Check for updates

DATA NOTE

REVISED The genome sequence of spotted medick, *Medicago arabica* (L.) Huds. (Fabaceae)

[version 2; peer review: 2 approved, 2 approved with reservations]

Maarten J. M. Christenhusz[1,2], Michael F. Fay[1,3], Ilia J. Leitch[1],
Royal Botanic Gardens Kew Genome Acquisition Lab,
Plant Genome Sizing collective, Darwin Tree of Life Barcoding 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

[1]Royal Botanic Gardens Kew, Richmond, England, UK
[2]Curtin University, Perth, Western Australia, Australia
[3]The University of Western Australia, Perth, Western Australia, Australia

v2 **First published:** 01 Mar 2024, **9**:116
https://doi.org/10.12688/wellcomeopenres.20996.1
Latest published: 04 Nov 2024, **9**:116
https://doi.org/10.12688/wellcomeopenres.20996.2

Abstract

We present a genome assembly from an individual *Medicago arabica* (the spotted medick; Tracheophyta; Magnoliopsida; Fabales; Fabaceae). The genome sequence is 515.5 megabases in span. Most of the assembly is scaffolded into 8 chromosomal pseudomolecules. The mitochondrial and plastid genome assemblies have lengths of 324.47 kilobases and 125.07 kilobases in length, respectively. Gene annotation of this assembly on Ensembl identified 24,619 protein-coding genes.

Keywords

Medicago arabica, spotted medick, genome sequence, chromosomal, Fabales

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ? ✓

	1	2	3	4
version 2 (revision) 04 Nov 2024		✓ view	? view	✓ view
version 1 01 Mar 2024	? view	? view		

1. **Marco Pessoa-Filho**, Brazilian Agricultural Research Corporation, Brasília, Brazil
2. **Xiaolong Lyu**, Zhejiang University, Hangzhou, China
3. **Yu Feng**, Chengdu Institute of Biology, Chinese Academy of Sciences, Sichuan, China
4. **Thomas Brazier**, University of Rennes, Rennes, France

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: Christenhusz MJM: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Fay MF**: Writing – Original Draft Preparation; **Leitch IJ**: 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 (206194) and the Darwin Tree of Life Discretionary Award (218328).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Christenhusz MJM *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: Christenhusz MJM, Fay MF, Leitch IJ *et al.* **The genome sequence of spotted medick, *Medicago arabi*ca (L.) Huds. (Fabaceae) [version 2; peer review: 2 approved, 2 approved with reservations]** Wellcome Open Research 2024, **9**:116 https://doi.org/10.12688/wellcomeopenres.20996.2

First published: 01 Mar 2024, **9**:116 https://doi.org/10.12688/wellcomeopenres.20996.1

REVISED **Amendments from Version 1**

In version 2 of this data note we have added information about the annotation of the genome on Ensembl. We have also specified that Hifiasm was run in primary mode, thus producing a primary and an alternate haplotype.

Any further responses from the reviewers can be found at the end of the article

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fabales; Fabaceae; Papilionoideae; 50 kb inversion clade; NPAAA clade; Hologalegina; IRL clade; Trifolieae; *Medicago*; *Medicago arabica* (L.) Huds. (NCBI: txid70936).

Background

Spotted medick, *Medicago arabica* (L.) Huds., is a winter-growing annual of the pea family, Fabaceae. It has creeping stems bearing trifoliate leaves, with each leaflet marked with dark purple spots. Yellow flowers appear in spring and early summer, followed by coiled, spiny seed pods that stick in the fur of animals (and clothes of humans), aiding their dispersal.

The species is native to the Mediterranean region, east to the Caucasus and Crimea, and is found along the Atlantic in western Europe, and north to Britain. It is frequently naturalised as a wool alien outside its natural range (e.g. in USA, Costa Rica, southern South America, Japan, China, Australia, New Zealand, Alps, Baltic States, Sweden, Ireland). In Britain, it has a predominantly southern and south-eastern distribution (e.g. Botanical Society of Britain and Ireland, 2024), being most common in southern England to the Midlands; it is much rarer in Wales and northern England and Scotland, where it occurs mostly along the coast. In England it is increasingly found in inland, lowland areas, but for reasons that are not known (OABIF, 2022; POWO, 2024). It grows in grassy places usually on light soils, and it can be found as a weed in lawns and in fields, roadside verges and rough ground.

Comparative genomics of *M. arabica*, *M. sativa* and other *Medicago* species could provide useful information on traits with agronomic potential for plant breeders. Like many other *Medicago* species, spotted medick is rich in a variety of saponins, with potential applications as antimicrobial compounds in agriculture and medicine (e.g. Avato *et al.*, 2006; Bialy *et al.*, 2004; Jarecka *et al.*, 2008; Tava *et al.*, 2009).

Medicago arabica is a diploid, with 16 chromosomes (e.g. Fyad-Lameche *et al.*, 2016). It is a relative of the important forage crop lucerne (alfalfa; *Medicago sativa* L.), which is a tetraploid ($2n$ = 32). Like many other *Medicago* species, spotted medick is rich in a variety of saponins with potential for use as antimicrobial compounds in agriculture and medicine (e.g. Avato *et al.*, 2006; Bialy *et al.*, 2004; Jarecka *et al.*, 2008; Tava *et al.*, 2009).

Here we present the first high-quality genome of *Medicago arabica* which will not only help shed light on the biochemical pathways involved in the biosynthesis of saponins, but may also be useful for comparative genomic studies with cultivated *Medicago* species and their wild relatives, providing useful information on traits of agronomic potential for plant breeders. For example, it joins the chromosome level genome assemblies available for three agriculturally important *Medicago* species comprising (i) alfalfa, *M. sativa* (Chen *et al.*, 2020) which is globally one of the highest yielding forage crops; (ii) the bur clover, *M. polymorpha* (Cui *et al.*, 2021), cultivated for its low lignin content which makes it particularly nutritious; and (iii) *M. ruthenica* (Wang *et al.*, 2021), a wild relative of *M. sativa* that is tolerant of environmental stress.

Genome sequence report

The genome was sequenced from a specimen of *Medicago arabica* (Figure 1) collected from Kingston Upon Thames, Surrey, UK (51.42, –0.31). Using flow cytometry, the genome size (1C-value) was estimated to be 0.62 pg, equivalent to 610 Mb. A total of 44-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 86-fold coverage in 10X Genomics read clouds was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 106 missing joins or mis-joins and removed 2 haplotypic duplications, reducing the scaffold number by 33.13%, and increasing the scaffold N50 by 24.91%.

The final assembly has a total length of 515.5 Mb in 107 sequence scaffolds with a scaffold N50 of 64.7 Mb (Table 1).

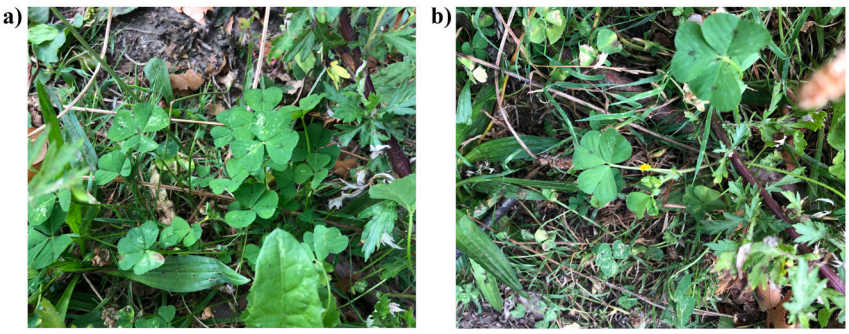

Figure 1. Photographs of the *Medicago arabica* (drMedArab1) specimen used for genome sequencing.

Table 1. **Genome data for *Medicago arabica*, drMedArab1.1.**

Project accession data		
Assembly identifier	drMedArab1.1	
Species	*Medicago arabica*	
Specimen	drMedArab1	
NCBI taxonomy ID	70936	
BioProject	PRJEB47317	
BioSample ID	SAMEA7521936	
Isolate information	drMedArab1	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	56.4	*≥ 40*
k-mer completeness	99.99%	*≥ 95%*
BUSCO**	C:98.8%[S:96.6%,D:2.2%], F:0.2%,M:1.0%,n:5,366	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.91%	*≥ 90%*
Sex chromosomes	None	*localised homologous pairs*
Organelles	Mitochondrial genome: 324.47 kb Plastid genome: 125.07 kb	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR6908000	
10X Genomics Illumina	ERR6688727, ERR6688725, ERR6688726, ERR6688728	
Hi-C Illumina	ERR6688404	
PolyA RNA-Seq Illumina	ERR6688729	
Genome assembly		
Assembly accession	GCA_946800305.1	
Accession of alternate haplotype	GCA_946800295.1	
Span (Mb)	515.5	
Number of contigs	235	
Contig N50 length (Mb)	6.5	
Number of scaffolds	107	
Scaffold N50 length (Mb)	64.7	
Longest scaffold (Mb)	76.24	
Genome annotation of assembly GCA_946800305.1 at Ensembl		
Number of protein-coding genes	24,619	
Number of non-coding genes	8,254	
Number of gene transcripts	40,979	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from Rhie *et al.* (2021).

** BUSCO scores based on the fabales_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/CAMPEK01/dataset/CAMPEK01/busco.

The snail plot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.91%) of the assembly sequence was assigned to 8 chromosomal-level scaffolds. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 2). Parts of the rRNA cluster on chromosome 1 at 24.5Mbp could not be uniquely placed and were submitted as unlocalised sequences of chromosome 1. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial and plastid genomes were also assembled and can be found as contigs within the multifasta file of the genome submission.

Genome annotation report

The Medicago arabica genome assembly (GCA_946800305.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes

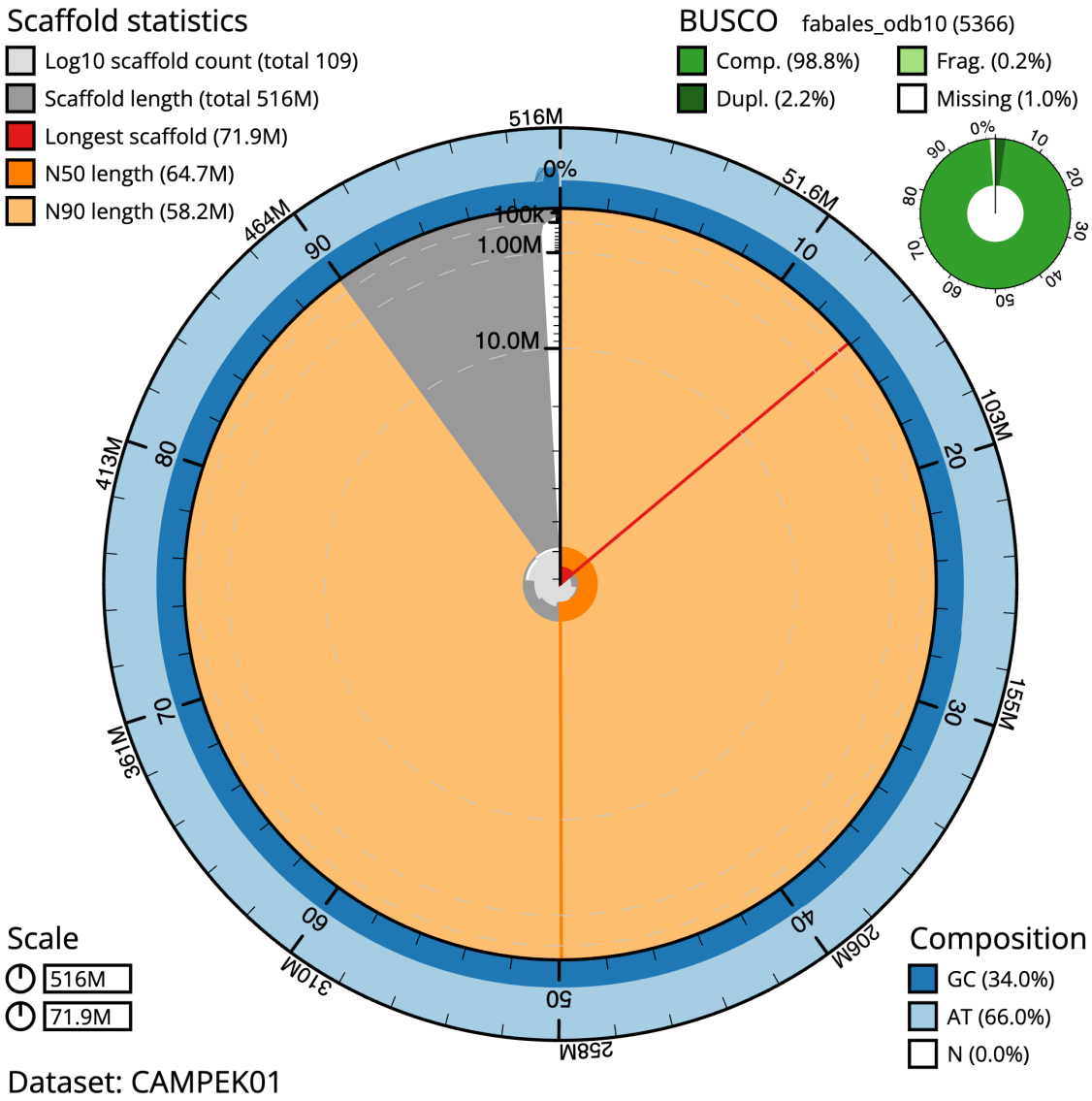

Figure 2. Genome assembly of *Medicago arabica*, drMedArab1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 bins around the circumference with each bin representing 0.1% of the 515,954,536 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (71,875,296 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (64,674,077 and 58,228,340 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the fabales_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CAMPEK01/dataset/CAMPEK01/snail.

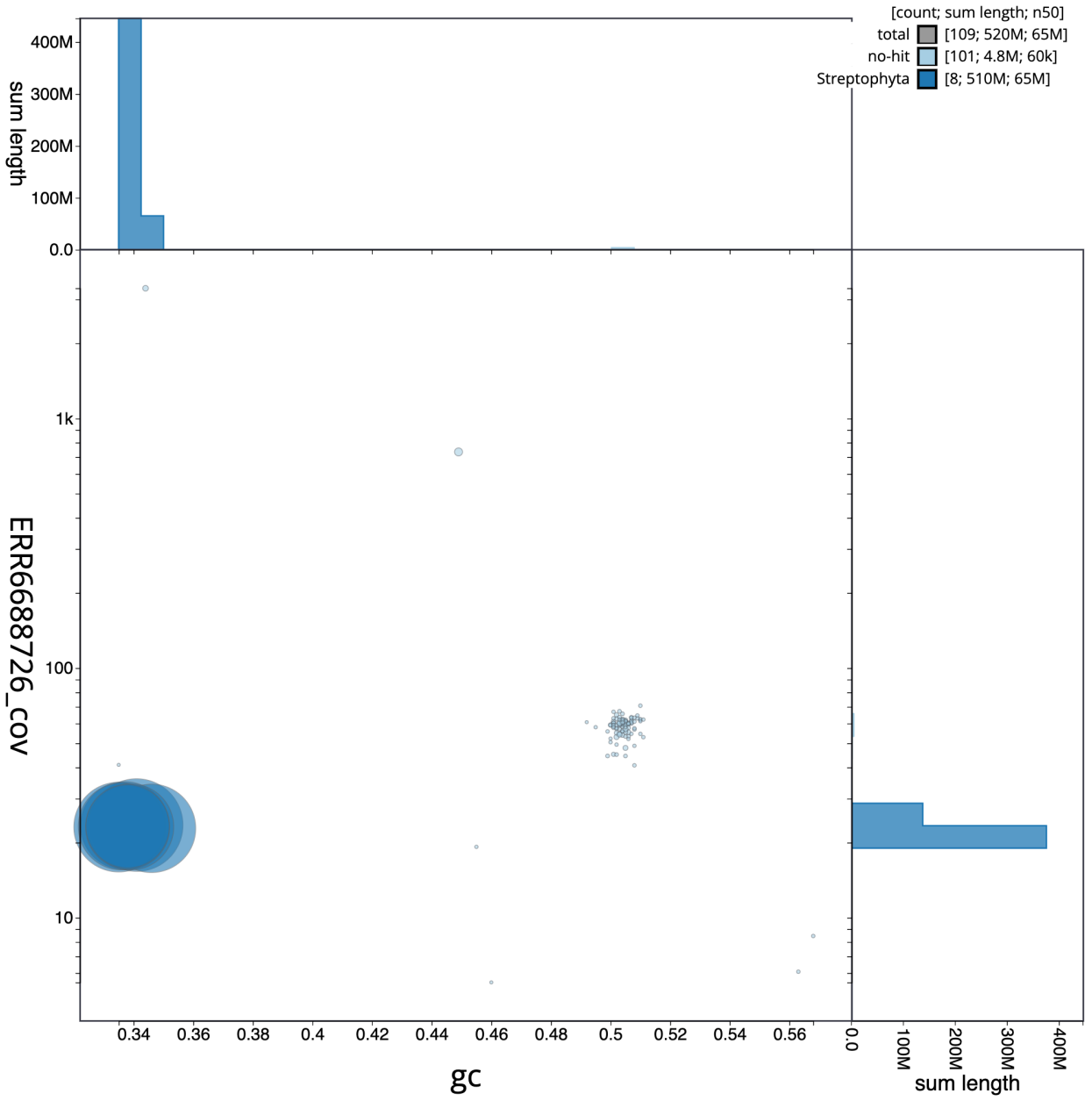

Figure 3. Genome assembly of *Medicago arabica*, drMedArab1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CAMPEK01/dataset/CAMPEK01/blob.

40,979 transcribed mRNAs from 24,619 protein-coding and 8,254 non-coding genes (Table 2; https://rapid.ensembl.org/Medicago_arabica_GCA_946800305.1/Info/Index). The average transcript length is 2,758.62. There are 1.25 coding transcripts per gene and 4.52 exons per transcript.

The estimated Quality Value (QV) of the final assembly is 56.4 with *k*-mer completeness of 99.99%, and the assembly has a BUSCO v5.3.2 completeness of 98.8% (single = 96.6%, duplicated = 2.2%), using the fabales_odb10 reference set (n = 5,366).

Metadata for specimens, barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/70936.

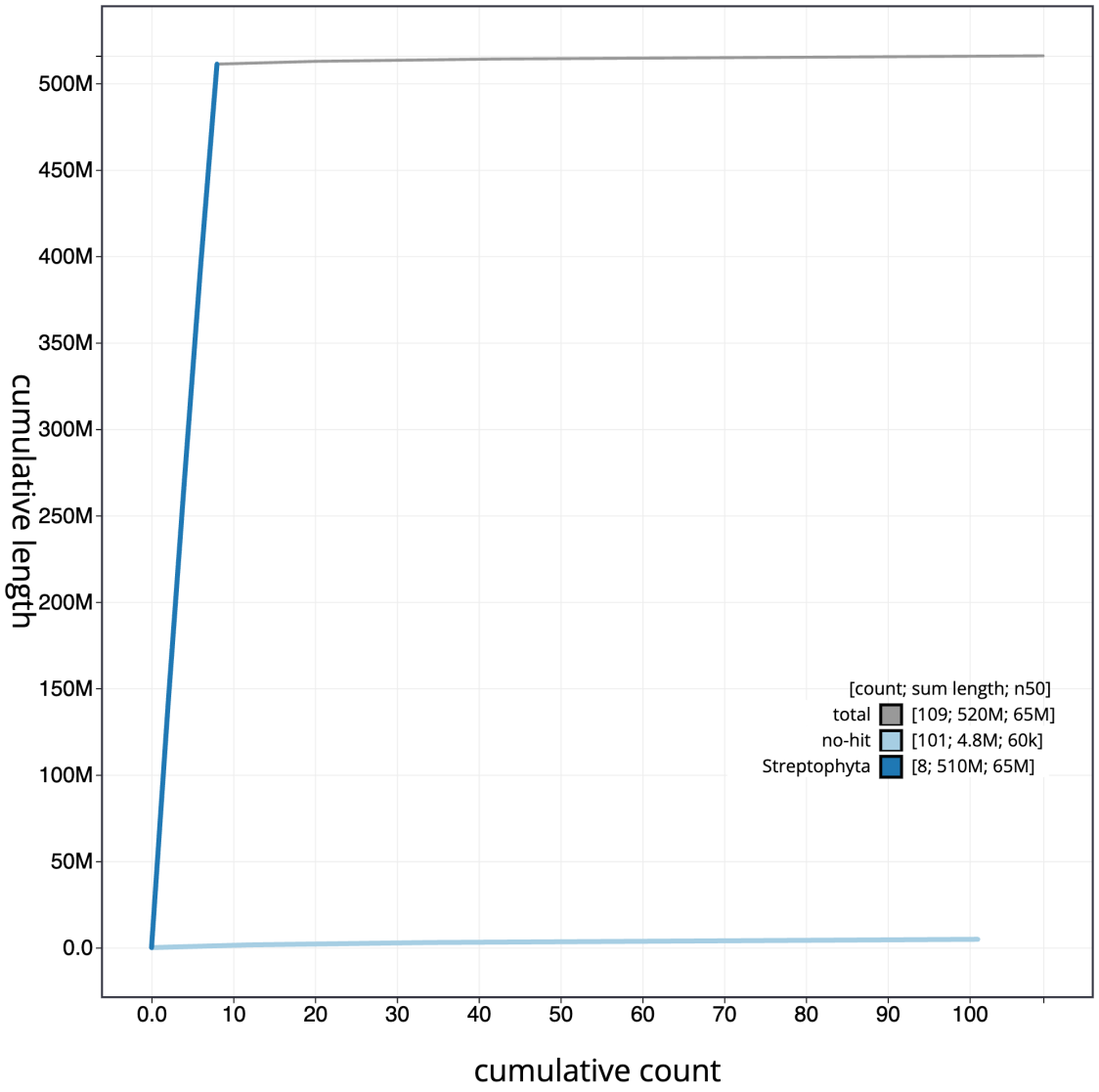

Figure 4. Genome assembly of *Medicago arabica*, drMedArab1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CAMPEK01/dataset/CAMPEK01/cumulative.

Methods

Sample acquisition, genome size estimation and nucleic acid extraction

A specimen of *Medicago arabica* (specimen ID KDTOL10027, ToLID drMedArab1) was collected from Canbury Gardens, Kingston Upon Thames, Surrey, UK (latitude 51.42, longitude –0.31) on 2020-08-10. The specimen was collected and identified by Maarten Christenhusz (Royal Botanic Gardens Kew), and then preserved by freezing at –80 °C.

The genome size was estimated by flow cytometry using the fluorochrome propidium iodide and following the 'one-step' method as outlined in Pellicer *et al.* (2021). For this species, the General Purpose Buffer (GPB) supplemented with 3% PVP and 0.08% (v/v) beta-mercaptoethanol was used for isolation of nuclei (Loureiro *et al.*, 2007), and the internal calibration standard was *Solanum lycopersicum* 'Stupiké polní rané' with an assumed 1C-value of 968 Mb (Dolezel *et al.*, 2007).

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample

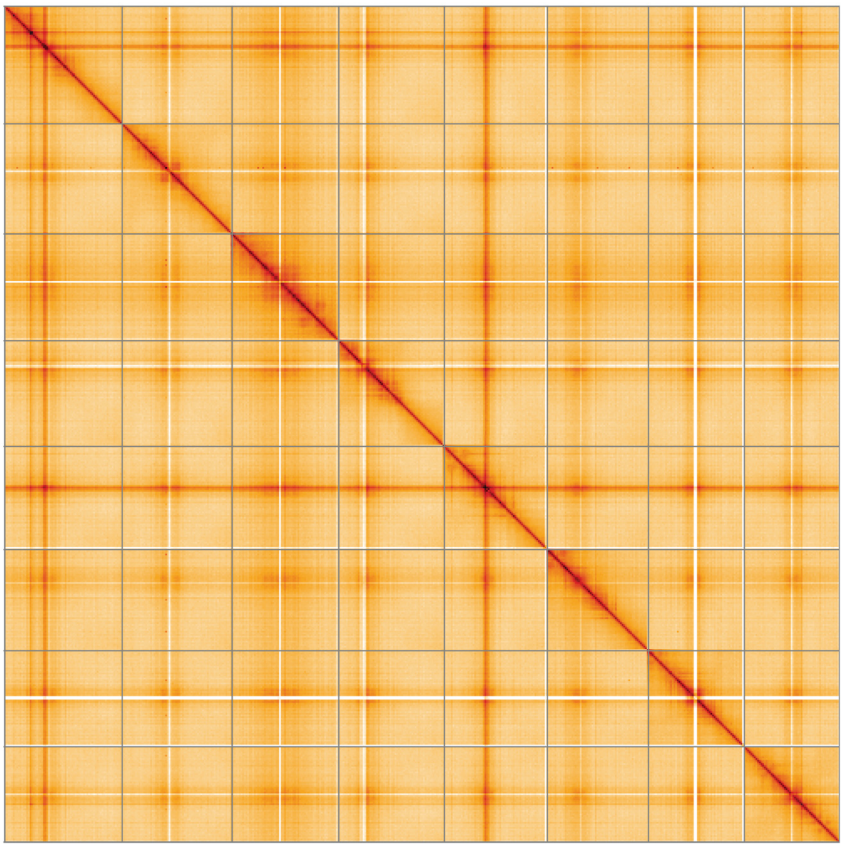

Figure 5. Genome assembly of *Medicago arabica*, drMedArab1.1: Hi-C contact map of the drMedArab1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=FKwiULypRCm3Y-gUQXShkQ.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Medicago arabica*, drMedArab1.

INSDC accession	Chromosome	Length (Mb)	GC%
OX326964.1	1	71.88	34.0
OX326965.1	2	67.34	33.5
OX326966.1	3	65.33	34.5
OX326967.1	4	64.67	33.5
OX326968.1	5	63.06	33.5
OX326969.1	6	61.9	34.0
OX326970.1	7	58.74	34.0
OX326971.1	8	58.23	34.0
OX326972.1	MT	0.32	45.0
OX326973.1	Pltd	0.13	34.5

preparation, the drMedArab1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). For sample homogenisation, leaf tissue was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *et al.*, 2023). HMW DNA was extracted using the Manual Plant MagAttract v2 protocol (Todorovic *et al.*, 2023a). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30 (Todorovic *et al.*, 2023b). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023): in brief, the method employs a 1.8X ratio of AMPure PB beads to sample 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 and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from leaf tissue of drMedArab1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol (do Amaral *et al.*, 2023). 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.

Protocols developed by the WSI Tree of Life core laboratory are publicly available on protocols.io (Denton *et al.*, 2023).

Sequencing

Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi), Illumina HiSeq 4000 (RNA-Seq) and Illumina NovaSeq 6000 (10X) instruments. Hi-C data were also generated from leaf tissue of drMedArab1 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) with the --primary option, and the primary contigs were used for the remainder of the assembly pipeline. Haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). One round of polishing was performed by aligning 10X Genomics read data to the assembly with Long Ranger ALIGN, calling variants with FreeBayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using SALSA2 (Ghurye *et al.*, 2019). The assembly was checked for contamination and corrected using the gEVAL system (Chow *et al.*, 2016) as described previously (Howe *et al.*, 2021). Manual curation was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and PretextView (Harry, 2022). The mitochondrial and chloroplast genomes were assembled using MBG (Rautiainen & Marschall, 2021) from PacBio HiFi reads mapping to related genomes. A representative circular sequence was selected for each from the graph based on read coverage.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury. FK (Rhie *et al.*, 2020). This work was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

Genome annotation

The Ensembl Genebuild annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Medicago arabica* assembly (GCA_946800305.1) in Ensembl Rapid Release at

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	3.5.2	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
FreeBayes	1.3.1-17-gaa2ace8	https://github.com/freebayes/freebayes
gEVAL	N/A	https://geval.org.uk/
Hifiasm	0.15.3	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Long Ranger ALIGN	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
MBG	1.0.13	https://github.com/maickrau/MBG
Merqury	MerquryFK	https://github.com/thegenemyers/MERQURY.FK
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
SALSA	2.2	https://github.com/salsa-rs/salsa
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0

the EBI. Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

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 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 here. 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 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: *Medicago arabica*. Accession number PRJEB47317; https://identifiers.org/ena.embl/PRJEB47317 (Wellcome Sanger Institute, 2022). The genome sequence is released openly for reuse. The *Medicago arabica* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. 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.

Author information

Members of the Royal Botanic Gardens Kew Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.4786680.

Members of the Plant Genome Sizing collective are listed here: https://doi.org/10.5281/zenodo.7994306.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893703.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.10066175.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.10043364.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.10066637.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013541.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

References

Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

Aken BL, Ayling S, Barrell D, *et al.*: **The ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
PubMed Abstract | Publisher Full Text | Free Full Text

Avato P, Bucci R, Tava A, *et al.*: **Antimicrobial activity of saponins from *Medicago* sp.: structure-activity relationship.** *Phytother Res.* 2006; **20**(6): 454–457.
PubMed Abstract | Publisher Full Text

Bialy Z, Jurzysta M, Mella M, *et al.*: **Triterpene saponins from aerial parts of *Medicago arabica* L.** *J Agric Food Chem.* 2004; **52**(5): 1095–9.
PubMed Abstract | Publisher Full Text

Botanical Society of Britain and Ireland: ***Medicago arabica* Distribution map**. *bsbi.org*. 2024; [Accessed 26 January 2024].
Reference Source

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

Chen H, Zeng Y, Yang Y, *et al.*: **Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa.** *Nat Commun.* 2020; **11**(1): 2494.
PubMed Abstract | Publisher Full Text | Free 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

Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL - a web-based browser for evaluating genome assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–2510.
PubMed Abstract | Publisher Full Text | Free Full Text

Cui J, Lu Z, Wang T, *et al.*: **The genome of *Medicago polymorpha* provides insights into its edibility and nutritional value as a vegetable and forage legume.** *Hortic Res.* 2021; **8**(1): 47.
PubMed Abstract | Publisher Full Text | Free Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io*. 2023.
Publisher Full Text

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | **Publisher Full Text**

do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana**. *protocols.io*. 2023.
Publisher Full Text

Dolezel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants using flow cytometry.** *Nat Protoc.* 2007; **2**(9): 2233–2244.
PubMed Abstract | **Publisher Full Text**

Fyad-Lameche FZ, Iantcheva A, Siljak-Yakovlev S, *et al.*: **Chromosome number, genome size, seed storage protein profile and competence for direct somatic embryo formation in Algerian annual *Medicago* species.** *Plant Cell Tissue Organ Cult.* 2016; **124**(3): 531–540.
Publisher Full Text

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing**. 2012; [Accessed 26 July 2023].
Reference Source

Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C links with assembly graphs for chromosome-scale assembly.** *PLoS Comput Biol.* 2019; **15**(8): e1007273.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps**. 2022; [Accessed 19 October 2022].
Reference Source

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *Gigascience.* Oxford University Press, 2021; **10**(1): giaa153.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Jarecka A, Saniewska A, Bialy Z, *et al.*: **The effect of *Medicago arabica*, *M. hybrida* and *M. sativa* saponins on the growth and development of *Fusarium oxysporum* Schlecht f. sp. *tulipae* apt.** *Acta Agrobot.* 2008; **61**(2): 147–155.
Publisher Full Text

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection**. *protocols.io*. 2023.
Publisher 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**

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**

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**

Narváez-Gómez JP, Mbye H, Oatley G, *et al.*: **Sanger Tree of Life Sample Homogenisation: Covaris cryoPREP® Automated Dry Pulverizer V.1.** *protocols.io*. 2023.
Publisher Full Text

OABIF: **Spotted medick *Medicago arabica* (L.) Huds**. 2022; [Accessed 26 January 2024].
Reference Source

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

POWO: **Plants of the world online**. Royal Botanic Gardens, Kew, 2024.
Reference Source

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**

Rautiainen M, Marschall T: **MBG: Minimizer-based Sparse de Bruijn Graph construction.** *Bioinformatics.* 2021; **37**(16): 2476–2478.
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.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | **Publisher Full Text**

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI**. *protocols.io*. 2023.
Publisher Full Text

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0)**. *Zenodo.* 2023a.
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev)**. *Zenodo.* 2023b.
Publisher Full Text

Tava A, Mella M, Avato P, *et al.*: **New triterpenic saponins from the aerial parts of *Medicago arabica* (L.) huds.** *J Agric Food Chem.* 2009; **57**(7): 2826–35.
PubMed Abstract | **Publisher Full Text**

Todorovic M, Oatley G, Denton A, *et al.*: **Sanger Tree of Life HMW DNA extraction: manual plant MagAttract v.2/3**. *protocols.io*. 2023a; [Accessed 3 January 2024].
Publisher Full Text

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for PacBio HiFi**. *protocols.io*. 2023b.
Publisher Full Text

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
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

Wang T, Ren L, Li C, *et al.*: **The genome of a wild *Medicago* species provides insights into the tolerant mechanisms of legume forage to environmental stress.** *BMC Biol.* 2021; **19**(1): 96.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wellcome Sanger Institute: **The genome sequence of the spotted medick, *Medicago arabica* (L.) Huds. 1762**. European Nucleotide Archive. [dataset], accession number PRJEB47317, 2022.

Open Peer Review

Current Peer Review Status: ? ✓ ? ✓

Version 2

Reviewer Report 28 November 2024

https://doi.org/10.21956/wellcomeopenres.25757.r112320

© **2024 Brazier T.** 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.

✓ **Thomas Brazier**
University of Rennes, Rennes, France

This Data Note presents the first genome assembly for the spotted medick, *Medicago arabica*. This new reference genome assembly is chromosome-scale, with 8 chromosomes corresponding to the expected haploid chromosome number and the mitochondrial and plastid genomes.

The ecology and genomics of the spotted medick are well described, and the importance of the study system is clearly stated.

DNA extraction followed publicly available protocols from the Wellcome Sanger Institute and sequencing was done according to manufacturer protocols (Pacbio HiFi, 10X Genomics read cloud, Hi-C). Genome assembly was performed with state-of-the-art bioinformatic tools with rigorous QC and manual curation and evaluation. Globally the material and methods were clearly presented.

The genome assembly is technically sound, and the QC is complete and appropriate. The final assembly covers 515 Mb of the 610 Mb estimated using flow cytometry (1C-value) and shows a high completeness. The results are sufficiently clear and detailed to evaluate that it is a good-quality genome assembly that will be an important reference genome for *Medicago* species.

I just have a few minor comments, mostly typos:

In the title, the 'ca' at the end of *Medicago arabica* is not italicised.

In the 'Background' section, the sentence "Like many other *Medicago* species, spotted medick is rich in a variety of saponins with potential for use as antimicrobial compounds in agriculture and medicine (e.g. Avato et al., 2006; Bialy et al., 2004; Jarecka et al., 2008; Tava et al., 2009)." is repeated in two consecutive paragraphs.

In the 'Genome Annotation Report' section, in the sentence "The Medicago arabica genome assembly (GCA_946800305.1) was annotated at the European Bioinformatics Institute (EBI)", *Medicago arabica* is not italicised.

In the sentence "The average transcript length is 2,758.62", please add the unit "2,758.62 bp"

Table 3. The gEVAL version is marked as 'N/A' since the website is not versioned. Maybe report the date (month/year) at which it was accessed in replacement of a proper version number.

In the 'Genome annotation, curation and evaluation' section, you could use 'MercuryFK' as on the official website instead of 'Mercury. FK'

Figure 5. The Hi-C contact map lacks axis ticks and labels.

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: Plant comparative genomics and population 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.

Reviewer Report 18 November 2024

https://doi.org/10.21956/wellcomeopenres.25757.r109992

© 2024 Feng 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.

?

Yu Feng
Chengdu Institute of Biology, Chinese Academy of Sciences, Sichuan, China

The manuscript reported high-quality a chrmosome scale assembly of the Medicago arabica genome, using Pacbio HiFi sequencing, 10X genomics and Hi-C library. There are still some points need to clarify:

1. The Hi-C contact map lacks axis ticks and numbers.

2. The number of annotated coding genes are much less than the other two reported Medicago genomes. Is there only one tissue (leaf) were used in the RNA-seq to annotate the genome? Usually we use more tissues (e.g. flower, root, seed) to cover as many expressed genes as possible.

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

Are the protocols appropriate and is the work technically sound?
Partly

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: Plant comparative genomics and population 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, however I have significant reservations, as outlined above.

Reviewer Report 15 November 2024

https://doi.org/10.21956/wellcomeopenres.25757.r109640

© 2024 Lyu X. 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.

Xiaolong Lyu
College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China

This article is approved for indexing in current form.

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: Plant science, genetics

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.

Version 1

Reviewer Report 11 September 2024

https://doi.org/10.21956/wellcomeopenres.23231.r95261

© 2024 Lyu X. 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.

?

Xiaolong Lyu
College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China

Christenhusz et al. present the first high-quality genome assembly of *Medicago arabica* (the spotted medick; Tracheophyta; Magnoliopsida; Fabales; Fabaceae).The genome assembly of *Medicago arabica* has important implications for the biochemical pathways of saponin compound synthesis, comparative genomics of cultivars and wild relatives, and agricultural breeding. A few points need to consider:

In this study, the mitochondrial and chloroplast genomes of this species were assembled from PacBio HiFi reads using the MBG software, with these reads mapped to the relevant genomes. According to the website https://github.com/maickrau/MBG, the MBG software has undergone multiple versions, but the specific version used in the study is not provided. Given that different versions of MBG software and their outputs can vary, it is recommended that the authors specify the exact version used.

Although this is the first high-quality genome assembly of *Medicago arabica*, it lacks functional and structural annotations. To make the genome more accessible for a wider audience, it is recommended to include the corresponding gene annotations.

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

Are the protocols appropriate and is the work technically sound?
Partly

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: Plant science, genetics

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, however I have significant reservations, as outlined above.

Reviewer Report 29 August 2024

https://doi.org/10.21956/wellcomeopenres.23231.r92783

© 2024 Pessoa-Filho M. 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.

?

Marco Pessoa-Filho
Embrapa Recursos Geneticos e Biotecnologia, Brazilian Agricultural Research Corporation, Brasília, DF, Brazil

The data note describes the assembly of the spotted medick genome (*Medicago arabica*), a species of the Fabaceae family native to the Mediterranean region and related to alfalfa and the bur clover.

PacBio HiFi reads were obtained (44-fold coverage) and used for assembly with hifiasm. 10X genomics read clouds (86-fold coverage) were used for polishing. Hi-C reads generated with the Arima2 kit were used to obtain chromosome scale scaffolds with SALSA2. Manual curation was carried with gEVAL, HiGlass and PretextView. Merqury was used to assess k-mer completeness and QV consensus quality. RNAseq data was obtained from leaf tissue, but annotation was not reported in the data note.

Raw data and assembly were deposited in INSDC databases and are publicly available.

The rationale for creating the dataset was clearly described. Protocols were appropriate and the work is technically sound. The datasets are clearly presented in a usable and accessible format.

More details on genome assembly, curation and evaluation could be provided in the Materials and Methods and would enrich the data note and allow its replication:

1) Was Hifiasm run with default parameters?

2) Considering that Hi-C data was available, was it used as input to Hifiasm for phasing?
3) What output was further used in the pipeline for curation and scaffolding? Was it the p_ctg? The hap1_ctg?
4) Considering that hifiasm already includes purging of haplotype duplications, why was purge_dups used? Was hifiasm run without haplotype duplication removal?
5) More details on the polishing step must be provided. Considering HiFi reads low error rates, how much of an improvement is seen in the final assembly by running this step?
6) What dataset was used as input to Merqury along with the assembly to assess k-mer completeness and QV consensus quality values?
7) Table 3 lists MerquryFK, not the original Merqury. This should be mentioned in the text.
8) Table 3 lists MitoHiFI but the text mentions MBG for mitochondrial and chloroplast genome assemblies. Which one was used?

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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant genomics, bioinformatics

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, however I have significant reservations, as outlined above.