



DATA NOTE

The genome sequence of the Alpine Milkvetch, *Astragalus*

alpinus L. (Fabales: Fabaceae)

[version 1; peer review: 3 approved]

Markus Ruhsam ¹, Peter M Hollingsworth ¹,
Royal Botanic Garden Edinburgh Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective, Plant Genome Sizing collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹Royal Botanic Garden Edinburgh, Edinburgh, Scotland, UK

V1 First published: 25 Jul 2025, 10:366
<https://doi.org/10.12688/wellcomeopenres.24576.1>
Latest published: 25 Jul 2025, 10:366
<https://doi.org/10.12688/wellcomeopenres.24576.1>

Abstract

We present a genome assembly of *Astragalus alpinus* (Alpine Milkvetch; Streptophyta; Magnoliopsida; Fabales; Fabaceae). The assembly consists of two haplotypes with total lengths of 766.58 megabases and 749.43 megabases. Most of haplotype 1 (98.55%) and of haplotype 2 (99.06%) are each scaffolded into 8 chromosomal pseudomolecules. The mitochondrial sequence has a length of 313.83 kilobases and the plastid genome assembly has a length of 123.2 kilobases.

Keywords

Astragalus alpinus, Alpine Milkvetch, genome sequence, chromosomal, Fabales



This article is included in the [Tree of Life gateway](#).

Open Peer Review

Approval Status

	1	2	3
version 1			
25 Jul 2025	view	view	view

1. **Qionghou Li**, Nanjing Agricultural University, Nanjing, China
2. **Oliver Rupp** , Justus-Liebig-University Giessen, giessen, Germany
3. **Yedomon Ange Bovys Zoclanclounon** , Rothamsted Research, Harpenden, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: **Ruhsam M:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Hollingsworth PM:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, <https://doi.org/10.35802/218328>].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Ruhsam M *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ruhsam M, Hollingsworth PM, Royal Botanic Garden Edinburgh Genome Acquisition Lab *et al.* **The genome sequence of the Alpine Milkvetch, *Astragalus alpinus* L. (Fabales: Fabaceae) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, 10:366 <https://doi.org/10.12688/wellcomeopenres.24576.1>

First published: 25 Jul 2025, 10:366 <https://doi.org/10.12688/wellcomeopenres.24576.1>

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetales; rosids; fabids; Fabales; Fabaceae; Papilionoideae; 50 kb inversion clade; NPAAA clade; Hologalegina; IRL clade; Galegeae; *Astragalus*; *Astragalus alpinus* L. (NCBI:txid20403)

Background

Astragalus alpinus or the Alpine milkvetch is a low, mat-forming perennial, 5 to 30 cm tall, with spreading rhizomes. It generally occurs in Arctic-alpine habitats on calcareous soils in short grassy or rocky places. Milkvetch is a widespread species with a circumpolar distribution, however, in the UK it is rare and restricted to four extant locations in the Scottish Highlands, ranging from altitudes of 650–700 m (Stace *et al.*, 2019; Stroh *et al.*, 2023). Grazing by sheep and deer have been recorded as a threat to the species in the UK, compounded by its presence on unstable rocks and soil substrates.

Astragalus alpinus is highly outcrossing and predominantly visited by bumblebees (Bergman *et al.*, 1996; Stenström & Bergman, 1998). Seed set in bagged flowers was shown to be less than 1%, suggesting an effective self-incompatibility system (Kudo *et al.*, 1999). Seed predation of developing seeds, mainly dipteran and lepidopteran larvae, was estimated to be about 10% (Kudo *et al.*, 1999).

Kudo and Molau (1999) reported that an alpine population had a larger mean flower size and a longer lasting period of anthesis compared to a subalpine population. Although there was no significant difference in ovule fertilisation rate and seed predation rate per inflorescence between the populations, the abortion rate of fertilised ovules was higher in the alpine population, which is likely to be because of resource limitation (Kudo *et al.*, 1999).

Astragalus alpinus has a basic chromosome number of $x = 8$ (Senn, 1938) and is reported to be either a diploid – $2n = 16$, based on North American (Spellenberg, 1976) and Slovakian (Májovský *et al.*, 1996) material – or a tetraploid species – $2n = 32$, based on Canadian material (Löve, 1982).

The genome of the Alpine Milk-vetch, *Astragalus alpinus*, was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomally complete genome sequence for *Astragalus alpinus*, based on a specimen from Ben Vrackie, Scotland, United Kingdom.

Methods

Sample acquisition, flow cytometry and DNA barcoding

A specimen of *Astragalus alpinus* (specimen ID EDTOL05685, ToLID drAstAlpi1; Figure 1) was used for genome sequencing. It was collected from Ben Vrackie, Scotland, UK (latitude 56.7468, longitude –3.7176) on 2023-06-27. The specimen was

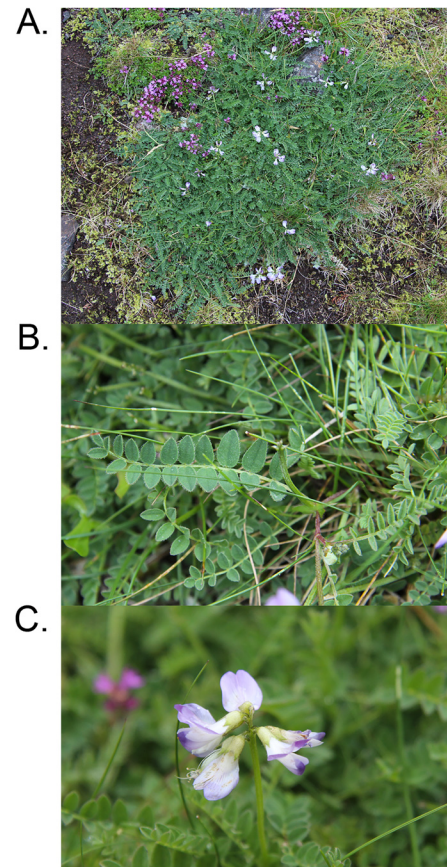


Figure 1. Photograph of the *Astragalus alpinus* (drAstAlpi1) specimen from which samples were taken for genome sequencing.

collected and identified by Markus Ruhsam (Royal Botanic Garden Edinburgh). The same specimen was used for RNA sequencing.

Metadata collection followed the recommended standards of the Darwin Tree of Life project (Lawiczak *et al.*, 2022). The herbarium voucher associated with the sequenced plant is E01358555 and is deposited in the herbarium of RBG Edinburgh (E) <https://data.rbge.org.uk/herb/E01358555>.

The genome size was estimated by flow cytometry following the ‘one-step’ method outlined in Pellicer *et al.* (2021) and using propidium iodide as the fluorochrome. Galbraith was used for isolation of nuclei (Loureiro *et al.*, 2007), and the internal calibration standard was *Petroselinum crispum* ‘Champion Moss Curled’ with an assumed 1C-value of 2 200 Mb (Obermayer *et al.*, 2002).

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). Part of the plant specimen was preserved in silica gel desiccant (Chase & Hills, 1991). DNA extracted

from the dried plant was amplified by PCR for standard barcode markers, with the amplicons sequenced and compared to public sequence databases including GenBank and the Barcode of Life Database (BOLD) (Ratnasingham & Hebert, 2007). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding are available on protocols.io.

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The drAstAlpi1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the leaf was homogenised by [cryogenic bead beating](#). HMW DNA was extracted using the [Automated Plant MagAttract v4](#) protocol. Sheared DNA was purified by [manual SPRI](#) (solid-phase reversible immobilisation), using (Pacific Biosciences) AMPure PB beads to eliminate shorter fragments and concentrate the DNA. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 15.3 ng/μL and a yield of 841.50 ng, with a fragment size of 16.2 kb. The 260/280 spectrophotometric ratio was 2.39, and the 260/230 ratio was 10.65.

RNA was extracted from leaf tissue of drAstAlpi1 in the Tree of Life Laboratory at the WSI using the [RNA Extraction: Automated MagMax™ mirVana protocol](#). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences) according to the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of

the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

Hi-C data were generated from the leaf tissue of drAstAlpi1 using the Arima-HiC v2 kit (Arima Genomics). Tissue was finely ground using the Covaris cryoPREP Dry Pulverizer (Covaris), and then subjected to nuclei isolation. Nuclei were isolated using a modified protocol based on the Qiagen QProteome Cell Compartment Kit (Qiagen), in which only the Lysis and CE2 buffers were used, with QIAshredder spin columns. After isolation, nuclei were fixed using formaldehyde to a final concentration of 2% to crosslink the DNA. The crosslinked DNA was then digested and biotinylated according to the manufacturer's instructions. A clean-up step was performed with SPRIselect beads before library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and the Qubit HS Assay Kit, following the manufacturer's instructions.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRIselect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRIselect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D5000 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were

performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X to generate 150-bp paired-end reads.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of k -mer counts ($k = 31$) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the k -mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020). The organelle genomes were assembled using OATK (Zhou *et al.*, 2024).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 132 breaks and 353 joins. Note that the order and orientation of the contigs on Chromosome 8 are of uncertain order and orientation. For haplotype 1, this is in the region 35.6–56.4 Mbp and for haplotype 2 it is from 33.5–53.4 Mbp. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextViewSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate k -mer completeness and assembly quality for both haplotypes using the k -mer databases ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Daneczek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified by querying the GoT database (Challis *et al.*, 2023). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastx

(Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with package management via Conda and Bioconda (Grüning *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

The genome of a specimen of *Astragalus alpinus* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 0.00 Gb (gigabases) from 0.00 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 749.51 Mb, with a heterozygosity of 0.75% and repeat content of 60.99% (Figure 2). Using flow cytometry, the genome size (1C-value) of the sample was estimated to be 0.96 pg, equivalent to 940.00 Mb. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 34× coverage. Hi-C sequencing produced 95.76 Gb from 634.15 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 766.58 Mb in 207 scaffolds, with 743 gaps, and a scaffold N50 of 95.36 Mb (Table 2).

Most of the assembly sequence (98.55%) was assigned to 8 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3).

The mitochondrial and plastid genomes were also assembled. These sequences are included as contigs in the multifasta file of the genome submission and as standalone records.

Assembly quality metrics

For haplotype 1, the estimated QV is 59.3, and for haplotype 2, 59.5. When the two haplotypes are combined, the assembly achieves an estimated QV of 59.4. The k -mer completeness is 81.22% for haplotype 1, 80.47% for haplotype 2, and 97.18% for the combined haplotypes (Figure 4).

BUSCO analysis using the fabales_odb10 reference set ($n = 5366$) identified 96.8% of the expected gene set (single = 92.3%, duplicated = 4.5%) for haplotype 1. The snail

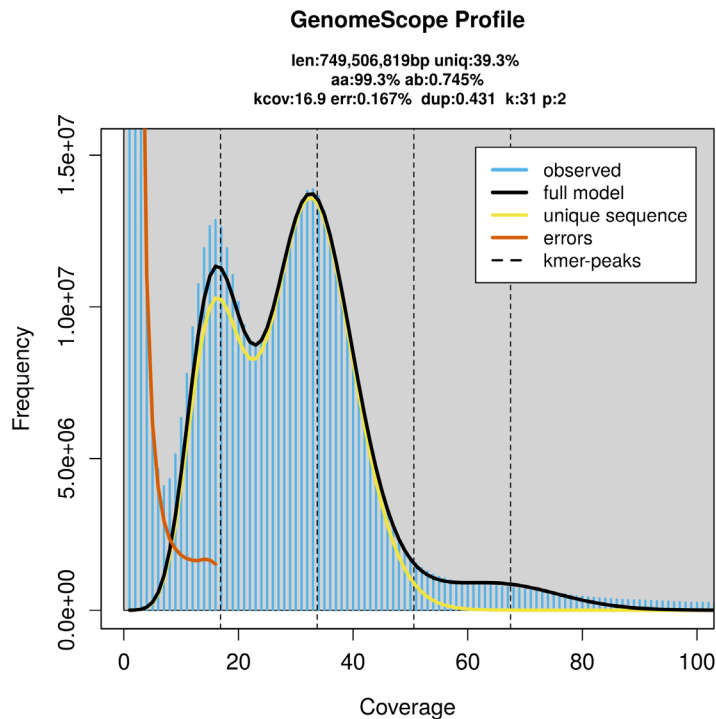


Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB73944.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	drAstAlpi1	drAstAlpi1	drAstAlpi1
Specimen ID	EDTOL05685	EDTOL05685	EDTOL05685
BioSample (source individual)	SAMEA114389508	SAMEA114389508	SAMEA114389508
BioSample (tissue)	SAMEA114389566	SAMEA114389566	SAMEA114389566
Tissue	leaf	leaf	leaf
Instrument	Sequel IIe	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR12760829	ERR12765202	ERR13962511
Read count total	0.00 million	634.15 million	76.20 million
Base count total	0.00 Gb	95.76 Gb	11.51 Gb

plot in [Figure 5](#) summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in [Figure 6](#) shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

[Table 4](#) lists the assembly metric benchmarks adapted from [Rhie et al. \(2021\)](#) the Earth BioGenome Project Report on Assembly

Standards [September 2024](#). The EBP metric calculated for the haplotype 1 is **6.C.Q59**, meeting the recommended reference standard.

Wellcome Sanger Institute – Legal and Governance
The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The

Table 2. Genome assembly statistics.

Assembly name	drAstAlpi1.hap1.1	drAstAlpi1.hap2.1
Assembly accession	GCA_964188245.1	GCA_964188115.1
Assembly level	chromosome	chromosome
Span (Mb)	766.58	749.43
Number of chromosomes	8	8
Number of contigs	950	898
Contig N50	1.78 Mb	1.61 Mb
Number of scaffolds	207	117
Scaffold N50	95.36 Mb	93.83 Mb
Longest scaffold length (Mb)	103.0	102.29
Organelles	Mitochondrial genome: 313.83 kb; Plastid genome: 123.2 kb	N/A

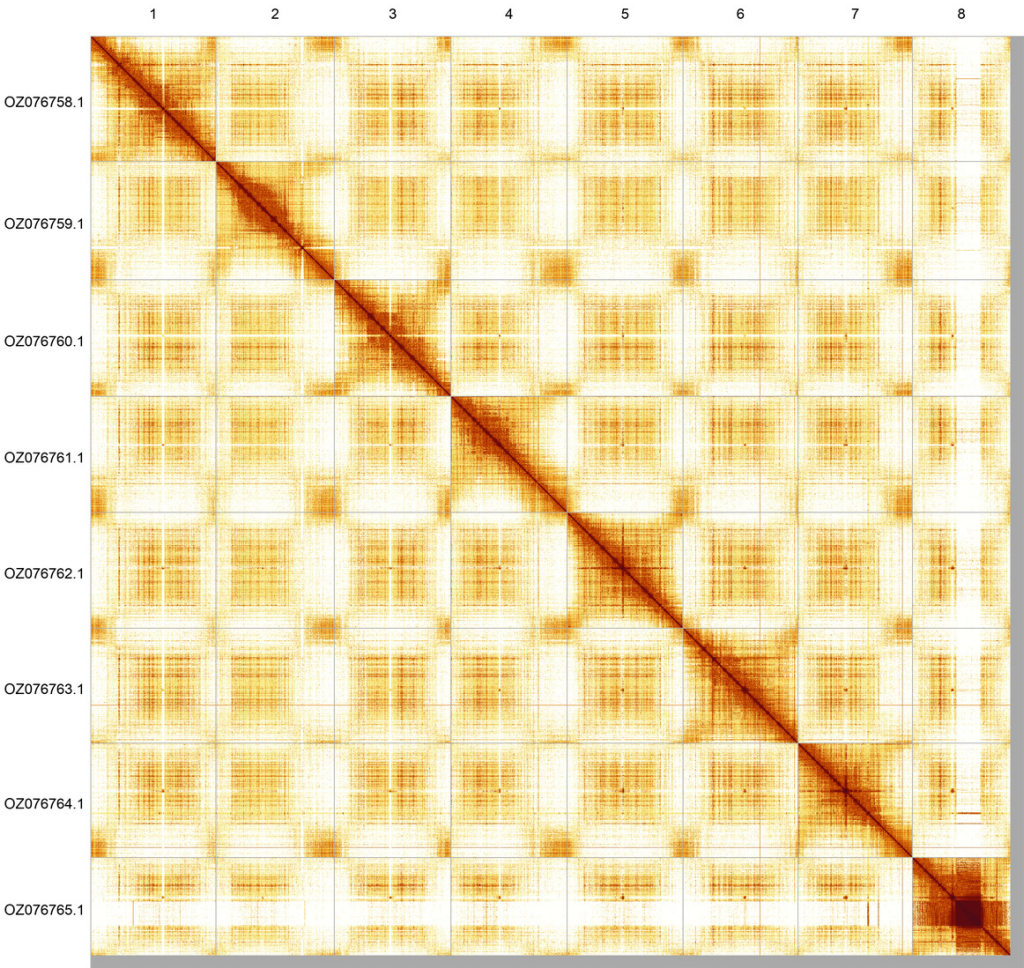


Figure 3. Hi-C contact map of the *Astragalus alpinus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in both haplotypes of the genome assembly of *Astragalus alpinus*, drAstAlpi1.

Haplotype 1				Haplotype 2			
INSDC accession	Name	Length (Mb)	GC%	INSDC accession	Name	Length (Mb)	GC%
OZ076758.1	1	103	37.50	OZ076816.1	1	102.29	37.50
OZ076759.1	2	97.23	37.50	OZ076818.1	2	96.27	37.50
OZ076760.1	3	95.77	37.50	OZ076817.1	3	97.32	38
OZ076761.1	4	95.36	37.50	OZ076822.1	4	91.48	37.50
OZ076762.1	5	95.10	38	OZ076820.1	5	92.67	38
OZ076763.1	6	94.53	37.50	OZ076819.1	6	93.83	37.50
OZ076764.1	7	93.88	37.50	OZ076821.1	7	92.26	37.50
OZ076765.1	8	80.63	41.50	OZ076823.1	8	76.24	41

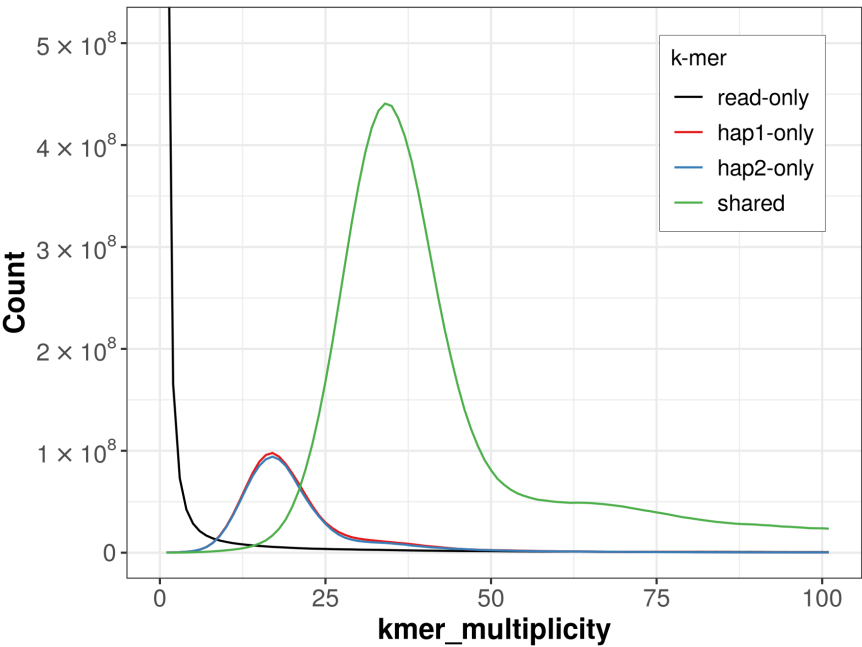


Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve (the homozygous peak) corresponds to *k*-mers shared by both haplotypes and the red and blue curves (the heterozygous peaks) show *k*-mers found only in one of the haplotypes.

submission of materials by a Darwin Tree of Life Partner is subject to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, which can be found in full on the [Darwin Tree of Life website](#). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set

out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The

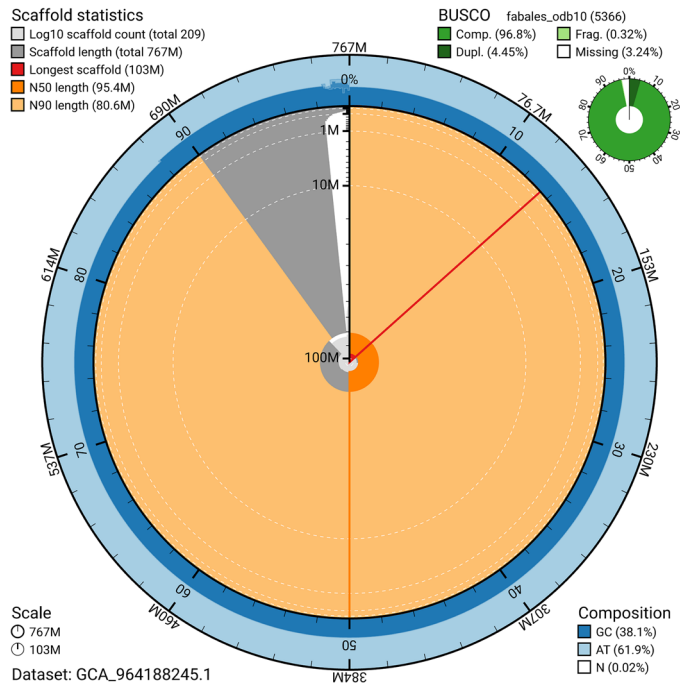


Figure 5. Assembly metrics for drAstAlpi1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the [BlobToolKit viewer](#).

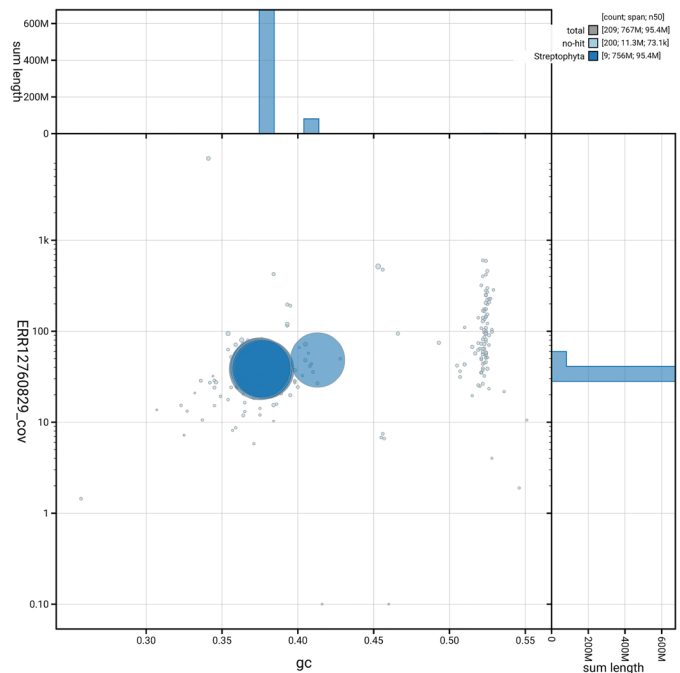


Figure 6. BlobToolKit GC-coverage plot for drAstAlpi1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the [BlobToolKit viewer](#).

Table 4. Earth Biogenome Project summary metrics for the *Astragalus alpinus* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q59	6.C.Q40
Contig N50 length	1.78 Mb	≥ 1 Mb
Scaffold N50 length	95.36 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 59.3; haplotype 2: 59.5; combined: 59.4	≥ 40
k-mer completeness	Haplotype 1: 81.22%; Haplotype 2: 80.47%; combined: 97.18%	≥ 95%
BUSCO	C:96.8% [S:92.3%; D:4.5%]; F:0.3%; M:2.9%; n:5 366	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.55%	≥ 90%

purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: *Astragalus alpinus*. Accession number [PRJEB73944](#). The genome sequence is released openly for reuse. The *Astragalus alpinus* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the [Ensembl](#)

pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in [Table 1](#) and [Table 2](#).

Pipelines used for genome assembly at the WSI Tree of Life are available at <https://pipelines.tol.sanger.ac.uk/pipelines>. [Table 5](#) lists software versions used in this study.

Author information

- Members of the [Royal Botanic Garden Edinburgh Genome Acquisition Lab](#)
- Members of the [Plant Genome Sizing collective](#)
- Members of the [Darwin Tree of Life Barcoding collective](#)
- Members of the [Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](#)
- Members of [Wellcome Sanger Institute Scientific Operations – Sequencing Operations](#)
- Members of the [Wellcome Sanger Institute Tree of Life Core Informatics team](#)
- Members of the [Tree of Life Core Informatics collective](#)
- Members of the [Darwin Tree of Life Consortium](#)

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/

Software	Version	Source
BlobToolKit	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MercuryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	https://github.com/lh3/minimap2
Oatk	0.9	https://github.com/c-zhou/oatk
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextView	N/A	https://github.com/sanger-tol/PretextView
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

- Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bergman P, Molau U, Holmgren B: **Micrometeorological impacts on insect**

- activity and plant reproductive success in an alpine environment, Swedish Lapland.** *Arctic Alpine Res.* 1996; **28**(2): 196–202.
[Publisher Full Text](#)
- Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoAT): a**

versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved]. *Wellcome Open Res.* 2023; **8**: 24.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Chase MW, Hills HH: **Silica gel: an ideal material for field preservation of leaf samples for DNA studies.** *Taxon.* 1991; **40**(2): 215–220.

[Publisher Full Text](#)

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.

[PubMed Abstract](#) | [Publisher Full Text](#)

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfstats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.

[Publisher Full Text](#)

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): g1aa153.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Kudo G, Molau U: **Variations in reproductive traits at inflorescence and flower levels of an arctic legume, *Astragalus alpinus* L.: comparisons between a subalpine and an alpine population.** *Plant Spec Biol.* 1999; **14**(3): 181–191.

[Publisher Full Text](#)

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Lawnczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.

[Publisher Full Text](#)

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Loureiro J, Rodriguez E, Dolezel J, *et al.*: **Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species.** *Ann Bot.* 2007; **100**(4): 875–888.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Löve Á: **IOPB chromosome number reports LXXV.** *Taxon.* 1982; **31**(2): 342–368.

[Publisher Full Text](#)

Májovský J, Uhríková A, Důbravcová Z: **Karyotaxonomic analysis of supramontane populations of species of the genus *Astragalus* in Slovakia.** *Oecologia Montana.* 1996; **5**(2): 87–92.

[Reference Source](#)

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.

[Reference Source](#)

Obermayer R, Leitch IJ, Hanson L, *et al.*: **Nuclear DNA C-values in 30 species double the familial representation in pteridophytes.** *Ann Bot.* 2002; **90**(2): 209–217.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Pellicer J, Powell RF, Leitch IJ: **The application of flow cytometry for estimating genome size, ploidy level endopolyploidy, and reproductive modes in plants.** In: P. Besse (ed), *Methods Mol Biol (Clifton, N.J.)*. New York, NY, 2021; **2222**: 325–361.

[PubMed Abstract](#) | [Publisher Full Text](#)

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ratnasingham S, Hebert PDN: **bold: the Barcode of Life Data System (<http://www.barcodinglife.org>).** *Mol Ecol Notes.* 2007; **7**(3): 355–364.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rhie A, Walenz BP, Koren S, *et al.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Senn HA: **Chromosome number relationships in the Leguminosae.** *Bibliographia Genetica.* 1938; **12**: 175–336.

Spellenberg R: **Chromosome numbers and their cytotoxic significance for North American *Astragalus* (Fabaceae).** *Taxon.* 1976; **25**(4): 463–476.

[Publisher Full Text](#)

Stace CA, Thompson H, Stace M: **New flora of the British Isles.** C&M Floristics, 2019.

Stenström M, Bergman P: **Bumblebees at an alpine site in Northern Sweden: temporal development, population size, and plant utilization.** *Ecography.* 1998; **21**(3): 306–316.

[Publisher Full Text](#)

Stroh PA, Walker KJ, Humphrey TA, *et al.*: **Plant Atlas 2020: mapping changes in the distribution of the British and Irish Flora.** Durham UK; Oxford UK: Botanical Society of Britain and Ireland; Princeton University Press, 2023.

[Publisher Full Text](#)

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019; 314–324.

[Publisher Full Text](#)

Zhou C, Brown M, Blaxter M, *et al.*: **Oatk: a *de novo* assembly tool for complex plant organelle genomes.** *bioRxiv.* 2024.

[Publisher Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:   

Version 1

Reviewer Report 10 September 2025

<https://doi.org/10.21956/wellcomeopenres.27085.r128949>

© 2025 Zoclanclounon Y. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Yedomon Ange Bovys Zoclanclounon 

Rothamsted Research, Harpenden, England, UK

The manuscript presents the sequencing and assembly of the nuclear, mitochondrial, and plastid genomes of *Astragalus alpinus*. This foundational genomic resource constitutes an important step toward understanding the genetic basis of the species' unique adaptations to Arctic-alpine environments and addressing its ongoing conservation challenges.

Overall, the manuscript is clearly written, well structured, and the results are appropriately supported by the analyses presented. The reported ~100× Hi-C coverage provides a robust foundation for a reliable chromosome-scale assembly of the species. The work is timely and of relevance to both the plant genomics and conservation biology communities.

My minor comments are as follows:

1. **Sequence data:** The HiFi read size and number of reads are reported as 0.00 Gb from 0.00 million reads, which appears to be an omission. A check of the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/view/ERX12134038>) indicates that only the BAM file has been deposited, while the corresponding FASTQ file is not yet available. Clarification and correction are needed.
2. **Transcriptome data:** Although a transcriptome dataset was generated, its use is not discussed in the manuscript (e.g., mapping rate onto the haplotype 1 assembly). In addition, the coverage is relatively low (~12.24×) and limited to a single tissue (leaf). While the scope of this report is not comparative genomics, increasing RNA-seq coverage across multiple tissues could provide valuable insights into the genetic basis of *A. alpinus*' ecological adaptations, and this possibility could be noted as a future perspective.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genome assembly, genome annotation, genome evolution, comparative genomics, multi-omics integration

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 30 August 2025

<https://doi.org/10.21956/wellcomeopenres.27085.r128950>

© 2025 Rupp O. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Oliver Rupp 

Justus-Liebig-University Giessen, giessen, Germany

This article presents the sequencing and assembly of the *Astragalus alpinus* genome. The data creation and data analyses follow the state-of-the art methods for plant genome assemblies, including PacBio Hifi and Hi-C sequencing, hifiasm assembly followed by automatic and manual assembly curation. The assembly quality was assessed by k-mer and orthology based methods like Merqury.FK or BUSCO. The methods are clearly presented and the quality of the data is shown. The pipeline and data are available at public databases and a gene annotation will be performed. Overall no major issues are to report.

The genome sequence report paragraph states "... generating 0.00 Gb (gigabases) from 0.00 million reads ...", this should be checked.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, genome assembly, RNAseq analyses

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 23 August 2025

<https://doi.org/10.21956/wellcomeopenres.27085.r128947>

© 2025 Li Q. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Qionghou Li

Nanjing Agricultural University, Nanjing, Jiangsu, China

Ruhsam et al. present the first chromosome-scale genome assembly of *Astragalus alpinus*. Using PacBio HiFi reads with Hi-C scaffolding, author report 96.8% BUSCO completeness and annotate the assembly with RNA-seq data. Overall, the manuscript reports two high-quality, haplotype-resolved chromosome-level assemblies. I have only some minor concerns:

1. Since the authors evaluated the assembly with Merqury. I recommend adding QV, k-mer completeness, and genome BUSCO completeness to Table 2.
2. The genome annotation was also performed, please also include in Table 2 the number of protein-coding genes, the number of transcripts, and gene-level BUSCO completeness.
3. Although the study reported PacBio HiFi sequencing, both Table 1 and the sentence—"The genome of a specimen of *Astragalus alpinus* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 0.00 Gb (gigabases) from 0.00 million reads"—report 0 data yield. Inspecting the ENA (EMBL-EBI) submission (ERR12760829: <https://www.ebi.ac.uk/ena/browser/view/ERR12760829>) shows the HiFi data are provided as BAM files, for which ENA does not automatically compute read counts. The authors should therefore report the total HiFi read count and total bases manually. If that is inconvenient, I suggest converting BAM to FASTQ with PacBio's bam2fastx (<https://github.com/PacificBiosciences/pbtk>) and then calculating the statistics with seqkit, and adding these values to draft.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genome assembly, genome evolution, comparative genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
