The origin of the octoploid cloudberry (*Rubus chamaemorus*) genome is the result of multiple and complex polyploidization events

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*]

Affiliations:

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 fairly recent polyploidization involving one divergent (β) and three closely related (α) subgenomes. Low coverage targeted sequencing data mapped to our assembly link the β-subgenome to a relative of *R. pedatus* or *R. lasiococcus,* while the α-subgenomes might derive from a putative auto-allohexaploid within the main *Rubus* clade. In conclusion, these results indicate that the cloudberry genome arose through multiple hybridization events, including recurrent allopolyploidy and possibly autopolyploidy.

Keywords: Earth Biogenome Project Norway, Haplotype-resolved, HiFi sequencing, Octoploid, Allopolyploidization, Autopolyploidization

Significance statement

Cloudberry cultivation has lagged due to unclear origin and genome structure. We present a haplotype-resolved, chromosome-scale assembly that resolves four homologs per ancestral chromosome into 28 pseudo-chromosomes per haplotype and shows a 4:1 correspondence with red raspberry (*Rubus idaeus*). Genomic analyses indicate a recent polyploidization in which three closely related subgenomes likely derive from a putative hexaploid within the main *Rubus* clade, associated with *R. arcticus* (Arctic raspberry), whereas the fourth, more divergent subgenome is linked to relatives of *R. pedatus* (strawberryleaf raspberry) and *R. lasiococcus* (dwarf bramble). These results clarify cloudberry's formation and provide a foundation for trait discovery and accelerated breeding.

Introduction

The cloudberry (*Rubus chamaemorus*) is a circumpolar boreal flowering plant species in the rose family (Urbaniak et al., 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 (see Figure 1). 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 with distinct unisexual individuals, each producing either male or female gametes (see Figure 1B). 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 (Hummer et al., 2016; Thompson, 1997). *R. chamaemorus* itself is an octoploid with eight genome copies. Evidence from nuclear and chloroplast phylogenies suggests a complex hybrid origin (Carter et al., 2019; Michael, 2006). 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*. However, based on ITS data *R. arcticus* was rejected as a possible progenitor. This is consistent with Carter et al. (2019), who placed *R. chamaemorus* in an early-diverging lineage outside the main *Rubus*

clade. In addition to interest in its complex polyploidy, *R. chamaemorus* has also received interest for potential cultivar development (Martinussen et al., 2013) as well as regional conservation efforts (Urbaniak et al., 2023).

Despite the potential for cultivar development, 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 assemble the first high-quality, haplotype-resolved, chromosome-scale genome of *R. chamaemorus* (Figure 2). 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. Comparison with other available genome data from *Rubus* species demonstrate that the octoploid *R. chamaemorus* has originated through multiple, and possibly recurrent, hybridizations, leading to the genome architecture of present-day cloudberries. The two haplotype-resolved assemblies provide a genomic foundation for future genetic and evolutionary studies of *R. chamaemorus.*

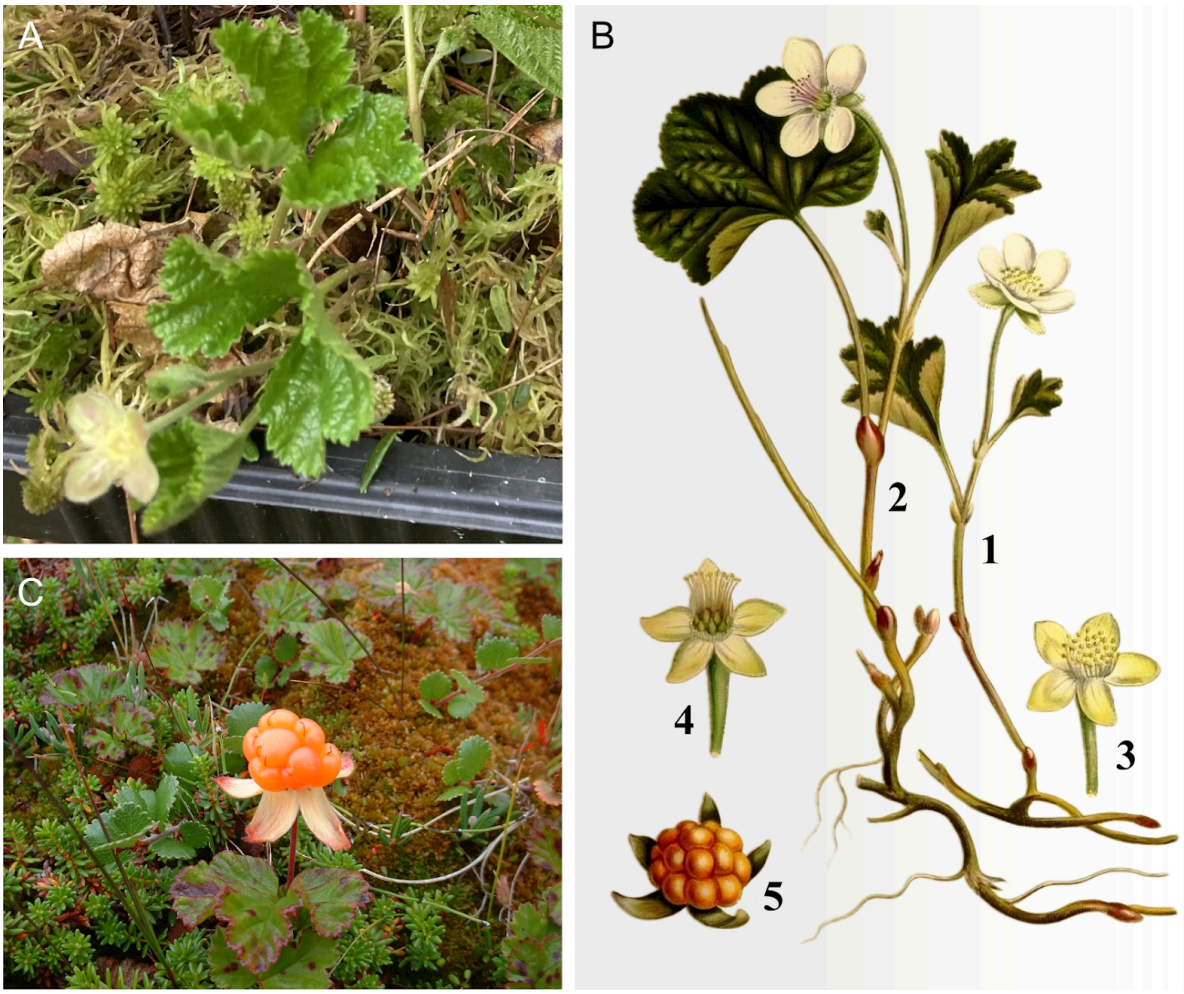

Figure 1: Illustration of dioecy in *Rubus chamaemorus* and the sequenced male specimen. *R. chamaemorus* is dioecious, with male and female flowers borne on separate plants. **A)** Our sample (a male plant) photographed late in the season (photo: Simen R. Sandve). **B)** Historical illustration: (1) male and (2) female plant in bloom; (3) male and (4) female flowers 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 1 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 1 and Figure 2). 28 pseudo-chromosomes were identified in both pseudo-haplotypes (ordered by homology to *Rubus idaeus* chromosomes first, and length in hap1 second, and then numbered 1–28, with the homologs in hap2 receiving the same number).

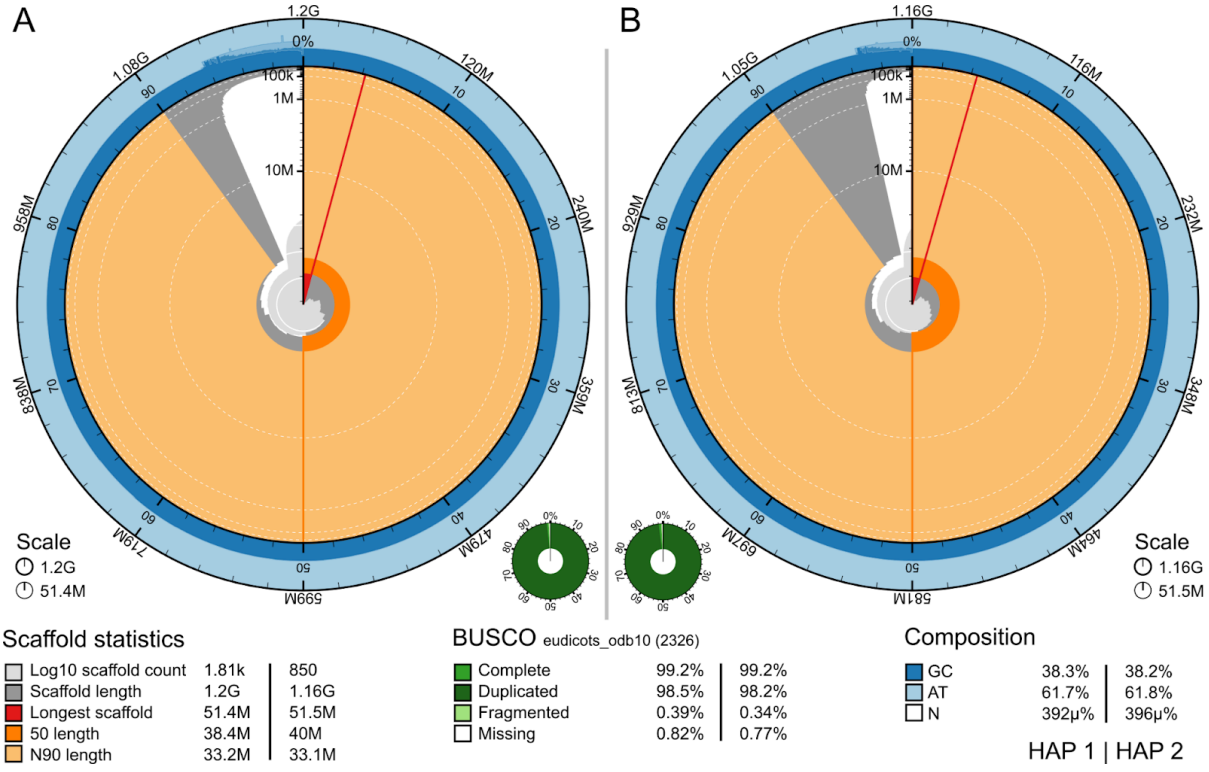

Figure 2: Metrics of the genome assemblies of *R. chamaemorus* hap1 and hap2. A) Hap1. **B)** Hap2. The BlobToolKit Snailplots show N50 metrics and BUSCO gene completeness. The two outermost bands of the circle signify GC versus AT composition at 0.1% intervals. Light orange shows the N90 scaffold length, 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.

Table 1: Genome data for *R. chamaemorus.* *Both MT1 and MT2 are 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.

Project accession data		
Species	*Rubus chamaemorus*	
Specimen	drRubCham1	
NCBI Taxonomy ID	57936	
BioProject	PRJEB98354	
BioSample ID	SAMEA120268085	
Isolate information	Male, leaf	
Raw data accessions		
PacBio HiFi reads	ERR15658311, ERR15658288	2 PACBIO_SMRT (Sequel IIe) runs: 4.66 M reads, 70.8 Gb
Hi-C Illumina reads	ERR15658302	1 ILLUMINA (Illumina NovaSeq 6000) run: 246 M pairs of reads, 74.3 Gb
Genome assembly metrics		
HiFi read coverage	59x	
Assembly accession	GCA_976986655	GCA_976986645
Assembly identifier	drRubCham1.1.h1	drRubCham1.1.h2

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 (plastid)	
Genome annotation metrics		
Number of protein-coding genes	74,132	70,692
Number of protein-coding genes with functional domain**	72,128	68,578
Number of protein-coding genes with gene names	54,204	50,835
BUSCO***	C:99.2%[S:0.7%,D:98.5%],F:0.4%,M:0.4%,n:2326,E:1.0%	C:99.2%[S:1.0%,D:98.2%],F:0.3%,M:0.4%,n:2326,E:0.8%

When compared to a k-mer database of the Hi-C reads, hap1 had a k-mer completeness of 96.0%, hap2 of 94.0%, and combined they have a completeness of 96.5%. 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)(Table 1). When comparing the two pseudo-haplotypes using minimap2, there are 1,775,487 SNP 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 74,132 and 70,692 protein-coding genes were annotated in hap1 and hap2, respectively (Table 1).

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. 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.

Using the phased assembly, we assessed genome-wide synteny between *R. chamaemorus* hap1 and *R. idaeus* (raspberry) – a closely related diploid species with a high-quality chromosome resolved assembly (Martinussen et al., 2013; Price et al., 2023). The synteny analysis 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 translocation 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 translocation to chromosomes 25–28, and a unique translocation 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 large 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.

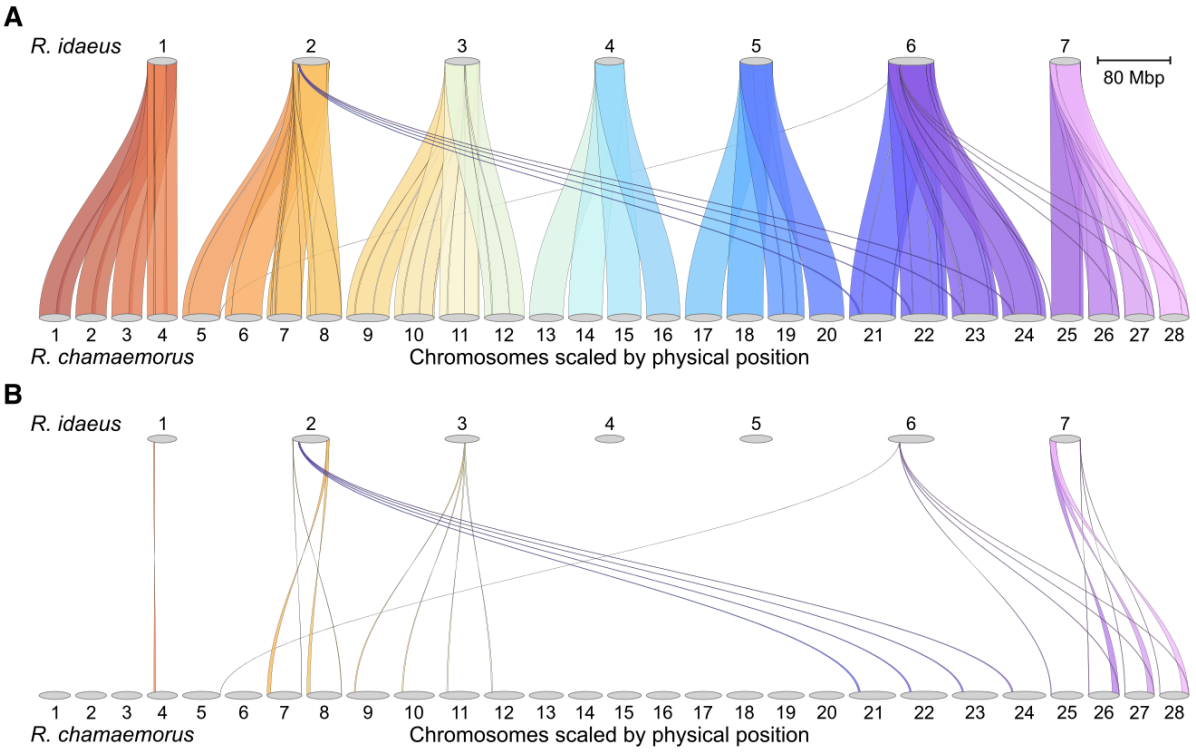

Figure 3: Synteny between *R. chamaemorus* (cloudberry) and *R. idaeus* (raspberry). Both panels are riparian plots generated using GENESPACE. Each *R. idaeus* chromosome aligns with four distinct *R. chamaemorus* hap1 chromosomes, consistent with an octoploid origin. Colors follow a continuous gradient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

Table 2: BUSCO duplicates in *R. chamaemorus* compared to *R. idaeus.* Number and percent of BUSCO genes found in *R. chamaemorus* in comparison to *R. idaeus*.

# BUSCO copies	*R. chamaemorus* hap 1	*R. chamaemorus* hap 2	*R. idaeus*
0 (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

Hap1 had 99.2% and hap2 99.2% complete BUSCO genes using the eudicots lineage set. Each pseudo-haplotype most commonly has 4 BUSCO copies, consistent with the known octoploid status of *R. chamaemorus* (Table 2). The next most frequent categories are 3 copies, likely reflecting gene loss (fractionation), and >4 copies, which may result from recent small scale 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 within *Rubus*.

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 (red, Figure 4A); the remaining 21 chromosomes, lacking such distinction, were collectively grouped as the α-subgenome(s) (grey, Figure 4A).

We then estimated Ks distributions using doubletrouble (Almeida-Silva & Van de Peer, 2025). The overall distribution showed a single, narrow peak 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 4C). 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.

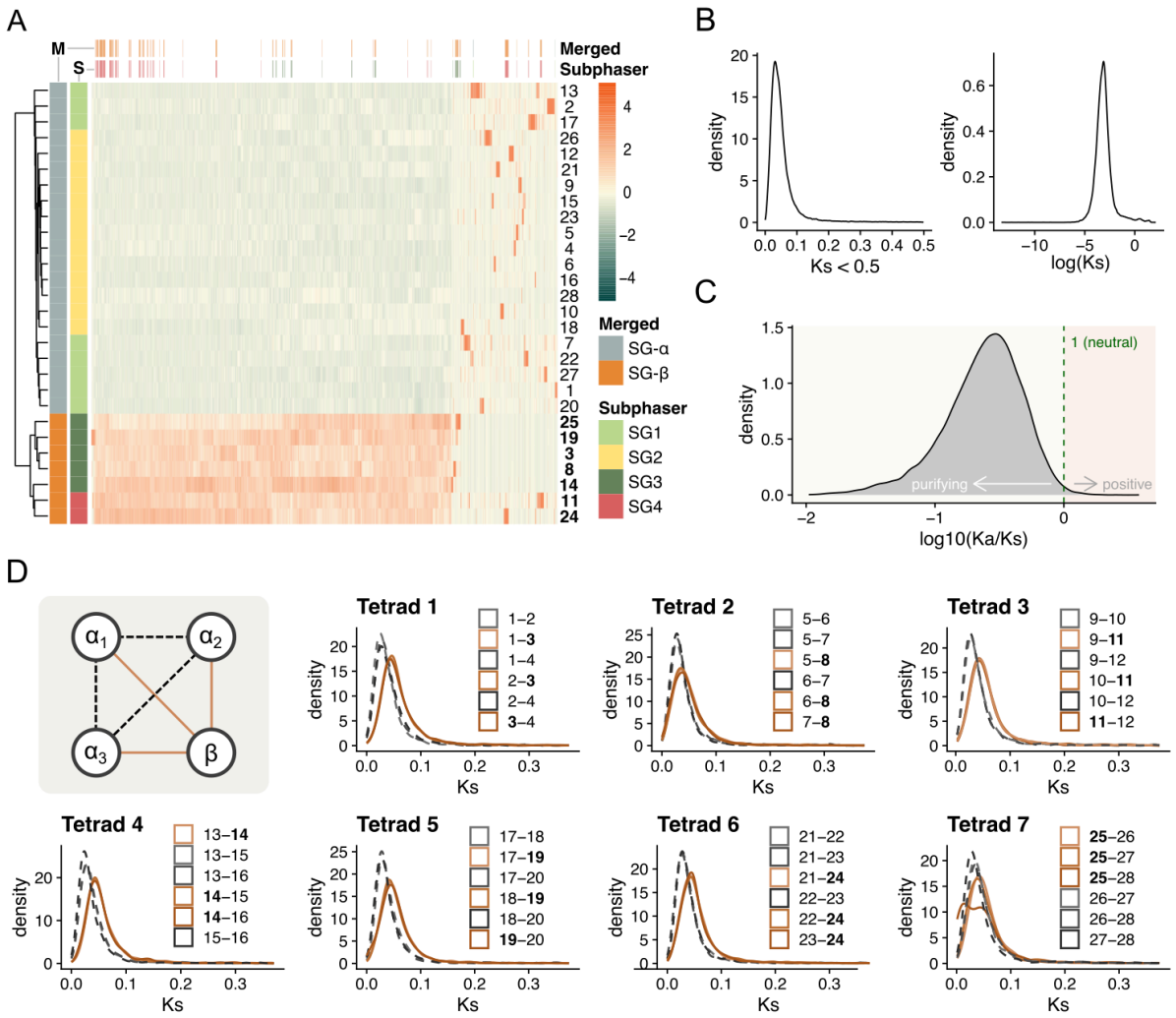

Figure 4: Subgenome structure inferred from k-mer and Ks analyses. A) Unsupervised hierarchical clustering of k-mer profiles. Top bar: k-mer subgenome specificity. Side bar: chromosome subgenome assignments. "Merged" labels reflect our inferred α/β subgenomes based on a stricter interpretation of the clustering. Heatmap shows Z-scored relative abundance of 10,000 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; artifact removal), 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 corresponding β-homolog. Stippled grey lines indicate comparisons among the three α-homologs.

To estimate the time since separation between the α- and the β-subgenomes, as well as within the α-subgenomes, we calculated divergence time (T) using T = Ks / (2 × substitution rate), assuming a constant substitution rate of 4.13×10^{-9} substitutions/site/year. The overall trend (Supplementary Figure 4) shows that the α- and β-subgenomes diverged approximately 5.83 Ma (95% CI: 5.80–5.87; median Ks = 0.0481), while the median duplication age within the α-subgenomes is around 4.17 Ma (95% CI: 4.14–4.21; median Ks = 0.0345); roughly 1.7 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 is an exception to the other tetrads, showing greater variability and less internal consistency especially between chromosomes 25, 26 and 28.

Given at least one hybridization event in the origin of *R. chamaemorus*, we assessed whether reads from candidate *Rubus* taxa map preferentially to α or β. *R. pedatus* and *R. lasiococcus* have previously been proposed as progenitor species of *R. chamaemorus* (Carter et al., 2019; Michael, 2006), while *R. arcticus* shares a GBSSI-1γ copy with *R. chamaemorus (Michael, 2006)*. Although genome assemblies are not available for these species, limited sequencing data is available (through Carter et al. (2019) and Kates et al. (2024); see Material and Methods). To minimize mapping ambiguity among closely related α subgenomes, we retained only the longest α representative per tetrad and quantified high-confidence alignments across α and β. Using each phylogenetic tree as a scaffold, we mapped $\log_2$ coverage bias to the reduced genome (full species lists and complete trees in Supplementary Figures 5–8).

In the Carter dataset, we observe the strongest β bias within the *R. pedatus–R. lasiococcus* clade, with weak β bias or near neutrality in other early-diverging *Rubus* lineages (Figure 5A; group 2). The main *Rubus* clade is predominantly α-biased in Carter, with a few neutral species nested within otherwise α-biased subclades (Supplementary Figures 5–6). When the signal is examined by tetrad, tetrad 7 (chr25–28; chr25 vs. 26) behaves idiosyncratically: *R. idaeus* exhibits an anomalously strong β bias (Figure 5B), and several otherwise weakly biased species show the same pattern specifically for this tetrad. Neither *R. idaeus* nor these other species show comparable β bias in any other tetrads.

In the Kates dataset, overall β bias is stronger than in Carter, shifting groups 5–8 to be neutral/slightly β-biased, with *R. pedatus* showing the strongest β bias and *R. arcticus* and associated clades (3–4) retaining their α-biased status (Supplementary Figures 7–8). While the overall bias in Kates is either neutral or β-biased, tetrad 6 stands out with its strong and consistent α-bias. Unlike in Carter there are spikes of β-bias in tetrads 4 and 5, as well as tetrad 7, which might be a result of fewer and less informative loci for *Rubus* in this dataset.

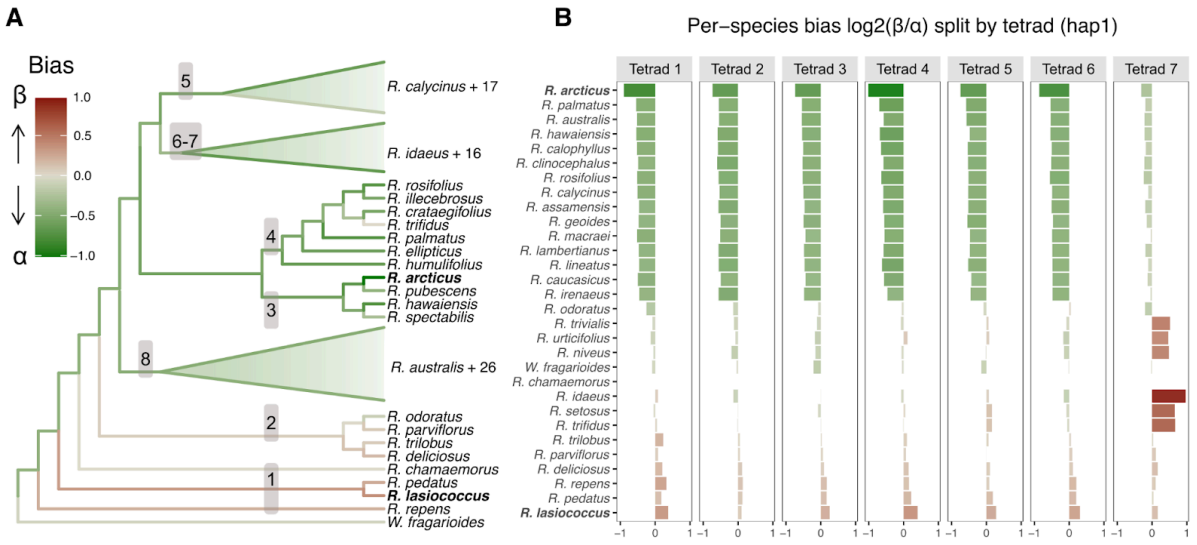

Figure 5: High-confidence read mappings (MAPQ ≥ 30) from Carter et al. (2019) depicting bias of *Rubus* species mapped to a reduced α and β subgenome set of *R. chamaemorus*. Reads from each species are mapped to *R. chamaemorus* haplotype one (hap1). Log_2 of the ratio of high-confidence reads mapped to subgenome β over subgenome α, with bias shifted so that *R. chamaemorus* is at 0. A value of 1 indicates a two-fold 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. **A)** *R. chamaemorus* centered bias (median across tetrads) projected onto the ASTRAL-II exon all-taxa tree; the tree is collapsed to show only relevant lineages. (Numbers denote Carter et al. groups; Fig. 2). The most α-biased and β-biased species are marked in bold. **B)** Bias stratified by chromosome pair (tetrad). Species are ordered from most α-biased (most negative) to most β-biased (most positive). Only the 15 most α-biased and the 15 least α-biased species are shown.

Subgenome evolution scenarios

Together, Figures 3–5 constrain the set of plausible polyploidization pathways leading to the octoploid *R. chamaemorus*. The subgenomes form two main lineages, α and β, which started diverging ~5.83 Ma. Within α, there are three sets of 7 (α_1–α_3), but k-mer composition and Ks profiles do not resolve them as distinct subgenomes. We therefore treat α_1–α_3 as interchangeable topologically and use the numbering only to track hypothetical compositions. Under the criterion that only one subgenome is distinguishable, the 15 theoretical topologies collapse to four end-state topologies (Figure 6A; putative scenarios depicted in Figure 6 B–F).

Figures 6B and 6C both show topology 1, in which an octoploid forms via two tetraploids, differing only by the mechanism of α polyploidization. In Scenario 1.1 (Figure 6B), three successive hybridizations; $\alpha_3 \times \beta$ and $\alpha_1 \times \alpha_2$, followed by $\alpha_3\beta \times \alpha_1\alpha_2$; yield four distinct subgenomes. In Scenario 1.2 (Figure 6C), α_1 first undergoes autotetraploidization ($\alpha_1 \rightarrow$

$\alpha_1\alpha_1$), α_2 hybridizes with β, and the merger $\alpha_1\alpha_1 \times \alpha_2\beta$ produces an octoploid with three distinct subgenomes.

Figures 6D–F depict topologies 2–4, each proceeding through a hexaploid intermediate and differing only in the order of β incorporation. Scenario 2 (Figure 6D) begins with autotetraploidization of α_1, followed by hybridization with α_2 ($\alpha_1\alpha_1 \times \alpha_2$), and a subsequent merger with β ($\alpha_1\alpha_1\alpha_2 \times \beta$), yielding an octoploid with three distinct subgenomes. In Scenario 3 (Figure 6E), β first merges with a tetraploid $\alpha_2\alpha_2$ to form a hexaploid, which then merges with α_1. In Scenario 4 (Figure 6F), β merges with one α (α_3) to form a tetraploid, then with a second α (α_2) to form a hexaploid, and finally with a third α (α_1) to produce an octoploid.

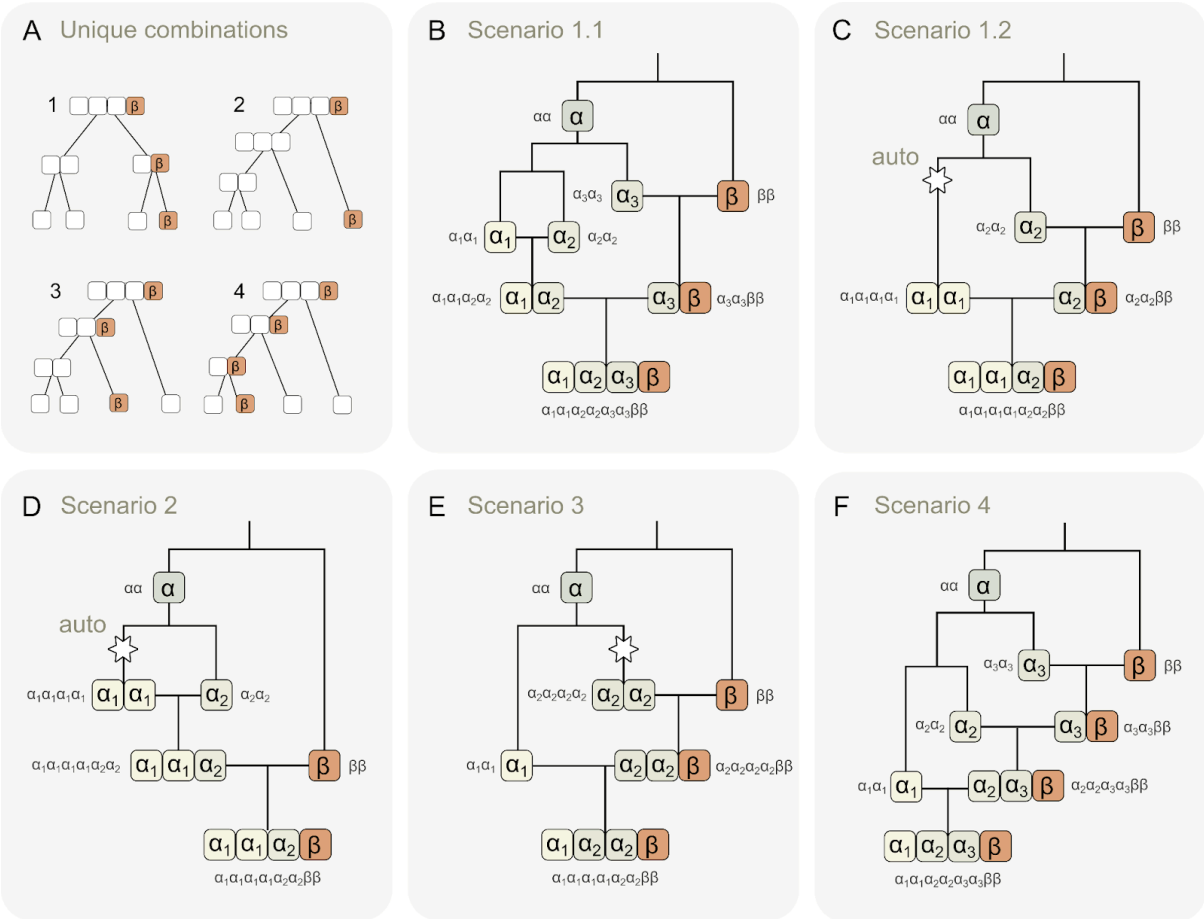

Figure 6. Scenarios for subgenome origin and polyploidization. A) When only one of four subgenomes is distinguishable, the 15 possible octoploid topologies collapse into four equivalence classes; One where the intermediates are tetraploids (Scenario 1), and 3 where a hexaploid is an intermediate (Scenarios 2–4). **B)** Scenario 1.1: three successive hybridizations ($\alpha_3 \times \beta$ and $\alpha_1 \times \alpha_2$, then $\alpha_3\beta \times \alpha_1\alpha_2$) yield four distinct subgenomes. **C)** Scenario 1.2: one autotetraploidization of α_1 ($\alpha_1 \rightarrow \alpha_1\alpha_1$), followed by $\alpha_2 \times \beta$ and $\alpha_1\alpha_1 \times \alpha_2\beta$, yields three distinct subgenomes. **D–F)** Scenarios 2–4: routes via a hexaploid intermediate that differ only in the order of β incorporation. Scenario 2: β joins after formation of $\alpha_1\alpha_1\alpha_2$. Scenario 3: β first merges with tetraploid $\alpha_2\alpha_2$ to form a hexaploid, which then merges with α_1. Scenario 4: β sequentially merges with three α lineages, forming a tetraploid, then a hexaploid, then the octoploid.

Discussion

Polyploidy is pervasive in *Rubus*, with 60–70% of species possessing chromosome numbers 4n or higher (Chen et al., 2020; Hummer et al., 2016; Thompson, 1997). Comparative synteny between *R. chamaemorus* and *R. idaeus* reveals a highly collinear 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). The lack of resolvable α subgenome structure suggests limited divergence among α copies, with subsequent homoeologous exchange erasing most differences, including accumulated repeats and transposable elements. Although such erasure permits multiple histories, Scenario 2, outlined in figure 6D, is the most parsimonious: an α autotetraploid merges with a closely related α diploid to form an

297 auto-allohexaploid with three α copies, which then merges with the more divergent β
298 diploid.
299
300 Using a reduced reference comprising one α-subgenome set (longest chromosome) and the
301 β-subgenome, we mapped short reads from Carter et al. (2019) and Kates et al. (2024) to
302 estimate α–β mapping bias. Since Carter et al. specifically targets *Rubus* and uses ~1,000
303 low-copy nuclear genes, in contrast to Kates et al. who use 100 loci across the nitrogen-fixing
304 clade only incidentally covering *Rubus*, we defer to Carter for the most accurate phylogeny
305 of *Rubus* and the most reliable results, treating Kates as a cross-check for mapping bias. Both
306 datasets recover the same extremes: strongest β affinity in the *R. pedatus–R. lasiococcus*
307 lineage and strongest α affinity within the main *Rubus* clade, specially Carter et al. (2019)
308 groups 3–4 with *R. arcticus*, where α-bias persists in Kates despite the overall β shift.
309
310 Although it has been proposed that both R. pedatus and *R. lasiococcus* are progenitors of *R.*
311 *chamaemorus* (Michael, 2006), our results indicate they may only account for the β
312 subgenome (Figure 5A). Based on the available chloroplast trees, it is likely the maternal
313 ancestor is more closely related to *R. pedatus* (Carter et al., 2019; Michael, 2006). *R. arcticus*
314 (group 3) exhibits the strongest α bias in Carter, followed by *R. palmatus in* the adjacent
315 group 4 (Figure 5). Taken together, these patterns are consistent with an unsampled or
316 extinct α-donor lineage along the stem leading to groups 3–4, rather than from any single
317 species present in the datasets.
318
319 Our estimate of ~5.83 Ma as the start of α–β divergence, which preceded the
320 polyploidization events leading to *R. chamaemorus,* should be regarded as highly uncertain,

as the substitution rates for herbaceous perennials found by Kay et al. (2006) ranges from 1.72×10^{-9} to 8.34×10^{-9} substitutions/site/year. Using the low estimate, the split between α- and β-subgenomes comes out to 14.1 Ma, which is more consistent with the time estimates of the split between *R. lasiococcus*–*R. pedatus* and the main *Rubus* clade estimated by Carter et al. (2019). It is also possible, given an evolutionary setting with frequent hybridizations and homoeologous exchanges over time, that divergence is underestimated.

Tetrad 7 (chromosomes 25–28) consistently departs from the otherwise clear α–β pattern across analyses. Only the K-mer analysis cleanly assigns chromosome 25 to β and chromosomes 26–28 to α, indicating unique sequence content on chromosome 25. In contrast, pairwise Ks within the tetrad do not follow the typical α–β divergence pattern: comparisons involving chromosome 25, especially with chromosome 28, are bimodal (Supplementary Figure 3), suggesting a mosaic history. Mapping-bias analysis further highlights this anomaly, with several species, most notably *R. idaeus*, showing a stark shift toward β only in this tetrad (Figure 5B). Hi-C and collinearity analyses provide no clear evidence of misjoins (Figure 3 and Supplementary Figure 1), and chromosome 25 uniquely lacks the large inversion present on chromosomes 26–28 (Figure 3B), which suggests a biological rather than assembly cause. Although data artifacts cannot be ruled out, these consistent irregularities across methods point to a complex evolutionary history for Tetrad 7.

Resolving these questions will require high-quality genome assemblies of the putative relatives of *R. chamaemorus*; including *R. pedatus*, *R. lasiococcus*, *R. arcticus*, and related taxa. More broadly, such assemblies are essential for disentangling polyploid origins,

identifying subgenome contributors, and enabling accurate comparative genomics across complex plant lineages.

The presented *R. chamaemorus* assemblies are a resource that 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.

Material and 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 (Figure 1A). 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.

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

372 Library preparation and sequencing for *de novo* assembly

373 Before PacBio HiFi library preparation, DNA was purified an additional time using AMPure PB
374 beads (1:1 ratio). Purified HMW DNA was sheared into an average fragment size of 15-20
375 kbp large fragments using the Megaruptor3 (Diagenode). A HiFi library was prepared
376 following the PacBio protocol for HiFi library preparation using the SMRTbell® Prep Kit 3.0.
377 The final HiFi library was size-selected with a 10 kbp cut-off using a BluePippin (Sage
378 Sciences) and sequencing was performed by the Norwegian Sequencing Centre on a PacBio
379 Sequel IIe instrument (Pacific Biosciences Inc). The library was sequenced on two 8M SMRT
380 cells using a Sequel II Binding kit 3.2 and Sequencing chemistry v2.0.
381
382 A Hi-C library was prepared using the Arima High Coverage HiC kit (Arima Genomics),
383 following the manufacturer's recommendations and starting with 1050 mg leaf tissue.
384 Chromatin was crosslinked prior to nuclei isolation as written in the user guide for plant
385 tissues (Doc A160163 v01). Final library quality was quantified using a Kapa Library
386 quantification kit for Illumina (Roche Inc.). The library was sequenced with other libraries on
387 an Illumina NovaSeq 6000 (Illumina Inc) with 2 × 150 bp paired end mode at the Norwegian
388 Sequencing Centre (https://www.sequencing.uio.no).

Genome assembly and curation, annotation and evaluation

A full list of relevant software tools and versions is presented in supplementary 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 et al., 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 et al., 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 et al., 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 et al., 2023) with default options. FCS-GX (Astashyn et al., 2024) was used to search for putative

412 contamination and contaminated sequences were removed. The mitochondrion was
413 searched for in contigs and reads using Oatk (Zhou et al., 2024). The assemblies were
414 manually curated using PretextView and Rapid curation 2.0. Chromosomes were identified
415 by mapping to *Rubus idaeus* JASCWY01 (Price et al., 2023) and *Fragaria vesca* v4 (Edger et
416 al., 2018) in addition to inspecting the Hi-C contact map in PretextView. We annotated the
417 genome assemblies using a pre-release version of the EBP-Nor genome annotation pipeline
418 (https://github.com/ebp-nor/GenomeAnnotation). First, AGAT
419 (https://zenodo.org/record/7255559) `agat_sp_keep_longest_isoform.pl` and
420 `agat_sp_extract_sequences.pl` were used on the *Arabidopsis thaliana* (TAIR10.1,
421 GCF_000001735.4) genome assembly and annotation to generate one protein (the longest
422 isoform) per gene. Miniprot (Li, 2023) was used to align the proteins to the curated
423 assemblies. UniProtKB/Swiss-Prot (UniProt Consortium, 2023) release 2023_02 in addition
424 to the Viridiplantae part of OrthoDB v11 (Kuznetsov et al., 2023) were also aligned
425 separately to the assemblies. RED (Girgis, 2015) was run via redmask
426 (https://github.com/nextgenusfs/redmask) on the assemblies to mask repetitive areas.
427 GALBA (Brůna, Li, et al., 2023; Buchfink et al., 2015; Hoff & Stanke, 2019; Li, 2023; Stanke et
428 al., 2006) was run with the *Arabidopsis* proteins using the miniprot mode on the masked
429 assemblies.
430
431 The `funannotate-runEVM.py` script from Funannotate was used to run EvidenceModeler
432 (Haas et al., 2008) on the alignments of TAIR10.1 proteins, UniProtKB/Swiss-Prot proteins,
433 Viridiplantae proteins and the predicted genes from GALBA. The resulting predicted proteins
434 were compared to the protein repeats that Funannotate distributes using DIAMOND blastp
435 and the predicted genes were filtered based on this comparison using AGAT. The filtered

436 proteins were compared to the UniProtKB/Swiss-Prot release 2023_02 using DIAMOND
437 (Buchfink et al., 2015) `blastp` to find gene names and InterProScan (Jones et al., 2014) was
438 used to discover functional domains. AGATs
439 `agat_sp_manage_functional_annotation.pl` was used to attach the gene names
440 and functional annotations to the predicted genes. EMBLmyGFF3 (Norling et al., 2018) was
441 used to combine the fasta files and GFF3 files into an EMBL format for submission to ENA.
442
443 All the evaluation tools have also been implemented in a pipeline, similar to assembly and
444 annotation (https://github.com/ebp-nor/GenomeEvaluation). Merqury (Rhie et al., 2020)
445 was used to assess the completeness and quality of the genome assemblies by comparing
446 them to the k-mer content of both the Hi-C reads and PacBio HiFi reads. BUSCO (Manni et
447 al., 2021) was used to assess the completeness of the genome assemblies by comparing
448 against the expected gene content in the eudicots lineage. Gfastats (Formenti et al., 2022)
449 was used to output different assembly statistics of the assemblies.
450
451 BlobToolKit and BlobTools2 (Laetsch & Blaxter, 2017), in addition to blobtk were used to
452 visualize assembly statistics. To generate the Hi-C contact map image, the Hi-C reads were
453 mapped to the assemblies using BWA-MEM (Li, 2013) using the same approach as above.
454 Finally, PretextMap was used to create a contact map which was visualized using
455 PretextSnapshot.
456
457 To characterize the differences between the two pseudo-haplotypes, we ran a genome
458 alignment using minimap2 (Li, 2017) on the homologous chromosomes. The resulting

459 alignment was processed with the minimap2-included paftools.js producing a report listing
460 the number of insertions, SNPs and indels between the two pseudo-haplotypes.
461
462 To demonstrate the duplicated nature of *R. chamaemorus*, we analyzed synteny between
463 hap1 and *Rubus idaeus* (red raspberry cv. 'Autumn Bliss') using the R package pipeline
464 GENESPACE (Lovell et al., 2022). *R. idaeus* was chosen as reference due to its close
465 phylogenetic relationship and availability of a high-quality chromosome-scale assembly
466 (Martinussen et al., 2013; Price et al., 2023). To improve orthogroup inference, we included
467 additional *Rubus*: *Rubus occidentalis* (v3.0 - black raspberry) and *Rubus argutus* ('Hillquist'
468 v1.0 - blackberry), as well as more distantly related Rosaceae members: *Fragaria vesca*
469 (v4.0.a1 - wild strawberry) and *Rosa chinensis* ('Old Blush' v2.0 - Chinese rose). Assemblies,
470 protein sequences and annotations were obtained from the Rosaceae database
471 (https://www.rosaceae.org/). We selected the longest protein isoform as the representative
472 for each gene, along with gene rank orders, to serve as input for GENESPACE. The pipeline
473 was run with the default method, which uses OrthoFinder (Emms & Kelly, 2019) to build
474 orthogroups and MCScanX (Wang et al., 2012) to detect collinear regions. All other
475 parameters were left at default. Synteny blocks between *R. chamaemorus* and *R. idaeus*
476 were then visualized using the plot_riparian function of the GENESPACE package.
477
478 Subgenome structure in *R. chamaemorus* was assessed using SubPhaser v1.2.6 (Jia et al.,
479 2022). The analysis was run independently on each haplotype, with default parameters and
480 the 7 sets of 4 homologous chromosomes (tetrads) defined via a configuration file.
481

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

487

488 To estimate divergence times, Ks values were filtered to exclude values greater than 2 to
489 minimize the effects of substitutional saturation. For each chromosome pair within each
490 tetrad, we calculated the median Ks and derived 95% confidence intervals using
491 non-parametric bootstrapping with 1,000 replicates, implemented using the boot package in
492 R. Comparisons with fewer than 10 valid Ks values were excluded. Divergence time (T) was
493 estimated using the formula: $T = Ks / (2 \times \mu)$, where μ is the substitution rate, set to a
494 constant 4.13×10^{-9} substitutions/site/year, a mean herbaceous annual and perennial
495 angiosperm rate reported by Kay et al. (2006). Confidence intervals for divergence time were
496 calculated by applying the same formula to the bootstrap-derived lower and upper bounds
497 of Ks.

498

499 To assess potential parental contributors to the *R. chamaemorus* subgenomes, we analyzed
500 mapping bias from available *Rubus* species (plus outgroup *Waldsteinia fragarioides*) using
501 two datasets: Carter et al. (2019) "Target Capture Sequencing Unravels *Rubus* Evolution"
502 (PRJNA510412) and Kates et al. (2024) "Shifts in Evolutionary Lability Underlie Independent
503 Gains and Losses of Root-Nodule Symbiosis in a Single Clade of Plants" (PRJNA1022323,
504 PRJNA1022030, PRJNA1022147, PRJNA1022025, PRJNA1022141, PRJNA1022138,
505 PRJNA1022032, PRJNA1022029, PRJNA1022027, PRJNA1022023, PRJNA1022015,

PRJNA1021620, PRJNA1021608). To reduce ambiguity among similar α copies, we retained only the longest α chromosome per tetrad together with the complete β subgenome and aligned reads to this reduced α+β reference with Bowtie2. We kept high-confidence alignments (MAPQ ≥ 30) to limit mismapping across homeologous regions, quantified bias between α and β chromosome pairs as the log2 ratio of read densities (reads per chromosome length) for β versus α [log2(β/α)], and centered each pair so that *R. chamaemorus* had a bias of 0. For each dataset, we projected median species estimates onto the accompanying phylogeny, rooting on *W. fragarioides*: for Carter et al. (2019) the ASTRAL-II exon all-taxa tree (Exon_alltaxa_AstralII.tre), and for Kates et al. (2024) Supplementary Dataset 3 (41467_2024_48036_MOESM6_ESM.txt).

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.

Funding

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

Acknowledgments

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). We used ChatGPT (temporary chat) for text revision and code refactoring. No AI tool processed study data or created figures; all analyses were performed by the authors.

Data Availability

All data are available in the European Nucleotide Archive (ENA). The umbrella project for EBP-Nor is BioProject PRJEB98354. Raw PacBio HiFi reads for the drRubCham1 sample (BioSample SAMEA120268085) are deposited under run accessions ERR15658288 and ERR15658311. Illumina Hi-C sequencing reads are available under run accession ERR15658302. The two assemblies are archived separately: pseudo-haplotype 1 under BioProject PRJEB98298 (assembly GCA_976986655) and pseudo-haplotype 2 under BioProject PRJEB98353 (assembly GCA_976986645)

Genome assemblies and gene annotations are also available at **10.5281/zenodo.15773906**

References

Almeida-Silva, F., & Van de Peer, Y. (2025). doubletrouble: an R/Bioconductor package for the identification, classification, and analysis of gene and genome duplications. *Bioinformatics (Oxford, England)*, *41*(2), btaf043.

Astashyn, A., Tvedte, E. S., Sweeney, D., Sapojnikov, V., Bouk, N., Joukov, V., Mozes, E., Strope, P. K., Sylla, P. M., Wagner, L., Bidwell, S. L., Brown, L. C., Clark, K., Davis, E. W., Smith-White, B., Hlavina, W., Pruitt, K. D., Schneider, V. A., & Murphy, T. D. (2024). Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biology*, *25*(1), 60.

Brůna, T., Aryal, R., Dudchenko, O., Sargent, D. J., Mead, D., Buti, M., Cavallini, A., Hytönen, T., Andrés, J., Pham, M., Weisz, D., Mascagni, F., Usai, G., Natali, L., Bassil, N., Fernandez, G. E., Lomsadze, A., Armour, M., Olukolu, B., … Worthington, M. (2023). A chromosome-length genome assembly and annotation of blackberry (*Rubus argutus*, cv. "Hillquist"). *G3 (Bethesda, Md.)*, *13*(2), jkac289.

Brůna, T., Li, H., Guhlin, J., Honsel, D., Herbold, S., Stanke, M., Nenasheva, N., Ebel, M., Gabriel, L., & Hoff, K. J. (2023). Galba: genome annotation with miniprot and AUGUSTUS. *BMC Bioinformatics*, *24*(1), 327.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60.

Carter, K. A., Liston, A., Bassil, N. V., Alice, L. A., Bushakra, J. M., Sutherland, B. L., Mockler, T. C., Bryant, D. W., & Hummer, K. E. (2019). Target capture sequencing unravels Rubus evolution. *Frontiers in Plant Science*, *10*, 1615.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175.

Chen, X., Li, J., Cheng, T., Zhang, W., Liu, Y., Wu, P., Yang, X., Wang, L., & Zhou, S. (2020). Molecular systematics of rosoideae (Rosaceae). *Plant Systematics and Evolution*, *306*(1), 1–12.

Edger, P. P., VanBuren, R., Colle, M., Poorten, T. J., Wai, C. M., Niederhuth, C. E., Alger, E. I., Ou, S., Acharya, C. B., Wang, J., Callow, P., McKain, M. R., Shi, J., Collier, C., Xiong, Z., Mower, J. P., Slovin, J. P., Hytönen, T., Jiang, N., … Knapp, S. J. (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.

574 Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative
575 genomics. *Genome Biology*, *20*(1), 238.
576 Formenti, G., Abueg, L., Brajuka, A., Brajuka, N., Gallardo-Alba, C., Giani, A., Fedrigo, O., & Jarvis, E. D.
577 (2022). Gfastats: conversion, evaluation and manipulation of genome sequences using assembly
578 graphs. *Bioinformatics (Oxford, England)*, *38*(17), 4214–4216.
579 Girgis, H. Z. (2015). Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the
580 genomic scale. *BMC Bioinformatics*, *16*(1), 227.
581 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman,
582 J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the
583 Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7.
584 Hoff, K. J., & Stanke, M. (2019). Predicting genes in single genomes with AUGUSTUS. *Et Al [Current*
585 *Protocols in Bioinformatics]*, *65*(1), e57.
586 Huang, T.-R., Chen, J.-H., Hummer, K. E., Alice, L. A., Wang, W.-H., He, Y., Yu, S.-X., Yang, M.-F., Chai,
587 T.-Y., Zhu, X.-Y., Ma, L.-Q., & Wang, H. (2023). Phylogeny of *Rubus* (Rosaceae): Integrating
588 molecular and morphological evidence into an infrageneric revision. *Taxon*, *72*(2), 278–306.
589 Hummer, K. E., Bassil, N. V., & Alice, L. A. (2016). *Rubus* ploidy assessment. *Acta Horticulturae*,
590 *1133*(1133), 81–88.
591 Jennings, S. N. (2018). Advances in *Rubus* Breeding. In *Raspberry* (pp. 17–28). Springer International
592 Publishing.
593 Jia, K.-H., Wang, Z.-X., Wang, L., Li, G.-Y., Zhang, W., Wang, X.-L., Xu, F.-J., Jiao, S.-Q., Zhou, S.-S., Liu, H.,
594 Ma, Y., Bi, G., Zhao, W., El-Kassaby, Y. A., Porth, I., Li, G., Zhang, R.-G., & Mao, J.-F. (2022).
595 SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific
596 k-mers. *The New Phytologist*, *235*(2), 801–809.
597 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell,
598 A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez,
599 R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification.

Bioinformatics (Oxford, England), *30*(9), 1236–1240.

Kates, H. R., O'Meara, B. C., LaFrance, R., Stull, G. W., James, E. K., Liu, S.-Y., Tian, Q., Yi, T.-S., Conde, D., Kirst, M., Ané, J.-M., Soltis, D. E., Guralnick, R. P., Soltis, P. S., & Folk, R. A. (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.

Kay, K. M., Whittall, J. B., & Hodges, S. A. (2006). A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. *BMC Evolutionary Biology*, *6*, 36.

Kokot, M., Dlugosz, M., & Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics (Oxford, England)*, *33*(17), 2759–2761.

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., & Zdobnov, E. M. (2023). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, *51*(D1), D445–D451.

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*, 1287.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997

Li, H. (2017). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*, 3094–3100.

Li, H. (2023). Protein-to-genome alignment with miniprot. *Bioinformatics (Oxford, England)*, *39*(1), btad014.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079.

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy

number variation across multiple genomes. *eLife*, *11*. https://doi.org/10.7554/eLife.78526

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654.

Martinussen, I., Uleberg, E., Sønsteby, A., Sønstebø, J. H., Grahamand, J., & Vivian-Smith, A. (2013). *Genomic survey sequences and the structure of the Rubus chamaemorus L. genome as determined by ddRAD tags*. https://www.academia.edu/31763032/Genomic_survey_sequences_and_the_structure_of_the_Rubus_chamaemorus_L_genome_as_determined_by_ddRAD_tags

Michael, K. (2006). *Clarification of basal relationships in Rubus (Rosaceae) and the origin of Rubus chamaemorus*. https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1253&context=theses

Norling, M., Jareborg, N., & Dainat, J. (2018). EMBLmyGFF3: a converter facilitating genome annotation submission to European Nucleotide Archive. *BMC Research Notes*, *11*(1), 584.

Price, R. J., Davik, J., Fernandéz Fernandéz, F., Bates, H. J., Lynn, S., Nellist, C. F., Buti, M., Røen, D., Šurbanovski, N., Alsheikh, M., Harrison, R. J., & Sargent, D. J. (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. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432.

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 245.

Sim, S. B., Corpuz, R. L., Simmonds, T. J., & Geib, S. M. (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, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*(1), 62.

Thompson, M. 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–D531.

Urbaniak, J., Kwiatkowski, P., & Pawlikowski, P. (2023). Genetic diversity and population structure of the endangered *Rubus chamaemorus* populations in Poland. *Global Ecology and Conservation*, *47*(e02633), e02633.

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B. S., Guo, H., Kissinger, J., & Paterson, A. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7), e49–e49.

Zhou, C., Brown, M., Blaxter, M., The Darwin Tree of Life Project Consortium, McCarthy, S. A., & Durbin, R. (2024). Oatk: a de novo assembly tool for complex plant organelle genomes. In *bioRxiv* (p. 2024.10.23.619857). https://doi.org/10.1101/2024.10.23.619857

Zhou, C., McCarthy, S. A., & Durbin, R. (2023). YaHS: yet another Hi-C scaffolding tool. *Bioinformatics (Oxford, England)*, *39*(1), btac808.