Towards whole plastome phylogeography: Resolving small genetic distances among European Arnica montana L. with the PlastidPipeline

Siddharth J. Annaldasula
Freie Universität Berlin
Manuel Vera
University of Santiago de Compostela
Adrián Casanova
University of Santiago de Compostela
Thomas Borsch
Freie Universität Berlin
Katja Reichel
katja.reichel@fu-berlin.de

Freie Universität Berlin

Article

Keywords: Arctic-alpine flora, Asteraceae, clonal evolution, heteroplasmy, Holarctic floristic region, pan-plastomics

Posted Date: August 1st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7219874/v1

Additional Declarations: No competing interests reported.

Abstract

Arnica montana L. is a medicinal plant endemic to Europe and declining in large parts of its range. Low sequence diversity in genetic marker regions caused biogeographic patterns and the phylogenetic relationship between *A. montana* and its Holarctic congeners to remain obscure. We therefore evaluated entire chloroplast genomes of *A. montana* and related species as complex markers. To reliably detect even small genetic distances, we developed the PlastidPipeline, which consistently and repeatably provides structurally standardized and annotated, ready-to-analyze plastomes from short-read sequences.

Complementing publicly available data, we obtained plastid genomes for eight *A. montana* accessions across Europe, four further species of *Arnica* and three outgroup species. All *A. montana* plastomes formed a well-supported clade. In contrast with the current circumscription of *A. montana* ssp. *atlantica*, both *A. montana* plastomes from the Iberian Peninsula were differentiated from those of Central and Northern Europe. Very low genetic distances among the Central European samples imply a common ancestor much more recent than the split from the Iberian lineage. Several plastome SNPs within *A. montana* further showed potential heteroplasmy. Rather than individual marker regions, phylogeographic studies in *Arnica* should use complete plastid genomes, including microstructural variation and intra-individual polymorphism in their analyses.

Introduction

Chloroplast DNA (cpDNA) sequences have long served as information source for phylogeographic studies in plants[1,2]. Research first focused on marker regions easy to amplify across a wide range of plant taxa[3,4]. Meanwhile, whole chloroplast genomes, also called plastid genomes or plastomes, have become available for an increasing number of species[5] (the GenBank database contains plastomes for > 8 000 species [02/2025]), mostly through assembly from high throughput, short-read sequence data[6,7]. Various biodiversity sequencing programs have facilitated this development, such as the PhyloAlps, DNAmark and PhyloNorway projects[8–12] aiming to provide reference plastome sequences for hundreds of species. While whole plastid genomes can be used to find and target marker regions which are polymorphic within specific taxa[13–15], directly comparing entire plastome sequences ("pan-plastomics") is also possible[16–19]. The latter is especially helpful if genetic distances between plastomes in the studied group are small, as expected for recent evolutionary radiations[20] or within-species phylogeographic comparisons. Yet, due to their sensitivity to even small errors, such comparisons require extra high standards for the correct assembly and further analysis of the sequence data.

Beside the nuclear and mitochondrial genomes, plastid genomes are a third, independently evolving genomic compartment in plant cells. Currently, their sequences are usually assembled from short-read whole genome sequencing (WGS) data representing a mixture of all three genomes, which, together with their specific structure and its inherent polymorphism, is a technical challenge. Plastomes typically exist in high copy numbers (hundreds to thousands)[21,22] within a single plant cell and, except in non-photosynthetic plants, harbor roughly 100 genes on 120–160 Kbp of DNA sequence[23]. They are traditionally depicted as circular molecules with a characteristic quadripartite structure, consisting of a long single copy region (LSC) and a short single copy region (SSC), separated by two inverted repeats (IR, or IRA/IRB for distinction). Depending on the orientation of the two single copy regions relative to each other, two functionally equivalent structural variants (known as haplotype A and B[24]) may co-occur in nature[25,26], in combination with various other structural morphs[27]. Accounting for this ambiguity in more or less sophisticated ways, a number of pipelines for the *de novo* assembly of plastid genomes from WGS data were developed over the years[28–30]. However, not all of them provide equally reliable results[31,32]. Consequently, the "linearized" plastid genome sequences deposited in nucleotide sequence databases may not just flip between structural variants (Supplementary Figs. S1.1-3), but also contain assembly artefacts such as spurious (e.g. duplicated or non-plastid) sequence stretches, within-plastome chimerism or gaps. Especially in studies involving closely related (con-generic/con-specific) specimens, published plastome sequences should therefore not be trusted blindly[31,32]. Still, if the raw sequence data is provided as well, re-assembly with standardized methods may ensure that biologically interpreted genetic differences are not merely technical artefacts.

Although chloroplast genomes are commonly pictured to evolve clonally and as a single linkage group[33] (excepting gene transfer[34,35]), dramatic differences in the level of polymorphism between different taxa and regions of the plastome have long been known[4,13]. This ranges from molecular marker regions otherwise used for genetic barcoding being too uniform for species identification in some groups[2,15], across major structural and functional mutations[24,36], to intra-individual plastome polymorphism attributed to various causes[37–40]. While it is often possible to find informative within-species plastid marker diversity as a basis for biogeographic studies[4], there also appear to be exceptions, where screening even a great number of classic plastid marker regions yields only (almost) uniform sequences. One example of such exceptionally low plastid marker diversity is *Arnica montana* L.[41–45].

The genus *Arnica* (Asteraceae) has 29 accepted species[41] and a circumboreal distribution, with centers of species diversity in the North American Cordillera (25 species) and the Appalachians (eight species)[46]. Polyploid, hybridizing *Arnica* species complexes are known from North America (e.g. *A. cordifolia*[47], but also *A. angustifolia*[48]). Yet, this does not apply to European *A. montana*: this particular lineage is diploid and geographically isolated from all other *Arnica* species. Even with its geographically closest congener, *A. angustifolia* in northern Scandinavia, no hybrids are known[49,50].

The range of *A. montana* extends from the western Iberian Peninsula to western Ukraine and from northern Italy to central Norway (Fig. 1). The plants grow in naturally open habitats in mountains and ecologically similar habitats in the lowlands[51,52]. Since almost a century, *A. montana* populations especially in the lowlands have declined dramatically, which can be attributed to several reasons including land use change, climate change as well as harvesting for medicinal use[44,50–52]. From the North-West of the Iberian Peninsula (Galicia and parts of Portugal), a subspecies *A. montana* ssp. *atlantica* has been described[53], but later questioned since its morphological basis proved unreliable[54]. However, as the distribution of the subspecies stated in the protologue roughly corresponds to the geographic distribution of two plastid haplotypes, based on *rps16* intron and *pafII-cemA* (= *ycf4-cemA*) spacer sequences[42,55,56], the taxon was upheld. Based on microsatellite length polymorphism, another study found strong differentiation between Galician and Central European *A.*

montana populations[57]. In general, the evolutionary relationships between *A. montana* populations in disjunct parts of its range, as well as between *A. montana* and other *Arnica* species, are currently not well known.

To test the power of using complete plastid genomes as complex markers for elucidating the evolutionary history of *A. montana*, we studied *Arnica* plastome diversity at different levels: At the inter-specific level, we compared the plastomes of three outgroup species with four arctic *Arnica* species and *A. montana*. At the intra-specific level, we compared the plastomes of eight *A. montana* accessions collected throughout Europe. To test the reliability of the intra-specific comparison, we also studied the intra-individual diversity of sequence reads from each sample mapping to sites showing polymorphism within *A. montana*. Particularly, we wanted to answer the following questions:

1. Is *A. montana* a monophyletic lineage within the genus *Arnica*?
2. Are there spatial patterns of genetic diversity within *A. montana*, especially linked to Quaternary climate shifts?
3. Which plastid genome polymorphisms can be reliably used to study migration history in *A. montana*?

Since we expected very low genetic distances between some samples, we had to exclude spurious polymorphism caused by technology as far as possible. We therefore developed the PlastidPipeline, a re-usable bioinformatic snakemake pipeline for the standardized and reproducible *de novo* generation of annotated plastid genomes from whole genome, short read sequencing data. This pipeline is introduced here along with the biological results.

Results

Pipeline

For each of the 23 datasets obtained from 15 samples (Table 1), we assembled a complete circular plastid genome sequence. Assembly parameters and statistics are given in Table 2 and Supplementary Table S2.1. Contrary to our expectation, the datasets acquired through PCR-free library preparation did not contain a lower percentage of duplicate reads; we therefore did not specifically filter these reads out. According to the backmapping step of the pipeline, the shares of plastome reads in the raw datasets (excluding downsampled data) ranged from 2.2% (Amon_BW) to 10.3% (Amon_MV), with a median of 5.5%. Both these libraries are from *A. montana* and were prepared in parallel from high-quality input DNA in the same lab, suggesting only individual or stochastic variation caused the difference. Plastomes with mean read depths between 100 and 1 000 still showed considerable variation in their read depth minimum: In Auna_AK, mean read depth is well over 100, yet local read depth at one site dropped as low as 15 (12.4%). With the exception of Vcar_CA and Amon_BB, samples with high mean read depths of > 1 000 had minimum:mean read depth ratios from 0.5 to 0.7, indicating more even read depth than in all others, where these ratios were below 0.5. Datasets of the herbarium material-derived libraries, particularly from the PhyloNorway project, were also generally smaller, so that effects of source material and sequencing effort are confounded.

Table 1
Overview of specimens included in this study. Sample names consist of a four-letter species code and the two-letter International Standards Organization (ISO) code for the country, or state/province of origin if multiple samples per country. cult. BG: cultivated in Botanic Garden, * no coordinates available, IDs for herbarium vouchers preceded by Herbarium Code and a colon.

Sample ID	Species	Lat	Long	Locality	Further IDs	Project	Reference
Amon_AG	*Arnica montana*	58.21	6.57	Norway, Agder, Rasvåg on Hidra Island	TROM:TROM_V_59412	PhyloNorway	8
Amon_DK	*Arnica montana*	55.65	8.17	Denmark, Syddanmark, Børsmose	DM447	DNAmark	10
Amon_MV	*Arnica montana*	54.39	12.70	Germany, Mecklenburg-Western Pommerania, Barth	AM09_09	BO Berlin	44 / this study
Amon_BB	*Arnica montana*	51.51	13.77	Germany, Brandenburg, Lauchhammer / cult. BG Berlin	B:B 10 1154693, AM0057	ArnicaSNP	this study
Amon_BW	*Arnica montana*	47.86	8.02	Germany, Baden-Württemberg, Feldberg	AM21_34	BO Berlin	44 / this study
Amon_CH	*Arnica montana*	46.39	7.42	Switzerland, Bern, Lenk	GR:GR003708, PHA000838	PhyloAlps	12
Amon_CT	*Arnica montana*	42.33	2.23	Spain, Catalonia, Vilallonga de Ter / cult. BG Barcelona	BC:BC-997190	rDNA-LIFE	58
Amon_GL	*Arnica montana* ssp. *atlantica*	43.28	-7.86	Spain, Galicia, Ponte Pedrido / cult. USC Lugo	AY5634	CONSERVARNICA	this study
Aang_FI	*Arnica angustifolia* ssp. *alpina*	70.62	29.81	Norway, Finnmark, Båtsfjord	TROM:TROM_V_962577	PhyloNorway	8
Agri_AK	*Arnica griscomii* ssp. *frigida*	63.05	-151.00	USA, Alaska, Denali National Park	TROM:TROM_V_300499	PhyloNorway	8
Ales_AK	*Arnica lessingii*	65.36	-146.41	USA, Alaska, White Mountains	TROM:TROM_V_300501	PhyloNorway	8
Auna_AK	*Arnica unalaschcensis*	54.16	-165.93	USA, Alaska, Akutan Island	TROM:TROM_V_300511	PhyloNorway	8
Ecan_CT	*Eupatorium cannabinum*	42.37	2.30	Spain, Catalonia, Setcases	BC:BC-997195	rDNA-LIFE	58
Gtul_MX	*Guardiola tulocarpus*	-	-	Mexico*	TEX:Panero 3062	Asteraceae Phylo-transcriptomics	59
Vcar_CA	*Venegasia carpesioides*	-	-	USA, California, Santa Barbara area*	JEPS:Baldwin 893	Asteraceae Phylo-transcriptomics	59

Table 2

Overview of read and assembly statistics for all datasets in this study. Reads – read length, Pairs – no. of raw read pairs, Filtered – no. of read pairs after qu with identical sequence in the data, R – GetOrganelle extension rounds for assembly, w – GetOrganelle wordsize for assembly, Mem. – memory needed fo assembly in minutes, Plastome – plastome length, LSC/SSC/IR – length of respective plastome part, Insert – mean insert size of assembled read pairs, Dep depth on the assembled plastome, Mapped – percentage of raw reads mapping to the assembled plastome.

Dataset	SRA run ID	Library prep	Reads [bp]	Pairs [M]	Filtered [M]	Dupl. [%]	R	w	Mem. [Gb]	Time [min]	Plastome [bp]	LSC [bp]	SSC [bp]	IR [bp]
Amon_AG	ERR8268641	with PCR	100	3.37	3.32	0.46	30	66	3.2	18	152 089	83 843	18 304	24 971
Amon_DK	SRR12432532	with PCR	150	8.08	7.56	2.14	27	65	38.0	213	152 090	83 844	18 304	24 971
Amon_MV	ERR15302247	with PCR	150	22.86	22.27	2.95	30	85	60.55	177	152 091	83 845	18 304	24 971
Amon_BB	ERR15310125	PCR-free	150	675.51	648.56	4.68	30	112	14.97	61	152 070	83 843	18 279	24 971
Amon_BB-d	*downsampled*	PCR-free	150	36.17	35.11	0.24	30	90	9.83	124	152 070	83 849	18 279	24 971
Amon_BW	ERR15302248	with PCR	150	18.30	17.69	2.08	30	112	17.08	111	152 094	83 849	18 303	24 971
Amon_CH	ERR14008899	with PCR	150	5.35	5.35	0.18	11	85	8.28	8	152 087	83 842	18 303	24 971
Amon_CT	SRR17032105	with PCR	150	4.77	4.39	2.51	30	84	14.24	41	151 989	83 750	18 297	24 971
Amon_GL	*aggregate*	PCR-free	150	361.65	354.78	2.29	30	112	16.58	96	151 999	83 757	18 300	24 971
Amon_GL-d	*downsampled*	PCR-free	150	224.83	220.94	1.88	40	100	29.1	142	151 999	83 716	18 300	24 971
Amon_GL-71	ERR15340092	PCR-free	150	189.51	186.21	1.88	30	112	16.58	99	151 999	83 757	18 300	24 971
Amon_GL-72	ERR15340095	PCR-free	150	192.20	189.22	2.13	30	112	15.98	89	151 999	83 757	18 300	24 971
Amon_GL-73	ERR15340093	PCR-free	150	190.09	185.11	1.89	30	112	15.73	95	151 999	83 757	18 300	24 971
Amon_GL-74	ERR15340094	PCR-free	150	184.84	181.84	2.79	30	112	16.52	96	151 999	83 757	18 300	24 971
Amon_GL-Y1	ERR15337219	PCR-free	150	60.38	59.15	2.22	30	112	15.83	89	151 999	83 757	18 300	24 971
Amon_GL-Y2	ERR15336770	PCR-free	150	61.16	59.93	2.19	30	112	3.97	23	151 999	83 757	18 300	24 971
Aang_FI	ERR5554746	with PCR	100	3.83	3.75	1.21	24	66	3.47	14	152 052	83 809	18 301	24 971
Agri_AK	ERR5529436	with PCR	100	4.45	4.39	1.00	30	65	3.97	23	151 981	83 733	18 306	24 971
Ales_AK	ERR5529317	with PCR	100	4.22	4.18	0.40	30	65	5.97	21	152 056	83 840	18 274	24 971
Auna_AK	ERR5529299	with PCR	100	6.12	6.06	0.35	12	85	10.18	12	152 008	83 716	18 262	25 016
Ecan_CT	SRR17032099	with PCR	150	3.99	3.74	3.10	30	112	6.21	20	151 384	83 024	18 312	25 024
Gtul_MX	SRR12917849	with PCR	150	22.40	21.55	1.56	17	111	12.15	52	151 368	83 219	18 163	24 993
Vcar_CA	SRR12917857	with PCR	150	16.57	15.72	1.69	30	110	10.85	70	152 351	83 866	18 326	25 037

While plastomes we assembled from different technical replicates were always identical, there were two indels between the assemblies for Amon_GL based on the original and downsampled datasets, respectively. Due to the plastid genome's much higher copy number per cell, plastome reads tend to be overrepresented in the data and thus get selectively depleted during frequency-dependent downsampling, as performed here (Table 2). With the frequencies of similar sequence plastome and nuclear genome reads closer to each other, extension-based assembly algorithms, such as employed in GetOrganelle, may be more prone to errors. Downsampling also did not speed up the assembly: Since GetOrganelle will only use as much of the data as needed to produce

contigs of the expected size, it may actually stop earlier in a large, but “plastome-biased” whole genome dataset. Other than that, we did not find correlations between input file characteristics (read pairs, percentage of reads mapping to plastome) and computational effort (memory use, wall-clock time).

Multiple runs with the same parameters consistently produced the same assembled sequences for our data, yet slight parameter changes sometimes had an effect: For the Vcar_CA dataset, assemblies with different pre-set numbers of assembly rounds (R = [10,15]) differed by 7 bp in the length of an A/T-homopolymer in the *atpE-ndhC* spacer. A comparison with the read depth graph produced by the PlastidPipeline, where the region in question shows a considerable drop, and with a previously published NOVOPlasty assembly[60] of the same dataset, in which the polyA/T-stretch is only 31 bp, not 85 or 78 bp long, suggests that this polymorphism may be an assembly artefact, e.g. from the faulty incorporation of similar sequence nuclear genome reads. For Gtul_MX, there was no difference between our assembly and the one previously published.

Plastomes

The assembled plastomes for the outgroup samples were larger and less uniform in size than within the genus *Arnica*, where all except the Auna_AK sample had exactly the same IR length. While SSC lengths of Central and Northern European (CNE) and Iberian (IB) *Arnica montana* samples were minimal, the LSCs of the Iberian samples were both at least 85 bp shorter. All *Arnica* plastomes contained the same 86 (80 unique) coding genes, 36 (29) tRNA genes and eight (four) rRNA genes. They did not contain *chlB, chlL, chlN* and *psaM*, as expected for angiosperms[61], but also lacked *trnE*-UUC, *trnP*-GGG and *trnR*-CCG. While the latter two tRNA genes for the transfer of proline and arginine have functional paralogs, the *Arnica* (and outgroup) plastomes appear to have entirely lost the ability to synthesize glutamic acid transfer RNA (*trnE*), possibly due to a gene transfer into the mitochondrial genome[62]. Non-canonical start codons were detected for *psbC*, *rps19* (both GTG) and *psbC, psbL* (both ACG). In Amon_CH and Gtul_MX, the *matK* gene had a premature stop codon, caused by a frameshift mutation due to a polyA/T length polymorphism and leading to a loss of 81 amino acids (16%) of the translated sequence.

Inter-specific variation

All *A. montana* plastomes we analysed belonged to a highly supported monophyletic lineage (Fig. 2c). Apart from the early divergence of Aleutian / NE Asian *A. unalaschcensis*, the relationships between the plastomes of European (*A. montana*) and the other Beringian (*A. griscomii, A. lessingii*) or circumpolar (*A. angustifolia*) *Arnica* species are not well supported. While Neighbor-Joining did not resolve them at all (Supplementary Fig. S1.4), all other phylogenetic inference methods depicted *A. lessingii* + (*A. montana* + (*A. griscomii* + *A. angustifolia*)) as the best-supported topology (Fig. 2c). Partitioning data into LSC, IR and SSC for RAxML did not change this result. Molecular clock analyses put the origin of *A. montana* at [2.33, 4.01] Ma (strict molecular clock) or [1.87, 4.34] Ma (relaxed molecular clock; Supplementary Fig. S1.5). As dating is based on secondary calibration with the split between the transcriptomes of *A. montana*, and *A. chamissonis* not included in this study, it should be taken with caution.

Polymorphism among the *Arnica* plastomes exists mostly in the LSC and SSC. Based on single nucleotide polymorphisms (SNPs) only, the most informative plastome regions for distinguishing *A. montana* from other *Arnica* species in this study were the *rps16* intron and *ndhF* gene, with two SNPs each (Fig. 2a). The *trnQ*-UUU gene, *trnY-rpoB* spacer, *trnG-trnR* spacer, *psbD* gene, *trnV*-UAC intron, *rbcL* gene, *trnI*-GAU intron, *ndhD* gene and *ccsA* gene each contained one SNP (Supplementary Table S2.2). This corresponds to two amino acid changes in the *rbcL* (Glu > Lys) and *ccsA* (His > Asn) amino acid sequences which are characteristic for *A. montana*. Apart from the *rps16* intron, none of these regions were previously studied in *Arnica*.

Intra-specific variation

The comparison of *A. montana* plastomes showed a very clear distinction between CNE and IB plastomes (Fig. 3b, d). Despite a significant Mantel test (r_M = 0.687, p = 0.007), this is not due to isolation by distance – at similar geographic distances of ca. 750–800 km, the pairwise genetic distance between CNE samples is smaller than between the two IB samples, and much smaller than between samples belonging to different groups (Fig. 3b). The observed pattern is also not in accordance with the proposed split between *A. montana* ssp. *atlantica* in Galicia and *A. montana* ssp. *montana* in the Eastern Pyrenees and the rest of Europe: at the previously studied *pafII-cemA* marker region, Amon_CT has the same, divergent genotype as Amon_GL, yet none of the samples in this study showed polymorphism in the *rps16* intron (IB: cHap1, CNE: cHap3 combined haplotypes[42]). The comparison with previously published sequences of plastid markers from further *A. montana* individuals, including one from an unspecified location in France[41], yielded no further informative characters. The marker sequences representing Central France and Belgium[45] were identical to those of the CNE plastomes.

Among the CNE samples, the Mantel test was not significant (r_M = 0.373, p = 0.069), but plotting distances on a map showed less genetic distance between samples towards the North, and less total polymorphism than between the two Iberian samples (Fig. 3a, d). This corresponds with the parsimony network (Fig. 3c) and in part with the phylogenetic tree (Fig. 2c), which is, however, based on an alignment simplified by removing multi-state indels. Genetic distances between CNE samples ranged from two to eleven, with the lowest genetic distance between the North German (Amon_MV) and Danish (Amon_DK) sample, collected at a direct geographic distance of ca. 320 km.

With two exceptions in the *rrn16* and *trnA*-UGC genes, polymorphism among the *A. montana* plastomes was concentrated in the single copy regions. Of 67 phylogenetically informative SNP loci across all *A. montana* samples, 62 clearly indicated the IB/CNE split. In comparison with the other *Arnica* species, the two Iberian samples showed the ancestral state for 33 SNPs, whereas for 25 SNPs only the CNE samples had genotypes identical with those outside *A. montana*. The regions containing most SNPs across all studied *A. montana* plastomes were the *matK* gene (3 SNPs) and 5’ adjacent part of the *trnK* intron (1), *trnS*-GCU-*trnC*-GCA spacer (3), *rpoC2* gene (3), *ycf1* gene (8), *ndhA* intron (4) and adjacent *ndhA* exon 2 (1), *ndhI-ndhG* spacer (4), *rpl32-trnL*-UAG spacer (3), *rpl32-ndhF* spacer (2) and adjacent *ndhF* gene (2). In the *rpoB, rpoC2, rbcL, rps11, rpl22, ycf1* and *ccsA* genes, a total of ten SNPs cause differences in the amino acid sequence between the CNE and IB *A. montana* samples, while a further five SNPs caused amino acid changes in single samples. Based on an amino acid similarity network[63], the coding sequence changes between IB and CNE plastomes (mean node distance: 4.1) cause greater change in amino acid function than those not associated with the split (mean node distance: 3.6; Supplementary Table S2.5).

Intra-individual read variation

We found low-frequency variant reads in all *A. montana* read datasets we analyzed. With the notable exception of one coding SNP in the 5'-part of the ycf1 gene, which overlaps into the inverted repeat, all internal polymorphisms were located in the single copy regions. Indels were usually associated with copy number variations in homopolymer stretches or microsatellites; variation for long indels (> 5 bp) or inversions was not detected by our method. In Amon_CH, the variant reads included a full-length version of the poly-A/T-stretch causing the matK frame shift mutation at a frequency of 2.5% (5 reads). For the two samples with the lowest genetic distance, Amon_DK data contained the Amon_MV variant of both divergent sites in low frequency (SNP: 5.9%, poly-A/T: 15.9%).

For our datasets acquired by various methods, the frequency distribution of alternate SNP alleles differed widely between samples (Fig. 4a): In the three datasets with the highest mean read depths, loci with alternate reads at 1–2% frequency were most common, this distribution shifting its peak to higher frequencies and spreading out with decreasing mean read depth. There was no clear correlation between the total number of detected internal polymorphisms and the number of input or plastome-mapping reads, suggesting factors other than constant polymerase or read error rates shape these distributions. The two technical replicates Amon_GL-Y1/Y2 had similar, though not identical patterns of internal polymorphism. The third-largest dataset Amon_MV, containing an exceptionally high proportion of plastome-mapping reads, had comparatively few internal polymorphisms. Apart from having the lowest mean read depth to begin with, we did not detect any patterns specific to the only herbarium specimen-derived dataset, Amon_AG. In the downsampled dataset Amon_BB-d, the shape of the frequency distribution reflected the low amount of plastome reads in the data, with many more apparently polymorphic loci detected. For all samples except Amon_BB-d and Amon_CT, three to seven out of the SNP loci with the highest minor allele frequencies were located at or near a local read depth peak in the atpI-atpH spacer (≈ 25 600–25 900 bp from the *psbA* end of the LSC).

Slightly more than half (36) of the sites at which we found SNPs between *A. montana* samples also showed within-sample polymorphism (Fig. 4b). Except for Amon_AG, where five sites only had read depths above 65, all internally polymorphic sites had read depths above 100, with at least four reads representing minor allele frequencies mostly between one and five percent, except in Amon_BB-d where they went up to 24.5% (Supplementary Table S2.4). There was no transition/transversion bias for these sites, yet we found a striking pattern: With one exception where a third allele was present, any alternate allele found within samples was also found at the intraspecific level. Moreover, and again with one exception, only samples with one of the two alleles, i.e. either IB or CNE samples, showed internal polymorphism, suggesting a directionality of the change. SNPs with polymorphism in either *A. montana* group were equally frequent in our data. The internally polymorphic loci supporting the IB/CNE split included the previously studied TC/GT double substitution in the *pafII-cemA* spacer, for which we found minor allele frequencies of 4.5% in Amon_DK and 10.3% in Amon_BW.

Discussion

Using whole plastid genomes for phylogeographic studies is both feasible and advisable, particularly in species with low plastome diversity such as *Arnica montana*. With our study, we demonstrated advantages of this approach, particularly increased resolution. We highlighted possible methodological pitfalls and strategies to avoid them, especially when genetic distances between plastomes are small.

Moving further towards findable, accessible, interoperable and reusable (FAIR)[64] plastome data, we introduced the PlastidPipeline as a tool to standardize and harmonize the *de novo* assembly and annotation of complete plastid genomes in a user-friendly "one-stop shop" pipeline, including both metadata annotation and quality checking steps in an easily adaptable and extensible snakemake framework. In contrast to nuclear genomes, where international networks establish reporting standards[65], and despite individual attempts[66,67], there currently are no universal conventions for plastid genomes and previously published plastomes should not be taken on trust[31]. To ease reuse and comparison, we propose LSC > IRA > SSC > IRB, with *rbcL* and *ndhF* in forward direction and the inverted repeats named in reading order, as a universal standard for submission to public databases, being aware that this is an artificial structure not reflective of the natural diversity of plastid genome conformations[24,40]. Though it can be adapted to fit individual projects' needs, this is also the standard output conformation of our pipeline. Comparing the PlastidPipeline with the existing plastaumatic pipeline[68], we exchanged NOVOPlasty for GetOrganelle as the standard assembly tool based on the better reproducibility of its results[31], included GeSeq as the standard annotation tool, and exchanged the generation of database-specific extra submission files with a backmapping step for quality control. Data requirements and potential technical issues with our pipeline are identical to those of the individual tools: Assuming 10% loss of reads in quality filtering, a plastome read content range of [0.02, 0.05] and a minimum:mean read depth ratio of [0.1, 0.5], 2.2 to 27.5 million 150 bp WGS read pairs should suffice to assemble a plastome ca. 150 000 bp long to a read depth of at least 100. We recommend truncating large input files, rather than downsampling reads by frequency[31]. GetOrganelle can fail to assemble complete plastomes, sub-optimally assemble homopolymer stretches[69] or erroneously incorporate nuclear genome reads. This may be remedied by varying the assembly parameters, checking for coverage drops in the backmapping data and, if possible, re-sequencing of doubtful sequence stretches. GeSeq may produce sub-optimal annotations[68], including a persistently wrong-sense IR annotation, or remnant duplicate features as multiple annotation approaches are used. Though our pipeline automatically catches many of these annotation issues, we recommend always manually checking its output. As no tool is perfect, we strongly encourage all future users of the PlastidPipeline to make any improvements to the current software which they see fit, while keeping backward compatibility and attribution.

In this study, we report the first fully assembled and annotated chloroplast genomes for eight samples of the European endemic *A. montana* (Fig. 2a), as well as for four specimens identified as other *Arnica* species (Fig. 1), and (re-)assemblies for three outgroup species (Table 1). We confirmed monophyly of *A. montana*, as well as the early branching of northeast Asian *A. unalaschcensis* within the genus (Fig. 2c, Supplementary Fig. S1.4)[41]. The sparsity of the currently available data does not allow further phylogenetic conclusions, e.g. with respect to an origin of the genus *Arnica* in arctic or subarctic western North America[70], temperate western North America[71] or the boreal northwest Pacific Rim[41]. For *A. montana*, its sister relationship to a clade comprising Beringian *A. griscomii* and circumpolar *A. angustifolia*, here represented by an individual from Norway, suggests it may have originated from southward

migration in Europe out of a circumpolar species group during the late Pliocene / early Pleistocene cooling[72–76] (Supplementary Fig. S1.5). Both *A. montana* and *A. angustifolia* thus most likely arrived independently in Europe. Sedimentary ancient DNA (*sed*aDNA) currently represents the only (sub-)fossil record of the genus, and confirms the presence of *Arnica sp.* in central Alaska since at least 15 000 years[77], on Svalbard at least 11 000 years ago[78] and in the Austrian Alps for at least 3 000 years[79]. However, it does not allow further retrospection in time or the discrimination of *Arnica* species. A comprehensive phylogenetic study of *Arnica* needs a better taxonomic and geographic representation, especially for wide-spread species such as *A. angustifolia*. Moreover, it should include nuclear genome information[80,81], for which two *A. montana* nuclear reference genomes currently being sequenced will be of great value.

Within *A. montana*, we found a deep split between our two IB and six CNE plastomes, a disproportionally large genetic distance between the IB samples and a northward decline in genetic distances within the CNE group. The tentative molecular dating of the IB-CNE split to the onset of the early Pleistocene climate oscillations[82,83] ([1.31, 3.21] Ma, Supplementary Fig. S1.5) suggests that *A. montana* spread throughout large parts of Europe (compare recent distribution, Fig. 3d) during an early glacial phase, but then had its distribution area permanently subdivided. Additional samples from southern France and the Pyrenees could help to better understand this pattern, which does, however, not coincide with the previously described subspecies. The comparatively large genetic distance between the two IB samples, despite their identical haplotypes at the previously studied *pafII-cemA* and *rps16* marker loci, lends further support to the chemotypic and genetic distinctiveness of populations in Galicia and northern Portugal[42,55–57], possibly caused by multiple independent refugia on the Iberian Peninsula[84,85]. Rather than classic glacial refugia[86], these were more likely warm-time refugia for boreo-alpine plants[87,88]. For the CNE samples, the genetic distance pattern (Fig. 3c, d) is consistent with northward recolonization, as previously suggested based on nuclear microsatellite genetic diversity[44,89]. Few parsimony informative characters and apparent homoplasy among the six plastomes make further interpretation difficult. With our data, we found no evidence for multiple refugia in CNE, as all six samples formed a shallow clade (Fig. 2c, 3a). The strong small-scale genetic structure observed in previous studies is thus more likely due to low dispersal/strong genetic isolation and, at least periodically, a high rate of clonal reproduction causing increased genetic drift[90]. Again, data for more individuals and a comparison with the nuclear genome would be needed to further elucidate relationships within both the CNE and IB groups. Our results once more highlight the importance of genetic studies to guide conservation, and of the effective protection and sustainable management of populations of *A. montana* throughout its range.

Both at the genus level and within *A. montana*, plastid genome polymorphisms were widely distributed mainly across the single copy regions, suggesting whole plastomes or nuclear genome data are most informative on migration history. Choosing only part of a genome to represent the evolutionary history of the whole is not trivial, since, beside technical or financial limitations and the general risk of homoplasy, the observed polymorphism needs to correspond with the diversity level under study. In our data, we found five regions (highlighted in Supplementary Table S2.2) with SNPs that characterize samples as *A. montana*, and with additional SNPs which could characterize samples of different geographic origin (IB/CNE). Remarkably, nearly all of these were coding regions, had only one or two SNPs at either diversity level and showed additional read polymorphism at their intra-specific, even amino-acid-changing, SNPs. Except for the *rbcL* gene, neither region had previously been proposed as a marker or barcoding region[2,4], or, to our knowledge, been sequenced in *Arnica*. Especially in taxa with low plastome diversity, comparing data from “standard” individually sequenced short chloroplast markers may thus give incomplete or misleading results, and should be avoided.

At more than half of the SNPs distinguishing IB and CNE plastomes, we found one of the alleles also present in low frequency in samples with the alternative allele, which could not generally be explained by mismapping or sequence-specific polymerase bias. Potential intra-individual polymorphism in chloroplast genome data has only recently started to gain attention[37–39,91]. Given the importance it has for understanding the evolution of widely-used genetic markers, its potential to explain apparent missense mutations, such as for the premature *matK* stop codon we observed in Amon_CH, and following the increasing number of pan-plastome studies analyzing differences between closely related samples[16–19], we expect more data on potential (transient) heteroplasmy to become available in the future[5]. So far, distinguishing technical artefacts from biological polymorphism is difficult: For our data, minor allele frequency spectra did not provide indications such as multiple peaks, and were apparently dependent on sequencing depth and the proportion of plastome vs. other reads within WGS datasets, making a quantitative comparison of potential heteroplasmy between samples even from the same experiment (e.g. Amon_MV and Amon_BW) statistically challenging. To better understand the significance of our observation, we suggest routinely inspecting backmapping data of plastid genome assemblies, such as provided by the PlastidPipeline, for read polymorphism.

Methods

Sequence data

The raw sequence data we used for this study was either acquired in our own labs, i.e. three samples of *A. montana* from Germany and one from NW Spain, obtained directly from the authors (PhyloAlps) or downloaded from GenBank or the European Nucleotide Archive (ENA). The geographical origin of *Arnica* samples relative to each species’ distribution area is displayed in Fig. 1, with more details given in Table 1. This data includes all currently (March 2025) publicly available WGS datasets for the genus *Arnica*.

Total genomic DNA came either from herbarium vouchers (PhyloNorway, Gtul_MX, Vcar_CA), flash-frozen material intended for nuclear genome sequencing (samples Amon_BB and Amon_GL) or silica-dried leaf samples (all others). Library preparation for Illumina sequencing usually involved additional PCR amplification of the sample (e.g. Illumina TruSeq Nano DNA Kit for Amon_BW and Amon_MV), except for the two nuclear genome samples where the Illumina TruSeq PCR-free kit was used. The German samples (Amon_BB, Amon_MV, Amon_BW) were sequenced on an Illumina HiSeq X at Macrogen Inc., Seoul, whereas the NW Spanish specimen (Amon_GL) was sequenced on an Illumina NovaSeq 6000 at the Spanish National Center of Genomic Analysis (Centro Nacional de Análisis Genómico, CNAG), Barcelona. Technical information about each read dataset, including read lengths (100 bp for PhyloNorway, 150 bp for all others) and read pair numbers, is given in Table 2.

Because of the size of the nuclear genome datasets (Amon_BB, Amon_GL), we also tested bioinformatically downsampled versions of the full data (Amon_BB-d, Amon_GL-d) obtained with bbnorm[92] and a target read depth of 200x (following a previous study[31]). For Amon_GL, data was available from multiple sequencing runs of the same sample, which we treat as technical replicates and distinguish by the last character of the run + lane identifiers, i.e. Amon_GL-71 to -74 for HFYM3DSX7_1 to _4 and Amon_GL-Y1 and -Y2 for HGM7JDMXY_1 and _2. The Amon_GL dataset was concatenated from part of lanes one and three of HFYM3DSX7.

PlastidPipeline

We developed the PlastidPipeline to create a cohesive pipeline for the reproducible *de novo* assembly and annotation of plastid genomes. It uses snakemake[93], a Python based workflow management system that allows for independent cross validation, adaptation and automation of data analysis. Due to their modular nature, snakemake pipelines easily allow input at different steps, or to resume the analysis with changed parameters starting from an intermediate result. Parameter settings for each step are fully configurable by the user through a configuration file, defaulting to the standard settings of the respective tool unless specified otherwise. For functionality not covered by the available software tools, e.g. structural standardization of the plastome sequence, the pipeline contains additional bash or Python scripts written by the authors, particularly using Biopython and Selenium[94–96]. The PlastidPipeline is available from GitHub (https://github.com/sannalda/plastidpipeline) and was developed and tested on FU Berlin's Curta HPC system[97].

A schematic overview of the PlastidPipeline is given in Fig. 5. Starting from the most common input, i.e. raw paired-end sequence reads in .fastq(.gz) file format, it typically performs the following steps: First, reads are trimmed and quality filtered using fastp[98]. Second, GetOrganelle[99] is used to assemble the plastid genome: Briefly, reads are first mapped to standard seed sequences from an embryophyte plastome database. Through iterative extension with SPAdes[100], more target-associated reads are added and contigs assembled *de novo*. Non-target contigs are filtered out and the remaining targeted organelle contigs are enchained into a complete circular plastid genome. Critical GetOrganelle parameters for the assembly are the number of extension rounds -R, which we doubled to 30 by default in the PlastidPipeline, and the "word size" -w, i.e. the length of sequence used to detect read similarity, which can be set for each pipeline run. Especially if the assembly fails, the contig graphs can be manually checked with Bandage[101] (here 0.9.0) and -R increased / -w decreased against fragmentation, or the inverse against entanglement.

The third step in the PlastidPipeline is to standardize the orientation of the two single copy regions and to annotate the plastome sequence: First, GeSeq[102] is used for a preliminary annotation, which identifies the four parts of the plastid genome and one gene each in the LSC and SSC, specified by the user along with their intended orientation. By default, these are the highly conserved chloroplast household genes *rbcL* in the LSC and *ndhF* in the SSC, both in forward orientation (i.e. haplotype A[24]). According to the linearized isomorph randomly produced by GetOrganelle (four possibilities, see Supplementary Fig. S1.1-3), the plastome parts are split up, potentially reversed and reassembled into the required form. Subsequently, GeSeq is called again to annotate the standardized plastome sequence, and a script is run to test for the completeness / duplication of annotations ("blatX/N" annotations being kept preferentially by default), verify the plausibility of coding sequence features and to correct potential mislabeling of the IRs. In a test on one sequence, GeSeq produced a more complete annotation of our data compared to its most popular command line alternative, PGA[103], though it is currently only available as a web application and thus cannot be directly integrated into an automated software pipeline. The PlastidPipeline therefore contains a script which simulates using the GeSeq web interface, setting parameters (e.g. including both "blat" and "Chloe" annotators) and uploading/downloading files according to the user's specifications. It also automatically corrects names and orientation of the IRs, removes duplicate annotations and produces a test report for the presence/absence of standard plastid genes[67], start/stop codons, reading frame lengths and translations of coding sequences.

The fourth step of the PlastidPipeline outputs backmapping information for the input reads, using Bowtie 2 and samtools[104,105]. This is intended for further quality checks, e.g. the detection of apparent heteroplasmy or regions with low or uneven support, for which a figure and a per-base read count file are produced. Finally, the fifth step of the pipeline optionally adds sample metadata, as specified in an additional input text file per sample, to the header of the final annotated .gb (Genbank) file the PlastidPipeline produces per input dataset.

Though the produced files are directly ready for use in further analyses (e.g. whole plastome alignment, phylogenies), we strongly recommend to first visually inspect them for any undetected errors. To submit the final plastid genomes to an INSDC database, the output Genbank files still have to be converted according to the respective database's requirements. Sequence submission is therefore not included in the PlastidPipeline. To format our data for submission to ENA, we used custom scripts including file conversion with Biopython[95] and the ENA webin-cli programmatic submission interface.

Data analysis

We visually inspected the plastid genomes produced by the PlastidPipeline in Geneious Prime 2025.0.3, using its MAFFT[106,107] interface to create preliminary alignments. The assembled sequences for Vcar_CA and Gtul_MX we compared with previous NOVOPlasty assemblies[60]. For detailed analyses, we then built three separate alignments at different levels: Excluding the technical replicates, we first aligned only the *A. montana* sequences ("*Arnica montana*" = AM alignment), then, keeping previous gaps, successively added first the other *Arnica* plastomes ("all *Arnica*" = AA alignment), and finally the three non-*Arnica* samples ("all samples" = AS alignment). At each step, we manually checked the alignment for undetected short indels or inversions (> 2 bp long), harmonized gap limits and removed the second IR (IRB) while, in AS, keeping variable sites at the beginning & end of the IRB to account for IR length variation in the non-*Arnica* outgroup. For the AM alignment which comprised the most similar sequences, we kept all polymorphic characters including gaps. We encoded these in two versions, either with indels/inversions as multistate characters (MSI) or by simple indel coding (SIC)[108]. For the AA alignment, we only kept gaps part of two-state characters, and removed all multistate characters involving gap "alleles", i.e. mostly length variable A/T-homopolymers, from the analysis. For the AS alignment, we removed all characters with gaps in at least one sample, keeping only unambiguously aligned single nucleotide polymorphisms (SNPs).

For all alignments, we calculated Pairwise Character Differences (PCD = “Hamming distance”) in Geneious, excluding and, where applicable, including indels / inversions, with gaps as a fifth character state.

We employed different criteria to infer phylogenetic trees, using the plugins for PAUP* 4.0a168 (Neighbor Joining and Parsimony)[109], MrBayes 3.2.6 (Bayesian Posterior Probability)[110], PhyML 3.3.20180621 (Maximum Likelihood, with different branch stability criteria)[111] and RAxML 8.2.11 (Maximum Likelihood, with different data partitioning schemes)[112] in Geneious. Based on the AS alignment, the *A. unalaschcensis* plastome consistently branched first in the *Arnica* clade (Supplementary Fig. S1.4). The Auna_AK plastome therefore served as outgroup for our inference of the *Arnica* phylogenetic tree based on the AA alignment. We used BEAST 2.7.7. with both a strict and a relaxed molecular clock model[113,114] to date a posterior probability-based phylogeny of all *Arnica* samples, visualizing the result with TreeAnnotator. Since it is currently the only available data, we set the range of estimates given in a recent Asteraceae phylotranscriptomics study[59] for the split between the *A. montana / A. chamissonis* transcriptomes, i.e. [7.73, 11.92] Ma, as uniform prior for the secondary calibration of the *A. unalaschcensis* branching point in our tree. All resulting divergence times are slight underestimates, as *A. chamissonis* belongs to the *Arnica* crown group[41] rather than being a sister lineage to all its congeners. Among *A. montana* plastomes, we employed the R package haplotypes 1.1.3.1 on the AM-MSI alignment to infer a parsimony network[115]. Additionally, we mapped all publicly available sequences of *A. montana* plastome marker regions (GenBank/ENA identifiers: AM690489, AM690504/27/40/67/82, AM690611/25/43/54, KF602249, KF679679/80/81/82, PV340595/96, PV344507/08) to our full sequence alignment. One previous plastome assembly for Amon_DK, marked as unverified (MT899992), was incomplete and not used in our analyses.

Individual WGS short-read datasets often contain alternative sequence reads[37–39,91], which may coincide with polymorphism at the population or phylogenetic levels. In the context of our study, such reads may have six sources: Partial degradation of the DNA templates, e.g. in old herbarium vouchers[116], PCR artefacts during library preparation, errors produced during the sequencing process itself[117], mismapping of similar sequences from other genomic compartments, e.g. related to nuclear-plastid transfers (NUPTs)[35], fading or incipient plastome polymorphism[39] or permanent heteroplasmy. Whereas mismapped sequences could be detectable through causing a local peak in read depth, the other sources are harder to distinguish. Since our study included datasets from diverse sources (herbarium vs. flash frozen material, PCR-free vs. PCR-based library preparation), comparing apparent intra-individual sequence polymorphism from different samples might shed some light on its origin and the reliability of the final plastome sequences. Based on the backmapping .bam files produced by the PlastidPipeline, we detected sequence variants with minor allele frequencies ≥ 0.01 for the *A. montana* plastomes (Amon_GL: Amon_GL-Y1, Amon_GL-Y2, Amon_BB: Amon_BB-d) with Geneious. We focused our analysis on SNPs, excluding the first 100 bases of the LSC, where mapping to the start of the linearized sequence is incomplete, and duplicate information from the IRB.

To run statistical analyses and draw diagrams, we used R 4.4.2 with the ggplot2, geosphere 1.5–18 and vegan 2.6–6.1 packages[118–121]. The pan-plastome map in Fig. 2a was modified from an OGDraw plot[122]. For geographic maps, we loaded and filtered observation data from gbif [123] with the rgbif 3.8.0 and CoordinateCleaner 3.0.1 R packages[124,125], combining it in QGIS 3.36 with background map data from www.naturalearth.org[126].

Declarations

Acknowledgements

We gratefully acknowledge the help of all people who provided additional data for our study, particularly Inger Greve Alsos, Sébastien Lavergne, José Panero, Daniel Vitales Serrano and Caifei Zhang. We warmly thank our colleagues, especially Rosa Romero and Manuel Antonio Rodríguez-Guitián for fieldwork in Galicia and, like Emma Behlke, Astrid de Mestier, Markus Dillenberger, Belén Escobari, Francisco Fajardo Gutiérrez, Max Fussan, Konstantina Koutroumpa and Jozefien van de Velde in Berlin, for helpful discussions and critical feedback. We are indebted to Sabine Scheel and Cristina Alonso for, respectively, preparing the German and Galician *Arnica* samples for sequencing, to Antonio Alfaro de Prado Martín for writing scripts for the upload to ENA, and to Virginia Duwe for information on her previous marker sequencing experiments. We thank the team of the Berlin Center for Genomics in Biodiversity Research (BeGenDiv) for lab space and support, Macrogen Inc. and the CNAG Barcelona for sequencing, the HPC Service of Zedat, FU Berlin, for computing time, and Pascal Lehwark and the Max-Planck-Institute in Potsdam-Golm for permission to re-use and remotely access GeSeq.

Funding declaration

This study was financed by BO Berlin core funding, the CONSERVARNICA contract from the Deputación de Lugo (Galicia, Spain) to Manuel Vera and an Initiative Grant from FU Berlin to Katja Reichel.

Author contribution statement

SJA conceived, implemented, tested and documented the PlastidPipeline, supported by KR, who contributed to programming and documentation, conceived the study, provided and acquired data, performed the biogeographic analyses and wrote a first draft of the manuscript. TB and MV participated in the interpretation of the results and, together with AC, provided data and critical comments. All authors contributed to the manuscript and read and approved its final version.

Competing interests

The authors declare no competing interests.

Data accessibility statement

All data included in this study are available from the INSDC databases under the study reference number PRJEB89801 and the reference numbers provided in the text, in Table 2 and in Supplementary Table S2.1. The PlastidPipeline and its documentation are available on Github, https://github.com/sannalda/plastidpipeline.

References

1. Linder, H. P. & Crisp, M. D. *Nothofagus* and Pacific biogeography. *Cladistics* **11**, 5–32 (1995).
2. Hollingsworth, P. M., Graham, S. W. & Little, D. P. Choosing and using a plant DNA barcode. *PLoS One* **6**, e19254 (2011).
3. Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol Biol* **17**, 1105–1109 (1991).
4. Shaw, J. *et al.* Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *Am J Bot* **101**, 1987–2004 (2014).
5. Wang, J. *et al.* Plant organellar genomes: much done, much more to do. *Trends Plant Sci* **29**, 754–769 (2024).
6. Tonti-Filippini, J., Nevill, P. G., Dixon, K. & Small, I. What can we do with 1000 plastid genomes? *Plant J* **90**, 808–818 (2017).
7. Twyford, A. D. & Ness, R. W. Strategies for complete plastid genome sequencing. *Mol Ecol Resour* **17**, 858–868 (2017).
8. Alsos, I. G. *et al.* The treasure vault can be opened: Large-scale genome skimming works well using herbarium and silica gel dried material. *Plants* **9**, 432 (2020).
9. Margaryan, A. *et al.* Mitochondrial genomes of Danish vertebrate species generated for the national DNA reference database, DNAmark. *Environ DNA* **3**, 472–480 (2021).
10. Chua, P. Y. S. *et al.* Towards the extended barcode concept: Generating DNA reference data through genome skimming of Danish plants. 2021.08.11.456029 Preprint at https://doi.org/10.1101/2021.08.11.456029 (2021).
11. Wang, Y. *et al.* Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature* **600**, 86–92 (2021).
12. Lavergne, S. *et al.* Towards a comprehensive barcoding and phylogenomic reference for the European arctic-alpine flora. Preprint at https://doi.org/10.21203/rs.3.rs-6136147/v1 (2025).
13. Korotkova, N., Nauheimer, L., Ter-Voskanyan, H., Allgaier, M. & Borsch, T. Variability among the most rapidly evolving plastid genomic regions is lineage-specific: Implications of pairwise genome comparisons in *Pyrus* (Rosaceae) and other angiosperms for marker choice. *PLoS One* **9**, e112998 (2014).
14. Song, F. Complete plastome sequencing resolves taxonomic relationships among species of *Calligonum* L. (Polygonaceae) in China. (2020).
15. Li, X. *et al.* Plant DNA barcoding: from gene to genome. *Biol Rev* **90**, 157–166 (2015).
16. Sielemann, K. *et al.* Complete pan-plastome sequences enable high resolution phylogenetic classification of sugar beet and closely related crop wild relatives. *BMC Genomics* **23**, 113 (2022).
17. Wang, J. *et al.* The Asian lotus (*Nelumbo nucifera*) pan-plastome: diversity and divergence in a living fossil grown for seed, rhizome, and aesthetics. *Ornam Plant Res* **2**, 1–10 (2022).
18. Go, S. *et al.* Pan-chloroplast genomes for accession-specific marker development in *Hibiscus syriacus*. *Sci Data* **11**, 246 (2024).
19. Sun, S.-S., Pan, Z.-Y., Fu, Y., Wang, S.-J. & Fu, P.-C. Rampant intraspecific variation of plastid genomes in *Gentiana* section *Chondrophyllae*. *Ecol Evol* **14**, e70239 (2024).
20. Escobari, B., Borsch, T., Quedensley, T. S. & Gruenstaeudl, M. Plastid phylogenomics of the Gynoxoid group (Senecioneae, Asteraceae) highlights the importance of motif-based sequence alignment amid low genetic distances. *Am J Bot* **108**, 2235–2256 (2021).
21. Sakamoto, W. & Takami, T. Chloroplast DNA dynamics: Copy number, quality control and degradation. *Plant Cell Physiol* **59**, 1120–1127 (2018).
22. Fernandes Gyorfy, M. *et al.* Nuclear–cytoplasmic balance: whole genome duplications induce elevated organellar genome copy number. *Plant J* **108**, 219–230 (2021).
23. Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F. & Quandt, D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* **76**, 273–297 (2011).
24. Wang, W. & Lanfear, R. Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants. *Genome Biol Evol* **11**, 3372–3381 (2019).
25. Palmer, J. D. Chloroplast DNA exists in two orientations. *Nature* **301**, 92–93 (1983).
26. Walker, J. F., Jansen, R. K., Zanis, M. J. & Emery, N. C. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. *Am J Bot* **102**, 1751–1752 (2015).
27. Bendich, A. J. Circular chloroplast chromosomes: The grand illusion. *Plant Cell* **16**, 1661–1666 (2004).
28. Bakker, F. T. *et al.* Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biol J Linn Soc Lond* **117**, 33–43 (2016).
29. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res* **45**, e18 (2017).
30. Freudenthal, J. A. *et al.* A systematic comparison of chloroplast genome assembly tools. *Genome Biol* **21**, 254 (2020).
31. Giorgashvili, E. *et al.* Software choice and sequencing coverage can impact plastid genome assembly – a case study in the narrow endemic *Calligonum bakuense*. *Front Plant Sci* **13**, 779830 (2022).

32. Wu, Z., Tembrock, L. R. & Ge, S. Are differences in genomic data sets due to true biological variants or errors in genome assembly: An example from two chloroplast genomes. *PLoS One* **10**, e0118019 (2015).
33. Maréchal, A. & Brisson, N. Recombination and the maintenance of plant organelle genome stability. *New Phytol* **186**, 299–317 (2010).
34. Wang, H., Liao, X., Tembrock, L. R., Yang, Z. & Wu, Z. Evaluation of intracellular gene transfers from plastome to nuclear genome across progressively improved assemblies for *Arabidopsis thaliana* and *Oryza sativa*. *Genes* **13**, 1620 (2022).
35. Zhang, G.-J. *et al.* Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. *Int J Mol Sci* **21**, 707 (2020).
36. de Mestier, A., Fálcon-Hidalgo, B. & Borsch, T. Plastid phylogenomics of the parietal clade of Malpighiales: ancestral genome synteny and congruent signal accross exons, introns and spacers in support of Salicaceae relationships. *Ann Bot* mcaf148 (2025) doi:10.1093/aob/mcaf148.
37. Scarcelli, N. *et al.* Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? *Mol Ecol Resour* **16**, 434–445 (2016).
38. Sun, S.-S. *et al.* Intra-individual heteroplasmy in the *Gentiana tongolensis* plastid genome (Gentianaceae). *PeerJ* **7**, e8025 (2019).
39. Khachaturyan, M., Santer, M., Reusch, T. B. H. & Dagan, T. Heteroplasmy is rare in plant mitochondria compared with plastids despite similar mutation rates. *Mol Biol Evol* **41**, msae135 (2024).
40. Lee, C., Ruhlman, T. A. & Jansen, R. K. Unprecedented intraindividual structural heteroplasmy in *Eleocharis* (Cyperaceae, Poales) plastomes. *Genome Biol Evol* **12**, 641–655 (2020).
41. Ekenäs, C., Baldwin, B. G. & Andreasen, K. A molecular phylogenetic study of *Arnica* (Asteraceae): Low chloroplast DNA variation and problematic subgeneric classification. *Systematic Botany* **32**, 917–928 (2007).
42. Vera, M. *et al.* Phylogeography and genetic variability of the *Arnica montana* chemotypes in NW Iberian Peninsula. *Silvae Genet* **63**, 293–300 (2014).
43. Van Rossum, F. & Raspé, O. Contribution of genetics for implementing population translocation of the threatened *Arnica montana*. *Conserv Genet* **19**, 1185–1198 (2018).
44. Duwe, V. K., Muller, L. A. H., Borsch, T. & Ismail, S. A. Pervasive genetic differentiation among Central European populations of the threatened *Arnica montana* L. and genetic erosion at lower elevations. *Perspect Plant Ecol Evol Syst* **27**, 45–56 (2017).
45. Van Rossum, F. *et al.* Taxonomic and genetic assessment of disjunct populations of the threatened plant *Arnica montana* (Asteraceae) from central France. *Perspect Plant Ecol Evol Syst* **67**, 125873 (2025).
46. Natural Resources Conservation Service. USDA Plants Database. *United States Department of Agriculture* https://plants.usda.gov (2025).
47. Kao, R. H. Origins and widespread distribution of co-existing Polyploids in *Arnica cordifolia* (Asteraceae). *Ann Bot* **101**, 145–152 (2008).
48. Downie, S. R. Morphological, cytological, and flavonoid variability of the *Arnica angustifolia* aggregate (Asteraceae). *Can J Bot* **66**, 24–39 (1988).
49. SLU Artdatabanken. Artfakta: slåttergubbe (*Arnica montana*). https://artfakta.se/taxa/219955 (2025).
50. Vikane, J. H., Rydgren, K., Jongejans, E. & Vandvik, V. Rainfall and temperature change drive *Arnica montana* population dynamics at the Northern distribution edge. *Oecologia* **191**, 565–578 (2019).
51. Hollmann, V. *et al.* From nutrients to competition processes: Habitat specific threats to *Arnica montana* L. populations in Hesse, Germany. *PLoS One* **15**, e0233709 (2020).
52. Stanik, N., Lampei, C. & Rosenthal, G. Summer aridity rather than management shapes fitness-related functional traits of the threatened mountain plant *Arnica montana*. *Ecol Evol* **10**, 5069–5078 (2020).
53. de Bolós y Vayreda, A. El *Arnica montana* L. en la Peninsula Ibérica. *Farmacognosia. Anales del Instituto José Celestino Mutis* **4**, 145–151 (1945).
54. Romero, R. *et al.* Estudio biométrico de *Arnica montana* L. (Asteraceae) y su relación con factores ambientales en el extremo occidental cantábrico. in *Actes del IX Colloqui Internacional de Botànica Pyrinieo-cantàbrica* (Ordino, Andorra, 2011).
55. Vera, M. *et al.* Living at the edge: population differentiation in endangered *Arnica montana* from NW Iberian Peninsula. *Plant Syst Evol* **306**, 44 (2020).
56. Bouza, C. *et al.* Genetic survey extension of the threatened Iberian *Arnica montana* L. revealed the presence of divergent plastid lineages and highly structured populations in northern Spain. *Genet Resour Crop Evol* **70**, 1677–1689 (2023).
57. Schmiderer, C., Torres-Londono, P., Lutz-Röder, A., Duwe, V. K. & Novak, J. *Arnica montana* subsp. *atlantica*: Really a subspecies? *Genet Resour Crop Evol* **65**, 1829–1838 (2018).
58. Pascual-Díaz, J. P., Garcia, S. & Vitales, D. Plastome diversity and phylogenomic relationships in Asteraceae. *Plants* **10**, 2699 (2021).
59. Zhang, C. *et al.* Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. *J Integr Plant Biol* **63**, 1273–1293 (2021).
60. Nguyen, T. L., Nguyen, H. D., Vu, N. H. & Vu, M. T. Plastome sequencing of *Gymnanthemum amygdalinum* and phylogenetic analysis of 17 newly assembled Asteraceae plastomes. *J Asia Pac Biodivers* **17**, 820–826 (2024).
61. Mohanta, T. K. *et al.* Gene loss and evolution of the plastome. *Genes (Basel)* **11**, 1133 (2020).
62. Damiano, F., Ceci, L. R., Siculella, L. & Gallerani, R. Transcription of two sunflower (*Helianthus annuus* L.) mitochondrial tRNA genes having different genetic origins. *Gene* **286**, 25–32 (2002).
63. Stephenson, J. D. & Freeland, S. J. Unearthing the root of amino acid similarity. *J Mol Evol* **77**, 159–169 (2013).
64. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
65. Böhne, A. *et al.* Contextualising samples: supporting reference genomes of European biodiversity through sample and associated metadata collection. *npj biodivers* **3**, 1–11 (2024).
66. Turudić, A. *et al.* Towards the well-tempered chloroplast DNA sequences. *Plants* **10**, 1360 (2021).

67. Qu, X.-J., Zou, D., Zhang, R.-Y., Stull, G. W. & Yi, T.-S. Progress, challenge and prospect of plant plastome annotation. *Front Plant Sci* **14**, (2023).
68. Chen, W., Achakkagari, S. R. & Strömvik, M. Plastaumatic: Automating plastome assembly and annotation. *Front Plant Sci* **13**, (2022).
69. Zhou, W. *et al.* Plastid genome assembly using long-read data. *Mol Ecol Resour* **23**, 1442–1457 (2023).
70. Maguire, B. A monograph of the genus *Arnica*. *Brittonia* (1943).
71. Baldwin, B. G., Wessa, B. L. & Panero, J. L. Nuclear rDNA evidence for major lineages of helenioid Heliantheae (Compositae). *Syst Bot* **27**, 161–198 (2002).
72. Jiménez-Moreno, G., Fauquette, S. & Suc, J.-P. Miocene to Pliocene vegetation reconstruction and climate estimates in the Iberian Peninsula from pollen data. *Rev Palaeobot Palynol* **162**, 403–415 (2010).
73. Khan, S. *et al.* Late Pliocene continental climate and vegetation variability in the Arctic-Atlantic gateway region prior to the intensification of Northern Hemisphere glaciations. *Palaeogeogr Palaeoclimatol Palaeoecol* **586**, 110746 (2022).
74. Tkach, N. V., Röser, M. & Hoffmann, M. H. Range size variation and diversity distribution in the vascular plant flora of the Eurasian Arctic. *Org DIvers Evol* **8**, 251–266 (2008).
75. Hoffmann, M. H. Not across the North Pole: Plant migration in the Arctic. *New Phytol* **193**, 474–480 (2012).
76. Eidesen, P. B. *et al.* Genetic roadmap of the Arctic: plant dispersal highways, traffic barriers and capitals of diversity. *New Phytol* **200**, 898–910 (2013).
77. Clarke, C. L. *et al.* Steppe-tundra composition and deglacial floristic turnover in interior Alaska revealed by sedimentary ancient DNA (*sed*aDNA). *Quat Sci Rev* **334**, 108672 (2024).
78. Voldstad, L. H. *et al.* A complete Holocene lake sediment ancient DNA record reveals long-standing high Arctic plant diversity hotspot in northern Svalbard. *Quat Sci Rev* **234**, 106207 (2020).
79. Zetter, S. *et al.* SedaDNA shows that transhumance of domestic herbivores has enhanced plant diversity over the Holocene in the Eastern European Alps. *Holocene* 09596836241307304 (2024) doi:10.1177/09596836241307304.
80. Ekenäs, C., Heidari, N. & Andreasen, K. *Arnica* (Asteraceae) phylogeny revisited using RPB2: Complex patterns and multiple d-paralogues. *Mol Phylogenet Evol* **64**, 261–270 (2012).
81. Ekenäs, C. *et al.* Secondary chemistry and ribosomal DNA data congruencies in *Arnica* (Asteraceae). *Cladistics* **25**, 78–92 (2009).
82. Frenzel, B. The Pleistocene vegetation of northern Eurasia. *Science* **161**, 637–649 (1968).
83. Szabó, P. *et al.* Pliocene - Early Pleistocene continental climate and vegetation in Europe based on stable isotope compositions of mammal tooth enamel. *Quat Sci Rev* **288**, 107572 (2022).
84. Gómez, A. & Lunt, D. H. Refugia within refugia: Patterns of phylogeographic concordance in the Iberian Peninsula. in *Phylogeography of Southern European Refugia* (eds. Weiss, S. & Ferrand, N.) 155–188 (Springer Netherlands, Dordrecht, 2007). doi:10.1007/1-4020-4904-8_5.
85. Feliner, G. N. Southern European glacial refugia: A tale of tales. *Taxon* **60**, 365–372 (2011).
86. Médail, F. & Diadema, K. Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *J Biogeogr* **36**, 1333–1345 (2009).
87. Gentili, R. *et al.* From cold to warm-stage refugia for boreo-alpine plants in southern European and Mediterranean mountains: the last chance to survive or an opportunity for speciation? *Biodiversity* **16**, 247–261 (2015).
88. Wootton, L. M. *et al.* The late rise of sky-island vegetation in the European Alps. *Nat Plants* 1–12 (2025) doi:10.1038/s41477-025-02001-0.
89. Capria, L., Liepelt, S., Eimert, K., Leyer, I. & Mosner, E. Neutral genetic diversity follows a latitudinal gradient in the endangered plant *Arnica montana* L.: a range-wide study. *Conserv Genet* (2023) doi:10.1007/s10592-023-01559-6.
90. Reichel, K., Masson, J.-P., Malrieu, F., Arnaud-Haond, S. & Stoeckel, S. Rare sex or out of reach equilibrium? The dynamics of F_{IS} in partially clonal organisms. *BMC Genet* **17**, 76 (2016).
91. Sabir, J. S. M. *et al.* Whole mitochondrial and plastid genome SNP analysis of nine date palm cultivars reveals plastid heteroplasmy and close phylogenetic relationships among cultivars. *PLoS One* **9**, e94158 (2014).
92. Bushnell, B. *BBTools*. https://sourceforge.net/projects/bbmap/ (2014).
93. Mölder, F. *et al.* Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).
94. Python Software Foundation. Python 3.13.2 Documentation. https://docs.python.org/3/index.html (2025).
95. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
96. The Selenium Browser Automation Project. Selenium 4.0 documentation. *Selenium* https://www.selenium.dev/documentation/ (2025).
97. Bennett, L., Melchers, B. & Proppe, B. Curta: A general-purpose high-performance computer at ZEDAT, Freie Universität Berlin. (2020) doi:10.17169/refubium-26754.
98. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
99. Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biology* **21**, 241 (2020).
100. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes *de novo* assembler. *Curr Protoc Bioinformatics* **70**, e102 (2020).
101. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: Interactive visualization of *de novo* genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
102. Tillich, M. *et al.* GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res* **45**, W6–W11 (2017).
103. Qu, X.-J., Moore, M. J., Li, D.-Z. & Yi, T.-S. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* **15**, 50 (2019).

104. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
105. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
106. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).
107. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
108. Simmons, M. P. & Ochoterena, H. Gaps as characters in sequence-based phylogenetic analyses. *Syst Biol* **49**, 369–81 (2000).
109. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. (Sinauer Associates, Sunderland, MA, 2002).
110. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
111. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321 (2010).
112. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
113. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88 (2006).
114. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019).
115. Aktas, C. *Haplotypes: Manipulating DNA Sequences and Estimating Unambiguous Haplotype Network with Statistical Parsimony*. https://cran.r-project.org/web/packages/haplotypes/index.html (2023).
116. Overballe-Petersen, S., Orlando, L. & Willerslev, E. Next-generation sequencing offers new insights into DNA degradation. *Trends Biotechnol* **30**, 364–368 (2012).
117. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3**, lqab019 (2021).
118. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2024).
119. Wickham, H. *Create Elegant Data Visualisations Using the Grammar of Graphics*. (Springer-Verlag, New York, 2016).
120. Hijmans, R. J. *geosphere: Spherical trigonometry*. https://CRAN.R-project.org/package=geosphere (2022).
121. Oksanen, J. *et al. Vegan: Community Ecology Package*. https://cran.r-project.org/web/packages/vegan/index.html (2025).
122. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res* **47**, W59–W64 (2019).
123. GBIF.org. GBIF Occurrence Download. https://doi.org/10.15468/dl.vy9cv6, https://doi.org/10.15468/dl.97g7cs, https://doi.org/10.15468/dl.p6yypa, https://doi.org/10.15468/dl.5zasby, https://doi.org/10.15468/dl.kx2ytu, https://doi.org/10.15468/dl.6tfvvk (2024).
124. Chamberlain, S. A. & Boettiger, C. R, Python, and Ruby clients for GBIF species occurrence data. Preprint at https://doi.org/10.7287/peerj.preprints.3304v1 (2017).
125. Zizka, A. *et al.* CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol Evol* **10**, 744–751 (2019).
126. QGIS Association. *QGIS Geographic Information System*. https://www.qgis.org/ (2024).

Figures

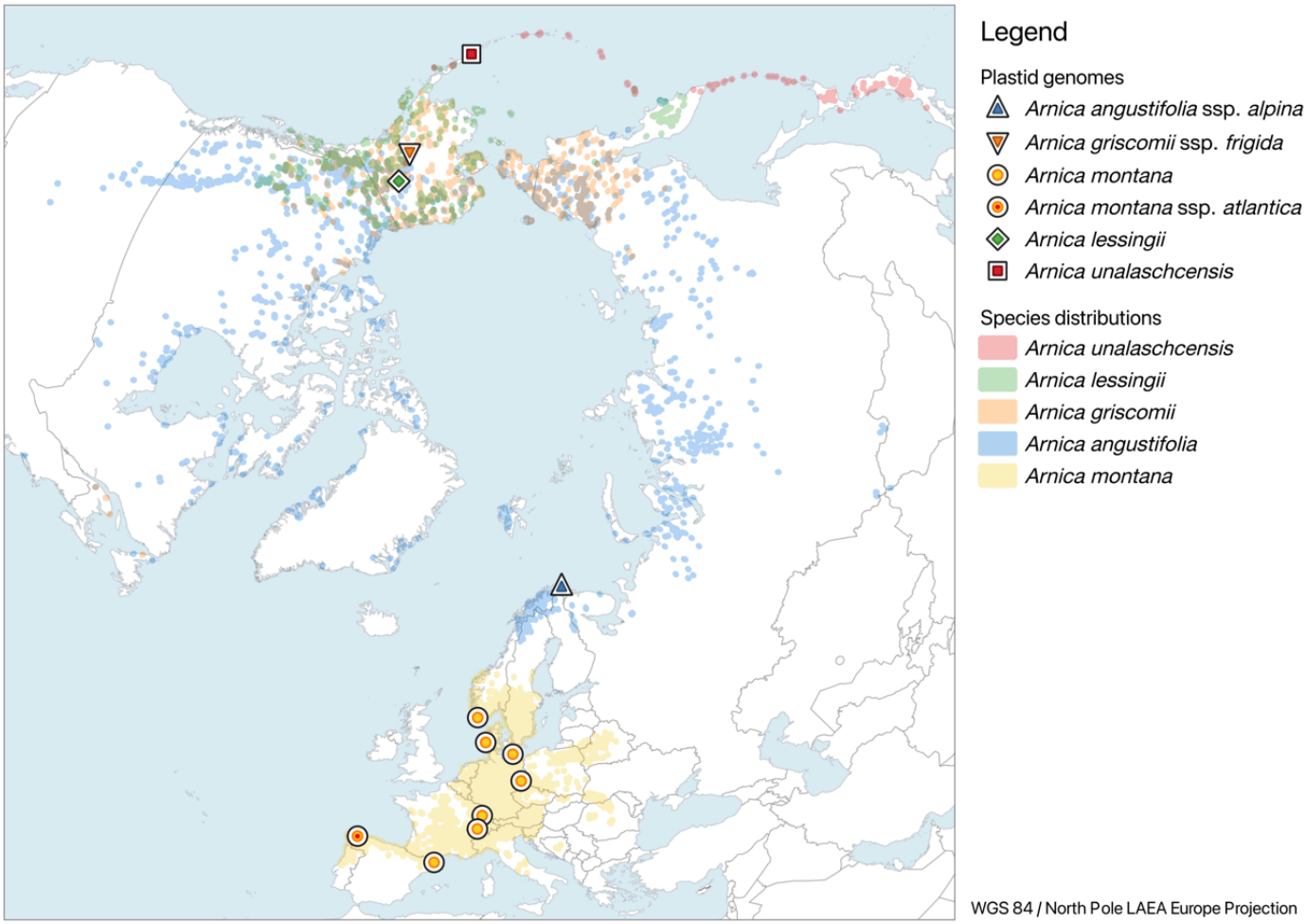

Figure 1

Sample origin and distribution map of *Arnica* species included in this study. Distribution areas of putative subspecies of *A. angustifolia* Vahl and *A. frigida* C.A. Mey. ex Iljin are not distinguished. Based on data from gbif.org (see text).

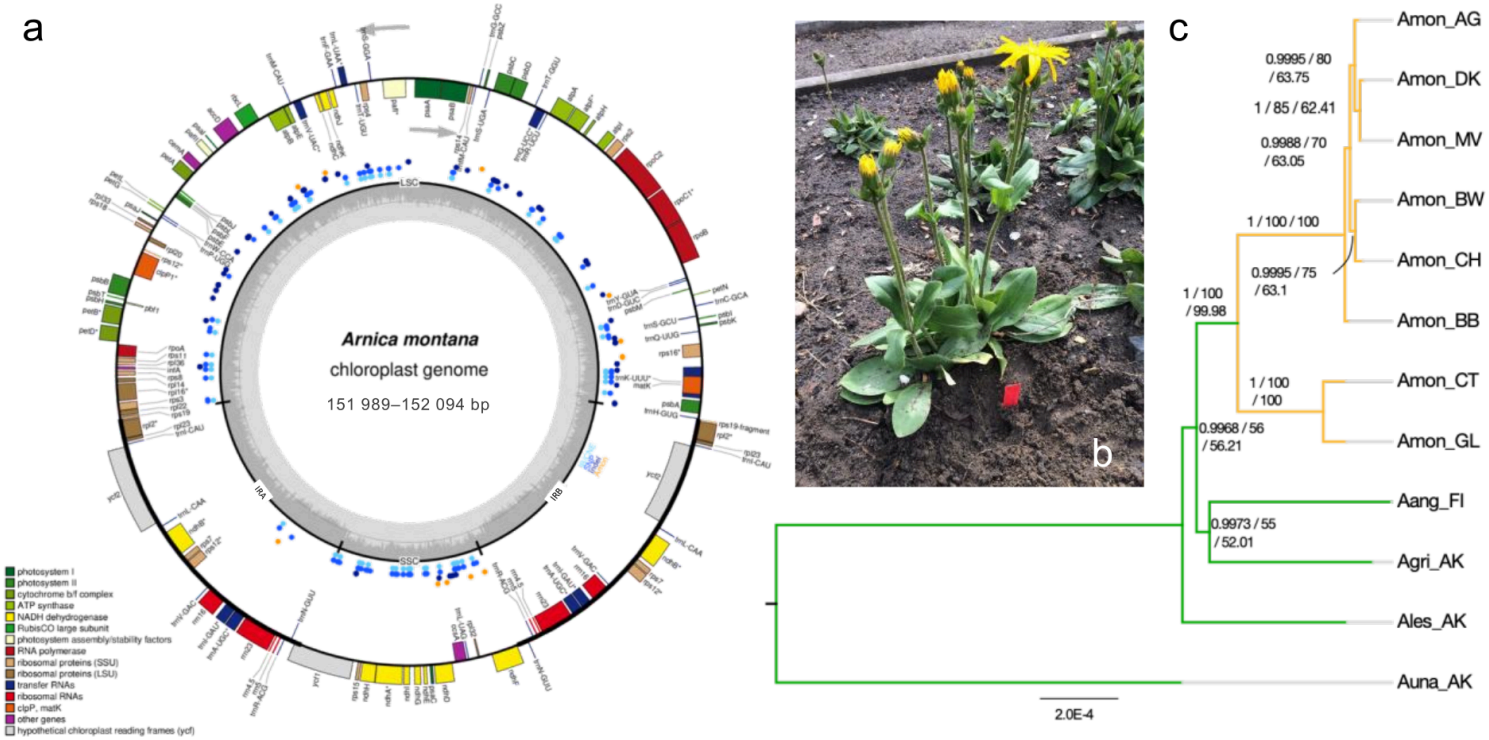

Figure 2

a. Pan-plastome map of *A. montana*. Species-characteristic SNPs (yellow), intraspecific indels (dark blue) and SNPs (mid blue), and geographic region-specific SNPs (light blue) marked by dots. b. *A. montana* habitus. c. Phylogram of the Bayesian consensus tree for all *Arnica* plastomes in this study. Based on SNPs and 2-state-indels only, multi-state indels removed. Tree topology is identical to the most strongly supported Parsimony and Maximum Likelihood trees. Branches labeled with posterior probability (MrBayes) and percentage support values (RAxML unpartitioned data / PAUP* parsimony) based on 10 000 bootstrap samples.

Genetic Distances	AS – all samples														
	AA – all *Arnica* samples**														
> SNP only	AM – *Arnica montana* samples*														
v SNP + indel	Amon_AG	Amon_DK	Amon_MV	Amon_BB	Amon_BW	Amon_CH	Amon_CT	Amon_GL	Aang_FI	Agri_AK	Ales_AK	Auna_AK	Vcar_CA	Gtul_MX	Ecan_CT
Amon_AG		3	2	4	4	5	73	73	114	107	109	317	1092	1672	1747
Amon_DK	6		1	5	5	6	74	74	115	108	110	318	1092	1672	1747
Amon_MV	6	2		4	4	5	73	73	114	107	109	317	1091	1671	1746
Amon_BB	7	9	9		4	5	71	71	112	105	107	315	1090	1670	1745
Amon_BW	10	10	8	11		3	73	73	114	107	109	317	1092	1672	1747
Amon_CH	9	11	10 / 11	10	10 / 11		72	72	115	108	110	318	1093	1671	1748
Amon_CT	108 / 114	109 / 116	108 / 116	108 / 113	109 / 116	109 / 113		14	112	102	104	319	1090	1672	1745
Amon_GL	110 / 117	110 / 117	109 / 117	110 / 116	111 / 119	111 / 116	25 / 27		110	102	104	319	1086	1672	1746
Aang_FI	137	138	137	137	138	139	134	136		113	120	323	1096	1682	1747
Agri_AK	127	128	127	127	128	129	123	125	138		112	323	1092	1666	1739
Ales_AK	133	134	133	133	134	135	129	131	151	142		321	1083	1662	1734
Auna_AK	381	382	381	381	382	383	384	386	392	389	395		1105	1677	1749
Vcar_CA														1448	1527
Gtul_MX	* MSI / SIC indel coding-based, if different								** 2-state indels only						2073
Ecan_CT															

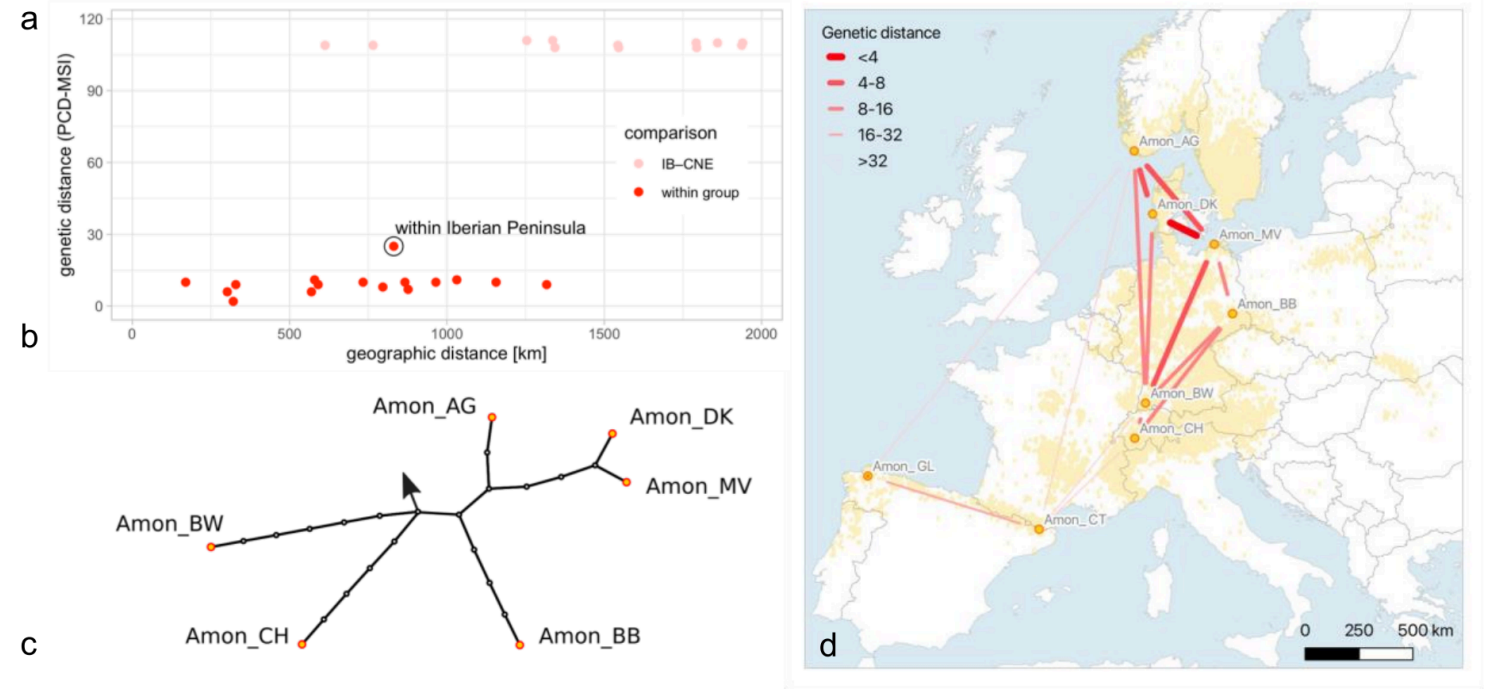

Figure 3

a. Heatmap of genetic distances between plastomes. Pairwise character distance (PCD) based on SNPs only (above diagonal) or both SNPs and indels/inversions (below diagonal). Indels coded with simple indel coding (SIC) or as multistate indels (MSI), where applicable. b. Mantel plot of geographic and genetic distances between all *A. montana* plastomes/samples. Distances within same geographic group, Iberian Peninsula (IB) or Central/Northern Europe (CNE) in red, between geographic groups in light red. c. Parsimony network of CNE *A. montana* plastomes. Yellow-red nodes observed, black nodes putative ancestral genotypes with one-step mutational distances, arrow connects to all other samples. d. Map and selected pairwise genetic distances between *A. montana* plastomes. Heavy/dark red lines correspond to low genetic distances. *A. montana* sampling sites and distribution area in yellow.

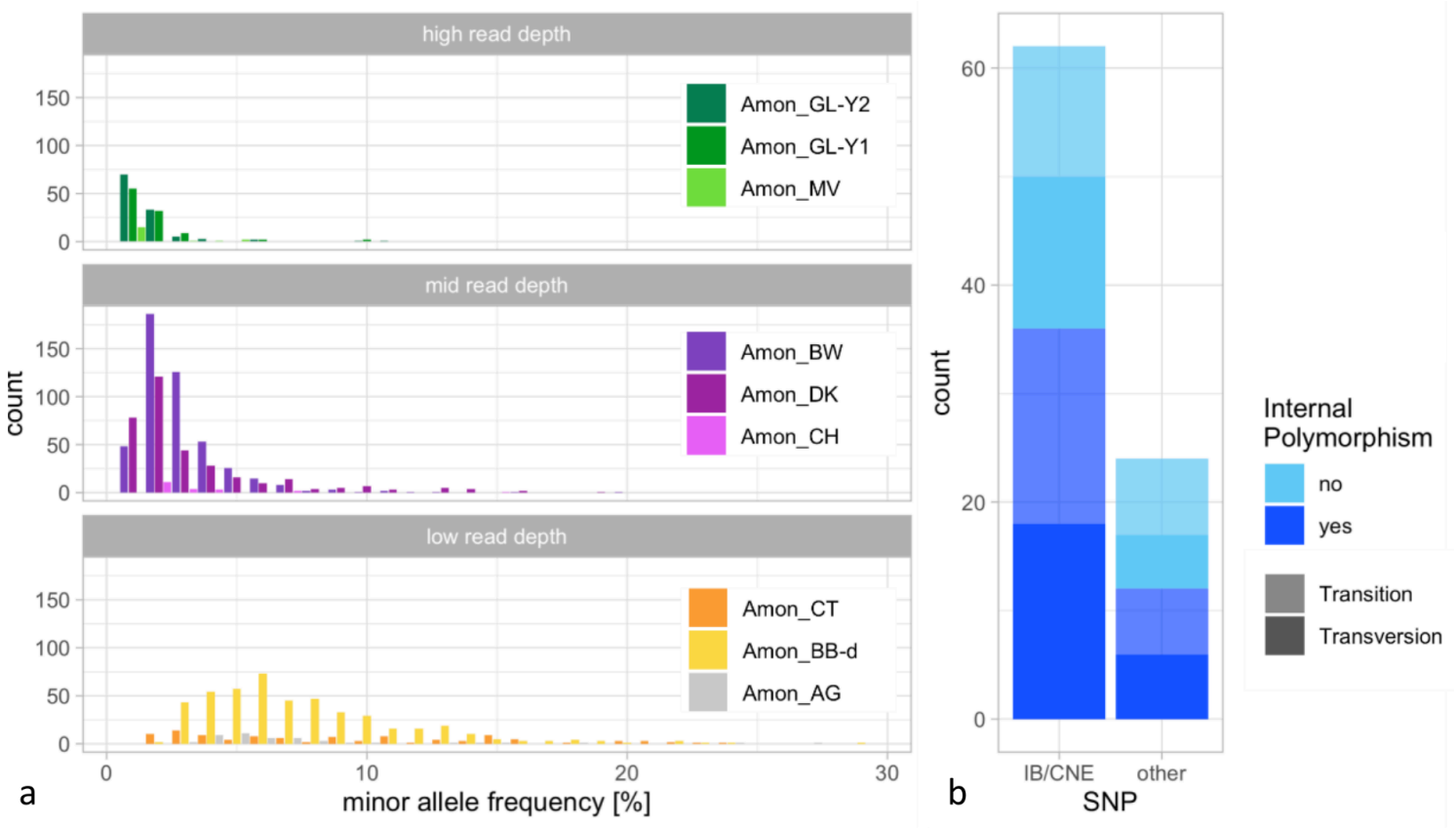

Figure 4

a. Histograms of minor allele frequencies at within-sample SNP sites for all studied *A. montana* plastomes. Samples grouped by plastome mean read depth: High read depth > 1000, mid read depth > 180, low read depth > 65. b. Within-sample polymorphism at sites of *A. montana* within-species SNPs. Loci categorized by association with the IB/CNE split (left / right), presence of internal polymorphism (dark blue / light blue) and type of SNP (transversion: solid, transition: faded).

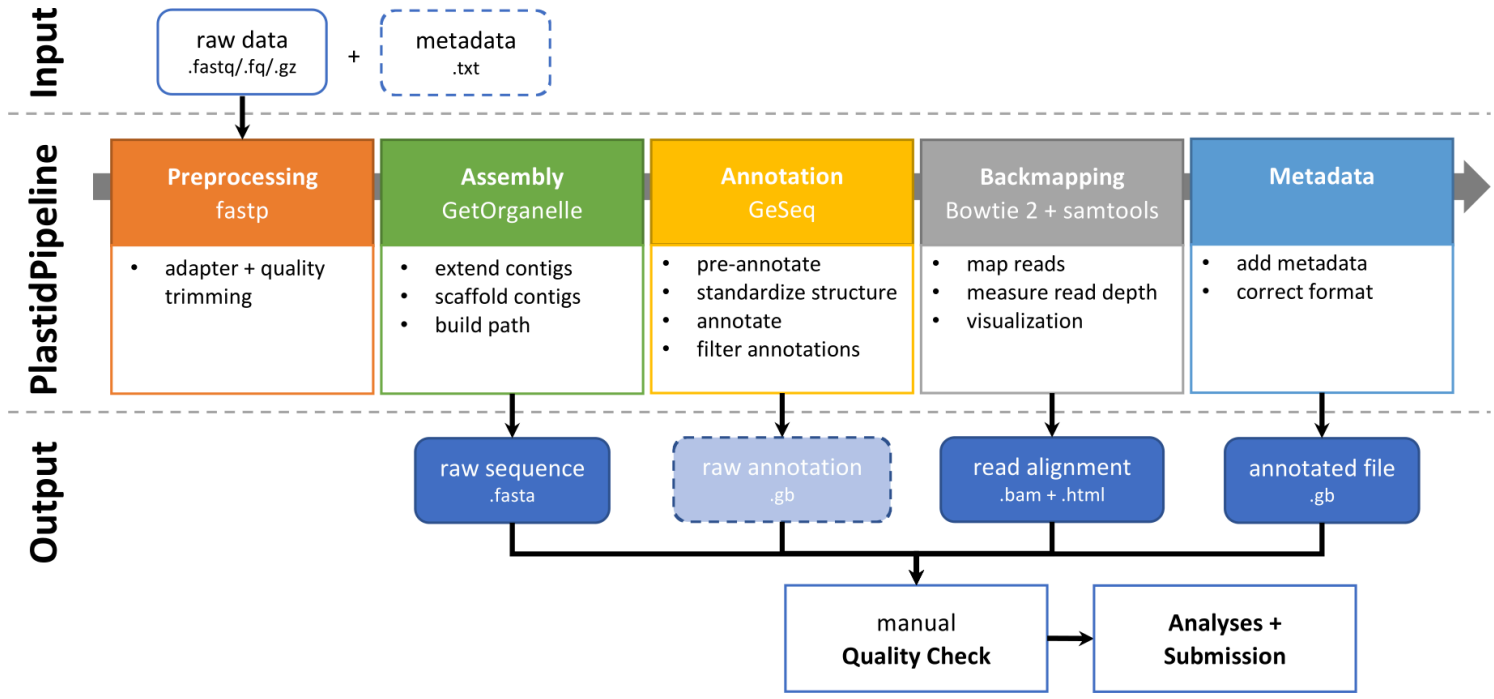

Figure 5

Schematic overview of the PlastidPipeline. Input, main output, subprocesses and main third-party tools used are specified for each step. More details are given in the pipeline's online documentation (https://github.com/sannalda/plastidpipeline).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Sup1Figures.pdf
- Sup2Tables.xlsx
- Sup3Alignments.zip