673598	0.5%	NA	NA
Highly diverged	17940	64258577	12.6%	57777646	10.8%
Tandem repeats	39	80644	0.0%	121033	0.0%

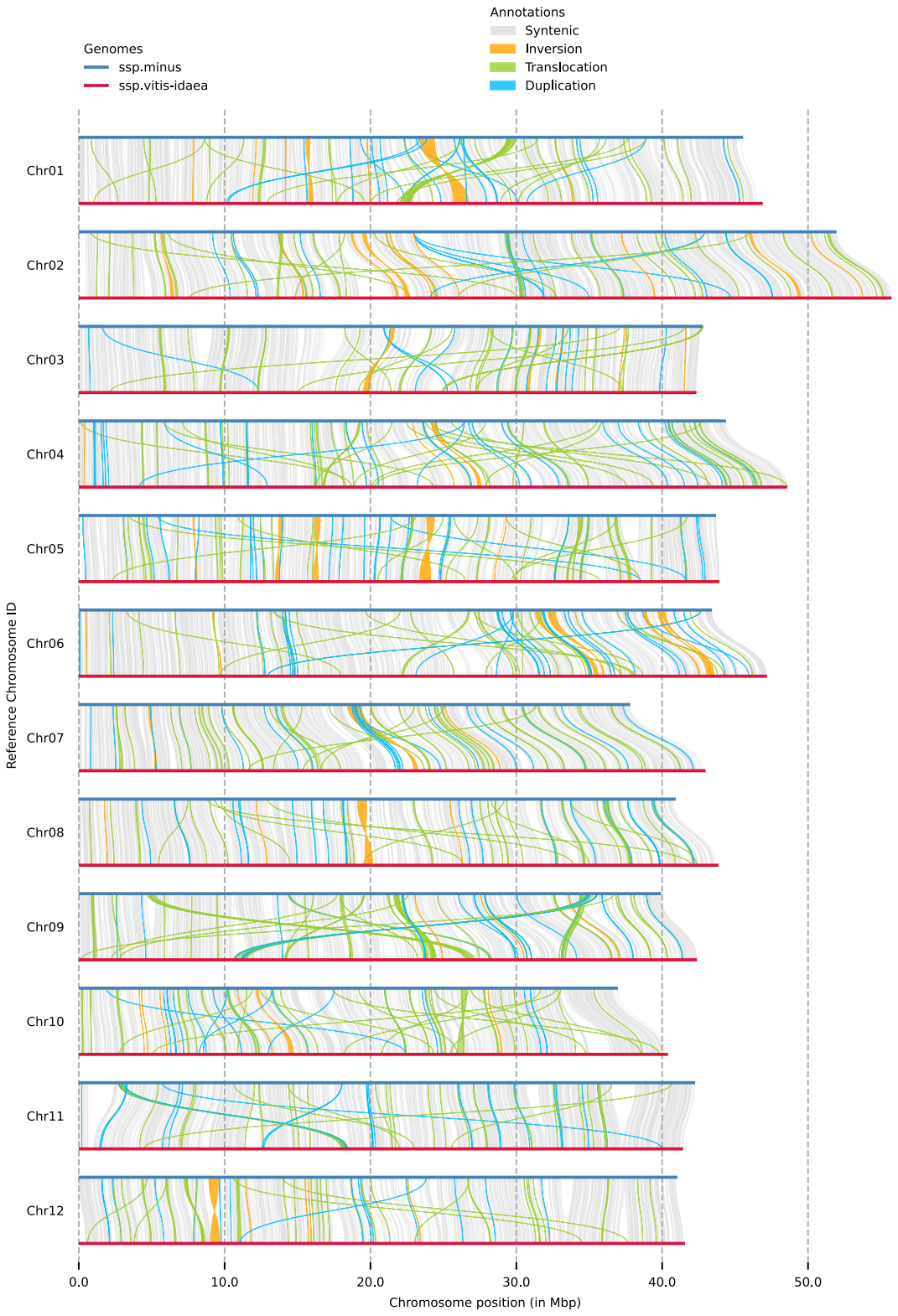

Figure 2.7: Alignment between two lingonberry subspecies. Horizontal lines represent chromosomes on each genome: blue is *V. vitis-idaea* ssp. *minus* (LW1), red is *V. vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy' (LC1). Structural variations are shaded in colours: grey for syntenic region; orange, inversion; green, translocation; light blue, duplication. Plots are made with plotsr (Goel and Schneeberger 2022).

2.3.4 Historical population size and origin of lingonberry

MSMC and PSMC are two population models that use the distribution pattern of heterozygous sites within an individual's genome to infer a coalescent history, including past effective population sizes (Ne). While PSMC is strictly a pairwise comparison, meaning it works only with a single diploid individual, MSMC can extend to more than two haplotypes (Schiffels and Wang 2020). In addition, MSMC provides a cross coalescent model to estimate when the two distinct populations coalesce back to the same population. This would indicate the approximate divergence time between the two compared populations.

Surprisingly, both PSMC and MSMC2 estimated an ongoing population bottleneck for both LC1 and LW1 populations (Figure 2.8, 2.9). Using a generation time estimate range of 5–10 years, LC1 and LW1 began declining in Ne around 0.8–1.7 MYA and 1.5–3.2 MYA. The minimum and maximum Ne observed for LW1 and LC1 was 35,148 at 0–0.156 MYA and 62,993 at 0–0.248 MYA, and 592,896 at 1.596–3.191 MYA and 636,105 at 0.858–1.716 MYA, respectively (Table 2.5). Plots shown below are models with generation time of 5 years as an example since modifying generation time to 10 years resulted in a time shift of ~0.8 MYA later without any change to the Ne. The cross coalescent time analysis with MSMC2 failed to produce a complete collapse of recombination patterns between the two subspecies, as indicated by the relative cross-coalescent rate not reaching 1 (Figure 2.10). However, the highest cross-coalescent rate corresponded to the time where the two populations seemed to merge in their Ne plots, (Figure 2.8, 2.9) around 10 MYA (Figure 2.10).

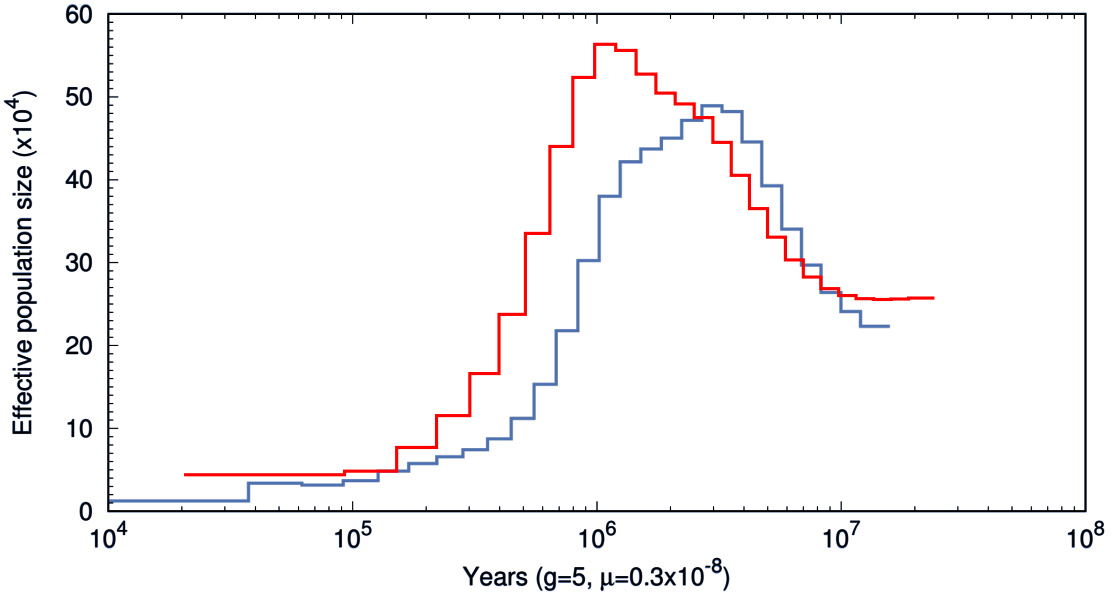

Figure 2.8: Past effective population size of lingonberry with PSMC. Effective population size (Ne) of a) *Vaccinium vitis*-idaea ssp. *minus* (blue; LW1) and b) *V. vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy' (red; LC1). Plots are generated with the generation time of 5 years and mutation rate of 3 x 10^{9} mutations/generation.

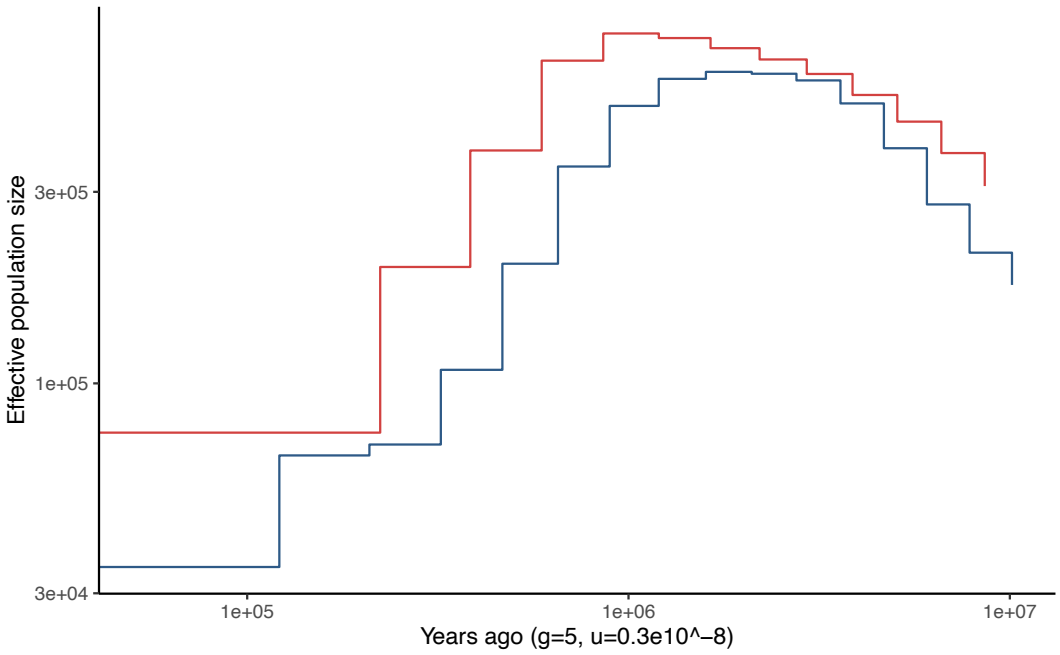

Figure 2.9: Past effective population size of lingonberry with MSMC2. Log-scaled effective population size (Ne) of *Vaccinium vitis*-idaea ssp. *minus* (blue; LW1) and *V. vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy' (red; LC1). Plots are generated to show up to 10 MYA, with the generation time of 5 years and mutation rate of 3 x 10^{9} mutations/generation.

Table 2.5: Maximum and minimum effective population sizes (Ne) and timing estimated with MSMC2. Mutation rate of $3x10^9$ substitutions per generation was based on *Arabidopsis thaliana* mutation rate estimate (Exposito-Alonso *et al.* 2018). The minimum generation time was determined based on the time it required for a lingonberry seed to grow into maturity (i.e. fully reproductive) in a field experiment (Hjalmarsson and Ortiz 1998).

Subspecies	Ne	Time (MYA)	generation time (years)	mutation rate (subs. per gen.)
ssp. *minus*	max. 592,896	1.596–3.191	5–10	3×10^9
	min. 35,148	0–0.156	5–10	3×10^9
ssp. *vitis-idaea*	max. 636,105	0.858–1.716	5–10	3×10^9
	min. 62,993	0–0.248	5–10	3×10^9

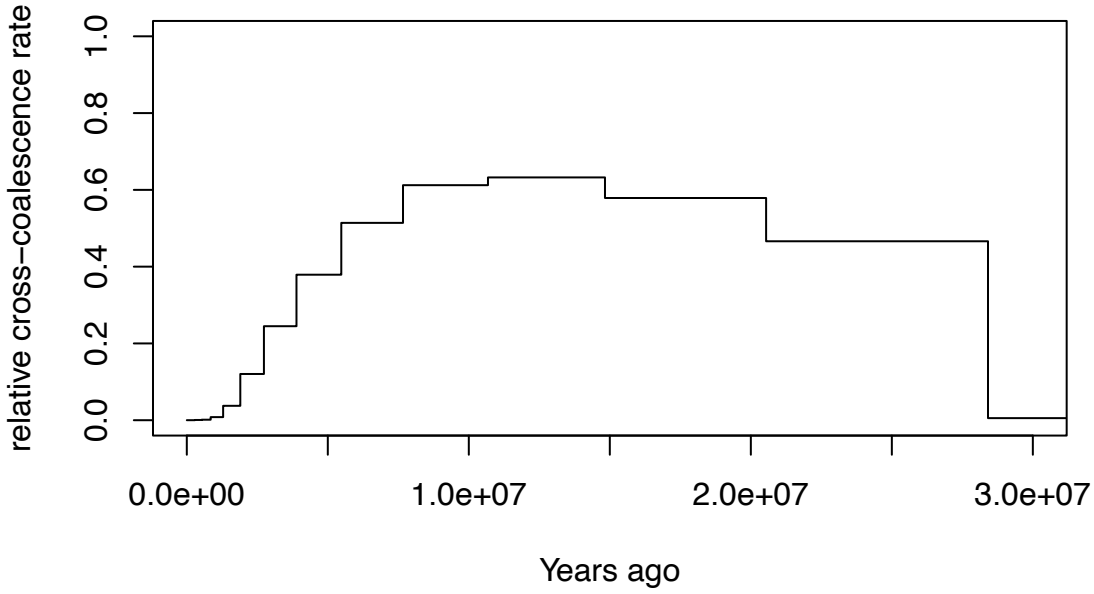

Figure 2.10: Cross-coalescence rate between lingonberry subspecies with MSMC. Cross-coalescence rate of 0 would indicate a completely separate, independent population, whereas 1 would indicate the compared two populations have the same recombination pattern meaning they share the common ancestry.

To test what could have led to the dramatic decline in the present-day effective population size as seen on those models, percent heterozygosity across the genome was plotted. The genome-wide heterozygosity percentage in LW1 was slightly lower than LC1 throughout the genome by 0.1–0.2% (mean %het: 0.24, 0.38, respectively). However, there was no evidence of recent inbreeding (inbreeding coefficient of 0 for both subspecies). Furthermore, the analysis of vcf files

with ROHan showed that only 0.0% or 0.3% was in a run of homozygosity for ssp. *minus* and ssp. *vitis-idaea*, respectively (Figure 2.11, 2.12).

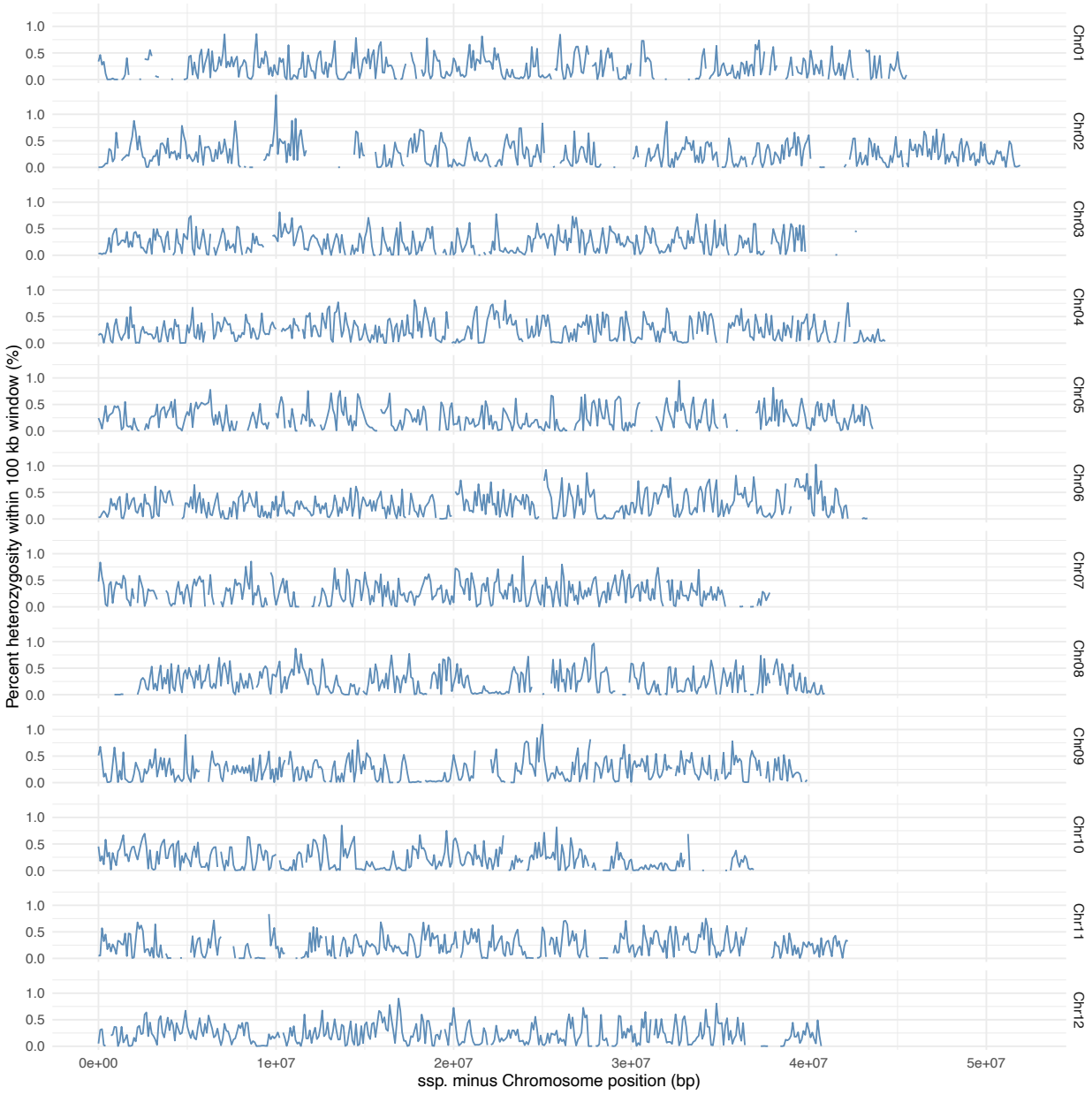

Figure 2.11: Percent heterozygosity across the *V. vitis-idaea* ssp. *minus* genome (LW1). Percent heterozygosity is calculated as the number of heterozygous calls detected divided by the number of total callable sites per 100 kbp window. Regions with less than 1000 callable sites were excluded from plotting. The absence of a datapoint/line thus indicates a region of missing data.

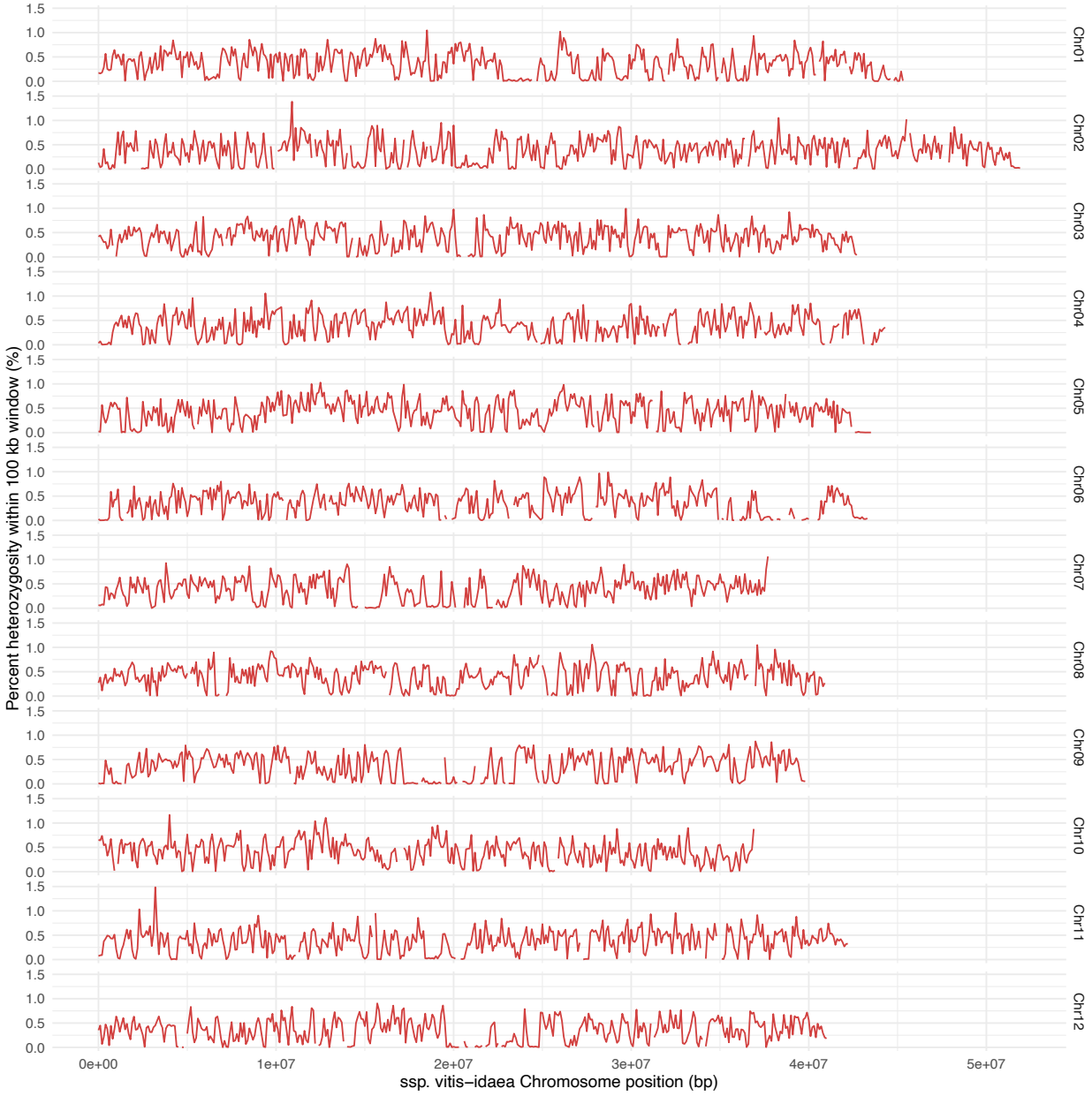

Figure 2.12: Percent heterozygosity across the *V. vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy' genome (LC1). Percent heterozygosity is calculated as the number of heterozygous calls detected divided by the number of total callable sites per 100 kbp window. Regions with less than 1000 callable sites were excluded from plotting. The absence of a datapoint/line thus indicates the region of missing data.

2.3.5 *Vaccinium* phylogenetics: species tree and gene trees

The protein sequence alignment across seven published *Vaccinium* species and two outgroup species with the *V. vitis-idaea* protein sequences generated in this study resulted in the total number of 351,892 genes analyzed, of which 323,463 were categorized into 30,569 orthogroups

by OrthoFinder (Emms and Kelly 2019). *Actinidia chinensis* protein and *Rhododendron williamsianum* protein sequences were used as outgroups based on previous phylogenetic studies with *Vaccinium* (Schlautman *et al.* 2017; Diaz-Garcia *et al.* 2021; Kawash *et al.* 2022). The mean orthogroup size was 10.6 genes and 6,021 orthogroups were shared by all the species, of which 249 were single-copy orthogroups. The species tree generated using a consensus method (STAG; Emms and Kelly 2018) produced a topology somewhat in agreement with previous studies (Figure 1.3b, c, 2.13). I found clades for cranberries (*V. microcarpum*, *V. oxycoccos*, *V. macrocarpon*) and blueberries (*V. darrowii*, *V. caesariense*, *V. corymbosum*), while bilberry (*V. myrtillus*) was identified as the closest relative of lingonberry (*V. vitis-idaea*). Gene concordance values, however, were generally low especially among species in the blueberry, bilberry, and lingonberry (ranging from 0.17–0.23). The node splitting lingonberry and bilberry from other blueberries had only 17.3% of gene trees supporting the topology.

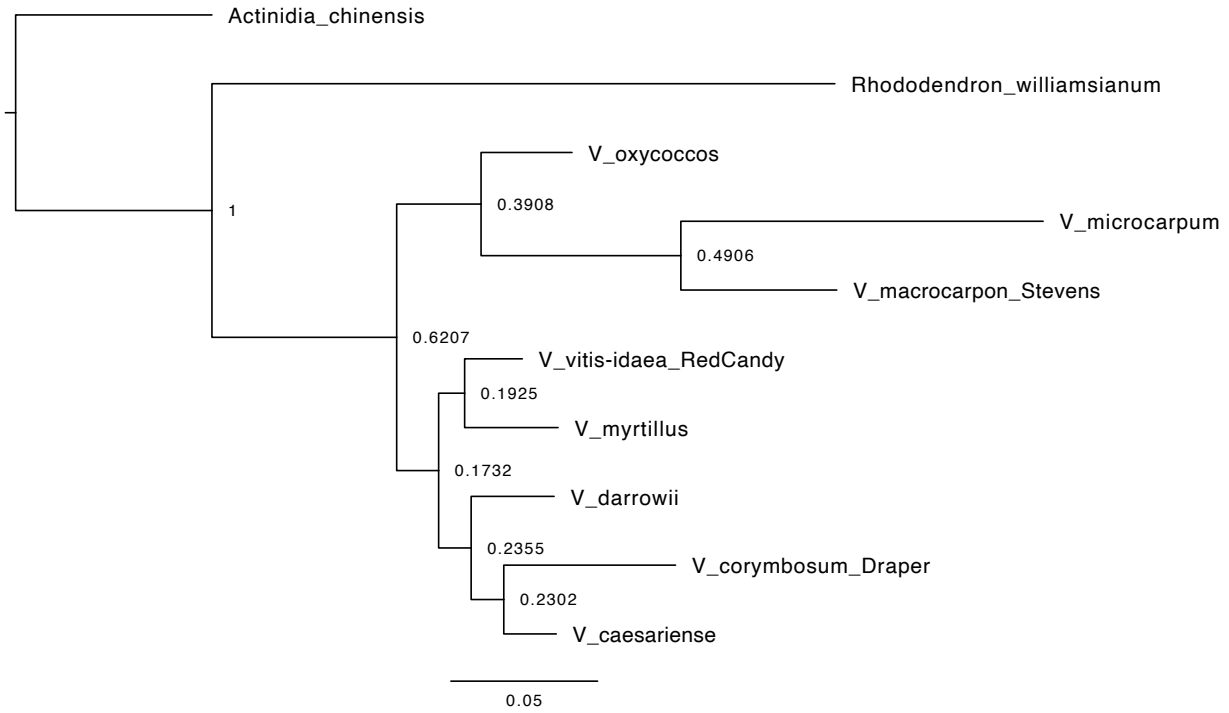

Figure 2.13: Phylogeny of *Vaccinium* species with OrthoFinder. Phylogram of genus *Vaccinium* was constructed based on amino acid sequences of orthologous protein data. Species tree was inferred by a consensus tree of all the input gene trees (single-copy and multi-copy genes) using STAG (Emms and Kelly 2018). A total of eight different *Vaccinium* species were included, with *Actinidia chinensis* (kiwi fruit) and *Rhododendron williamsianum* (azalea) as the outgroups. Numbers at the node indicate gene concordance value (0–1).

To further confirm the above phylogenetic relationship, a total of 249 strictly single-copy orthologues sequences identified using OrthoFinder were extracted and aligned individually. Then the maximum likelihood species tree was constructed using a concatenation method (Minh *et al.* 2020b; Figure 2.14). For this, bootstrap values were calculated at each node using 1000 replicates, and it was found to be mostly close to 100 except one node involving blueberry species scoring only 64. In fact, the only discrepancy between this tree and the previous tree was this blueberry group where one of the wild progenitors of commercial blueberry species (*V. caesariense*) and the primary haplotype of commercial blueberry (*V. corymbosum*) were more closely related than the other wild progenitor (*V. darrowii*) in the OrthoFinder tree. The gene concordance factors were comparable or slightly higher for this concatenated species tree than the OrthoFinder tree.

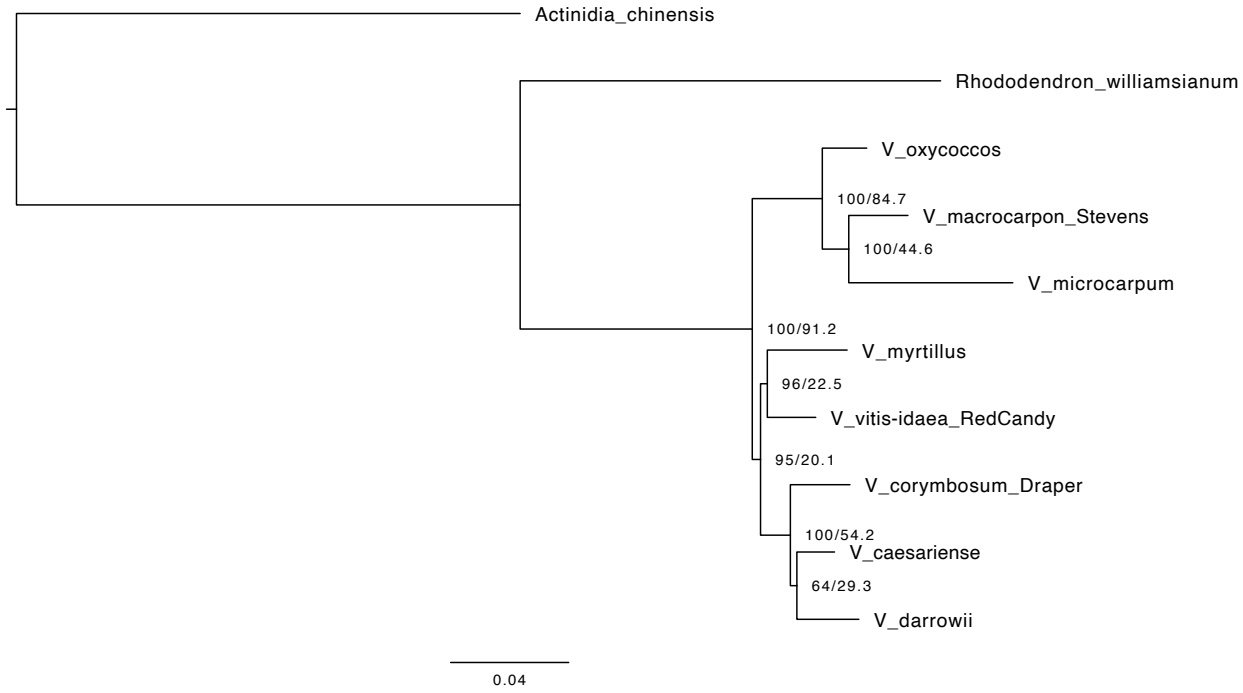

Figure 2.14: Phylogeny of *Vaccinium* species based on single-copy orthologues. Multiple sequence alignment was performed on single-copy orthologues using MAFFT, then the maximum likelihood species tree and gene trees were built using IQ-TREE2. Nodes are labelled with bootstrap values (1000 replicates)/gene concordance factor (0–100).

As a third approach, 2,226 conserved BUSCO genes were extracted and aligned individually to estimate the species tree (Figure 2.15). For this approach, ASTRAL was used to infer a species tree, which takes individual gene trees as input and finds a species topology that reconciles the variation in the gene tree topology (Sayyari and Mirarab 2016; Zhang *et al.* 2018). The results showed that the general classifications (cranberries and blueberries) were well supported by gene trees, indicated by the node splitting cranberries apart from the rest of the species having 91.67% gene concordance factor, as well as 83.32% for cranberry clustering, and 60.33% for blueberry clustering. The two lingonberry subspecies clustered together in the tree as expected, with a gene concordance factor of 72.79%. However, roughly 10% of the gene trees supported alternate topology where *V. vitis-idaea* ssp. *minus* was more closely related to bilberry (*V. myrtillus*) than *V. vitis-idaea* ssp. *vitis-idaea* or vice versa (data not shown). Some low support basal nodes were also apparent including the node splitting blueberries with lingonberry and bilberry, and bilberry with lingonberry (gene concordance factor of 26.65% and 24.54%, respectively; >50% of gene trees supporting alternate branching patterns).

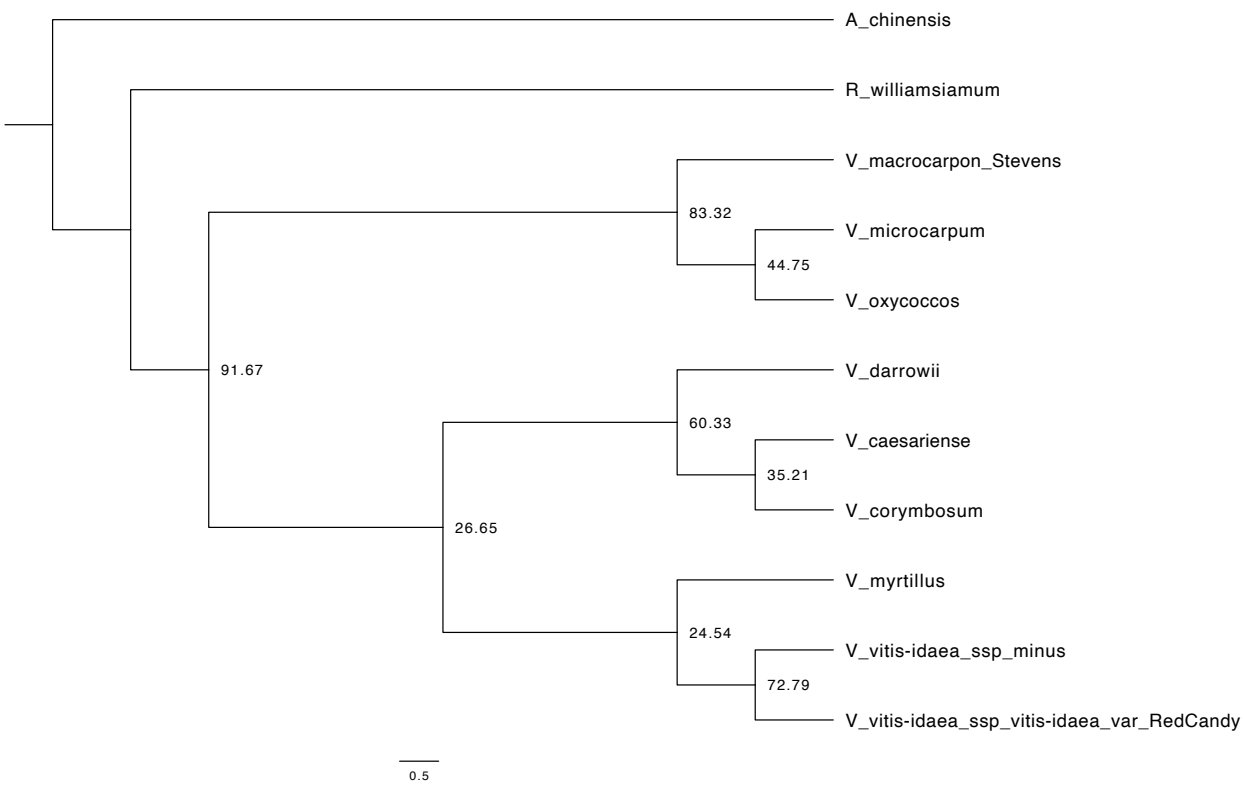

Figure 2.15: Phylogeny of *Vaccinium* species using BUSCO genes. All the single-copy BUSCO genes found from running BUSCO analysis on each genome assembly (--genome, eudicots_odb10) were extracted. Only genes with more than three species aligned were retained. Multiple sequence alignment was performed using MAFFT, then the individual gene trees were built using IQ-TREE, and the most concordant species tree based on all gene trees was built with Astral III. Nodes are labelled with gene concordance factor (0–100). The tree branches were proportionally transformed for visualization in FigTree.

The overall topology of different berry groups observed was consistent across three methods. I find that lingonberry is most closely related to bilberry and is more closely related to blueberry species than cranberry species.

2.4 Discussion

2.4.1 Genome assembly and annotation

2.4.1.1 Draft genome assembly quality

Using a hybrid approach combining the long-read sequencing (ONT) and short-read sequencing (Illumina) data, the two lingonberry subspecies genomes (*Vaccinium vitis-idaea* ssp. *vitis-idaea* var. 'Red Candy'; LC1, and *Vaccinium vitis-idaea* ssp. *minus*; LW1) were *de novo* assembled into 757 contigs of the total length 548.004 Mbp (contig N50 = 1.170 Mbp, BUSCO (Complete) = 96.6%), and 696 contigs of the total length 518.642 Mbp (contig N50 = 1.400 Mbp, BUSCO (Complete) = 96.8%), respectively. The assembled genome size was consistent with the previous flow cytometry study predicting ~550 Mbp (Redpath *et al.* 2022). The latter wild subspecies assembly, LW1, was slightly smaller in size than expected from flow cytometry data but considering the high BUSCO score and the difference in subspecies designation, my assemblies are both reasonably complete. Compared to the short-read only assemblies which generally do not reach N50 of 1 Mbp, my ONT-based assemblies are significantly more contiguous (Rhie *et al.* 2021), and my assembly statistics are comparable to many draft genome assemblies of similar size (e.g., Hamilton, Vaillancourt, Wood, & Buell, 2023; Marrano et al., 2020; C. Wu et al., 2021; Y. Zhang et al., 2023). However, recently published reference-quality genomes are often more contiguous at the contig-level and are further *de novo* scaffolded into chromosomes using physical linkage approaches like Hi-C and optical mapping (e.g., contig N50 = 29.5 Mbp by Nakandala *et al.*, 2023; contig N50 = 15.4 Mbp by Wu *et al.*, 2023; contig N50 = 7.57 Mbp by Zhang *et al*., 2023). Moreover, in well-studied crops like blueberry and sweet orange, haplotype-resolved assemblies are available allowing in-depth inheritance analysis (Mengist *et al.* 2023; Wu *et al.* 2023). The two lingonberry genomes created in this study are therefore not at the quality initially planned but can serve as a reference genome to identify genes, polymorphic genetic markers and compare with related species.

There are several reasons why the assembly was not chromosome-level, and there are ways to improve future assembly efforts. First, the high error rate of the raw long-read data can make the assembly difficult. This is because during the assembly process, the sequencing error creates many possible paths that result in bubbles on the assembly graph, similar to how highly heterozygous genomes behave (Liu *et al.* 2021). This is usually compensated by high read depth as read error is usually not replicated and so they can be identified by their low depth, but when errors are biased towards certain parts of the genome (repetitive, non-unique, homopolymer sites), they can be troublesome (Delahaye and Nicolas 2021). Even though ONT has improved its basecalling accuracy over time (88.7% of reads were >Q10 in R10 vs. 69.4% in R9.4), a total of ~17.4X and ~7.1X of sequenced reads were removed because they did not pass the Q10 threshold. The duplex reads had an average Q score of 25–30, indicating an error rate of one base pair in one thousand and equally accurate to the much shorter Illumina reads. But only about 6–12% of the reads were duplex reads and the rest were simplex reads with Q16–18 (Appendix B). Given the increases in read accuracy for ONT data even over the course of this project and planned technological developments that increase the duplex read rate, error-prone reads will be less of an issue with future ONT genome assemblies.

Aside from higher accuracy, longer reads can also improve assembly contiguity. When assembling a human genome, the contig N50 was doubled by inputting an extra 5X coverage of ultra-long reads, defined as reads >100 kbp (Jain *et al.* 2018). Similarly, the previous reference genome of *Arabidopsis thaliana* has been improved with the additional ultra-long reads (Wang *et al.* 2022). My raw reads averaged ~20 kbp in length, with 1,408 ultralong reads (~0.326X) for LW1 and 1,739 ultralong reads (~0.381X) for LC1, which could be improved further to collect longer reads. This would require gentler extraction and handling of the HMW DNA as well as more rigorous filtering of short DNA fragments. One possible way my assembly could be improved without performing further sequencing, is to re-analyze using more advanced basecalling algorithms, such as Bonito (https://github.com/nanoporetech/bonito). This method improves

basecalling accuracy by producing a species-specific model (Silvestre-Ryan and Holmes 2021). Unfortunately, this requires existing genomic resources as a truth-set for the model development, which is not possible now, but may be possible for future lingonberry genomes.

Since lingonberry shares the diploid karyotype with other *Vaccinium* species (2n = 24, Redpath et al., 2022) and the most closely related assembled genome is from bilberry (*V. myrtillus*), I mapped the lingonberry contigs to the bilberry chromosome-level assembly as the reference genome (Wu *et al.* 2021) to anchor them onto 12 chromosomes. More than 98% of my contigs were scaffolded onto chromosomes (LC1: 98.9%, LW1: 98.5%) with a scaffold N50 of 43.867 Mbp and 42.799 Mbp, respectively, allowing a wide range of downstream analyses using these assemblies as a reference genome. The caveat to this reference-based scaffolding of the genome is that the final chromosome-scale assembly does not necessarily represent the real genome structure of lingonberry. This is because the true structural variations such as large insertions or deletions, duplications, translocations, or inversions, could be rearranged during scaffolding following my methods. Ragtag, an automated homology-based scaffolding tool I used for lingonberry scaffolding (Alonge *et al.* 2019), makes decisions on the orientation and order of individual contigs in my lingonberry assembly based on its alignment to the bilberry genome. The consequence of this is a possibility of contig misplacement if there were true structural variations existing between lingonberry and bilberry genome. That being said, a recent study in *Eucalyptus* scaffolded ONT genomes on congeneric reference genome to study genome structure evolution and found that a very small proportion of synteny breakpoints were at contig joins, as might be expected if scaffolding is inducing false rearrangements (Ferguson *et al.* 2023). Taken together, my scaffolded genome is not a true unbiased representation of the lingonberry genome structure but is likely very close to it. Future efforts could generate an unbiased scaffolding using Hi-C or optical mapping and additionally test for the amount of bias introduced by scaffolding to a related reference genome.

2.4.1.2 Gene/TE annotation

Conceptually, there are three main approaches to gene annotation. The first approach is intrinsic or *ab initi*o predictions where the statistical models determine the most probable combinations of genetic structures (e.g., exons, introns, promoters) based solely on the nucleotide sequence in the genome (Majoros *et al.* 2004). The second is extrinsic or evidence-based approach where some form of external evidence such as proteins, cDNA, or assembled transcripts, are used to predict gene locations and structures on the genome. The third approach is a combined method in which the pipeline uses both the *ab initio* statistical models without external evidence and the mappings of external evidence as ways to predict gene structures (Ejigu and Jung 2020). Because of the pros and cons of using either intrinsic or extrinsic approach alone, the baseline methods applied in genome annotation have been pipelines that use a combined approach, and there are several automated programs to do that (e.g., MAKER, BRAKER, Comparative Augustus). MAKER pipeline has been most commonly used among the non-model species genome assembly and annotation projects, including the commercial blueberry and cranberry (Colle *et al.* 2019; Diaz-Garcia *et al.* 2021). It annotates the genes by first masking the repetitive regions in the genome, aligning RNAseq to the genome as evidence of transcription, and lastly doing *ab initio* gene predictions based on the stored evidence using multiple gene finders incorporated in the program (Campbell *et al.* 2014). More recently, BRAKER2 is claimed to perform better over MAKER and was used in wild blueberry and bilberry assembly projects (Wu *et al.* 2021; Yu *et al.* 2021). This pipeline runs by first training the *ab initio* gene predictor with existing protein database the users provide, then the predictor uses this information to look for gene structures in the genome by scanning the whole sequence. Lastly and optionally, RNAseq or transcripts assembly can be used to filter unlikely genes after the *de novo* prediction (Brůna *et al.* 2021).

However, it is recognized that above-mentioned automated pipelines can frequently create false positives and errors in annotations such as deletion/insertion in exons that can lead to frameshifts,

requiring the use of downstream filtering (Gabriel *et al.* 2021; Vuruputoor *et al.* 2022). Moreover, the initial training involved in the first step of the *ab initio* predictions depends on the quality and availability of the comprehensive and relevant protein dataset (Brůna *et al.* 2021). For non-model species that do not have established sets of verified genes, a comparative study suggests that the annotation based on only RNAseq evidence is the best in terms of reducing the errors and finding only the true genes on the genome (unpublished work from Freedman, et al. 2023 at https://github.com/harvardinformatics/GenomeAnnotation). Thus, to facilitate gene annotation on the lingonberry assembly with least possible errors, I chose to perform only the extrinsic approach that uses genome-guided predictions from aligned RNAseq data of the same species.

RNA sequencing was performed on four different tissue types: leaf, rhizome, flower, berry. Eventually, the LC1 assembly was annotated with 27,243 genes (BUSCO (Complete) = 91.4%) which was relatively fewer than the expected 30–40,000 genes in the existing literatures on *Vaccinium* species (Wu *et al.* 2021; Cui *et al.* 2022; Mengist *et al.* 2023). In addition, the BUSCO score for the protein-coding genes was lower than that of the genome assembly (96.5%). The missing ~5% BUSCO genes may be a result of transcript misassembly. Because BUSCO genes are conserved genes, this means they are likely expressed by default in any flowering plants (Simão *et al.* 2015; Manni *et al.* 2021), it is doubtful to think that the RNAseq from four different tissue types could not capture their presence. The fact that both methods do not detect the same BUSCO genes highlights the limitations to genome annotation even for high quality genomes, and highly conserved genes. In fact, different methods often do not identify identical genes for the same reference genome (Weisman *et al.* 2022). This is likely because gene annotation is not a solved problem and because a “gene” is a biologically blurry concept. Genes constantly turn over in evolutionary time scale from *de novo* gene birth to pseudogenization and defining a hard boundary between gene and non-gene is impossible (Benovoy and Drouin 2006). In conclusion, my gene annotation provides a useful and reasonable representation of the gene content of

lingonberry, but like all non-model organisms, should be treated as a hypothesis rather than an absolute truth.

2.4.1.3 Flavonoid biosynthesis evolution

In this study, 51 putative flavonoid pathway related genes composed of 20 distinct enzymes/structural gene categories were identified through orthology to the tetraploid commercial blueberry genome (Colle *et al.* 2019). Lingonberry has been described to contain cyanidin-derived anthocyanins in the berries (Brown *et al.* 2012; Amundsen *et al.* 2021), and anthocyanins are what makes the blue/red pigmentation in *Vaccinium* berries (Albert *et al.* 2023). Moreover, anthocyanins and the related flavonoids are the major targets of breeding due to their health benefits (Edger et al., 2022). Therefore, the genes involved in phenolic production are of industry interest.

While there has been effort to build QTL maps associating genomic regions to increased anthocyanin production in commercial blueberry and cranberry (Diaz-Garcia *et al.* 2018; Montanari *et al.* 2022), the genetic basis for anthocyanin biosynthesis in lingonberry is relatively understudied. Previous studies have found varying levels of flavonoids across cultivars and within populations, and showed significant effects of genotypes on the total anthocyanin content (Debnath and Sion 2009; Vilkickyte *et al.* 2022). Alam et al. (2018) attempted to associate SNPs with the total anthocyanin content in lingonberry using 1,586 loci but observed significant variability within samples and highlighted possible environmental influences. Since anthocyanins serve physiological roles in protection against abiotic stresses (Albert *et al.* 2022, 2023), whether individual genotypes associate with the biochemical profiles rather than the environmental conditions is hard to distinguish. The QTL study that specifically targeted the increased anthocyanin production in blueberry suggested candidate genes including BADH acyltransferase and UDP glucose:flavonoid 3-O-glucosyl transferase (UFGT) to be highly correlated with the increased anthocyanin profile (Montanari *et al.* 2022). I was able to annotate four copies of UFGT

in the lingonberry genome, one of which was highly expressed in red berries (STRG.15162 on chromosome 4; Figure 2.4). The genomic resource created in my study could be used to find such orthologues and provide a starting point to develop a set of lingonberry-specific markers that could be useful to accelerate the breeding efforts by encouraging marker-assisted selection. Additionally, I have identified candidate flavonoid synthesis genes, and combined with gene expression from multiple tissues I have found genes that are likely to be involved specifically in berry flavonoids.

For instance, phenylalanine ammonia-lyase (PAL), the first committed step in making anthocyanins following the essential amino acid phenylalanine, was highly expressed in red berries but relatively low in white and green berries. Similarly, one copy of TT19 was expressed very highly in red berries but not in green or white (Figure 2.3). Transparent testa 19 (TT19) is described to function as glutathione S-transferase in allowing the accumulation of anthocyanins in the vacuoles in *Arabidopsis* (Kitamura *et al.* 2004). It would therefore make sense to have high TT19 expression in the storage tissue where the anthocyanins accumulate, for example, berries. My results are in line with the lingonberry transcriptome study from *V. vitis-idaea* var. 'Sunna' (Tian *et al.* 2020), including what the authors have pointed out about the odd discovery of F3'5'H. They mention that because the delphinidin-derived anthocyanins (with 3' and 5' OH groups) have not been detected in lingonberry, it is unexpected to find the F3'5'H gene in lingonberry fruit transcriptome. There were two putative copies of F3'5'H in my annotation, and one of them was detected in all tissue types sampled. One possibility is that the produced delphinidin-derived molecules are not the preferred storage form in lingonberry and thus they exist for a short period in a transitory phase, making it impossible to be detected when measured in berries (Tian *et al.* 2020). Another explanation could be that they are only primarily made in non-berry tissue like leaves and flowers, which could result in the observed expression pattern (Figure 2.3, 2.4). While this enzyme is critical in making complex anthocyanins that produce blue or purple hues in fruits, it is an intermediate step for making the downstream flavonoids other than anthocyanins in the

intricate metabolic network, such as proanthocyanidins (Figure 2.3). Proanthocyanidins have been described to perform critical defence roles against herbivores; particularly, in immature berries, they provide unpalatable flavours to deter herbivore ingestion (Albert *et al.* 2022).

Although studies on flavonoid related genes tend to focus on their health outcomes, my result have some interesting findings regarding gene evolution. Looking at the expression profiles by tissue types (Figure 2.3), some genes are present in multiple copies, and sometimes they include active/inactive forms, or are expressed in a tissue-specific manner. It is biologically relevant to find flavonoid genes expressed in leaf tissue since numerous flavonoids have been quantified and described in metabolic profiles of lingonberry leaves (Liu *et al.* 2014; Ferlemi and Lamari 2016; Vilkickyte and Raudone 2021). On the other hand, their expression in rhizome was not recognized before. Considering various physiological roles of flavonoids in plants (Albert *et al.* 2022), the fact that most of the genes downstream of the chalcone synthase (CHS) including CHI, FHT, F3'H, LAR, ANR, and OMT are more abundant in rhizome/leaves suggests that flavonoids play roles in vegetative tissues possibly related to stress tolerance or rhizobia interactions (Albert *et al.* 2022). However, follow-up functional studies are needed to confirm their physiological roles, particularly in the rhizome.

Gene duplication/deletion is a well-recognized molecular mechanism that drives the creation of novelty in genome evolution. The fact that there are tissue-specific copies of the same enzyme such as 4CL, CHS, and FLS (Figure 2.3, 2.4), may suggest that they have been neofunctionalized or subfunctionalized. It has been described that the FLS gene, for example, has three paralogs in *Vaccinium*, one of which is almost perfectly conserved at protein-level sequences across species while the other two are not (Pucker *et al.* 2020). In lingonberry, the highly conserved functional copy is likely STRG.13258 as seen in the sequence alignment and the other two copies are categorized as paralogs to each other and not under the same orthogroup as the conserved copy (data not shown). This could be an example of neofunctionalization as FLS is a central

enzyme in flavonol production (Pucker *et al.* 2020); the duplicated copy may have developed new functions such as altered specificity to their substrates (Albert *et al.* 2023). Alternatively, subfunctionalization is possible as the expression level of STRG.13258 and STRG.5589 are almost complementing each other when looking at the tissue localization (Figure 2.4). Patterns of such gene duplications compared to the closely related *Vaccinium* species may provide interesting insights into how they evolved and why they are tissue specific.

2.4.2 Species and subspecies origin of lingonberry

2.4.2.1 Phylogenetic relationship with other *Vaccinium* species

The phylogeny of *Vaccinium* has been a struggle to resolve. Frequent polyploid formation, hybrid species formation (Lyrene *et al.* 2003), as well as ongoing introgression from closely related species (Bjedov *et al.* 2015) all contribute to this issue. While classifying them into sections may help categorize them by morphology or habitat types, the current *Vaccinium* phylogeny consists of potential polyphyletic groups (Edger et al., 2022). To resolve this issue, previous works have applied several molecular markers but found conflicting phylogenetic patterns depending on which markers were used (see Figure 1.3). For example, one of the earliest attempts combining the morphological and molecular phylogenetics used only two chloroplast genes and indicated that lingonberry and cranberry are sister to each other, while blueberry and bilberry are more distant (Kron *et al.* 2002b). One other study conducted more recently using 507 SSR markers developed for cranberry also supported this topology (Rodriguez-Bonilla *et al.* 2019). But comparatively more studies agree on an alternative topology where lingonberry is more closely related to blueberries than cranberries, and is sister to bilberry and other bog blueberries (Schlautman *et al.* 2017; Kim *et al.* 2020; Fahrenkrog *et al.* 2022). Phylogenetic inference performed on recent genome assemblies has focused beyond the genus and not resolved intra-genus relationships within *Vaccinium*. The blueberry and cranberry assembly studies both looked at the commercial vs. wild progenitor (Diaz-Garcia *et al.* 2021; Cui *et al.* 2022) to search for useful genetic materials for crossing, rather than disentangling the *Vaccinium* phylogeny. Little research

has been done on the evolutionary relationships of *Vaccinium* species using genome-scale molecular markers.

My study aimed to update our understanding of *Vaccinium* phylogeny using the whole genome data and found that the overall species tree topology based on three different methods was consistent with previous studies (Schlautman *et al.* 2017; Kim *et al.* 2020; Fahrenkrog *et al.* 2022) in which lingonberry (*V. vitis-idaea*) is closest to bilberry (*V. myrtillus*), followed by blueberries (*V. corymbosum, V. darrowii, V. caesariense*) and cranberries (*V. macrocarpon, V. oxycoccos, V. microcarpum*) (Figure 2.13, 2.14, 2.15). The advantage of using genome-wide molecular markers is that it has the power to look at the genealogy of many orthologous genes individually, then find the most congruent species topology based on the topology of all gene trees. When speciation is recent, or ancestral branches are short (i.e. past rapid speciation), it is common for gene trees to disagree with the overall species tree due to incomplete lineage sorting (Coyne and Orr 2004). The species trees from my work have presented fairly low gene concordance factors (< 30%) for nodes among closely related species, indicating the history of incomplete lineage sorting or possibly introgression.

Lingonberry is sometimes referred as ‘mountain cranberry’ or ‘alpine cranberry’ (Penhallegon 2006) due to their red berry colours and their similar taste profile. Using the phylogenetic relationship drawn from this study, we can ask whether the ancestral plant had red or blue berries or neither. Both scenarios require two steps of colour change (Figure 2.16); thus, the colour of the ancestral state is ambiguous. However, considering the complexity of blue/purple colouration (i.e., decorated sugar and hydroxylation patterns on anthocyanins), convergent evolution of red berries is arguably more plausible. Berries are culturally and economically important, and there have been active research on elucidating the mechanism of anthocyanin composition and biosynthesis (Albert *et al.* 2023), which affects berry skin colouration. Bringing the evolutionary standpoint to metabolic regulation might reveal the underlying molecular basis for anthocyanin

and the surrounding flavonoid production in berries. Particularly, adding species closely related to lingonberry and bilberry could address and resolve those questions.

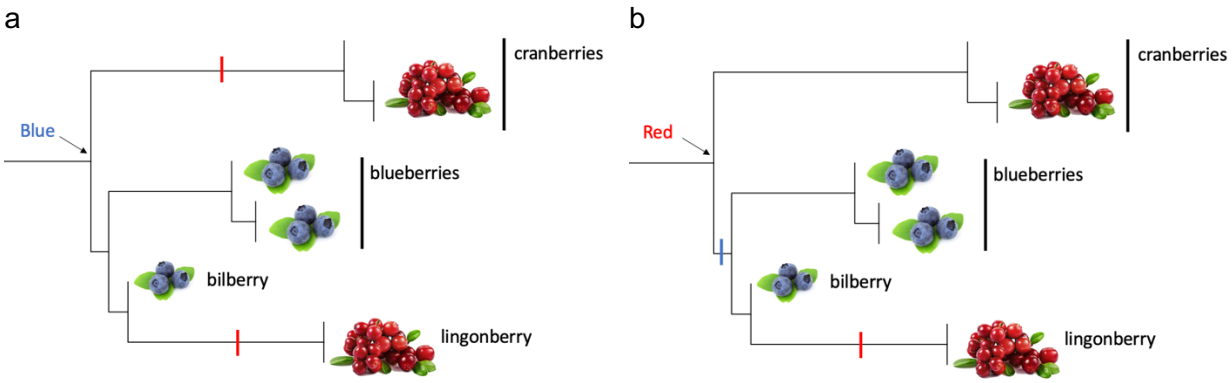

Figure 2.16: Berry colour evolution in *Vaccinium*. Hypothetical evolution of colour morphology if the ancestral population had a) blue or b) red berries.

The split between blueberry species and bilberry/lingonberry has only ~25% gene concordance factor, which indicates that there is significant incomplete lineage sorting or introgression. The habitat types of lingonberry and bilberry are similar, and they are representative subarctic dwarf shrubs usually found to live in proximity in natural populations (Bjedov *et al.* 2015; Gailīte *et al.* 2020). The ability to occasionally form hybrid populations has been also observed in Nordic populations, which could also introduce frequent gene flow. But considering the diversity of *Vaccinium*, bilberry may not be the sister species to lingonberry. My current analysis is missing several possible candidates. Species like bog blueberry (*V. uliginosum*) is shown to be the closest relative in chloroplast genomes (Kim *et al.* 2020), and they share their habitats in Canadian Arctic (Boulanger-Lapointe *et al.* 2020). Species in the section Pyxothamnus (e.g., *V. ovatum*, *V. meridionale, V. floribundum*) could also be potentially closely related to lingonberry according to their positionality in Rodriguez-Bonilla *et al*. (2019). Given the large range and difference in geography between the two lingonberry subspecies, it is possible they hybridize with different species and have different introgression patterns. In the flora of the Canadian Arctic, *V. myrtillus* is not described but *V. uliginosum* is (Aiken, *et al.*, 2007), while in Sweden and other Nordic

nations, both species are described to dominate the alpine (Auffret *et al.* 2010; Gailīte *et al.* 2020). Using the genomes developed here as a reference, future researchers could test this hypothesis.

2.4.2.2 Genomic divergence between North American and European subspecies

The concept of subspecies in plants has been historically defined as local races or groups of morphologically identifiable populations within the same species depending on their geographical origin (Grant 1981; Mallet 2013). Systematically speaking, this subspecies classification comes below the species classification, meaning they can interbreed readily in the hybrid zones even though they could sometimes produce slightly maladapted offspring. Because of their not-yet fully speciated status, subspecies can also be described as incipient species (Mallet 2013). With this definition, it seems reasonable to call the European and North American lingonberries as separate subspecies entities.

The two lingonberry subspecies are defined by geographic origin and morphological differences (Figure 1.2b, Table 1.1) but it is unclear whether this represents a sharp division between subspecies, or a gradual change across its circumpolar range. Even though morphological differences are reasonably well-recognized, the actual geographical distributions of the subspecies are vague, and they span all continents in the Northern hemisphere (Figure 1.1). If the ancestral population of the lingonberry subspecies was geographically isolated between Europe and North America, the two genomes could have accumulated variations independently over a long enough time such that they may show reduced hybrid fitness when compared to each other when in secondary contact. Nonetheless, lingonberry populations in these intermixing areas possibly the case in Russia largely lack genetic or genomic data. Future studies should fill this gap by sampling across the circumpolar range, and compare to the reference genomes created here as the two extremes of the subspecies spectrum. The whole genome alignment showed that the two lingonberry subspecies have the average pairwise divergence of 0.86% and no particular windows were notably different in divergence (Figure 2.5). The syntenic blocks identified between

the two genomes covered 51.1% and 55.3% (Table 2.3), which indicated a good colinear relationship considering 45.82% of the genome is repetitive or unalignable (Table 2.1). Relatively even genomic divergence is consistent with isolation without gene flow, which is expected based on their geographic separation.

One way of estimating divergence times between populations is using the statistic D_a (Nei and Li 1979). It accounts for the amount of variation segregating within populations and uses that to normalize the amount of variation that has accumulated since the populations diverged, using the formula $D_a = D_{XY} - \frac{(\pi_x + \pi_y)}{2}$. In this case, D_{XY} represents the sequence divergence between the two subspecies genomes and π is the proportion of heterozygosity in each genome. I can then calculate a divergence time in generations by dividing D_a by twice the average mutation rate (3×10^{-9}). This estimate puts the divergence time between subspecies at ~1 MYA. This estimate of divergence time is preliminary, due to the minimal sample size, and does not account for other factors such as variation in mutation rate or generation time. Population-level sampling is needed to better understand the amount of diversity within and between the lingonberry subspecies.

2.4.2.3 Subspecies origin – impacts of repeated glacial cycles during Pleistocene

Between 0.9-2.4 MYA, the Earth had the major ice ages in the Quaternary period. During this time, an Arctic ice cap was established, and the northern ice sheets repeatedly advanced and receded in ~41,000 years cycles (Hewitt 2000). Species that existed prior to these glaciation events inevitably experienced an extreme change in climate, especially subarctic or arctic-alpine plants, as they were living on the southern edges of the ice sheets (Hewitt 2000). Hultén (1937) noted that the presence of Beringia (= Bearing Land Bridge, the bridge between Alaska and Eurasia) during the Pleistocene (1–2 MYA) acted as a refugium for most contemporary subarctic plant species; Beringia once bridged the two continents, North America and Eurasia, and became the place of origin of plants that have subsequently expanded their range into the circumpolar region. Empirical studies support this theory, showing that those subarctic species that underwent

repeated fragmentation and reformation in their ancestral populations form a genetic structure that divides contemporary populations into five main groups: Siberian, Beringian, Canadian, east Canadian/west Greenlandic, and east Greenlandic/ west Scandinavian, among which Beringian population tends to have the highest genetic diversity (Eidesen *et al.* 2013). This can be explained by the species radiation from Beringia starting at the end of Pleistocene, as proposed by Hultén (1937), because the populations in other parts such as Europe and Canada have undergone a period of population bottlenecking, resulting in small effective population size, and thus show limited genetic diversity within the geographical populations.

Based on my MSMC analysis (Figure 2.9, Table 2.5), both European and North American lingonberry subspecies were at their maximum Ne around 1–3 MYA and have been declining since then, indicating a major population bottleneck. Given its current range, lingonberry has likely undergone repetitive range contractions followed by expansion due to ice sheets advancing and receding. The survivors from the glacial refugia on the south ends of the ice sheets would have a small population size initially, leading to population bottlenecks post-glaciation. Lingonberry was included in a previous genetic study that particularly looked at this effect of glaciation on genetic structure (Eidesen *et al.* 2013), which found that lingonberry followed the Beringian glacial refugia model in the Pleistocene. Another study within Europe and Norwegian nations (Garkava-Gustavsson *et al.* 2005) also found genetic structure, where samples from Sweden, Finland, Norway, Estonia, and Russia resulted in a significant genetic difference between countries. Because of rapid colonization in the circumpolar range post-glaciation, the patterns of genetic variation within countries are usually heterogeneous in wild lingonberry populations, as shown in the wild Canadian populations (ssp. *minus*) (Debnath 2007a; Debnath and Sion 2009) and Swedish populations (ssp*. vitis-idaea*) (Persson and Gustavsson 2001; Garkava-Gustavsson *et al.* 2005). Nevertheless, the impact of glacial refugia other than Beringia is also possible such as those in Japanese populations (Ikeda *et al.* 2015). However, for my samples from Quebec, Canada (LW1), and Netherlandish variety ‘Red Candy’ (LC1), the Ne maxima at 1-3 MYA could

correspond to the refugia in Beringia. Noneetheless, it should be noted that the timing of my models is limited by the confidence in the generation time and mutation rate parameters. For lingonberry, both parameters are undefined and are approximate estimates based on indirect inference from the literature (Ritchie 1955; Exposito-Alonso *et al.* 2018); thus, the timing of events should be interpreted with caution.

An alternative explanation for this Ne decline towards the present could be that the individuals I used for this study were inbred and generally low in genetic diversity, resulting in smaller effective population size. This has been observed in continuous and widespread contemporary populations like western redcedar, where the recent range expansion is seen as having been at the cost of losing genetic diversity through inbreeding (Shalev *et al.* 2022). However, this is unlikely for lingonberry based on the heterogeneous genetic structure within countries, as mentioned above. Moreover, selfing is not a primary reproductive strategy for lingonberry since it is documented to preferentially produce more berries when pollinated by insects (Guillaume and Jacquemart 1999; Nuortila *et al.* 2002), which would logically lead to higher seed dispersal and reproductive fitness. Nevertheless, it can self-fertilize, so it is possible that the specific individual I used for sequencing in this study was from a highly inbred population. This is likely not the case though because I found no evidence for long runs of homozygosity, which would be characteristic of recent inbreeding (Figure 2.11, 2.12).

Taken together, the parallel declines in Ne for both subspecies is consistent with a hypothesis that repeated glaciation driven bottlenecks reduced diversity despite its current expansive range.

2.5 Conclusion

This study defined the subspecies divergence in lingonberry at the whole-genome scale, for the first time, using genome assemblies. With the sequence variants detection and genome annotations, basic genomic knowledge was built for lingonberry including genome-wide heterozygosity estimates, its phylogenetic position in the genus, and genes involved in flavonoid biosynthesis pathway. The data generated in this study will facilitate future work, such as generation of genetic markers for breeding and analysis of population structure across the species range. Further, the results encouraged future scientists in the field to address novel hypotheses regarding not only the evolution of lingonberry, but also the evolution of diverse edible berries in the genus *Vaccinium*.

Chapter 3: Concluding remark

Lingonberry is a culturally and economically important berry-bearing crop that has growing potential for domestication and potential use in medicine or natural products. Despite continuing efforts in defining biochemical and metabolic profiles across different cultivars and wild populations, genomic studies of lingonberry are far behind compared to other berry crops. Genomic resources accelerate the breeding efforts by providing a molecular tool to select progeny with desirable traits when crossed or even before being crossed (e.g., marker-assisted breeding, genomic selection). Additionally, genomics provides a way to answer evolutionary questions about lingonberry's past and future evolution.

Vaccinium encompasses diverse morphologically distinct species shaped by the frequent polyploidization, hybrid formation, and introgression (Vander Kloet 1988; Lyrene *et al.* 2003; Redpath *et al.* 2022). Polyploidization doubles the amount of genetic material, which fuels evolutionary changes by introducing space to create diversity (Heslop-Harrison *et al.* 2023). In *Vaccinium*, polyploidization is sometimes responsible for a new species formation such as the commercial tetraploid blueberry (Colle *et al.* 2019; Mengist *et al.* 2023), and other times it merely represents the spontaneous formation of populations that differ in ploidy but are not necessarily separate species like tetraploid lingonberry in Japan (Wakui and Kudo 2021). But at the same time, intraspecies polyploidization could be an adaptive strategy for an organism under stressful or changing conditions like in the case of arctic or alpine plants (Brochmann *et al.* 2004; Wakui and Kudo 2021). Future work could use genomics to quantify the distribution of tetraploid lingonberry across the range and determine whether it's responsible for latitudinal variations, as predicted. Furthermore, it could address whether tetraploid lingonberry has repeatedly evolved (e.g., Soltis *et al.* 2004; Redpath *et al.* 2022), or originated from a single polyploidy event.

Some species in *Vaccinium* are sub-arctic plants and are survivors of the repeated glacial cycles including lingonberry (*V. vitis-idaea*) and bog blueberry (*V. uliginosum*) (Eidesen *et al.* 2013), and

my model suggests these glacial cycles had large effects on lingonberry populations. However, they thrive in modern times, and are predicted to further expand their range under climate change (Egorova 2020; Hirabayashi *et al.* 2022). The rapid range expansion after glaciation raises many questions about how and whether lingonberry populations are adapted to their local habitat. This will become increasingly important as climate change causes plant species to shift their range north.

Aside from polyploidization, novel adaptive materials can also be introduced into the species genome by gene duplications through other phenomena like TEs or repeats. In fact tandem repeats are described to be frequent drivers of adaptive evolution in a subtropical blueberry (Cui *et al.* 2022). Gene duplications create genetic redundancy and an opportunity to diversify genes through sub- or neofunctionalization. This may have occurred in the flavonoid related genes in lingonberry which show differential expression across tissues. Tissue specific expression of duplicated genes could be explored further to elucidate the physiological functions of those enzymes in the flavonoid pathway which is known to have various roles in stress tolerance (Albert *et al.* 2022, 2023), or they could possibly interact with other pathways to perform completely different functions. Phylogenetic analysis implies that berry color has changed multiple times in the genus. Given the tight regulatory network of anthocyanin biosynthesis with the berry skin colour and complexity of its regulation (Zhao *et al.* 2019; Diaz-Garcia *et al.* 2021), studying the metabolic system with evolutionary perspective may advance our understanding.

My research created a valuable resource to initiate future studies on alpine/sub-arctic plant evolution and diversity. Looking forward, lingonberry could be developed as a model species for persistence and evolution in arctic and alpine species under climate change. Follow-up studies on lingonberry population-level genomics will clarify answers to some of these unapproached questions and hypotheses raised in this study and reveal the evolutionary footprint of this charismatic plant.

References

Agriculture and Agri-food Canada, 2023 LINGONBERRY Red treasures from the North–A guide to growingn lingonberries. Lingonberry book English.

Ahokas, H., 1971 Notes on polyploidy and hybridity in *Vaccinium* species. Ann. Bot. Fenn. 8: 254–256.

Aiken, S.G., Dallwitz, M.J., Consaul, L.L., McJannet, C.L., Boles, R.L., Argus, G.W., Gillett, J.M., Scott, P.J., Elven, R., LeBlanc, M.C., Gillespie, L.J., Brysting, A.K., Solstad, H., and Harris, J. G., 2007 Flora of the Canadian Arctic Archipelago.

Alam, Z., J. Roncal, and L. Peña-Castillo, 2018 Genetic variation associated with healthy traits and environmental conditions in Vaccinium vitis-idaea. BMC Genomics 19: 1–13.

Albert, N. W., M. Iorizzo, M. F. Mengist, S. Montanari, J. Zalapa *et al.*, 2023 *Vaccinium* as a comparative system for understanding of complex flavonoid accumulation profiles and regulation in fruit. Plant Physiol. 1–15.

Albert, N. W., D. J. Lafferty, S. M. A. Moss, and K. M. Davies, 2022 Flavonoids–flowers, fruit, forage and the future. J. R. Soc. New Zeal. 53: 5.

Alonge, M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin *et al.*, 2019 RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 20: 224.

Amundsen, M., K. Aaby, I. Martinussen, A. L. Hykkerud, and L. Jaakola, 2021 Effects of geographic origin and environmental conditions on anthocyanin profile of wild Norwegian lingonberries (Vaccinium vitis-idaea L.), pp. 30 in *Book of Abstracts in XII International Vaccinium Sympodium*, International Society for Horticultural Science; Dalhouse University, Halifax, Nova Scotia.

Andrews, S., 2019 A quality control tool for high throughput sequence data.

Arctic Lingonberry, 2022 Finnish Minist. Agric. For.

Arnason, R., 2023 Scientist sets out to increase lingonberry acres. West. Prod.

Auffret, A. G., E. Meineri, H. H. Bruun, R. Ejrnæs, and B. J. Graae, 2010 Ontogenetic niche shifts in three Vaccinium species on a sub-alpine mountain side. Plant Ecol. Divers. 3: 131–139.

Van der Auwera, G., and B. O'Connor, 2020 Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).

Belichenko, O., V. Kolosova, K. Jernigan, O. Belichenko, V. Kolosova *et al.*, 2022 Diachronic and cultural variations in Chukchi ethnobotany. Etudes Inuit Stud. 45: 315–340.

Belton, J. M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan *et al.*, 2012 Hi-C: A comprehensive technique to capture the conformation of genomes. Methods 58: 268–276.

Benovoy, D., and G. Drouin, 2006 Processed pseudogenes, processed genes, and spontaneous mutations in the *Arabidopsis* Genome. J. Mol. Evol. 62: 511–522.

Bjedov, I., D. Obratov-Petković, D. Mišić, B. Šiler, and J. M. Aleksić, 2015 Genetic patterns in range-edge populations of *Vaccinium* species from the central balkans: Implications on conservation prospects and sustainable usage. Silva Fenn. 49:.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics btu170:.

Boulanger-Lapointe, N., 2017 Importance of berries in the Inuit biocultural system: A multidisciplinary investigation in the Canadian North.

Boulanger-Lapointe, N., G. H. R. Henry, E. Lévesque, A. Cuerrier, S. Desrosiers *et al.*, 2020 Climate and environmental drivers of berry roductivity from the forest-tundra ecotone to the high Arctic in Canada. Arct. Sci. 6: 529–544.

Bourque, G., K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov *et al.*, 2018 Ten things you should know about transposable elements. Genome Biol. 19: 199.

Brochmann, C., A. K. Bysting, I. G. Alsos, L. Borgen, H. H. Grundt *et al.*, 2004 Polyploidy in arctic plants. Biol. J. Linn. Soc. 82: 521–536.

Brown, P. N., C. E. Turi, P. R. Shipley, and S. J. Murch, 2012 Comparisons of large (*Vaccinium macrocarpon* Ait) and small (*Vaccinium oxycoccos* L., *Vaccinium vitis-idaea* L.) cranberry in British Columbia by phytochemical determination, antioxidant potential, and metabolomic profiling with chemometric analysis. Planta Med. 78: 630–640.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR genomics Bioinforma. 3: lqaa108.

Bushnell, B., 2014 BBMap: A Fast, Accurate, Splice-Aware Aligner, in Conference: 9th Annual Genomics of Energy & Environment Meeting, Walnut Creek, CA, US.

Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014 Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinforma. 48: 4.11.1-4.11.39.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross *et al.*, 2008 MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18: 188–196.

Chen, Y., F. Nie, S.-Q. Xie, Y.-F. Zheng, Q. Dai *et al.*, 2021 Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat. Commun. 12: 60.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li, 2021 Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18: 170–175.

Chor, B., D. Horn, N. Goldman, Y. Levy, and T. Massingham, 2009 Genomic DNA k-mer spectra: models and modalities. Genome Biol. 10: R108.

Colantonio, V., L. F. V Ferrão, D. M. Tieman, N. Bliznyuk, C. Sims *et al.*, 2022 Metabolomic selection for enhanced fruit flavor. Proc. Natl. Acad. Sci. 119: e2115865119.

Colle, M., C. P. Leisner, C. M. Wai, S. Ou, K. A. Bird *et al.*, 2019 Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. Gigascience 8: 1–15.

Coyne, J. A., and A. H. Orr, 2004 Species: Reality and concepts, pp. 8–54 in *Speciation*, edited by Sinauer. Oxford University Press, Sunderland, MA.

Cuerrier, A., 2011 *The Botanical Knowledge of the Inuit of Kangiqsujuaq, Nunavik*. Avataq Cultural Institute, Inukjuak, QC.

Cui, F., X. Ye, X. Li, Y. Yang, Z. Hu *et al.*, 2022 Chromosome-level genome assembly of the diploid blueberry Vaccinium darrowii provides insights into its subtropical adaptation and cuticle synthesis. Plant Commun. 3:.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan *et al.*, 2021 Twelve years of SAMtools and BCFtools. Gigascience 10: 1–4.

Debnath, S. C., 2007a Inter simple sequence repeat (ISSR) to assess genetic diversity within a collection of wild lingonberry (*Vaccinium vitis-idaea* L.) clones. Can. J. Plant Sci. 87: 337–344.

Debnath, S. C., 2007b Strategies to propagate Vaccinium nuclear stocks for the Canadian berry industry. Can. J. Plant Sci. 87: 911–922.

Debnath, S. C., and U. Arigundam, 2020 In vitro propagation strategies of medicinally important berry crop, lingonberry (*Vaccinium vitis-idaea* L.). Agronomy 10: 1–19.

Debnath, S. C., and M. Sion, 2009 Genetic diversity, antioxidant activities, and anthocyanin contents in lingonberry. Int. J. Fruit Sci. 9: 185–199.

Delahaye, C., and J. Nicolas, 2021 Sequencing DNA with nanopores: Troubles and biases. PLoS One 16:.

Diaz-Garcia, L., L. F. Garcia-Ortega, M. González-Rodríguez, L. Delaye, M. Iorizzo *et al.*, 2021 Chromosome-level genome assembly of the American cranberry (*Vaccinium macrocarpon* Ait.) and its wild relative *Vaccinium microcarpum*. Front. Plant Sci. 12: 1–12.

Diaz-Garcia, L., B. Schlautman, G. Covarrubias-Pazaran, A. Maule, J. Johnson-Cicalese *et al.*, 2018 Massive phenotyping of multiple cranberry populations reveals novel QTLs for fruit anthocyanin content and other important chemical traits. Mol. Genet. Genomics 293: 1379–1392.

Douglas, G. W., and D. V. Meidinger, 2002 *Illustrated Flora of British Columbia, Volume 8: General Summary, Maps and Keys.* (J. Pojar, Ed.). B.C. Ministry of Sustainable Resource Management and B.C. Ministry of Forests., Victoria.

Duitama, J., 2023 Phased genome assemblies, pp. 273–286 in *Haplotyping: Methods and Protocols*, edited by B. A. Peters and R. Drmanac. Springer US, New York, NY.

Dutheil, J., S. Gaillard, and E. Stukenbrock, 2014 MafFilter: a highly flexible and extensible multiple genome alignment files processor. BMC Genomics 15:.

Edger, P. P., 2023 *Vaccinium* pangenome project.

Edger, P. P., M. Iorizzo, N. V Bassil, J. Benevenuto, L. F. V Ferrão *et al.*, 2022 There and back again; historical perspective and future directions for *Vaccinium* breeding and research studies . Hortic. Res. 9:.

Egorova, N. Y., 2020 Influence of ecological factors on the population-ontogenetic parameters of *Vaccinium vitis-idaea* L. in forest ecosystems of the European northeast of Russia. Contemp. Probl. Ecol. 13: 656–662.

Ehlenfeldt, M. K., and J. R. Ballington, 2018 *Vaccinium corymbodendron* Dunal as a bridge between taxonomic sections and ploidies in *Vaccinium* : A work in progress, pp. 15 in *North American Blueberry Research and Extension Workers Conference*,.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle *et al.*, 2009 Real-time DNA sequencing from single polymerase molecules. Science. 323: 133–138.

Eidesen, P. B., D. Ehrich, V. Bakkestuen, I. G. Alsos, O. Gilg *et al.*, 2013 Genetic roadmap of the Arctic: plant dispersal highways, traffic barriers and capitals of diversity. New Phytol. 200: 898–910.

Ejigu, G. F., and J. Jung, 2020 Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology (Basel). 9: 1–27.

Ellinghaus, D., S. Kurtz, and U. Willhoeft, 2020 LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:.

Emms, D. M., and S. Kelly, 2019 OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 20: 1–14.

Emms, D. M., and S. Kelly, 2018 STAG: Species tree inference from all genes. bioRxiv 267914.

Emms, D. M., and S. Kelly, 2017 STRIDE: Species tree root inference from gene duplication Events. Mol. Biol. Evol. 34: 3267–3278.

Exposito-Alonso, M., C. Becker, V. J. Schuenemann, E. Reiter, C. Setzer *et al.*, 2018 The rate and potential relevance of new mutations in a colonizing plant lineage. PLOS Genet. 14: e1007155.

Fahrenkrog, A. M., G. Matsumoto, K. Toth, S. Jokipii-Lukkari, H. M. Salo *et al.*, 2022 Chloroplast genome assemblies and comparative analyses of major Vaccinium berry crops. bioRxiv 2022.02.23.481500.

Ferguson, S., A. Jones, K. Murray, B. Schwessinger, and J. O. Borevitz, 2023 Interspecies genome divergence is predominantly due to frequent small scale rearrangements in Eucalyptus. Mol. Ecol. 32: 1271–1287.

Ferlemi, A. V., and F. N. Lamari, 2016 Berry leaves: an alternative source of bioactive natural products of nutritional and medicinal Value. antioxidants 5:.

Ferrão, L. F. V., R. R. Amadeu, J. Benevenuto, I. de Bem Oliveira, and P. R. Munoz, 2021 Genomic selection in an outcrossing autotetraploid fruit crop: lessons from blueberry breeding. Front. Plant Sci. 12:.

Ferrão, L. F. V., T. S. Johnson, J. Benevenuto, P. P. Edger, T. A. Colquhoun *et al.*, 2020 Genome-

wide association of volatiles reveals candidate loci for blueberry flavor. New Phytol. 226: 1725–1737.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark *et al.*, 2020 RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. 117: 9451–9457.

Gabriel, L., K. J. Hoff, T. Brůna, M. Borodovsky, and M. Stanke, 2021 TSEBRA: transcript selector for BRAKER. BMC Bioinformatics 22: 566.

Gailīte, A., A. Gaile, and D. E. Ruņģis, 2020 Genetic diversity and structure of wild vaccinium populations-*V. myrtillus, V. vitis-idaea* and *V. uliginosum* in the Baltic States. Silva Fenn. 54: 1–20.

Garkava-Gustavsson, L., H. A. Persson, H. Nybom, K. Rumpunen, B. A. Gustavsson *et al.*, 2005 RAPD-based analysis of genetic diversity and selection of lingonberry (*Vaccinium vitis-idaea* L.) material for ex situ conservation. Genet. Resour. Crop Evol. 52: 723–735.

Gavrielatos, M., K. Kyriakidis, D. A. Spandidos, and I. Michalopoulos, 2021 Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly. Mol. Med. Rep. 23: 1–12.

Goel, M., and K. Schneeberger, 2022 plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics btac196:.

Goel, M., H. Sun, W. B. Jiao, and K. Schneeberger, 2019 SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 20: 1–13.

Grant, V., 1981 Populations and Races, in *Plant Speciation*, Columbia University Press, New York, NY.

Guillaume, P., and A.-L. Jacquemart, 1999 Early-inbreeding depression in *Vaccinium myrtillus* and *V. vitis-idaea*. Protoplasma 208: 107–114.

Gustavsson, B. A., 2001 Genetic variation in horticulturally important traits of fifteen wild lingonberry *Vaccinium vitis-idaea* L. populations. Euphytica 120: 173–182.

Haas, B., 2023 https://github.com/TransDecoder/TransDecoder.

Hamilton, J. P., B. Vaillancourt, J. C. Wood, and C. R. Buell, 2023 Chromosome-scale assembly of the Verbenaceae species queen's wreath (*Petrea volubilis* L.). BMC Genomic Data 24: 14.

Hancock, J., P. Lyrene, C. Finn, N. Vorsa, and G. Lobos, 2008 Blueberries and cranberries, pp. 115–150 in *Temperate Fruit Crop Breeding: Germplasm to Genomics*, edited by J. F. Hancock. Springer Netherlands, Dordrecht.

Hare, E. E., and J. S. Johnston, 2011 Genome size determination using flow cytometry of propidium iodide-stained nuclei. Methods Mol. Biol. 772: 3–12.

Harris, T. D., P. R. Buzby, H. Babcock, E. Beer, J. Bowers *et al.*, 2008 Single-molecule DNA sequencing of a viral genome. Science. 320: 106–109.

Heslop-Harrison, J. S. (Pat), T. Schwarzacher, and Q. Liu, 2023 Polyploidy: its consequences and enabling role in plant diversification and evolution. Ann. Bot. 131: 1–10.

Hewitt, G., 2000 The genetic legacy of the Quaternary ice ages. Nature 405: 907–913.

Hirabayashi, K., S. J. Murch, and L. A. E. Erland, 2022 Predicted impacts of climate change on wild and commercial berry habitats will have food security, conservation and agricultural implications. Sci. Total Environ. 845: 157341.

Hjalmarsson, I., and R. Ortiz, 1998 Effect of genotype and environment on vegetative and reproductive characteristics of lingonberry (*Vaccinium vitis-idaea* L.). ACTA Agric. Scand. Sect. B-SOIL PLANT Sci. 48: 255–262.

Hossain, M. Z., E. Shea, M. Daneshtalab, and J. T. Weber, 2016 Chemical analysis of extracts from newfoundland berries and potential neuroprotective effects. Antioxidants 5:.

Hu, J., J. Fan, Z. Sun, and S. Liu, 2020 NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36: 2253–2255.

Hultén, E., 1937 *Outline of the history of arctic and boreal biota during the Quarternary period: their evolution during and after the glacial period as indicated by the equiformal progressive areas of present plant species*.

Ikeda, H., Y. Yoneta, H. Higashi, P. B. Eidesen, V. Barkalov *et al.*, 2015 Persistent history of the bird-dispersed arctic–alpine plant *Vaccinium vitis-idaea* L. (Ericaceae) in Japan. J. Plant Res. 128: 437–444.

Illumina, 2022 Coverage depth recommendations.

Illumina, 2011 Quality scores for next-generation sequencing: assessing sequencing accuracy using Phred queality scoring. Qual. scores next-generations Seq. 1–2.

Isaak, C. K., J. C. Petkau, H. Blewett, O. Karmin, and Y. L. Siow, 2017 Lingonberry anthocyanins protect cardiac cells from oxidative-stress-induced apoptosis. Can. J. Physiol. Pharmacol. 95: 904–910.

Jaakola, L., A. M. Pirttilä, M. Halonen, and A. Hohtola, 2001 Isolation of high quality RNA from bilberry (*Vaccinium myrtillus* L.) fruit. Appl. Biochem. Biotechnol.–Part B Mol. Biotechnol. 19: 201–203.

Jacquemart, A.-L., and J. D. Thompson, 1996 Floral and pollination biology of three sympatric *Vaccinium* (Ericaceae) species in the upper ardennes, Belgium. Can. J. Bot. 74: 210–221.

Jain, M., S. Koren, K. H. Miga, J. Quick, A. C. Rand *et al.*, 2018 Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36: 338–345.

Jun, W. D., S. Dierking, and W. D. Beerenobst, 1993 European *Vaccinium* species. Acta Hortic. 299–304.

Katoh, K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:.

Kawash, J., K. Colt, N. T. Hartwick, B. W. Abramson, N. Vorsa *et al.*, 2022 Contrasting a reference

cranberry genome to a crop wild relative provides insights into adaptation, domestication, and breeding. PLoS One 17: e0264966.

Kim, D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, 2019 Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37: 907–915.

Kim, Y., J. Shin, D. R. Oh, D. W. Kim, H. S. Lee *et al.*, 2020 Complete chloroplast genome sequences of *Vaccinium bracteatum* Thunb., *V. vitis-idaea* L., and *V. uliginosum* L. (Ericaceae). Mitochondrial DNA Part B Resour. 5: 1843–1844.

Kitamura, S., N. Shikazono, and A. Tanaka, 2004 TRANSPARENT TESTA 19 is involved in the accumulation of both anthocyanins and proanthocyanidins in *Arabidopsis*. Plant J. 37: 104–114.

Vander Kloet, S. P., 1988 *The Genus Vaccinium in North America* (M. Wolf, S. Rudnitski, & F. Smith, Eds.). Res. Branch Agric. Can. Publ. 1828.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37: 540–546.

Kondo, M., S. L. Mackinnon, C. C. Craft, M. D. Matchett, R. A. R. Hurta *et al.*, 2011 Ursolic acid and its esters: Occurrence in cranberries and other Vaccinium fruit and effects on matrix metalloproteinase activity in DU145 prostate tumor cells. J. Sci. Food Agric. 91: 789–796.

Koren, S., B. Walenz, K. Berlin, J. Miller, and A. Phillippy, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res.

Kowalska, K., 2021 Lingonberry (*Vaccinium vitis-idaea* L.) fruit as a source of bioactive compounds with health-promoting effects—a review. Int. J. Mol. Sci. 22:.

Kron, K. A., W. S. Judd, P. F. Stevens, D. M. Crayn, A. A. Anderberg *et al.*, 2002a Phylogenetic classification of Ericaceae: Molecular and morphological evidence. Bot. Rev. 68: 335–423.

Kron, K. A., E. A. Powell, and J. L. Luteyn, 2002b Phylogenetic relationships within the blueberry tribe (Vaccinieae, Ericaceae) based on sequence data from MATK and nuclear ribosomal ITS regions, with comments on the placement of Satyria. Am. J. Bot. 89: 327–336.

Kumar, S., G. Stecher, M. Suleski, and S. Blair Hedges, 2017 TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. 34: 1812–1819.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lander, E. S., and M. S. Waterman, 1988 Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231–239.

Lawrence, K., 2022 Advances in duplex basecalling.

Levene, H. J., J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead *et al.*, 2003 Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 299: 682–686.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr.

Li, H., 2016 Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32: 2103–2110.

Li, H., 2018 Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li, H., 2021 New strategies to improve minimap2 alignment accuracy. Bioinformatics 37: 4572–4574.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Liu, P., A. Lindstedt, N. Markkinen, J. Sinkkonen, and J. Suomela, 2014 Characterization of metabolite profiles of leaves of bilberry (*Vaccinium myrtillus* L.) and lingonberry (*Vaccinium vitis-idaea* L.). J. Agric. Food Chem. 62: 12015–12026.

Liu, H., S. Wu, A. Li, and J. Ruan, 2021 SMARTdenovo: a de novo assembler using long noisy reads. Gigabyte.

Lyrene, P. M., N. Vorsa, and J. R. Ballington, 2003 Polyploidy and sexual polyploidization in the genus *Vaccinium*. Euphytica 133: 27–36.

Mai, U., and S. Mirarab, 2018 TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19: 272.

Majoros, W. H., M. Pertea, and S. L. Salzberg, 2004 TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20: 2878–2879.

Mallet, J., 2013 Subspecies, semispecies, superspecies. Encycl. Biodivers. Second Ed. 45–48.

Mallory, C., and S. Aiken, 2012 *Common Plants of Nunavut*. Inhabit Media Inc., Iqaluit, NU, Toronto, ON.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov, 2021 BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. 38: 4647–4654.

Marrano, A., M. Britton, P. A. Zaini, A. V Zimin, R. E. Workman *et al.*, 2020 High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome. Gigascience 9: giaa050.

Maxam, A. M., and W. Gilbert, 1977 A new method for sequencing DNA. Proc. Natl. Acad. Sci. United States Am. 74: 560–564.

Mengist, M. F., H. Bostan, D. De Paola, S. J. Teresi, A. E. Platts *et al.*, 2023 Autopolyploid inheritance and a heterozygous reciprocal translocation shape chromosome genetic behavior in tetraploid blueberry (*Vaccinium corymbosum*). New Phytol. 237: 1024–1039.

Minh, B. Q., M. W. Hahn, and R. Lanfear, 2020a New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams *et al.*, 2020b IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol.

Biol. Evol. 37: 1530–1534.

Misikangas, M., A. M. Pajari, E. Päivärinta, S. I. Oikarinen, J. Rajakangas *et al.*, 2007 Three Nordic berries inhibit intestinal tumorigenesis in multiple intestinal neoplasia/+ mice by modulating β-catenin signaling in the tumor and transcription in the mucosa. J. Nutr. 137: 2285–2290.

Moerman, D. E., 2010 *Native American Food Plants–An Ethnobotanical Dictionary*. Timber Press, Portland, London.

Montanari, S., S. Thomson, S. Cordiner, C. S. Günther, P. Miller *et al.*, 2022 High-density linkage map construction in an autotetraploid blueberry population and detection of quantitative trait loci for anthocyanin content. Front. Plant Sci. 13: 1–15.

Muoki, R. C., A. Paul, A. Kumari, K. Singh, and S. Kumar, 2012 An improved protocol for the isolation of RNA from roots of tea (*Camellia sinensis* (L.) O. Kuntze). Mol. Biotechnol. 52: 82–88.

Nakandala, U., A. K. Masouleh, M. W. Smith, A. Furtado, P. Mason *et al.*, 2023 Haplotype resolved chromosome level genome assembly of *Citrus australis* reveals disease resistance and other citrus specific genes. Hortic. Res. 10: uhad058.

Nei, M., and W. H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U. S. A. 76: 5269–5273.

Nestby, R., A. L. Hykkerud, and I. Martinussen, 2019 Review of botanical characterization, growth preferences, climatic adaptation and human health effects of Ericaceae and Empetraceae wild dwarf shrub berries in boreal, alpine and arctic areas. J. Berry Res. 9: 515–547.

Nguyen, L. S., H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, 2015 IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol. 32: 268–274.

Nuortila, C., J. Tuomi, and K. Laine, 2002 Inter-parent distance affects reproductive success in two clonal dwarf shrubs, *Vaccinium myrtillus* and *Vaccinium vitis-idaea* (Ericaceae). Can. J. Bot. 80: 875–884.

Nurk, S., S. Koren, A. Rhie, M. Rautiainen, A. V Bzikadze *et al.*, 2022 The complete sequence of a human genome. Science. 376: 44–53.

Olver, L., 1999 The Food Timeline. Agric. Hist. Soc.

Onali, T., A. Kivimäki, M. Mauramo, T. Salo, and R. Korpela, 2021 Anticancer effects of lingonberry and bilberry on digestive tract cancers. Antioxidants 10: 1–17.

Ou, S., and N. Jiang, 2019 LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA 10: 48.

Ou, S., and N. Jiang, 2018 LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 176: 1410–22.

Ou, S., W. Su, Y. Liao, K. Chougule, J. R. A. Agda *et al.*, 2019 Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome

Biol. 20: 1–18.

Oxford Nanopore, 2021 Oxford Nanopore tech update: new duplex method for Q30 nanopore single molecule reads, PromethION 2, and more. Oxford Nanopore News.

Oxford Nanopore, 2020 R10.3: the newest nanopore for high accuracy nanopore sequencing – now available in store. Oxford Nanopore News.

Penhallegon, R. H., 2006 Lingonberry production guide for the pacific northwest. PNW 583-E.

Penhallegon, R. H., 2009 Lingonberry yields in the pacific northwest. Acta Hortic. 810: 223–228.

Persson, H. A., and B. A. Gustavsson, 2001 The extent of clonality and genetic diversity in lingonberry (*Vaccinium vitis-idaea* L.) revealed by RAPDs and leaf-shape analysis. Mol. Ecol. 10: 1385–1397.

Pertea, M., G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell *et al.*, 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33: 290–295.

Pockrandt, C., M. Alzamel, C. S. Iliopoulos, and K. Reinert, 2020 GenMap: ultra-fast computation of genome mappability. Bioinformatics 36: 3687–3692.

Pucker, B., I. Irisarri, J. de Vries, and B. Xu, 2022 Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quant. Plant Biol. 3:.

Pucker, B., F. Reiher, and H. M. Schilbert, 2020 Automatic identification of players in the flavonoid biosynthesis with application on the biomedicinal plant *Croton tiglium*. Plants 9:.

Raudone, L., G. Vilkickyte, L. Pitkauskaite, R. Raudonis, R. Vainoriene *et al.*, 2019 Antioxidant activities of *Vaccinium vitis-idaea* L. leaves within cultivars and their phenolic compounds. Molecules 24:.

Redpath, L. E., R. Aryal, N. Lynch, J. A. Spencer, A. M. Hulse-Kemp *et al.*, 2022 Nuclear DNA contents and ploidy levels of North American *Vaccinium* species and interspecific hybrids. Sci. Hortic. (Amsterdam). 297: 110955.

Renaud, G., K. Hanghøj, T. S. Korneliussen, E. Willerslev, and L. Orlando, 2019 Joint estimates of heterozygosity and runs of homozygosity for modern and ancient samples. Genetics 212: 587–614.

Rhie, A., S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti *et al.*, 2021 Towards complete and error-free genome assemblies of all vertebrate species. Nature 592: 737–746.

Rhie, A., B. P. Walenz, S. Koren, and A. M. Phillippy, 2020 Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21: 245.

Ritchie, J. C., 1955 *Vaccinium Vitis-Idaea* L . J. Ecol. 43: 701–708.

Roach, M. J., S. A. Schmidt, and A. R. Borneman, 2018 Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19: 460.

Rodriguez-Bonilla, L., J. Rohde, D. Matusinec, and J. Zalapa, 2019 Cross-transferability analysis

of SSR markers developed from the American cranberry (*Vaccinium macrocarpon* Ait.) to other Vaccinium species of agricultural importance. Genet. Resour. Crop Evol. 66: 1713–1725.

Rosindell, J., and L. J. Harmon, 2012 OneZoom: a fractal explorer for the tree of life. PLOS Biol. 10: e1001406.

Ruan, J., and H. Li, 2020 Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17: 155–158.

Sanger, F., 1975 The Croonian Lecture , 1975: Nucleotide sequences in DNA. Proc. R. Soc. London Ser. B 191: 317–333.

Sayyari, E., and S. Mirarab, 2016 Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33: 1654–1668.

Schiffels, S., and K. Wang, 2020 MSMC and MSMC2: the multiple sequentially Markovian coalescent. Methods Mol. Biol. 2090: 147–166.

Schlautman, B., G. Covarrubias-Pazaran, D. Fajardo, S. Steffan, and J. Zalapa, 2017 Discriminating power of microsatellites in cranberry organelles for taxonomic studies in *Vaccinium* and Ericaceae. Genet. Resour. Crop Evol. 64: 451–466.

Shalev, T. J., O. G. El-Dien, M. M. S. Yuen, S. Shengqiang, S. D. Jackman *et al.*, 2022 The western redcedar genome reveals low genetic diversity in a self-compatible conifer. Genome Res. 32: 1952–1964.

Shi, J., and C. Liang, 2019 Generic repeat finder: a high-sensitivity tool for genome-wide de novo repeat detection. Plant Physiol. 180: 1803–1815.

Silvestre-Ryan, J., and I. Holmes, 2021 Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. Genome Biol. 22: 38.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Simpson, J. T., 2014 Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30: 1228–1235.

Simpson, J. T., and M. Pop, 2015 The Theory and practice of genome sequence assembly. Annu. Rev. Genomics Hum. Genet. 16: 153–172.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, 2014 Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15: 121–132.

Soltis, D. E., P. S. Soltis, J. C. Pires, A. Kovarik, J. A. Tate *et al.*, 2004 Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genomic and genetic comparisons. Biol. J. Linn. Soc. 82: 485–501.

Soza, V. L., D. Lindsley, A. Waalkes, E. Ramage, R. P. Patwardhan *et al.*, 2019 The *Rhododendron* genome and chromosomal organization provide insight into shared whole-genome duplications across the heath family (Ericaceae). Genome Biol. Evol. 11: 3353–

3371.

Staden, R., 1979 A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 6: 2601–10.

Stang, E. J., M. . Anderson, S. Pan, and J. Klueh, 1993 Lingonberry cultural management research in wisconsin, USA. Acta Hortic. 327–333.

Stanke, M., O. Schöffmann, B. Morgenstern, and S. Waack, 2006 Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62.

Statistics Canada. Area, production, and farm gate value of marketed fruits, 2022.

Stoler, N., and A. Nekrutenko, 2021 Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma. 3: lqab019.

Su, W., X. Gu, and T. Peterson, 2019 TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize Genome. Mol Plant 12: 447–60.

Sultana, N., G. Menzel, T. Heitkam, K. K. Kojima, W. Bao *et al.*, 2020 Bioinformatic and molecular analysis of satellite repeat diversity in *Vaccinium* genomes. Genes (Basel). 11:.

Sun, J., R. Li, C. Chen, J. D. Sigwart, and K. M. Kocot, 2021 Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 376: 1825 20200160.

Tang, W., X. Sun, J. Yue, X. Tang, C. Jiao *et al.*, 2019 Chromosome-scale genome assembly of kiwifruit *Actinidia eriantha* with single-molecule sequencing and chromatin interaction mapping. Gigascience 8: 4 giz027.

The Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

Tian, Y., Z. Ma, H. Ma, Y. Gu, Y. Li *et al.*, 2020 Comparative transcriptome analysis of lingonberry (Vaccinium vitis-idaea) provides insights into genes associated with flavonoids metabolism during fruit development. Biotechnol. Biotechnol. Equip. 34: 1252–1264.

Tirmenstein, D., 1991 Vaccinium vitis-idaea. Fire Eff. Inf. Syst.

Torkamaneh, D., B. Boyle, and F. Belzile, 2018 Efficient genome-wide genotyping strategies and data integration in crop plants. Theor. Appl. Genet. 131: 499–511.

Turner, N. J., 1978 *Food Plants of British Columbia Indians–Part 2/Interior Peoples*. British Columbia Provincial Museum, Victoria, BC.

Turner, N. J., 1975 *Food Plants of British Columbia Indians Part 1/Coastal Peoples*. British Columbia Provincial Museum, Victoria, BC.

Turner, N. J., 2004 *Plants of Haida Gwaii*. Sononis Press, Winlaw, BC.

Turner, N. J., L. J. Luczaj, P. Migliorini, A. Pieroni, A. L. Dreon *et al.*, 2011 Edible and tended wild

plants, traditional ecological knowledge and Agroecology. CRC. Crit. Rev. Plant Sci. 30: 198–225.

U.S. Department of Agriculture Natural Resources Conservation Service, 2021 The PLANTS Database.

U.S. Department of Agriculture National Agricultural Statistics Service, 2021 Noncitrus fruits and nuts 2020 summary. Noncitrus Fruits Nuts 2020 Summ. 107.

Vaara, M., O. Saastamoinen, and M. Turtiainen, 2013 Changes in wild berry picking in Finland between 1997 and 2011. Scand. J. For. Res. 28: 586–595.

Varshney, R. K., R. Terauchi, and S. R. McCouch, 2014 Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLOS Biol. 12: e1001883.

Vaser, R., I. Sović, N. Nagarajan, and M. Šikić, 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746.

Vilkickyte, G., V. Motiekaityte, R. Vainoriene, and L. Raudone, 2022 Promising cultivars and intraspecific taxa of lingonberries (*Vaccinium vitis-idaea* L .): profiling of phenolics and triterpenoids. J. Food Compos. Anal. 114:.

Vilkickyte, G., and L. Raudone, 2021 Phenological and geographical effects on phenolic and triterpenoid content in *Vaccinium vitis-idaea* L. leaves. Plants 10:.

Vuruputoor, V. S., D. Monyak, K. C. Fetter, C. Webster, A. Bhattarai *et al.*, 2022 Welcome to the big leaves: best practices for improving genome annotation in non-model plant genomes. bioRxiv 2022.10.03.510643.

Wakui, A., and G. Kudo, 2021 Ecotypic differentiation of a circumpolar Arctic-alpine species at mid-latitudes: Variations in the ploidy level and reproductive system of *Vaccinium vitis-idaea*. AoB Plants 13: 1–13.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963.

Wang, J., K. Chen, Q. Ren, Y. Zhang, J. Liu *et al.*, 2021 Systematic comparison of the performances of de novo genome assemblers for Oxford Nanopore technology reads From piroplasm . Front. Cell. Infect. Microbiol. 11:.

Wang, B., X. Yang, Y. Jia, Y. Xu, P. Jia *et al.*, 2022 High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long reads. Genomics Proteomics Bioinformatics 20: 4–13.

Waterhouse, R. M., F. Tegenfeldt, J. Li, E. M. Zdobnov, and E. V Kriventseva, 2013 OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 41: D358–D365.

Weisman, C. M., A. W. Murray, and S. R. Eddy, 2022 Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. Curr. Biol. 32: 2632-2639.e2.

Wenger, A. M., P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall *et al.*, 2019 Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37: 1155–1162.

Whibley, A., J. L. Kelley, and S. R. Narum, 2021 The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. Mol. Ecol. Resour. 21: 641–652.

Wu, C., C. Deng, E. Hilario, N. W. Albert, D. Lafferty *et al.*, 2021 A chromosome-scale assembly of the bilberry genome identifies a complex locus controlling berry anthocyanin composition. Mol. Ecol. Resour. 1–16.

Wu, B., Q. Yu, Z. Deng, Y. Duan, F. Luo *et al.*, 2023 A chromosome-level phased genome enabling allele-level studies in sweet orange: a case study on citrus Huanglongbing tolerance. Hortic. Res. 10: uhac247.

Xiong, W., L. He, J. Lai, H. . Dooner, and C. Du, 2014 HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc. Natl. Acad. Sci. United States Am. 111: 10263–8.

Xu, Z., and H. Wang, 2007 LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35:.

Yang, L., M. Li, M. Shen, S. Bu, B. Zhu *et al.*, 2022 Chromosome-level genome assembly and annotation of the native Chinese wild blueberry *Vaccinium bracteatum*. Fruit Res. 2:.

Yoshida, K., D. Ma, and C. P. Constabel, 2015 The MYB182 protein down-regulates proanthocyanidin and anthocyanin biosynthesis in Poplar by repressing both structural and regulatory flavonoid genes . Plant Physiol. 167: 693–710.

Yu, J., A. M. Hulse-Kemp, E. Babiker, and M. Staton, 2021 High-quality reference genome and annotation aids understanding of berry development for evergreen blueberry (*Vaccinium darrowii*). Hortic. Res. 8:.

Zhang, S., J. Chen, C. Zhang, S. Zhang, X. Zhang *et al.*, 2023a Insights into identifying resistance genes for cold and disease stresses through chromosome-level reference genome analyses of *Poncirus polyandra.* Genomics 115: 110617.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab, 2018 ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19: 153.

Zhang, Y., Y. Wei, J. Meng, Y. Wang, S. Nie *et al.*, 2023b Chromosome-scale de novo genome assembly and annotation of three representative *Casuarina* species: *C. equisetifolia, C. glauca,* and *C. cunninghamiana*. Plant J. 114: 1490–1505.

Zhao, Y., M.-C. Li, M. M. Konaté, L. Chen, B. Das *et al.*, 2021 TPM, FPKM, or normalized counts? a comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. J. Transl. Med. 19: 269.

Zhao, M., J. Li, L. Zhu, P. Chang, L. Li *et al.*, 2019 Identification and characterization of MYB-bHLH-WD40 regulatory complex members controlling anthocyanidin biosynthesis in blueberry fruits development. Genes (Basel). 10: 1–11.

Zhu, L., Y. Zhang, Y. Li, H. Wang, G. Shen *et al.*, 2022 Inhibitory effect of lingonberry extract on

HepG2 cell proliferation, apoptosis, migration, and invasion. PLoS One 17: e0270677.

Appendix A: Sequencing data summary tables

Table A1: Oxford Nanopore (ONT) MinION reads output from this study.

RUN	SAMPLE ID	Genus	Species	Flow Cell ver.	FLOW CELL TYPE	LIB KIT	# USED (how many times)	# PORES	MinKNOW ver.	DATE sequenced	DURATION (h)	MODE of sequencing	DATA (fast5, GB)	est. BASE PAIR (Gbp)	>Q10 (Gbp)	N50 (kbp)	% duplex	NOTES
1	RedCandy_SREXS_01_May5_2022	Vaccinium	vitis-idaea var. Red Candy	R9.4	FLO-MIN106D	SQK-LSK110	0	1200	21.11.7	05-May-22	72	NA	137	17.2	12.3	20.5	NA	
2	RedCandy_SREXS_02_May10_2022	Vaccinium	vitis-idaea var. Red Candy	R9.4	FLO-MIN106D	SQK-LSK110	1	3021	21.11.7	10-May-22	48	NA	33	4.14	2.6	21.2	NA	
3	RedCandy_SREXS_03_May_13_2022	Vaccinium	vitis-idaea var. Red Candy	R9.4	FLO-MIN106D	SQK-LSK110	2	1551	21.11.7	13-May-22	48	NA	7.9	0.981	0.6	20.9	NA	
4	RedCandy_Takara_04_SREXS_1_29Sept2022	Vaccinium	vitis-idaea var. Red Candy	R10.4	FLO-MIN114	SQK-LSK114	0	1700	22.05.5	29-Sep-22	48	NA	76	8.34	7.27	18.8	8.7	
5	RedCandy_Takara_04_SREXS_02_Oct_3_2022	Vaccinium	vitis-idaea var. Red Candy	R10.4	FLO-MIN114	SQK-LSK114	1	6881	22.05.5	03-Oct-22	48	NA	67	7.43	6.49	24.1	7.1	
6	RedCandy_Takara_04_SREXS_03_Oct_6_2022	Vaccinium	vitis-idaea var. Red Candy	R10.4	FLO-MIN114	SQK-LSK114	2	6001	22.08.9	06-Oct-22	12	accurate (250bp)	31	2.57	2.33	21.6	9	stopped by accident (memory outage)
7	minus_Takara_01_SREXS_01_Oct6_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	0	1200	22.08.9	06-Oct-22	12	accurate (250bp)	36	2.95	2.75	21.9	12.1	stopped by accident (memory outage)
8	RedCandy_Takara_04_SREXS_03_Oct_7_2022	Vaccinium	vitis-idaea var. Red Candy	R10.4	FLO-MIN114	SQK-LSK114	2	6001	22.08.9	07-Oct-22	24	accurate (250bp)	39	3.04	2.69	18.9	6.8	restarted with loaded flow cell
9	minus_Takara_01_SREXS_01_1_Oct7_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	0	1200	22.08.9	07-Oct-22	24	accurate (250bp)	11	0.778	0.697	22.0	7.9	restarted with loaded flow cell
10	RedCandy_Takara_04_SREXS_03_Oct_8_2022	Vaccinium	vitis-idaea var. Red Candy	R10.4	FLO-MIN114	SQK-LSK114	3	~300	22.08.9	08-Oct-22	96	accurate (250bp)	15	1.13	0.981	18.5	4.5	

11	minus_Takara_02_SRE_01_25Oct2022_1	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	0	120 0	22.08.9	25-Oct-22	72	accurate (250bp)	89	6.53	5.97	22.37	11
12	minus_Takara_02_SRE_01_25Oct2022_2	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	1	400	22.08.9	25-Oct-22	72	accurate (250bp)	44	3.22	3.12	24.78	10.4
13	minus_Takara_02_SRE_01_6Dec_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	1	618	22.10.7	06-Dec-22	48	accurate (250bp)	43	3.12	2.61	24.2	9.1
14	minus_Takara_02_SRE_01_9Dec_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	0	149 0	22.10.7	09-Dec-22	68	accurate (250bp)	113	8.03	6.74	23.5	9
15	minus_Takara_02_nsSRE_01_9Dec_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	2	189	22.10.7	09-Dec-22	72	accurate (250bp)	23	1.7	1.45	24.3	6.6
16	minus_Takara_02_nsSRE_01_12Dec_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	1	793	22.10.7	12-Dec-22	45	accurate (250bp)	56	3.91	3.51	23.4	8.9
17	minus_Takara_02_nsSRE_01_14Dec_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	2	286	22.10.7	14-Dec-22	48	accurate (250bp)	19.7	1.34	1.19	22	7.6
18	minus_Takara_02_nsSRE_01_16Dec_2022	Vaccinium	vitis-idaea ssp. minus	R10.4	FLO-MIN114	SQK-LSK114	3	179	22.10.7	16-Dec-22	72	accurate (250bp)	9.5	0.644	0.548	23.1	6.0
												Total:	**Red Candy**	**405.9**	**44.831**	**35.261**	**-**	**-**
													minus	**444.2**	**32.222**	**28.585**	-	-
												Average:	**Red Candy**	50.74	5.60	4.41	**20.56**	**7.22**
													minus	44.42	3.22	2.86	**23.16**	**8.86**

Table A2: Illumina reads output from this study.

RUN	SAMPLE ID	Genus	Species	PLATFORM	READ TYPE	LIB KIT	Target reads per sample (million)	DATE sequenced	RAW DATA (GB)	FILTERED DATA (GB)
1	RedCandy_leaf_RNA	Vaccinium	vitis-idaea var. Red Candy	Ilumina NovaSeq	PE150 bp	PolyA+ mRNA	50	09-Mar-23	52.47	51.83
2	RedCandy_rootstalk_RNA	Vaccinium	vitis-idaea var. Red Candy	Ilumina NovaSeq	PE150 bp	PolyA+ mRNA	50	09-Mar-23	46.44	45.84
3	RedCandy_flower_RNA	Vaccinium	vitis-idaea var. Red Candy	Ilumina NovaSeq	PE150 bp	PolyA+ mRNA	50	09-Mar-23	76.76	75.84
4	RedCandy_berry_RNA	Vaccinium	vitis-idaea var. Red Candy	Ilumina NovaSeq	PE150 bp	PolyA+ mRNA	50	09-Mar-23	79.20	78.29
5	RedCandy_DNA	Vaccinium	vitis-idaea var. Red Candy	Ilumina NovaSeq	PE150 bp	PCR-Free Genome	75	13-Feb-23	36.27	35.32
6	RedCandy_DNA	Vaccinium	vitis-idaea var. Red Candy	Ilumina NovaSeq	PE150 bp	PCR-Free Genome	-	20-Mar-23	48.97	47.46
7	minus_DNA	Vaccinium	vitis-idaea ssp. minus	Ilumina NovaSeq	PE150 bp	PCR-Free Genome	75	13-Feb-23	72.59	70.67
							Total:	**Red Candy (DNA)**	**85.24**	**82.79**
								minus (DNA)	**72.59**	**70.67**
								Red Candy (RNA)	**254.86**	**251.80**

Appendix B: Raw reads quality on ONT MinION

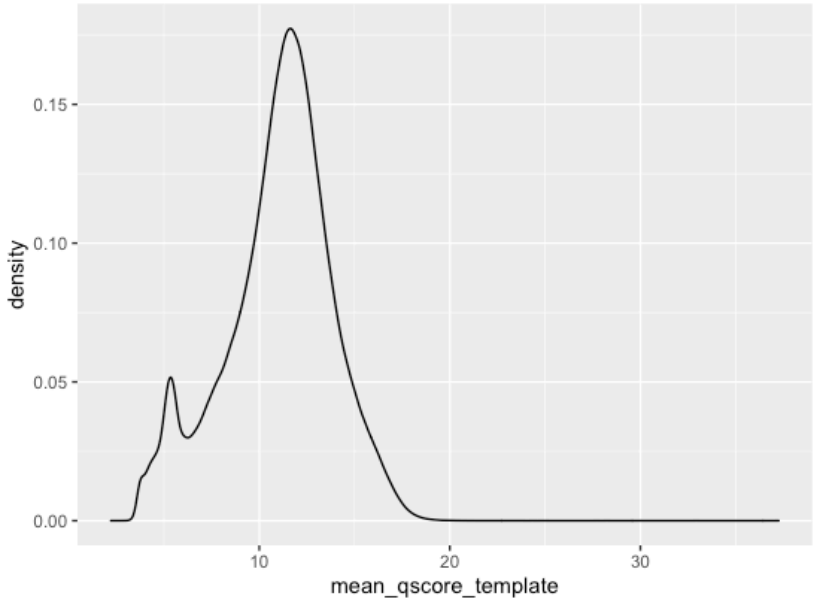

Figure B1: Raw reads Qscore distribution for MinION R9.4 results. The plot is generated with the combined *V. vitis-idaea* var. 'Red Candy' raw reads data before quality filtering. Basecall was done with guppy v6.1.2+e0556ff with "sup" model.

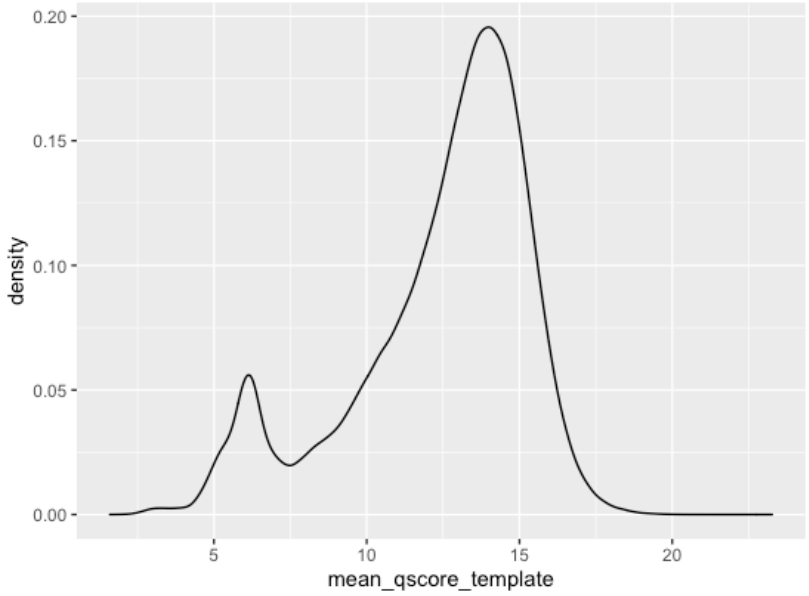

Figure B2: Example Qscore distribution for a MinION R10 run with "slow" transversion speed mode simplex reads only. The plot is generated with the first run on *V. vitis-idaea* var. 'Red Candy' raw reads data on R10 before quality filtering. Basecall was done with Guppy v6.1.2+e0556ff with "sup" model, and duplexed reads were removed to generate this plot.

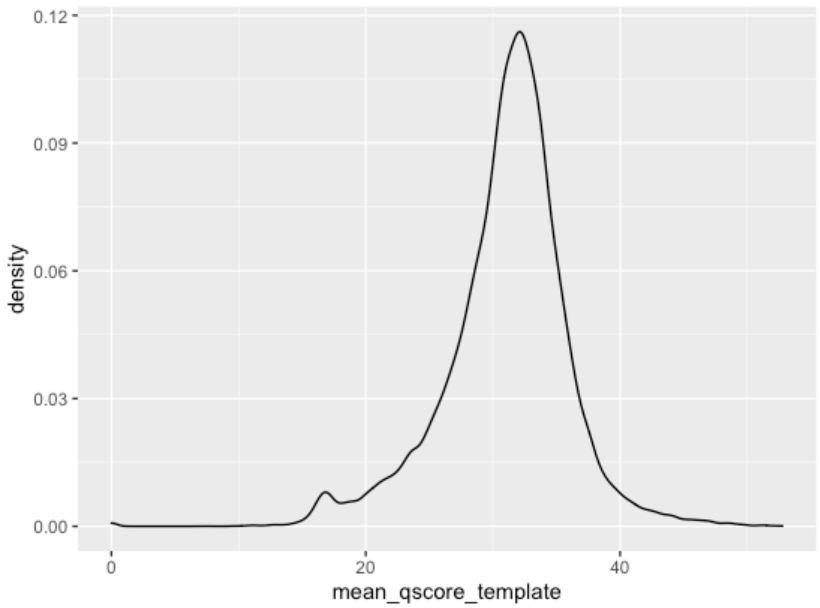

Figure B3: Example Qscore distribution for a MinION R10 run with "slow" transversion speed mode, duplex reads only. The plot is generated with the first run on *V. vitis-idaea* ssp. *minus* raw reads data on R10 before quality filtering. Basecall was done with Guppy Duplex-basecalling pipeline v6.3.8+d9e0f64, and only duplexed reads are plotted.