Chromosome-level assembly of the octoploid cloudberry (*Rubus chamaemorus*)

Authors: Marius A. Strand[1,2], Thu-Hien To[2], Ole K. Tørresen[1], Sarah Fullmer[1], Morten Skage[1], Ave Tooming-Klunderud[1], Simen Rød Sandve[2], Kjetill S. Jakobsen[1]*
Addresses:
1) Dept of Biosciences, Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, Norway; 2) CIGENE, Norwegian University of Life Sciences
*corresponding author

Abstract

We describe a chromosome-level genome assembly from an individual male plant of the cloudberry (*Rubus chamaemorus*). The haplotype-resolved assemblies contain one pseudo-haplotype spanning 1198 megabases and one pseudo-haplotype spanning 1161 megabases. Most of these two assemblies, 93.57% and 96.55% respectively, are each scaffolded into 28 pseudo-chromosomes. Both assemblies show high completeness, with the same BUSCO completeness score of 99.2%. Most BUSCO genes are duplicated in both pseudo-haplotypes, in line with the polyploid nature of the cloudberry genome. The assemblies contain 74,132 and 70,692 predicted protein-coding genes, respectively. Analysis of repetitive sequences classified ~60% of each haplotype as repeats. Comparative synteny with red raspberry (*Rubus idaeus*) reveals a 4:1 chromosome correspondence, supporting an octoploid origin. Ks distributions and k-mer clustering indicate a recent polyploidization involving one divergent (β) and three closely related (α) subgenomes. Genome skimming data mapped to our assembly link the α-subgenomes to a putative hexaploid lineage related to *Rubus arcticus*, and the β-subgenome to *Rubus pedatus*, previously proposed as the maternal ancestor.

Keywords: Earth Biogenome Project Norway, haplotype-resolved, HiFi sequencing, octoploid, polyploidisation

Main Text

Introduction

The cloudberry (*Rubus chamaemorus*) is a circumpolar boreal flowering plant species in the rose family (Urbaniak, Kwiatkowski, and Pawlikowski 2023). Within the rose family (subfamily Rosoideae) *Rubus* is a diverse genus with an estimated number of species ranging from 250 - 1000 (Huang et al. 2023) found on all continents except Antarctica. Many species of *Rubus* are known for their aggregate fruits such as blackberries, raspberries, dewberries and several hybrids. Aggregate fruits of *R. chamaemorus* are golden-yellow, soft and juicy, rich in vitamin C, ellagic acid, citric acid and anthocyanins, giving it a distinctive tart taste. They are used in jams, desserts, and liqueurs, and are considered a delicacy, particularly in Nordic countries. Despite great demand, *R. chamaemorus* is not commonly cultivated and is primarily harvested as a wild plant.

Unlike most *Rubus* species, *R. chamaemorus* has biparental (dioecious) reproduction implying distinct unisexual individuals, each producing either male or female gametes. Dioecy is a strategy to avoid self-fertilization and promote outcrossing. Within *Rubus,* around 60-70% of the species are polyploid ranging from 4x up to 14x (Thompson 1997; Hummer, Bassil, and Alice 2016). *R. chamaemorus* is an octoploid with eight genome copies. Evidence from nuclear and chloroplast phylogenies suggests a complex hybrid origin (Michael 2006; Carter et al. 2019). Michael (2006) found cpDNA support for *R. pedatus* as the maternal ancestor, and nuclear GBSSI-1γ sequences in *R. chamaemorus* revealed two divergent copies: one clustering with *R. lasiococcus*, the other with *R. arcticus*. This is consistent with Carter et al. (2019), who placed these species in an early-diverging lineage outside the main Rubus clade. *R. chamaemorus* has also received interest for potential cultivar development (Martinussen et al. 2013) as well as regional conservation efforts (Urbaniak, Kwiatkowski, and Pawlikowski 2023).

Despite this interest, as well as the demand as a delicacy, no genome assembly has been made available to date. Here, we have used long-read HiFi (PacBio) sequencing combined with long range chromosomal contact mapping by Hi-C, to create the first high-quality chromosome-level haplotype-resolved genome assembly of *R. chamaemorus*. The resulting genome consists of two haplotype-resolved assemblies spanning 1198 Mb and 1161 Mb, respectively, with 93.57% and 96.55% of the estimated genome scaffolded into 28 pseudo-chromosomes. Gene annotation identified 74,132 and 70,692 protein-coding genes, respectively. The two haplotype-resolved assemblies provide a genomic foundation for future genetic and evolutionary studies of *R. chamaemorus.*

Methods

Sample acquisition and DNA extraction

The sample was derived from an adult male plant collected in a forested area of Ås, Norway (59.670983,10.789649) on the 29th of May 2022. Sampling of this plant material does not require a specific permit under Section 15 of the Norwegian Nature Diversity Act, provided that it does not threaten the viability of the population and is not otherwise restricted.

DNA isolation for PacBio long read sequencing was performed using Circulomics Nanobind CBB BIG DNA kit and protocol according to manufacturer's recommendations, including treatment with EtOH removal buffer (Circulomics, now PacBio company). Quality check of amount, purity and integrity of isolated DNA was performed using Qubit BR DNA quantification assay kit (Thermo Fisher), Nanodrop (Thermo Fisher), and Fragment Analyser (DNA HS 50kb large fragment kit, Agilent Tech.).

Library preparation and sequencing for *de novo* assembly

Before PacBio HiFi library preparation, DNA was purified an additional time using AMPure PB beads (1:1 ratio). Purified HMW DNA was sheared into an average fragment size of 15-20 kbp large fragments using the Megaruptor3 (Diagenode). A HiFi library was prepared following the PacBio protocol for HiFi library preparation using the SMRTbell® Prep Kit 3.0. The final HiFi library was size-selected with a 10 kbp cut-off using a BluePippin (Sage Biosciences) and sequencing was performed by the Norwegian Sequencing Centre on a PacBio Sequel IIe instrument. The library was sequenced on two 8M SMRT cells using a Sequel II Binding kit 3.2 and Sequencing chemistry v2.0.

A Hi-C library was prepared using the Arima High Coverage HiC kit, following the manufacturer's recommendations and starting with 1050 mg leaf tissue. Chromatin was crosslinked prior to nuclei isolation as written in the user guide for plant tissues (Doc A160163 v01). Final library quality was quantified using a Kapa Library quantification kit for Illumina (Roche Inc.). The library was sequenced with other libraries on an Illumina NovaSeq 6000 with 2*150 bp paired end mode at the Norwegian Sequencing Centre.

Genome assembly and curation, annotation and evaluation

A full list of relevant software tools and versions is presented in Table 1. We assembled the species using a pre-release of the EBP-Nor genome assembly pipeline (https://github.com/ebp-nor/GenomeAssembly). KMC (Kokot, Dlugosz, and Deorowicz 2017) was used to count k-mers of size 32 in the PacBio HiFi reads, excluding k-mers occurring more than 10,000 times. GenomeScope (Ranallo-Benavidez, Jaron, and Schatz 2020) was run as part of the pipeline on the k-mer histogram output from KMC and was included in the methods for completeness, but as it does not support ploidy above 6, we do not report its results. Ploidy level was investigated using Smudgeplot (Ranallo-Benavidez, Jaron, and Schatz 2020). HiFiAdapterFilt (Sim et al. 2022) was applied on the HiFi reads to remove possible remnant PacBio adapter sequences. The filtered HiFi reads were assembled using Hifiasm (Cheng et al. 2021) with Hi-C integration resulting in a pair of haplotype-resolved assemblies, pseudo-haplotype one (hap1) and pseudo-haplotype two (hap2). Unique k-mers

in each assembly/pseudo-haplotype were identified using meryl (Rhie et al. 2020) and used to create two sets of Hi-C reads, one without any k-mers occurring uniquely in hap1 and the other without k-mers occurring uniquely in hap2.

K-mer filtered Hi-C reads were aligned to each scaffolded assembly using BWA-MEM (Li 2013) with `-5SPM` options. The alignments were sorted based on name using samtools (Li et al. 2009) before applying samtools fixmate to remove unmapped reads and secondary alignments and to add mate score, and samtools markdup to remove duplicates. The resulting BAM files were used to scaffold the two assemblies using YaHS (Zhou, McCarthy, and Durbin 2023) with default options. FCS-GX (Astashyn et al. 2024) was used to search for putative contamination and contaminated sequences were removed. The mitochondrion was searched for in contigs and reads using Oatk, varying the expected coverage parameter (Zhou et al. 2024). The assemblies were manually curated using PretextView and Rapid curation 2.0. Chromosomes were identified by mapping to *Rubus idaeus* JASCWY01 (Price et al. 2023) and *Fragaria vesca* v4 (Edger et al. 2018) in addition to inspecting the Hi-C contact map in PretextView.

Table 1. Software tools: versions and sources.

Software tool	**Version**	**Source**
BlobToolKit	4.3.2	https://github.com/blobtoolkit/blobtoolkit
blobtk	0.5.8	https://github.com/blobtoolkit/blobtk
BUSCO	5.4.7	https://gitlab.com/ezlab/busco
Hifiasm	0.19.8	https://github.com/chhylp123/hifiasm
KMC	3.2.1	https://github.com/refresh-bio/KMC
GenomeScope	2.0	https://github.com/tbenavi1/genomescope2.0
HiFiAdapterFilt	2.0.1	https://github.com/sheinasim/HiFiAdapterFilt
PretextView	1.0.1	https://github.com/sanger-tol/PretextView
PretextMap	0.1.9	https://github.com/sanger-tol/PretextMap
PretextSnapshot	0.0.4	https://github.com/sanger-tol/PretextSnapshot

meryl	1.3.0	https://github.com/marbl/meryl
BWA-MEM	0.7.17	https://github.com/lh3/bwa
samtools	1.17	https://github.com/samtools/samtools
YaHS	1.2a.2	https://github.com/c-zhou/yahs
FCS-GX	0.4.0	https://github.com/ncbi/fcs
Merqury	1.3	https://github.com/marbl/merqury
AGAT	1.4.0	https://github.com/NBISweden/AGAT
OatK	1.0	https://github.com/c-zhou/oatk
miniprot	0.13	https://github.com/lh3/miniprot
GALBA	1.0.9	https://github.com/Gaius-Augustus/GALBA
RED	2018.09.10	https://github.com/BioinformaticsToolsmith/Red
Funannotate	1.8.17	https://github.com/nextgenusfs/funannotate
EvidenceModeler	2.1.0	https://github.com/EVidenceModeler/EVidenceModeler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
InterProScan	5.62-94.0	https://www.ebi.ac.uk/interpro/search/sequence/
EMBLmyGFF3	2.2	https://github.com/NBISweden/EMBLmyGFF3
Earl Grey	4.1.1	https://github.com/TobyBaril/EarlGrey
Rapid curation 2.0	964d17e997e00c69f25940cf96d3658bda631147	https://github.com/Nadolina/Rapid-curation-2.0,
GENESPACE	1.3.1	https://github.com/jtlovell/GENESPACE

We annotated the genome assemblies using a pre-release version of the EBP-Nor genome annotation pipeline (https://github.com/ebp-nor/GenomeAnnotation). First, AGAT (https://zenodo.org/record/7255559) `agat_sp_keep_longest_isoform.pl` and `agat_sp_extract_sequences.pl` were used on the *Arabidopsis thaliana* (TAIR10.1, GCF_000001735.4) genome assembly and annotation to generate one protein (the longest isoform) per gene. Miniprot (Li 2023) was used to align the proteins to the curated assemblies. UniProtKB/Swiss-Prot (UniProt Consortium 2023) release 2023_02 in addition to the Viridiplantae part of OrthoDB v11 (Kuznetsov et al. 2023) were also aligned separately to the assemblies. RED (Girgis 2015) was run via redmask (https://github.com/nextgenusfs/redmask) on the assemblies to mask repetitive areas. GALBA (Brůna, Li, et al. 2023; Buchfink, Xie, and Huson 2015; Hoff and Stanke 2019; Li 2023; Stanke et al. 2006) was run with the *Arabidopsis* proteins using the miniprot mode on the masked assemblies.

The `funannotate-runEVM.py` script from Funannotate was used to run EvidenceModeler (Haas et al. 2008) on the alignments of TAIR10.1 proteins, UniProtKB/Swiss-Prot proteins, Viridiplantae proteins and the predicted genes from GALBA. The predicted proteins were compared to the protein repeats that Funannotate distributes using DIAMOND blastp and the filtered based on this comparison using AGAT. The filtered proteins were compared to the UniProtKB/Swiss-Prot release 2023_02 using DIAMOND (Buchfink, Xie, and Huson 2015) `blastp` to find gene names and InterProScan (Jones et al. 2014) was used to discover functional domains. AGATs `agat_sp_manage_functional_annotation.pl` was used to attach the gene names and functional annotations to the predicted genes. EMBLmyGFF3 (Norling, Jareborg, and Dainat 2018) was used to combine the fasta files and GFF3 files into a EMBL format for submission to ENA.

All the evaluation tools have also been implemented in a pipeline, similar to assembly and annotation (https://github.com/ebp-nor/GenomeEvaluation). Merqury (Rhie et al. 2020) was used to assess the completeness and quality of the genome assemblies by comparing them to the k-mer content of both the Hi-C reads and PacBio HiFi reads. BUSCO (Manni et al. 2021) was used to assess the completeness of the genome assemblies by comparing against the expected gene content in the eudicots lineage. Gfastats (Formenti et al. 2022) was used to output different assembly statistics of the assemblies.

BlobToolKit and BlobTools2 (Laetsch and Blaxter 2017), in addition to blobtk were used to visualize assembly statistics. To generate the Hi-C contact map image, the Hi-C reads were mapped to the assemblies using BWA-MEM *(Li 2013)* using the same approach as above. Finally, PretextMap was used to create a contact map which was visualized using PretextSnapshot.

To characterize the differences between the two pseudo-haplotypes, we ran a genome alignment using minimap2 (Li 2017) on the homologous chromosomes. The resulting alignment was processed with the minimap2-included paftools.js producing a report listing the number of insertions, SNPs and indels between the two pseudo-haplotypes.

To demonstrate the duplicated nature of *R. chamaemorus*, we analyzed synteny between hap1 and *Rubus idaeus* (red raspberry cv. 'Autumn Bliss') using the R package pipeline GENESPACE (Lovell et al. 2022). *R. idaeus* was chosen as reference due to its close phylogenetic relationship and availability of a high-quality chromosome-scale assembly (Martinussen et al. 2013; Price et al. 2023). To improve orthogroup inference, we included additional *Rubus*: *Rubus occidentalis* (v3.0 - black raspberry) and *Rubus argutus* ('Hillquist' v1.0 - blackberry), as well as more distantly related Rosaceae members: *Fragaria vesca* (v4.0.a1 - wild strawberry) and *Rosa chinensis* ('Old Blush' v2.0 - Chinese rose). Assemblies, protein sequences and annotations were obtained from the Rosaceae database (https://www.rosaceae.org/). We selected the longest protein isoform as the representative for each gene, along with gene rank orders, to serve as input for GENESPACE. The pipeline was run with the default method, which uses OrthoFinder (Emms and Kelly 2019) to build orthogroups and MCScanX (Wang et al. 2012) to detect collinear regions. All other parameters were left at default. Synteny blocks between *R. chamaemorus* and *R. idaeus* were then visualized using the plot_riparian function of the GENESPACE package.

Subgenome structure in *R. chamaemorus* was assessed using Subphaser v1.2.6 (Jia et al. 2022). The analysis was run independently on each haplotype, with default parameters and the 7 sets of 4 homologous chromosomes (tetrads) defined via a configuration file.

The Doubletrouble R-package (Almeida-Silva and Van de Peer 2025) was used to calculate pairwise Ks values between segmental duplicates (i.e. duplicated genes in conserved syntenic blocks) in the hap1 assembly. Since we could not reliably group chromosomes into four subgenomes based on shared k-mer signatures found by Subphaser (Figure 4A), pairwise Ks values were instead stratified by chromosome tetrads.

To estimate divergence times, Ks values were filtered to exclude values greater than 2 to minimize the effects of substitutional saturation. For each chromosome pair within each tetrad, we calculated the median Ks and derived 95% confidence intervals using non-parametric bootstrapping with 1,000 replicates, implemented using the boot package in R. Comparisons with fewer than 10 valid Ks values were excluded. Divergence time (T) was estimated using the formula: $T = Ks / (2 \times \mu)$, where μ is the neutral mutation rate, assumed to be a constant 6.5×10^{-9} substitutions per site per year (Ossowski et al. 2010). Confidence intervals for divergence time were calculated by applying the same formula to the bootstrap-derived lower and upper bounds of Ks.

To investigate potential parental contributors to the Rubus chamaemorus subgenomes, we analyzed read mapping patterns from closely related diploid species. Raw Illumina paired-end reads from *Rubus lasiococcus*, *Rubus pedatus, Rubus arcticus, and Rubus saxatilis* were obtained from Kates et al. (2024) and Carter et al. (2019) (Supplementary table 4). To reduce mapping ambiguity among the similar α subgenomes, we kept only the longest α chromosome per tetrad, along with the full β subgenome. Reads were mapped to the reduced α+β genome using Bowtie2, and only high-confidence alignments (MAPQ ≥30) were retained to minimize uncertainty from mismapping across homeologous regions.

The text in Methods and parts of Results is based on a template we use for all the species we publish in the EBP-Nor project.

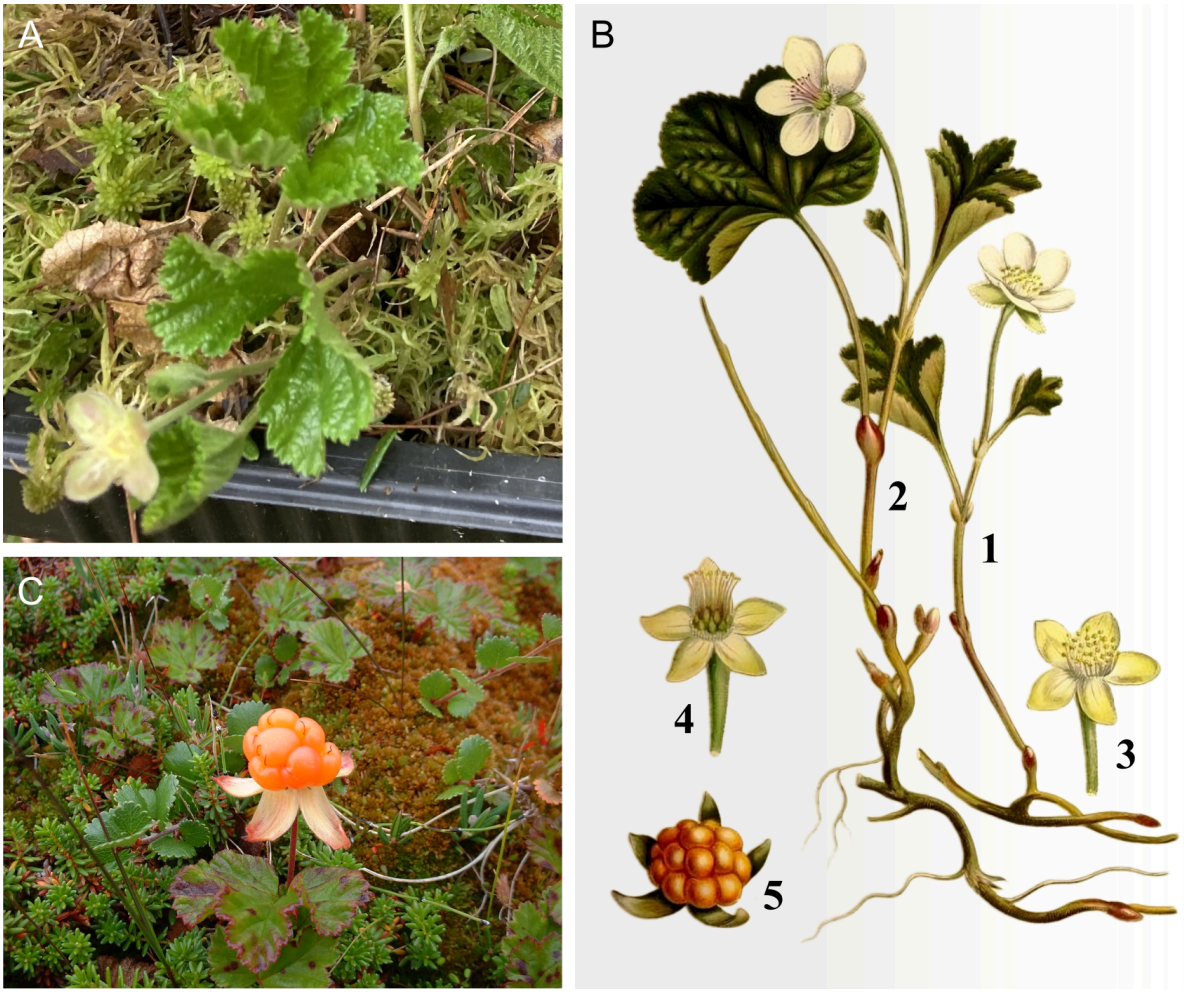

Figure 1: Illustration of dioecy in *Rubus chamaemorus* and the sequenced male specimen. *R. chamaemorus* is dioecious, with male and female flowers borne on separate plants. A) Our sample (a male plant) photographed late in the season (photo: Simen R. Sandve). B) Historical illustration: (1) male and (2) female plant in bloom; (3) and (4), male and female flower with petals removed; (5) calyx and aggregate fruit. (adapted from *Bilder ur Nordens Flora*, author: C.A.M. Lindman, via https://runeberg.org/nordflor/311.html, public domain). C) Female plant with ripe aggregate fruit (image from Wikipedia, author: Philipum, public domain).

Results

De novo genome assembly and annotation

The genome from *R. chamaemorus* was assembled from a total of 59-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 62-fold coverage in Arima Hi-C reads resulting in two haplotype-separated assemblies. The final assemblies have total lengths of 1198 Mb and 1161 Mb (Table 2 and Figure 2), respectively. Pseudo-haplotypes one (hap1) and two (hap2) have scaffold N50 size of 38.4 Mb and 40.0 Mb, respectively, and contig N50 of 29.4 Mb and 33.1 Mb, respectively (Table 2 and Figure 2). 28 pseudo-chromosomes were

identified in both pseudo-haplotypes. Chromosomes were first grouped by homology to the seven *Rubus idaeus* chromosomes, then ordered by length in hap1 within each group, and numbered 1–28. The corresponding homologs in hap2 were assigned the same numbers.

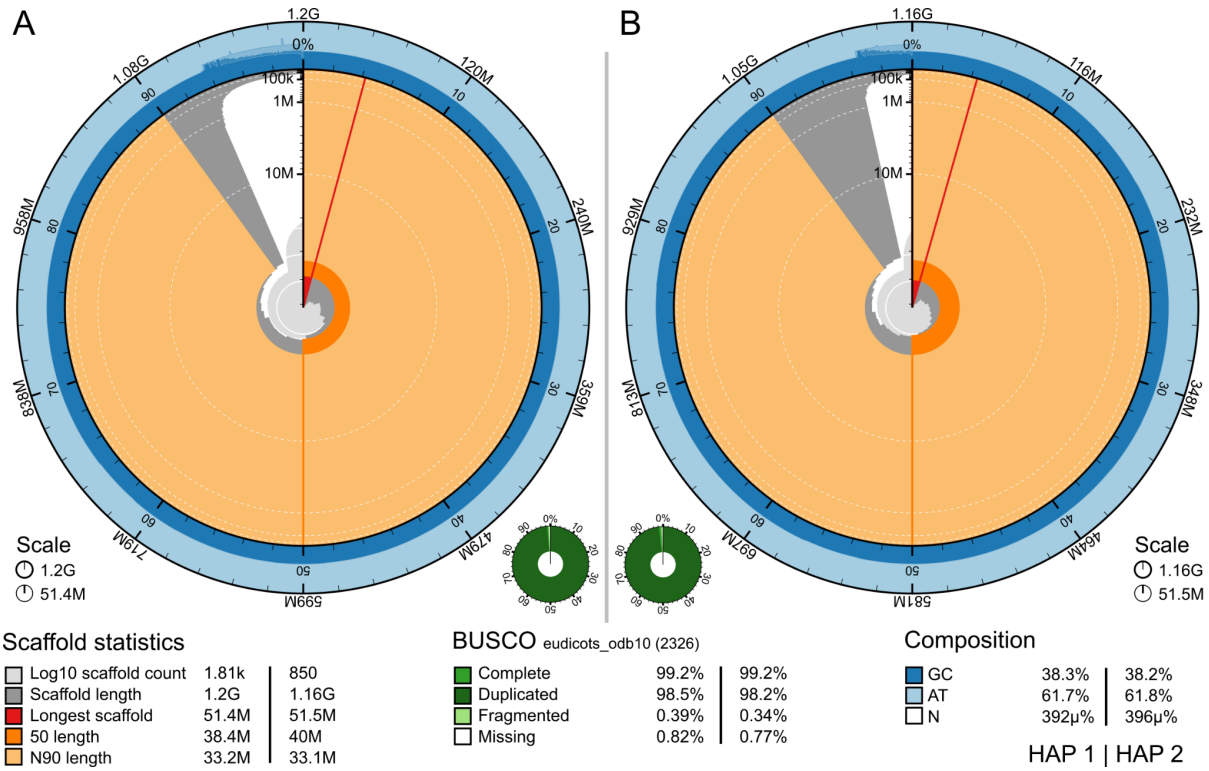

Figure 2: Metrics of the genome assemblies of *R. chamaemorus* hap1 and hap2. A) Hap1. B) Hap2. The BlobToolKit Snailplots show N50 metrics and BUSCO gene completeness. The two outermost bands of the circle signify GC versus AT composition at 0.1% intervals. Light orange shows the N90 scaffold length, while the deeper orange is N50 scaffold length. The red line shows the size of the largest scaffold. All the scaffolds are arranged in a clockwise manner from the largest to the smallest and are shown in darker gray with white lines at different orders of magnitude, while the light gray shows cumulative count of scaffolds.

Project accession data		
Species	*Rubus chamaemorus*	
Specimen	drRubCham1	
NCBI Taxonomy ID	57936	
Isolate information	Male, leaf	
Available data (DOI)	10.5281/zenodo.15773906	
Raw data accessions		
PacBio HiFi reads	To be submitted to ENA	2 PACBIO_SMRT (Sequel IIe) runs: 4.66 M reads, 70.8 Gb
Hi-C Illumina reads	To be submitted to ENA	1 ILLUMINA (Illumina NovaSeq 6000)

		run: 246 M pairs of reads, 74.3 Gb
Genome assembly metrics		
HiFi read coverage	59x	
Assembly accession	Pending	Pending
Assembly identifier	drRubCham1.1.hap1	drRubCham1.1.hap2
Span (Mb)	1198	1161
Number of contigs	1840	874
Contig N50 length (Mb)	29.4	33.1
Longest contig (Mb)	51.3	51.5
Number of gaps	25	24
Number of scaffolds	1815	850
Scaffold N50 length (Mb)	38.4	40.0
Longest scaffold (Mb)	51.4	51.5
Consensus quality (QV) compared to Hi-C (compared to HiFi)	45.9 (58.5)	47.0 (60.3)
Both assemblies	46.4 (59.3)	
k-mer completeness (%) compared to Hi-C (compared to HiFi)	96.0 (82.7)	94.0 (82.6)
Both assemblies	96.5 (98.6)	
Percentage of assembly mapped to chromosomes	93.6%	96.6%
Sex chromosomes	None	
Organelles*	MT1/MT2, PL	
Genome annotation metrics		
Number of protein-coding genes	74,132	70,692
Number of protein-coding genes	72,128	68,578

Table 2: Genome data for *R. chamaemorus.* *Both MT1 and MT2 likely represent partial representations of the full *R. chamaemorus* mitochondria. **Number of genes annotated with a functional domain as found by InterProScan. *** BUSCO scores based on the eudicots BUSCO set using v5.4.7. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

# BUSCO copies	*R. chamaemorus* hap 1	*R. chamaemorus* hap 2	*R. idaeus*
Missing	10 (0.43)	10 (0.43)	32 (1.38)
1	26 (1.12)	32 (1.38)	2143 (90.76)
2	66 (2.84)	58 (2.49)	172 (7.39)
3	467 (20.08)	461 (19.82)	10 (0.43)
4	1648 (70.85)	1659 (71.32)	1 (0.04)
>4	109 (4.69)	106 (4.56)	0
Total	2326	2326	2326

Table 3: Busco duplicates in *R. chamaemorus* compared to *R. idaeus.* Number and percent of BUSCO genes found in *R. chamaemorus* in comparison to *R. idaeus*.

Hap1 had 99.2% and hap2 99.2% complete BUSCO using the eudicots lineage set. Each pseudo-haplotype most commonly has 4 BUSCO copies, consistent with the known octoploid status of *R. chamaemorus* (Table 3). The next most frequent categories were 3 copies, likely reflecting gene loss (fractionation), and >4 copies, which may result from recent duplications or assembly noise. The presence of 2-copy BUSCOs (7.39%) in diploid *R. idaeus* suggests that some genes in the eudicot BUSCO set may not be strictly single-copy in this lineage.

When compared to a k-mer database of the Hi-C reads, hap1 had a k-mer completeness of 82.7%, hap2 of 82.6%, and combined they have a completeness of 98.6%. Further, hap1 had an assembly consensus quality value (QV) of 45.9 and hap2 of 47.0, where a QV of 40 corresponds to one error every 10,000 bp, or 99.99% accuracy compared to a k-mer database of the Hi-C reads (QV 58.5 and 60.3, respectively, compared to a k-mer database of the HiFi reads). When comparing the two pseudo-haplotypes using minimap2, there were 1,775,487 SNPs differences (2.098 % of the aligned sequence), 149,334 deletions in hap2 compared to hap1 ranging from 1 bp to more than 1000 bp and 149,763 insertions from 1

bp to more than 1000 bp in size (Supplementary Table 2). A total of 75,132 and 70,692 protein-coding genes were annotated in hap1 and hap2, respectively (Table 2).

The Hi-C contact maps show clear separation of chromosomes into homologous sets (Supplementary Figure 1). When sorted by similarity rather than size, each pseudo-haplotype reveals seven distinct groups of four.

A total of 74,132 and 70,692 protein-coding genes were annotated in hap1 and hap2, respectively. Repeat masking with RED (Girgis 2015) identified a high proportion of repetitive content, typical for plant genomes: 60.1% in hap1 and 59.6% in hap2.

Synteny analysis between *R. chamaemorus* and *R. idaeus* - the closest related diploid species with a high quality chromosome resolved assembly (Price et al. 2023; Martinussen et al. 2013) - revealed a consistent 4:1 chromosome correspondence, supporting an octoploid origin with largely conserved subgenome structure (Figure 3). While most *R. idaeus* chromosomes align directly with four *R. chamaemorus* homologs, several minor deviations indicate historical structural rearrangements. Chromosome 2 in *R. idaeus* shows near-complete synteny with *R. chamaemorus* chromosomes 5–8, but also a small syntenic link to chromosomes 21–24 (Figure 3B). A similar pattern is seen for *R. idaeus* chromosome 6, which aligns primarily with chromosomes 21–24 but also shows a small link to chromosomes 25–28, and a unique syntenic connection from *R. idaeus* chromosome 6 to *R. chamaemorus* chromosome 5. *R. idaeus* chromosome 3 has a smaller central region syntenic across all four homologs of *R. chamaemorus* chromosomes 9–12). There is also evidence of larger inversions. A large central inversion is the simplest explanation for the synteny break between R. idaeus chromosome 2 and chromosomes 7–8, and a similar inversion pattern is observed between chromosome 7 and chromosomes 26–28 (Figure 3B). *R. chamaemorus* chromosome 25 is structurally distinct, lacking this inversion.

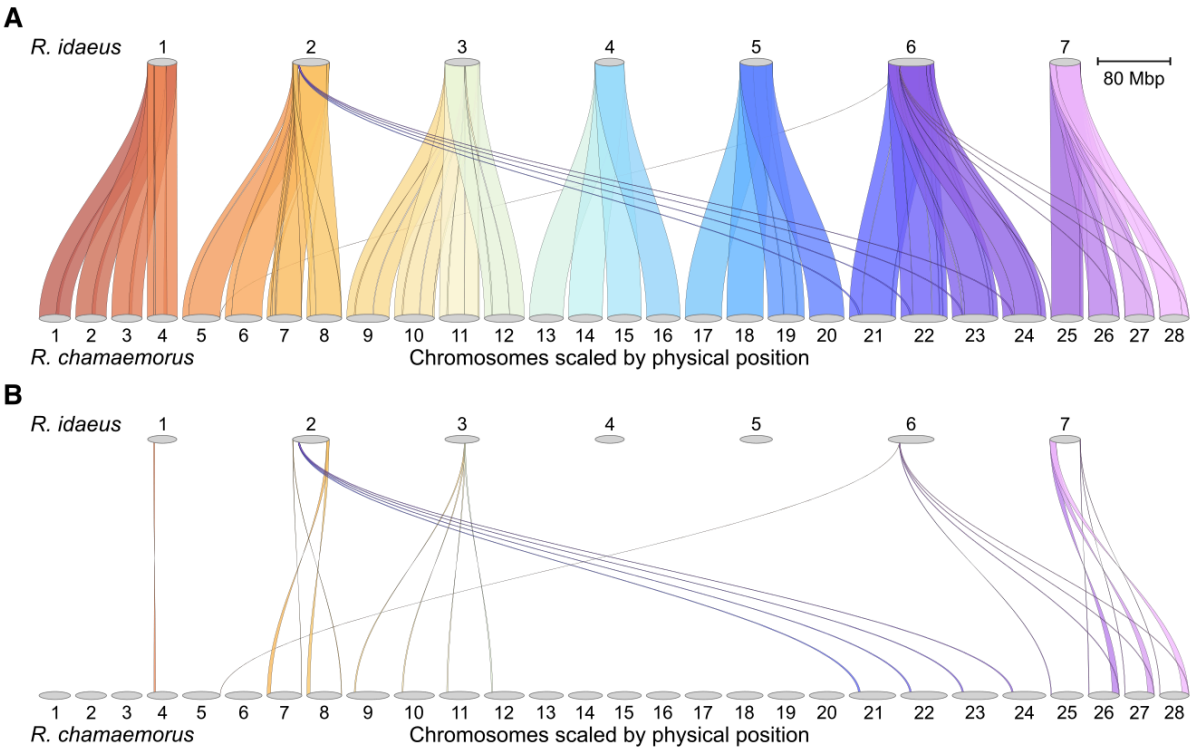

Figure 3: Synteny between *R. chamaemorus* and *R. idaeus*. Both panels are riparian plots generated using GENESPACE. Each *R. idaeus* chromosome aligns with four distinct *R. chamaemorus* hap1 chromosomes, consistent with an octoploid origin. Colors follow a continuous gradient for visual clarity. A) Full syntenic alignment. B) Same plot with conserved collinear blocks manually removed in Inkscape to highlight structural differences (e.g., inversions, translocations).

To explore the evolutionary history of polyploidization events in *R. chamaemorus*, we first used SubPhaser (Jia et al. 2022) to identify subgenome-specific k-mers and cluster homoeologous chromosomes into putative subgenomes (Figure 4A). The analysis identified one distinct subgenome of 7 chromosomes, hereafter termed the β-subgenome; the remaining 21 chromosomes, lacking such distinction, were collectively grouped as the α-subgenome(s).

We then estimated Ks distributions using doubletrouble (Almeida-Silva and Van de Peer 2025). The overall distribution showed a single, narrow peak at with only a slight shoulder on the right (Figure 4B). Ka/Ks ratios indicate strong purifying selection across most gene pairs, consistent with functional constraint and subgenome stability (Figure 5C). Stratifying the analysis by tetrad (7 sets of 4 syntenic chromosomes), 6 of the 7 showed a consistent pattern: one chromosome had higher pairwise Ks values relative to the other three (Figure 4D, Supplementary Table 3). These chromosomes correspond to the β-subgenome identified in the k-mer analysis. The 7th tetrad displayed a more complex pattern.

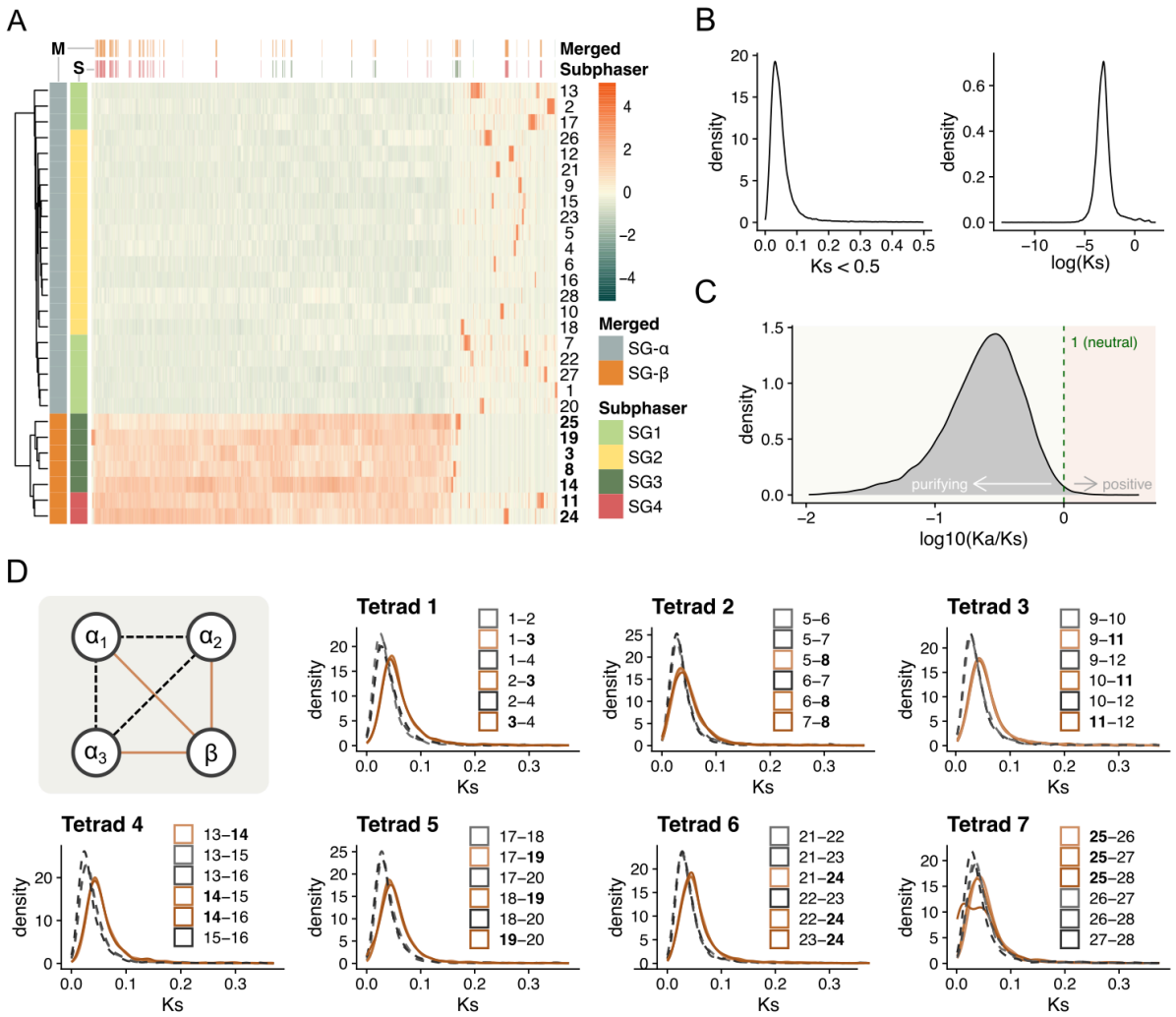

Figure 4: Subgenome structure inferred from k-mers and Ks. A) Unsupervised hierarchical clustering of k-mer profiles. Top bar: k-mer subgenome specificity. Side bar: chromosome subgenome assignments. "Merged" labels reflect our inferred α/β subgenomes based on a stricter interpretation of the clustering. Heatmap shows Z-scored relative abundance of 10k randomly sampled k-mers. B) Ks distribution of segmental duplicates (as defined by DoubleTrouble) across all chromosomes. Left: linear scale; right: log10-transformed Ks. C) Ka/Ks distribution (filtered for Ks > 0.05, Ka/Ks > 0.01), showing most duplicates are under strong purifying selection. D) Ks distributions of segmental duplicates within each of the seven syntenic pseudo-chromosome tetrads. The leftmost panel shows pairwise comparisons among the four homologs in each tetrad. Solid orange lines indicate comparisons between each α-homolog and the β-homolog. Stippled grey lines indicate comparisons among the three α-homologs.

To estimate the time since separation between the β-subgenome and the α-subgenomes, as well as within the α-subgenomes, we calculated divergence time (T) using the formula T = Ks / (2 × mutation rate), assuming a mutation rate of 6.5×10^{-9} substitutions/site/year. The overall trend (Supplementary Figure 4) shows that the β- and α-subgenomes diverged approximately 3.70 MYA (95% CI: 3.68–3.73; median Ks = 0.0481), while the median duplication age within the α-subgenomes is around 2.65 MYA (95% CI: 2.63–2.67; median Ks = 0.0345); roughly 1 million years later. Across tetrads, Ks values between α-subgenome chromosomes vary moderately but show no consistent clustering pattern, suggesting that divergence within the α-subgenome is relatively symmetric and lacks a clear hierarchy.

Tetrad 7 was an exception, showing greater variability and less internal consistency especially between chr26, chr25 and chr28.

R. pedatus, *R. lasiococcus* have been proposed as putative progenitor species of *R. chamaemorus* (Michael 2006; Carter et al. 2019), while *R. arcticus* shares a GBSSI-1γ copy with *R. chamaemorus (Michael 2006)*. *R. saxatilis,* which is known to hybridize with *R. arcticus (NatureGate, n.d.)*, was included due to its potential relevance based on this relationship. Although genome assemblies are not available for these species, limited sequencing data is available (Through Kates et al. (2024) and Carter et al. (2019); See Methods). To reduce ambiguity among the similar α-subgenomes, we selected one α-representative per tetrad based on length. We then quantified high-confidence alignments across the α- and β-subgenomes.

Using genome skimming data from Kates et al. (2024), we find that *R. arcticus* consistently shows stronger mapping to the reduced α-subgenome across all seven tetrads (Figure 5). This pattern is visible in the total read counts (Figure 5A), but is more evident in the log-transformed mapping ratios (β/α)(Figure 5B). In contrast, *R. lasiococcus*, *R. pedatus*, and *R. saxatilis* preferentially map to the β-subgenome in tetrads 1–6, and are nearly neutral in tetrad 7.

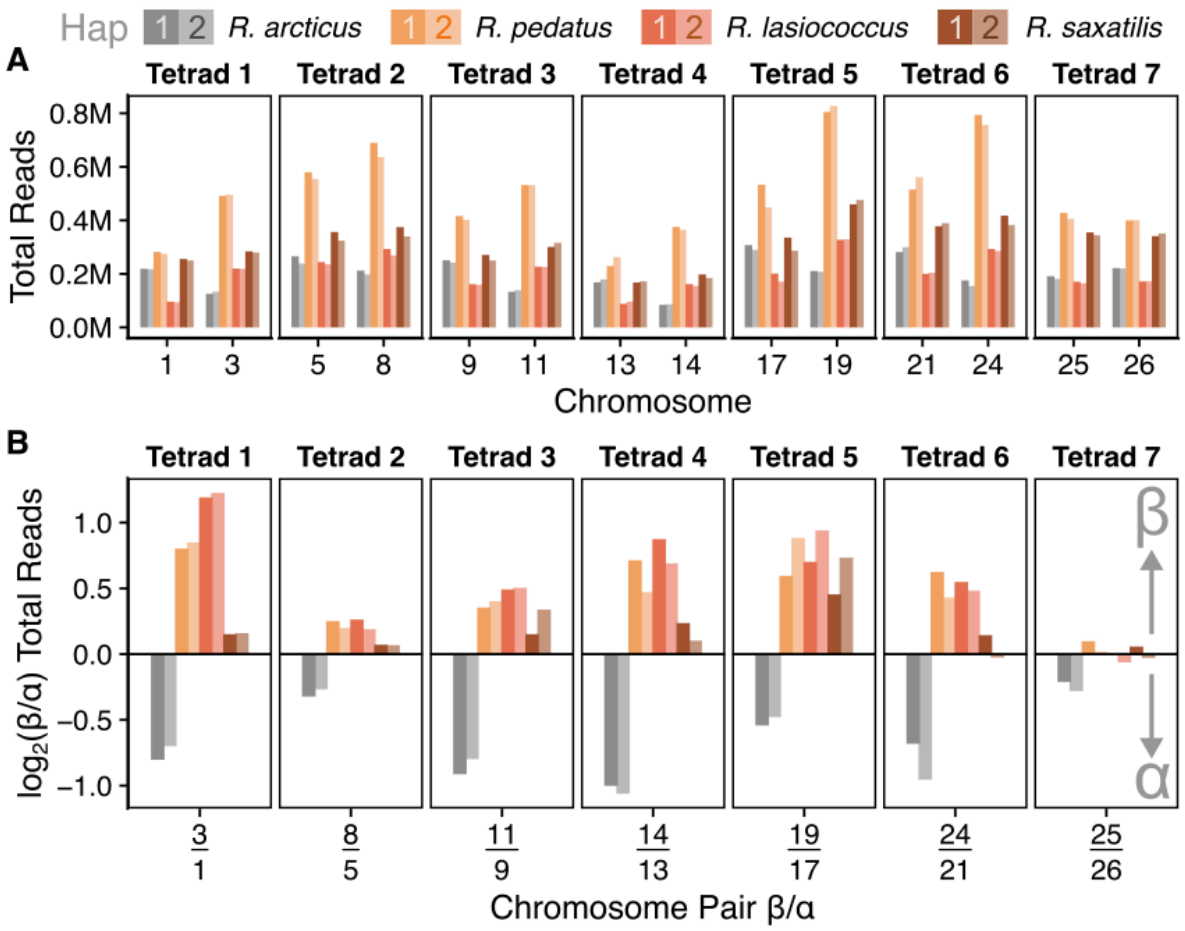

Figure 5: High-confidence read mappings (MAPQ ≥ 30) from Kates et al. 2024: *R. lasiococcus*, *R. pedatus*, *R. arcticus*, and *R. saxatilis* to the α and β subgenomes of *R. chamaemorus*. Reads from each species are

mapped to both *R. chamaemorus* haplotypes (hap1 and hap2). A) Total number of reads per species per chromosome per haplotype. B) Log_2 of the ratio of high-confidence reads mapped to subgenome β over subgenome α. Positive values indicate preferential mapping to subgenome β; negative values indicate bias toward subgenome α. A value of 1 indicates a twofold mapping bias toward subgenome β. Chromosomes from subgenome α: 1, 5, 9, 13, 17, 21, 26; from subgenome β: 3, 8, 11, 14, 19, 24, 25.

Subgenome Evolution Scenarios

Figure 6 illustrates three non-exhaustive possible scenarios of subgenome origin involving different combinations of hybridization and autotetraploidization events among closely related ancestral diploid lineages (α_1, α_2, α_3, and β). Each scenario represents a possible pathway to the formation of an octoploid with either three or four distinct subgenomes. Scenario 1 involves three successive hybridizations: $\alpha_3 \times \beta$ and $\alpha_1 \times \alpha_2$, followed by a cross between the resulting hybrids ($\alpha_3\beta \times \alpha_1\alpha_2$), leading to an octoploid with four distinct subgenomes. Scenario 2 includes one autotetraploidization of α_1 (forming $\alpha_1\alpha_1$), followed by hybridizations $\alpha_2 \times \beta$ and then $\alpha_1\alpha_1 \times \alpha_2\beta$, resulting in an octoploid with three different subgenomes. Scenario 3 begins with an autotetraploidization of α_1, followed by hybridization with α_2 ($\alpha_1\alpha_1 \times \alpha_2$), and a subsequent hybridization with the β-subgenome ($\alpha_1\alpha_1\alpha_2 \times \beta$), yielding an octoploid with three subgenomes, where all α-derived subgenomes are highly homogeneous due to their close evolutionary origin.

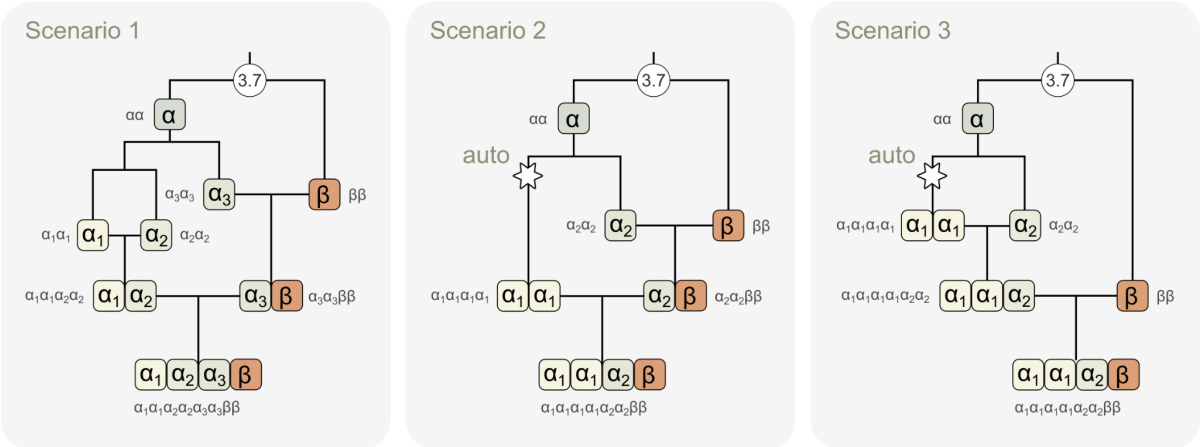

Figure 6: Possible scenarios of subgenome origin and polyploidisation events. For all scenarios, the α- and β-genomes split 3.7 MYA. Scenario 1: Three successive hybridizations (α_3 x β and α_1 x α_2 followed by $\alpha_3\beta$ x $\alpha_1\alpha_2$) creates 4 different subgenomes. Scenario 2: One autotetraploidization and two hybridizations (α_2 x β, followed by $\alpha_1\alpha_1$ x $\alpha_2\beta$) result in three different subgenomes. Scenario 3: An initial autotetraploidization is followed by hybridization with the original diploid ($\alpha_1\alpha_1$ x α_2), then a second hybridization with a more distant β-genome species ($\alpha_1\alpha_1\alpha_2$ x β) results in three different subgenomes where two are highly homogeneous.

Discussion

We present two chromosome-level, pseudo-haplotype resolved genome assemblies for R. chamaemorus, each spanning ~1.2 Gb with scaffold N50s >38 Mb. *R. chamaemorus* forms a distinct and basal lineage within *Rubus*, one of the most phylogenetically complicated angiosperm genera (Huang et al. 2023). These assemblies provide a genomic resource for comparative analyses within *Rubus* and broader studies of Rosaceae evolution and polyploidy.

Polyploid origin of *R. chamaemorus*

Polyploidy is common in Rubus, with 60–70% of species possessing chromosome numbers 4n or higher (Thompson 1997; Hummer, Bassil, and Alice 2016; Chen et al. 2020). Comparative synteny between *R. chamaemorus* and *R. idaeus* reveals a highly conserved subgenome structure, with four homologs corresponding to each of the seven basal chromosomes typical of Rubus.

K-mer clustering and Ks analyses suggest an evolutionary history in which three of the four subgenomes originate from closely related species or populations (α), while the fourth is more divergent (β) (Figure 4). Pairwise Ks values are uniformly low, suggesting a recent polyploid origin, with divergence between α- and β-subgenomes estimated at ~3.70 MYA (molecular clock rate 6.5×10^{-9} substitutions/site/year). Ks values among α-subgenome chromosomes indicate later divergence (median Ks = 0.0345, ~2.65 MYA), consistent with a shared or highly similar evolutionary origin, rather than three clearly distinct lineages. The overall narrow Ks distributions across tetrads point to rapid genome mergers and limited post-polyploid divergence. Segmental duplicates show low Ka/Ks ratios, consistent with widespread purifying selection.

Tetrad 7 deviates from the clear α–β subgenome pattern seen in other tetrads, showing inconsistent Ks signals and no clear subgenome separation. Ks distributions between chromosomes 25 and 28 show a distinct double peak (Supplementary Figure 3), indicating a unique evolutionary history between these two chromosomes, though misassembly cannot be ruled out. However, k-mer clustering robustly assigns chr25 to the β-subgenome, indicating that this deviation likely reflects biological complexity rather than assembly error.

Together, the Ks and k-mer analyses point to a recent allopolyploid origin involving one divergent β-subgenome and a set of closely related α-subgenomes. The lack of strong substructure among α chromosomes argues against scenarios requiring multiple distinct α-genome donors. Scenario 3 in Figure 5, involving α-genome autotetraploidization followed by hybridization with another α- and a β-genome, best explains the observed patterns among our selected scenarios. Scenarios 1 and 2 are less consistent with the data, as they imply greater divergence among α-subgenomes than observed.

The *Rubus* species *R. lasiococcus and R. pedatus* were first proposed by Michael (2006) as possible progenitors of *R. chamaemorus,* but our results suggest they are unlikely to have contributed to the α-subgenomes. Read mapping from *R. lasiococcus* and *R. pedatus* shows a clear preference for the β-subgenome, while the previously dismissed *R. arcticus* preferentially maps to the α-subgenomes (Figure 5B). Sparse coverage is likely to introduce noise, but β/α patterns remain consistent across tetrads. Tetrad 7 notably deviates from consistent subgenome patterns, with *R. pedatus* shifting slightly towards α and *R. lasiococcus* exhibiting reduced β-bias, likely reflecting ancestral recombination or introgression events. Targeted sequencing data from Carter et al. (2019) recapitulate the same trends for *R. lasiococcus, R. pedatus, and R. arcticus,* further validating the inferred subgenome relationships (Supplementary Figure 5). An exception is *R. saxatilis* which shows the weakest β-bias across most tetrads in the genome skimming data (Figure 5), while in the targeted sequencing data it exhibits an oscillating pattern across tetrads in the targeted dataset: alternating weak β and α signals, followed by a clear β-enrichment in tetrad 7

(Supplementary Figure 5). However, as this dataset includes only ~1,000 targeted low-copy nuclear genes, it may not fully reflect ancestry.

The α-subgenome likely stems from an ancestral hexaploid species related to *R. arcticus*, while the β-subgenome originates from a lineage closely related to *R. pedatus*. Previous work by Michael (2006) identified two divergent nuclear GBSSI-1γ copies in *R. chamaemorus*: one clustering with *R. arcticus*, the other with *R. lasiococcus*. However, Michael rejected *R. arcticus* as a direct progenitor based on ITS data, which placed it outside the clade containing *R. chamaemorus*, *R. pedatus*, and *R. lasiococcus*. High similarity among α-derived subgenomes suggests successive autopolyploidization and intraspecific hybridization events, culminating in an octoploid genome with a largely homogeneous α component. Plastid evidence implicates *R. pedatus* as the maternal donor (Michael 2006); together this data point to an hexaploid *R. arcticus* as the likely non-maternal contributor. Inconsistencies in Tetrad 7 may reflect additional introgression, homoeologous exchange, or incomplete lineage sorting, though the underlying cause remains uncertain. Resolving this fully will require high-quality genome assemblies of the putative relatives of R. chamaemorus. More broadly, such assemblies are essential for disentangling polyploid origins, identifying subgenome contributors, and enabling accurate comparative genomics across complex plant lineages.

Genomic value for cultivar improvement

The presented assemblies can support cultivar development. Knowledge of the polyploid genome structure, combined with a high-quality assembly and annotation, can accelerate breeding in *R. chamaemorus*, where conventional approaches are limited by dioecy and slow maturation (Martinussen et al. 2013). One promising target trait is primocane-fruiting (fruiting on first-year stems), which occurs naturally in many *Rubus* species, including raspberries and some blackberries (Jennings 2018). In blackberry, the diploid 'Hillquist' (*R. argutus*) enabled identification of candidate flowering-time genes for this trait (Brůna, Aryal, et al. 2023). Similar strategies may be explored in *R. chamaemorus*, using genomic data to identify flowering-related loci and reduce the long breeding cycle.

Funding

This project was funded by the Research Council of Norway project grant no 326819 (The Earth Biogenome Project Norway) to KSJ.

Acknowledgements

This project received data management and infrastructure support from ELIXIR Norway, supported by the Research Council of Norway's grant 270068, the University of Bergen, the University of Oslo, the Arctic University of Norway in Tromsø, the Norwegian University of Science and Technology and the Norwegian University of Life Sciences: NMBU. The authors acknowledge support from the National Infrastructure for High Performance Computing and resources provided by Sigma2 as well as Data Storage in Norway (project NN8013K) for computational work. The Norwegian Sequencing Centre generated the sequencing data used in this project (http://sequencing.uio.no).

Author contributions

Marius A. Strand: Writing - original draft, Formal analysis, Visualization, Writing review and editing. **Thu-Hien To**: Writing - original draft, Formal analysis, Visualization. **Ole K. Tørresen:** Writing - review and editing. **Sarah Fullmer:** Writing – review & editing. **Morten Skage:** Investigation. **Ave Tooming-Klunderud:** Investigation. **Simen Rød Sandve:** Resources, Writing - review and editing. **Kjetill S. Jakobsen:** Funding acquisition, Writing - review and editing.

Competing interests

The authors declare no competing interests.

Data Availability

Raw PacBio and Hi-C sequencing data will be submitted to ENA and made publicly available once accession numbers are assigned. Genome assemblies and gene annotations are available at **10.5281/zenodo.15773906**

References

Almeida-Silva, Fabricio, and Yves Van de Peer. 2025. "Doubletrouble: An R/Bioconductor Package for the Identification, Classification, and Analysis of Gene and Genome Duplications." *Bioinformatics (Oxford, England)* 41 (2): btaf043.

Astashyn, Alexander, Eric S. Tvedte, Deacon Sweeney, Victor Sapojnikov, Nathan Bouk, Victor Joukov, Eyal Mozes, et al. 2024. "Rapid and Sensitive Detection of Genome Contamination at Scale with FCS-GX." *Genome Biology* 25 (1): 60.

Brůna, Tomáš, Rishi Aryal, Olga Dudchenko, Daniel James Sargent, Daniel Mead, Matteo Buti, Andrea Cavallini, et al. 2023. "A Chromosome-Length Genome Assembly and Annotation of Blackberry (*Rubus Argutus*, Cv. 'Hillquist')." *G3 (Bethesda, Md.)* 13 (2): jkac289.

Brůna, Tomáš, Heng Li, Joseph Guhlin, Daniel Honsel, Steffen Herbold, Mario Stanke, Natalia Nenasheva, Matthis Ebel, Lars Gabriel, and Katharina J. Hoff. 2023. "Galba: Genome Annotation with Miniprot and AUGUSTUS." *BMC Bioinformatics* 24 (1): 327.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60.

Carter, Katherine A., Aaron Liston, Nahla V. Bassil, Lawrence A. Alice, Jill M. Bushakra, Brittany L. Sutherland, Todd C. Mockler, Douglas W. Bryant, and Kim E. Hummer. 2019. "Target Capture Sequencing Unravels *Rubus* Evolution." *Frontiers in Plant Science* 10 (December):1615.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. "Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm." *Nature Methods* 18 (2): 170–75.

Chen, Xun, Jinlu Li, Tao Cheng, Wen Zhang, Yanlei Liu, Ping Wu, Xueying Yang, Ling Wang, and Shiliang Zhou. 2020. "Molecular Systematics of Rosoideae (Rosaceae)." *Plant Systematics and Evolution* 306 (1): 1–12.

Edger, Patrick P., Robert VanBuren, Marivi Colle, Thomas J. Poorten, Ching Man Wai, Chad E. Niederhuth, Elizabeth I. Alger, et al. 2018. "Single-Molecule Sequencing and Optical Mapping Yields an Improved Genome of Woodland Strawberry (*Fragaria Vesca*) with Chromosome-Scale Contiguity." *GigaScience* 7 (2): 1–7.

Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for

Comparative Genomics." *Genome Biology* 20 (1): 238.
Formenti, Giulio, Linelle Abueg, Angelo Brajuka, Nadolina Brajuka, Cristóbal Gallardo-Alba, Alice Giani, Olivier Fedrigo, and Erich D. Jarvis. 2022. "Gfastats: Conversion, Evaluation and Manipulation of Genome Sequences Using Assembly Graphs." *Bioinformatics (Oxford, England)* 38 (17): 4214–16.
Girgis, Hani Z. 2015. "Red: An Intelligent, Rapid, Accurate Tool for Detecting Repeats de-Novo on the Genomic Scale." *BMC Bioinformatics* 16 (1): 227.
Haas, Brian J., Steven L. Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E. Allen, Joshua Orvis, Owen White, C. Robin Buell, and Jennifer R. Wortman. 2008. "Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments." *Genome Biology* 9 (1): R7.
Hoff, Katharina J., and Mario Stanke. 2019. "Predicting Genes in Single Genomes with AUGUSTUS." *Et Al [Current Protocols in Bioinformatics]* 65 (1): e57.
Huang, Ti-Ran, Jian-Hui Chen, Kim E. Hummer, Lawrence A. Alice, Wen-He Wang, Yi He, Sheng-Xiang Yu, et al. 2023. "Phylogeny of *Rubus* (Rosaceae): Integrating Molecular and Morphological Evidence into an Infrageneric Revision." *Taxon* 72 (2): 278–306.
Hummer, K. E., N. V. Bassil, and L. A. Alice. 2016. "*Rubus* ploidy Assessment." *Acta Horticulturae* 1133 (1133): 81–88.
Jennings, Sophia Nikki. 2018. "Advances in *Rubus* Breeding." In *Raspberry*, 17–28. Cham: Springer International Publishing.
Jia, Kai-Hua, Zhao-Xuan Wang, Longxin Wang, Guang-Yuan Li, Wei Zhang, Xiao-Ling Wang, Fang-Ji Xu, et al. 2022. "SubPhaser: A Robust Allopolyploid Subgenome Phasing Method Based on Subgenome-Specific K-Mers." *The New Phytologist* 235 (2): 801–9.
Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics (Oxford, England)* 30 (9): 1236–40.
Kates, Heather R., Brian C. O'Meara, Raphael LaFrance, Gregory W. Stull, Euan K. James, Shui-Yin Liu, Qin Tian, et al. 2024. "Shifts in Evolutionary Lability Underlie Independent Gains and Losses of Root-Nodule Symbiosis in a Single Clade of Plants." *Nature Communications* 15 (1): 4262.
Kokot, Marek, Maciej Dlugosz, and Sebastian Deorowicz. 2017. "KMC 3: Counting and Manipulating K-Mer Statistics." *Bioinformatics (Oxford, England)* 33 (17): 2759–61.
Kuznetsov, Dmitry, Fredrik Tegenfeldt, Mosè Manni, Mathieu Seppey, Matthew Berkeley, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2023. "OrthoDB v11: Annotation of Orthologs in the Widest Sampling of Organismal Diversity." *Nucleic Acids Research* 51 (D1): D445–51.
Laetsch, Dominik R., and Mark L. Blaxter. 2017. "BlobTools: Interrogation of Genome Assemblies." *F1000Research* 6 (July):1287.
Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.
———. 2017. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (August):3094–3100.
———. 2023. "Protein-to-Genome Alignment with Miniprot." *Bioinformatics (Oxford, England)* 39 (1): btad014.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics (Oxford, England)* 25 (16): 2078–79.
Lovell, John T., Avinash Sreedasyam, M. Eric Schranz, Melissa Wilson, Joseph W. Carlson, Alex Harkess, David Emms, David M. Goodstein, and Jeremy Schmutz. 2022. "GENESPACE Tracks Regions of Interest and Gene Copy Number Variation across Multiple Genomes." *eLife* 11 (September). https://doi.org/10.7554/eLife.78526.
Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021.

"BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes." Edited by Joanna Kelley. *Molecular Biology and Evolution* 38 (10): 4647–54.

Martinussen, Inger, Eivind Uleberg, Anita Sønsteby, Jørn Hernik Sønstebø, Julie Grahamand, and Adam Vivian-Smith. 2013. "Genomic Survey Sequences and the Structure of the *Rubus Chamaemorus* L. Genome as Determined by ddRAD Tags." https://www.researchgate.net/profile/Inger-Martinussen/publication/256504036_Genomic_survey_sequences_and_the_structure_of_the_Rubus_chamaemorus_L_genome_as_determined_by_ddRAD_tags/links/004635232f0ce7b8b4000000/Genomic-survey-sequences-and-the-structure-of-the-Rubus-chamaemorus-L-genome-as-determined-by-ddRAD-tags.pdf.

Michael, Karen. 2006. "Clarification of Basal Relationships in *Rubus* (Rosaceae) and the Origin of *Rubus Chamaemorus*." Masters Theses & Specialist Projects., Western Kentucky University. https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1253&context=theses.

NatureGate. n.d. "Stone Bramble, Rubus Saxatilis." NatureGate. Accessed July 7, 2025. https://luontoportti.com/en/t/314/stone-bramble.

Norling, Martin, Niclas Jareborg, and Jacques Dainat. 2018. "EMBLmyGFF3: A Converter Facilitating Genome Annotation Submission to European Nucleotide Archive." *BMC Research Notes* 11 (1): 584.

Ossowski, Stephan, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M. Clark, Ruth G. Shaw, Detlef Weigel, and Michael Lynch. 2010. "The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis Thaliana." *Science (New York, N.Y.)* 327 (5961): 92–94.

Price, R. Jordan, Jahn Davik, Felicidad Fernandéz Fernandéz, Helen J. Bates, Samantha Lynn, Charlotte F. Nellist, Matteo Buti, et al. 2023. "Chromosome-Scale Genome Sequence Assemblies of the 'Autumn Bliss' and 'Malling Jewel' Cultivars of the Highly Heterozygous Red Raspberry (*Rubus Idaeus* L.) Derived from Long-Read Oxford Nanopore Sequence Data." *PloS One* 18 (5): e0285756.

Ranallo-Benavidez, T. Rhyker, Kamil S. Jaron, and Michael C. Schatz. 2020. "GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes." *Nature Communications* 11 (1): 1432.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. 2020. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21 (1): 245.

Sim, Sheina B., Renee L. Corpuz, Tyler J. Simmonds, and Scott M. Geib. 2022. "HiFiAdapterFilt, a Memory Efficient Read Processing Pipeline, Prevents Occurrence of Adapter Sequence in PacBio HiFi Reads and Their Negative Impacts on Genome Assembly." *BMC Genomics* 23 (1): 157.

Stanke, Mario, Oliver Schöffmann, Burkhard Morgenstern, and Stephan Waack. 2006. "Gene Prediction in Eukaryotes with a Generalized Hidden Markov Model That Uses Hints from External Sources." *BMC Bioinformatics* 7 (1): 62.

Thompson, Maxine M. 1997. "Survey of Chromosome Numbers in *Rubus* (Rosaceae: Rosoideae)." *Annals of the Missouri Botanical Garden. Missouri Botanical Garden* 84 (1): 128.

UniProt Consortium. 2023. "UniProt: The Universal Protein Knowledgebase in 2023." *Nucleic Acids Research* 51 (D1): D523–31.

Urbaniak, Jacek, Paweł Kwiatkowski, and Paweł Pawlikowski. 2023. "Genetic Diversity and Population Structure of the Endangered *Rubus Chamaemorus* Populations in Poland." *Global Ecology and Conservation* 47 (e02633): e02633.

Wang, Yupeng, Haibao Tang, Jeremy D. DeBarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-Ho Lee, et al. 2012. "MCScanX: A Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity." *Nucleic Acids Research* 40 (7): e49–e49.

Zhou, Chenxi, Max Brown, Mark Blaxter, The Darwin Tree of Life Project Consortium, Shane A. McCarthy, and Richard Durbin. 2024. "Oatk: A de Novo Assembly Tool for Complex Plant

Organelle Genomes." *bioRxiv*. https://doi.org/10.1101/2024.10.23.619857.
Zhou, Chenxi, Shane A. McCarthy, and Richard Durbin. 2023. "YaHS: Yet Another Hi-C Scaffolding Tool." *Bioinformatics (Oxford, England)* 39 (1): btac808.

List of figures and tables:

Figure 1: Illustration of dioecy in Rubus chamaemorus and the sequenced male specimen.

Figure 2: Metrics of the genome assemblies of *R. chamaemorus* hap1 and hap2.

Figure 3: Synteny between *R. chamaemorus* and *R. idaeus.*

Figure 4: Subgenome structure inferred from k-mers and Ks.

Figure 5: High-confidence read mappings (MAPQ ≥ 30) from Kates et al. 2024: R. lasiococcus, R. pedatus, R. arcticus, and R. saxatilis to the α and β subgenomes of R. chamaemorus.

Figure 6: Possible scenarios of subgenome origin and polyploidisation events.

Table 1. Software tools: versions and sources.

Table 2: Genome data for *R. chamaemorus*.

Table 3: Busco duplicates in *R. chamaemorus* compared to *R. idaeus.*