scientific data

OPEN

DATA DESCRIPTOR

A chromosome-scale reference genome of grasspea (*Lathyrus sativus*)

Marielle Vigouroux[1], Petr Novák[2], Ludmila Cristina Oliveira[2], Carmen Santos[3], Jitender Cheema[1,4], Roland H. M. Wouters[1], Pirita Paajanen[1], Martin Vickers[1], Andrea Koblížková[2], Maria Carlota Vaz Patto[3], Jiří Macas[2], Burkhard Steuernagel[1], Cathie Martin[1] & Peter M. F. Emmrich[1,5] ✉

Grasspea (*Lathyrus sativus* L.) is an underutilised but promising legume crop with tolerance to a wide range of abiotic and biotic stress factors, and potential for climate-resilient agriculture. Despite a long history and wide geographical distribution of cultivation, only limited breeding resources are available. This paper reports a 5.96 Gbp genome assembly of grasspea genotype LS007, of which 5.03 Gbp is scaffolded into 7 pseudo-chromosomes. The assembly has a BUSCO completeness score of 99.1% and is annotated with 31719 gene models and repeat elements. This represents the most contiguous and accurate assembly of the grasspea genome to date.

Background & Summary

Grasspea (*Lathyrus sativus* L.) is a legume crop valued for its resilience in the face of environmental stress, including drought, flooding and salinity[1]. The crop has been cultivated for at least 8000 years[2,3], and has been widely distributed around parts of Europe, Asia and Africa, although most present-day cultivation takes place in South Asia and the highlands of Ethiopia and Eritrea[1,4].

Grasspea is a diploid species with seven chromosome pairs and predominantly autogamous reproduction[5]. Two genome assemblies of *Lathyrus sativus* have been published to date, for the genotypes Pusa-24[6] and LS007[7]. The new reference genome assembly of LS007 which we present here represents a major advance in completeness, contiguity and accuracy of assembly and can serve as a reference genome for future research on grasspea. The material used for sequencing had undergone 6 generations of single-seed descent to ensure a low degree of heterozygosity.

This *de-novo* assembly was based on Pacific Biosciences HiFi long reads, scaffolded to chromosome scale using HiC-data previously used in assembling the LS007 draft genome[7]. Repeat elements in this assembly were annotated using a combination of *de novo* repeat identification and similarity searches to previously published repeat domain[8] and class II transposon databases[7,9,10]. The distribution patterns of selected satellite repeats visualized by fluorescence *in situ* hybridization (FISH) were used to assign pseudomolecules to specific chromosomes. The positions of the centromeres in the assembly were determined by ChIP-seq with CENH3-specific antibodies. The repeat-masked assembly was annotated using the Braker3 pipeline, using previously published RNA-seq data[7], and gene hints from the ODB11 Viridiplantae set and the *Pisum sativum* ZW6 annotation[11] resulting in 31,719 high confidence gene annotations. The workflow used to assemble and annotate this genome is shown in Fig. 1.

This reference genome is suitable for comparative genomics analyses regarding legume evolution, as a basis for genome wide association studies and for the identification of candidate genes for reverse-genetics approaches, enabling accelerated crop improvement in grasspea and a genetic characterisation of grasspea water stress tolerance mechanisms[12] to inform the breeding of other legume crops.

[1]John Innes Centre, Norwich Research Park, Colney Lane, Norwich, NR4 7UH, UK. [2]Institute of Plant Molecular Biology, Biology Centre CAS, Branisovska 31, Ceske Budejovice, CZ, 37005, Czech Republic. [3]Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, Oeiras, 2780-157, Portugal. [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, Cambridge, United Kingdom. [5]Norwich Institute for Sustainable Development, School of International Development, University of East Anglia, Norwich, NR4 7TJ, UK. ✉e-mail: p.emmrich@uea.ac.uk

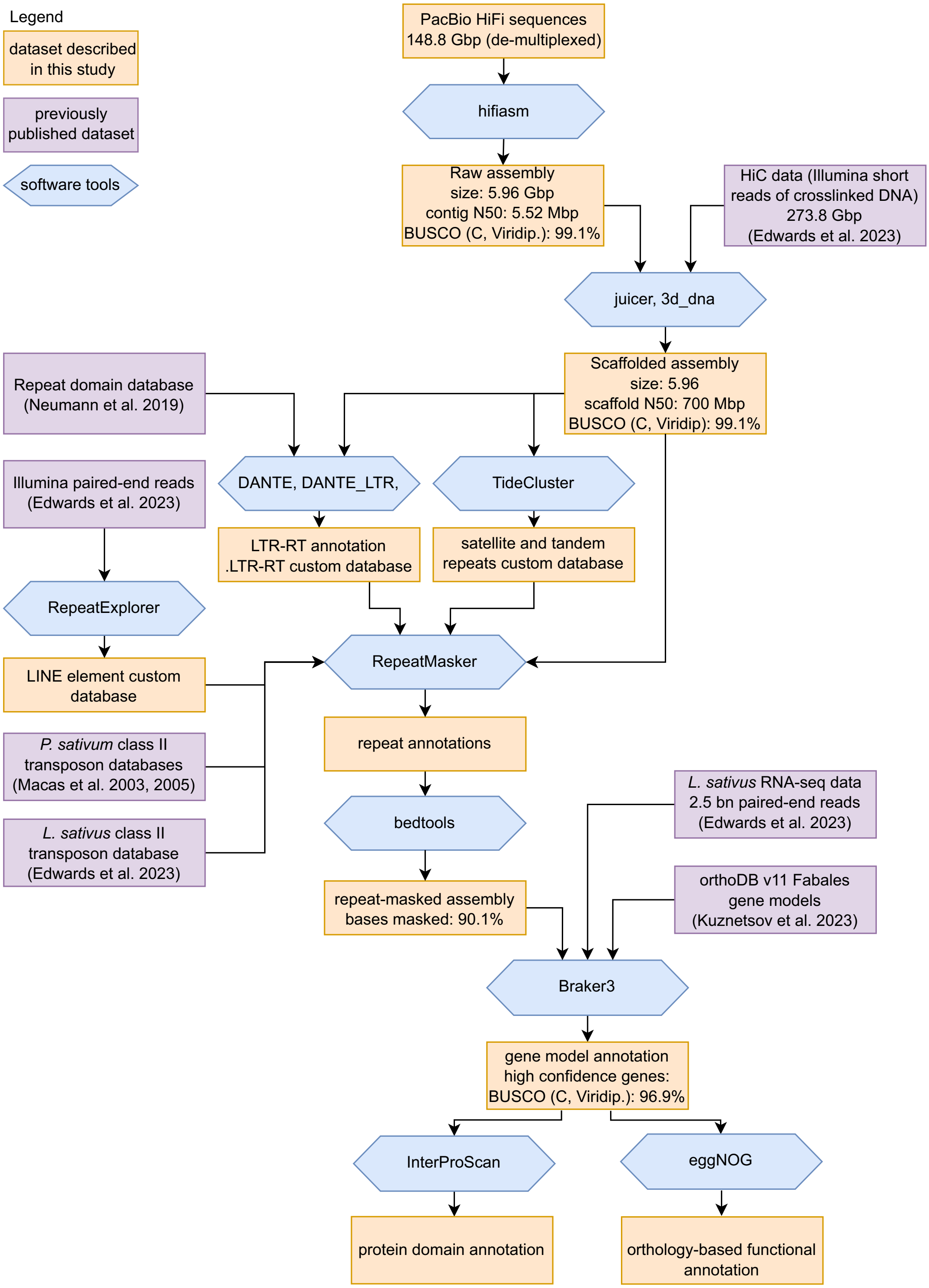
Legend
dataset described in this study
previously published dataset
software tools
PacBio HiFi sequences 148.8 Gbp (de-multiplexed)
hifiasm
Raw assembly size: 5.96 Gbp contig N50: 5.52 Mbp BUSCO (C, Viridip.): 99.1%
HiC data (Illumina short reads of crosslinked DNA) 273.8 Gbp (Edwards et al. 2023)
juicer, 3d_dna
Scaffolded assembly size: 5.96 scaffold N50: 700 Mbp BUSCO (C, Viridip): 99.1%
Repeat domain database (Neumann et al. 2019)
DANTE, DANTE_LTR,
TideCluster
Illumina paired-end reads (Edwards et al. 2023)
LTR-RT annotation .LTR-RT custom database
satellite and tandem repeats custom database
RepeatExplorer
LINE element custom database
RepeatMasker
P. sativum class II transposon databases (Macas et al. 2003, 2005)
repeat annotations
L. sativus RNA-seq data 2.5 bn paired-end reads (Edwards et al. 2023)
bedtools
orthoDB v11 Fabales gene models (Kuznetsov et al. 2023)
L. sativus class II transposon database (Edwards et al. 2023)
repeat-masked assembly bases masked: 90.1%
Braker3
gene model annotation high confidence genes: BUSCO (C, Viridip.): 96.9%
InterProScan
eggNOG
protein domain annotation
orthology-based functional annotation

Fig. 1 Workflow of genome assembly and annotation. BUSCO scores are given as percentage of complete (single-copy and duplicated) BUSCOs.

Methods

DNA extraction and HiFi sequencing. For PacBio long-read sequencing, 10 tubes of ~0.1 gram of young, fresh leaf tissue were collected in 1.5-ml low-bind Eppendorf tubes and snap-frozen in liquid nitrogen. Frozen leaf tissue was ground using a mortar and pestle and homogenized and washed in sorbitol buffer[13]. High molecular weight DNA was extracted using the Illustra Nucleon PhytoPure kit (Cytiva, RPN8510) following the manufacturer's protocol. Solutions were transferred using wide-bore pipette tips to circumvent the shearing of DNA. DNA concentration was determined using the Qubit broad-range assay. The purity of each extraction was assessed using a NanoDrop spectrophotometer (Thermo Fisher) based on A260nm/A280nm (1.8–2.0) and

A260nm/A230nm (1.8–2.2) absorbance ratios, and by comparing the NanoDrop concentration estimate to the Qubit estimate, mQubit/mNanoDrop ratio close to 1:1.5[14]. The length of extracted DNA molecules was assessed using a TapeStation System (Agilent). The samples that passed QC were combined to a total amount of 35 µg (quantified by Qubit) and sent to Earlham Institute for a final QC on the Femto fragment analyzer (Agilent), library preparation for PacBio HiFi and sequencing on the Sequel IIe system (PacBio). The library preparation for PacBio HiFi used the ELF2 fraction (~20Kb) after size selection using the SageELF instrument (Sage Sciences). Sequencing and de-multiplexing produced a sequence yield of 148.8 Gbp, in HiFi reads with a mean read length of 16,390 bp.

Long-read assembly. Reads from 9 PacBio SMRT cells were de-multiplexed, resulting in a total of 148.8 Gbp of HiFi reads. These were assembled into contigs using hifiasm[15] version 0.16.1, all parameters default except -f 38. Haplotype-collapsed assembly (outputfile including substring "bp.p_ctg" was used for further processing). Blobtools (version 1.1.1)[16,17] was used to inspect contig assembly for contamination, with short-read data from PRJEB33571 (run accessions ERR3453988, ERR3453989, ERR3453990) and the NCBI nucleotide collection (downloaded 21/Oct/2022) as input data. Short-read data were mapped to contigs using bwa (version 0.7.17)[18] and sorted SAM files using samtools (version 1.9)[19].

HiC scaffolding. The assembly was scaffolded using Hi-C data (NCBI: SRX19210597) and the same procedure as detailed in Edwards *et al.*[7] Briefly, it followed the Juicer[20] (version 1.6) and the 3D-DNA[21] (release 201008-cb63403) pipeline followed by manual curation using Juicebox[22] (version 2.13.07). The contact map resulting from manual curation is shown in Fig. 2.

Merqury v1.3 was used to assess assembly quality of present assembly as well as previously published assembly. Meryl (v. 1.4.1) was used to build a kmer library of Illumina short read data sets of LS007 (PRJEB33571, accession numbers ERR3453988, ERR3453989, ERR3453990).

Fluorescence *in situ* hybridization. The metaphase chromosomes used for FISH were prepared from synchronized root tip meristems[23] using the air-dry dropping method[24]. The oligonucleotide probes (Supplementary table 1) were designed according to the sequences of seven abundant satellite DNA families that had been previously identified[23,25]. Probes were labeled with rhodamine red-X during synthesis (Integrated DNA Technologies, Leuven, Belgium). FISH was performed[26], with hybridization and wash temperatures adjusted to account for AT/GC content and stringency of hybridization, allowing for 10–20% mismatches. Slides were counterstained with 4',6-diamidino-2-phenylindole (DAPI) in Vectashield mounting medium (Vector Laboratories, Burlingame, CA, USA) and examined using a Zeiss AxioImager.Z2 microscope with an Axiocam 506 monocamera. Images were captured and processed using ZEN 3.2 software (Carl Zeiss GmbH).

Chromatin-Immunoprecipitation sequencing. ChIP experiments were performed with native chromatin as described previously[27] using custom antibodies that specifically recognize one of the two variants of the CENH3 proteins found in *Pisum* and *Lathyrus* species[28–30]. The P43 antibody was raised against the CENH3-1 variant using the peptide sequence "GRVKHFPSPSKPAASDNLGKKKRRCKPGTKC"[27]. The CENH3-2 variant was detected with antibody P60 raised against the peptide "QTPRHARENQERKKRRNKC"[31]. DNA fragments were purified from the immunoprecipitated samples, and the corresponding control samples (Input; digested chromatin not subjected to immunoprecipitation) were sequenced on the Illumina platform (Admera Health, NJ, USA) in paired-end, 150 bp mode. The reads were quality-filtered and trimmed using Trimmomatic[32] (minimum allowed length = 100 nt), resulting in 82–99 million forward reads per sample, which were mapped to assembly using Bowtie2 version 2.5.1[33] with options -p 64 -U. Subsequent analysis was performed using either the full output of the Bowtie2 program, or the output with all multimapped reads filtered out. Filtering of multimapped reads was performed using Sambamba version 1.0.0[34] with the options "-F [XS] = = null and not unmapped and not duplicate". Regions with statistically significant ChIP/Input enrichment ratio were identified by comparing ChIP and Input mapped reads using the epic2 program[35], with the parameter "--bin-size 200".

Repeat masking and annotation. Tandem repeats and satellites were annotated using TideCluster[36], a wrapper for TideHunter[37]. Satellite repeats with a monomer size ranging from 40 to 3 kbp and a minimum array length of 5 kbp were annotated using the default TideCluster settings. Satellites with a monomer size between 10 to 39 bp and a minimum array length of 5 kbp were identified using TideCluster with parameters -T "-p 10 -P 39 -c 5 -e 0.25" -m 5000.

LTR retrotransposons (LTR-RT) were annotated using DANTE v0.1.8[38] and the DANTE_LTR v0.2.3.2 pipeline[39] on the RepeatExplorer Galaxy server[40]. The sequences of the identified LTR-RT elements were used to create a custom library of LTR-RT elements using the "dante_ltr_to_library" script from the DANTE_LTR repository[39].

A custom library of Class II transposable elements was obtained using RepeatExplorer clustering procedure 1[41] on unassembled Illumina paired-end reads. Contigs corresponding to Class II retrotransposons with a minimum read depth of 5 reads and a minimum length of 100 bp were obtained using tools on the RepeatExplorer Galaxy server. A custom library of LINE elements was created by extracting regions with LINE protein coding domains identified by DANTE, along with the upstream and downstream 4 kb flanking regions. The extracted genomic sequences were split into 100 nt fragments and analyzed by RepeatExplorer clustering. Contigs corresponding to LINE elements with a read depth of at least 3 reads and a minimum length of 150 nt were converted into a custom library. Consensus sequences of rRNA gene arrays including intergenic spacer sequences were fully reconstructed from the RepeatExplorer contigs.

All custom libraries were concatenated and used as a library for RepeatMasker[42] search. The RepeatMasker search was performed on the RepeatExplorer Galaxy server with options "-xsmall -no_is -e ncbi". All regions

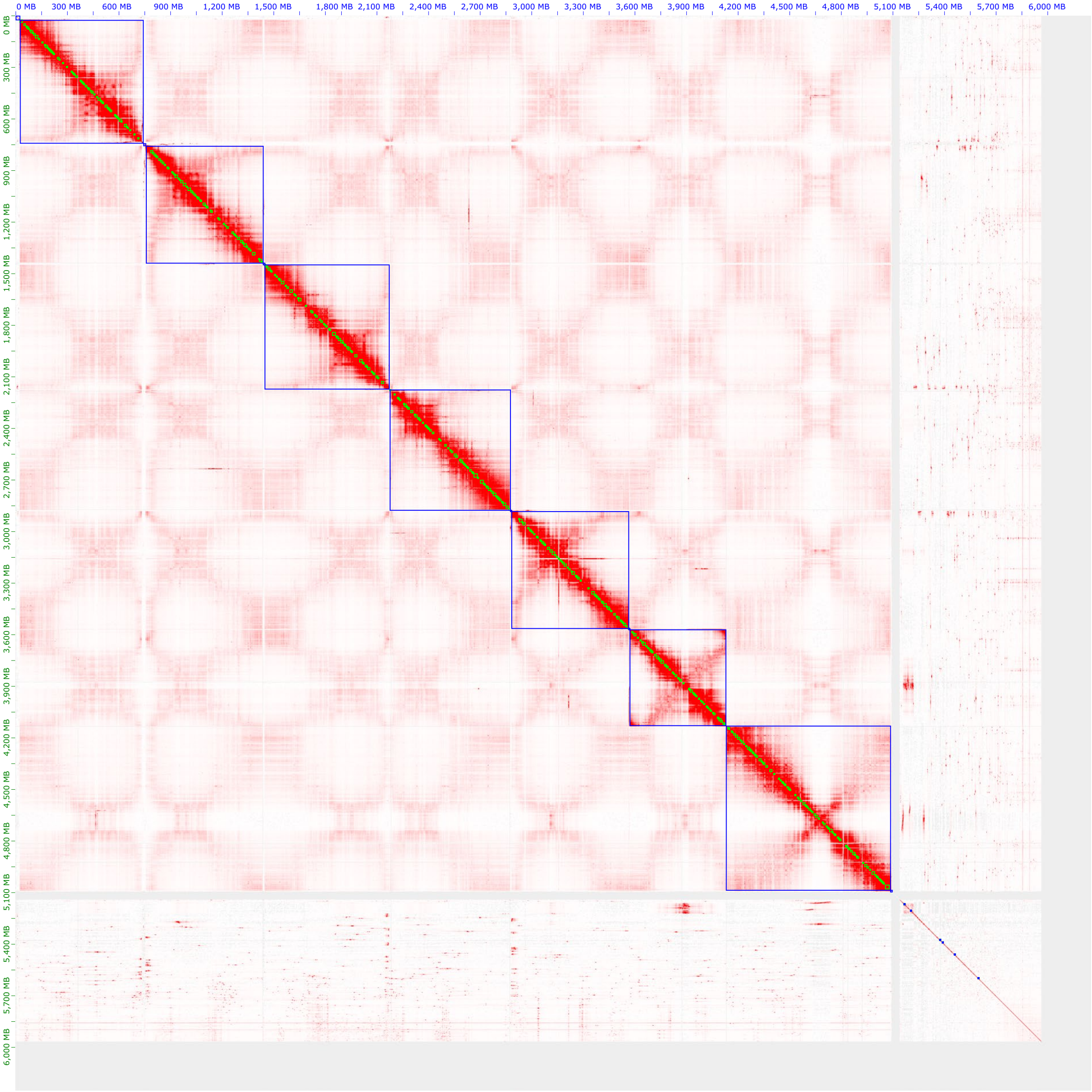

Fig. 2 HiC contact map following manual curation of scaffolds. Chromosome-scale scaffolds are shown in blue boxes, prior to reordering shown in Table 2.

annotated as mobile elements with RepeatMasker based on custom library search which overlapped with satellite repeats annotated by TideCluster were removed from the annotation using bedtools[43] with command "bedtools subtract".

The resulting GFF3 was then merged with the DANTE annotation using a custom R script[44]. The classification of mobile elements in the annotation files corresponds to the classification system used in the REXdb database[8].

For the final repeat-masking process, all of the above repeat annotation GFF3 files were consolidated. We merged the annotated regions into a single BED file using the bedtools merge tool[43].

Gene model annotation. We used Braker3 (3.0.0)[45,46] for gene annotation, which uses mapped RNA-seq data and a protein database as inputs to annotate gene models. We mapped RNA-Seq data from Edwards *et al.* (PRJNA929208) to our scaffolded assembly using hisat2 and used OrthoDB11/Viridiplantae.fa as the protein database. Braker3 was run with default parameters using both inputs.

Comparative mapping. A comparative mapping analysis was performed between the LS007 scaffolded assembly and two genetic linkage maps developed for two RILs populations: (1) the *L. sativus* RAIPUR-4 x LS87-124-4-1[47] and (2) the phylogenetically close *L. cicera* BGE023542 x BGE008277[48].

		EIv1 (2020)[60]	Rpb (2020[60], scaff. 2023[7])	Rajarammohan *et al.*[6]	this assembly
	grasspea accession	LS007	LS007	Pusa-24	LS007
Assembly	Sequencing technology	Illumina paired end	PromethION nanopore + Illumina PE	Illumina paired end + PacBio Sequel	PacBio HiFi
	total length	8.12 Gbp	6.22 Gbp	3.81 Gbp	5.96 Gbp
	of which N's	1920 Mbp	0 Mbp	5.44 Mbp	2.93 Mbp
	contig N50	0.006 Mbp	0.156 Mbp	0.078 Mbp	5.52 Mbp
	Number of contigs	669,893	162,994	80,744	8,449
Scaffolding	Scaffolding approach	Illumina long mate pair	HiC (partial)	aligned to Caméor v1a[49]	HiC
	scaffold N50	0.06 Mbp	363 Mbp (partial)	421 Mbp	700 Mbp
	total size of top 7 scaffolds	not chromosome-level	2.51 Gbp	3.21 Gbp	5.03 Gbp
Merqury	Illumina QV	41.9	15.7	18.0	42.4
	Illumina kmer completeness	92.7	77.6	48.6	90.0
BUSCO	Complete-ness, Viridip-lantae	86.4%	89.8%	98.3%	99.1%
	Complete-ness, Fabales	79.9%	82.6%	96.0%	97.4%

Table 1. Comparison of benchmarks of this assembly with previous grasspea assemblies. BUSCO scores are given for "Complete" (single-copy and duplicated) BUSCOs.

L. sativus LS007 chromosome	*L. sativus* LS007 HiC scaffold	*L. sativus* Linkage Group	*L. cicera* Linkage Group	*P. sativum* Caméor chromosome	*L. sativus Pusa-24* chromosome
Lschr1	13	I, V and VIII	II	chr1LG6 and chr5LG3	Ls_pschr1 and Ls_pschr5
Lschr2	12	II	V	chr2LG1	Ls_pschr2
Lschr3	2	III	VI and VIII	chr3LG5	Ls_pschr3
Lschr4	4	IV	VII and IX	chr4LG4	Ls_pschr4
Lschr5	8	IX and X	III	chr5LG3	Ls_pschr5
Lschr6	10	VI	I	chr6LG2	Ls_pschr6
Lschr7	6	VII	IV	chr7LG7	Ls_pschr7

Table 2. Syntenic relationships between LS007 assembly with *Pisum sativum* Caméor assembly[49], *Lathyrus sativus* Pusa-24 assembly[6] and *Lathyrus sativus* RAIPUR-4 x LS87-124-4-1 and *Lathyrus cicera* BGE023542 x BGE008277 genetic linkage maps.

The *L. sativus* RAIPUR-4 x LS87-124-4-1 linkage map was constructed using DArTseq-based SNPs and silicoDArT markers (microarray dominant markers), whereas *L. cicera* BGE023542 x BGE008277 linkage map contains not only DArTseq-based SNPs and silicoDArT markers, but also E-SSR (Expressed-simple sequence repeats), E-SNPs and ITAPs (intron targeted amplified polymorphism) markers. For a more comprehensive comparative analysis, the genomic sequences of the mapped markers on these two linkage maps were aligned against the LS007 assembly, and to the *L. sativus* Pusa-24 assembly[6] and the *P. sativum* cv. Caméor v1a assembly[49] using the BLASTn tool (e-value < 1e-5) from the OmicsBox v2.0 software[50]. BLAST results were further investigated for identification and removal of markers with multiple BLAST hits of identical probability of alignment (based on bit score, percentage of similarity and e-values) to different genomic regions in the genome assemblies.

Synteny between the genetic position/order of markers in the *L. sativus* RAIPUR-4 x LS87-124-4-1 and the *L. cicera* BGE023542 x BGE008277 linkage groups (LGs) and their corresponding physical position on the *L. sativus* and *P. sativum* assemblies was examined using Strudel visualization software[51]. The order rearrangement of *L. sativus* and *L. cicera* LGs was performed according to the assemblies in study.

Source data. This study makes use of the following previously published datasets:

- orthoDB v11[52] databases of orthologs
- Illumina short reads of cross-linked genomic DNA for HiC-scaffolding[7] https://identifiers.org/ena.embl:SRP419926, (SRA Run ID SRR23266411), FASTQ
- Illumina paired end reads of LS007 genomic DNA https://identifiers.org/ena.embl:ERP116375 (run accessions ERR3453988, ERR3453989, ERR3453990), FASTQ[7]
- grasspea RNA-seq data[7] https://identifiers.org/ena.embl:SRP419926. 7 tissues of genotype LSWT11, libraries GSM7008672 through GSM7008681, and drought/well-watered samples of whole roots and whole shoots of genotypes LS007 and Mahateora, libraries GSM7008683 through GSM7008706, FASTQ
- repeat databases[7,9,10]
- Genetic linkage maps developed for two RILs populations: (1) *L. sativus* RAIPUR-4 x LS87-124-4-1[47] and (2) *L. cicera* BGE023542 x BGE008277[48]

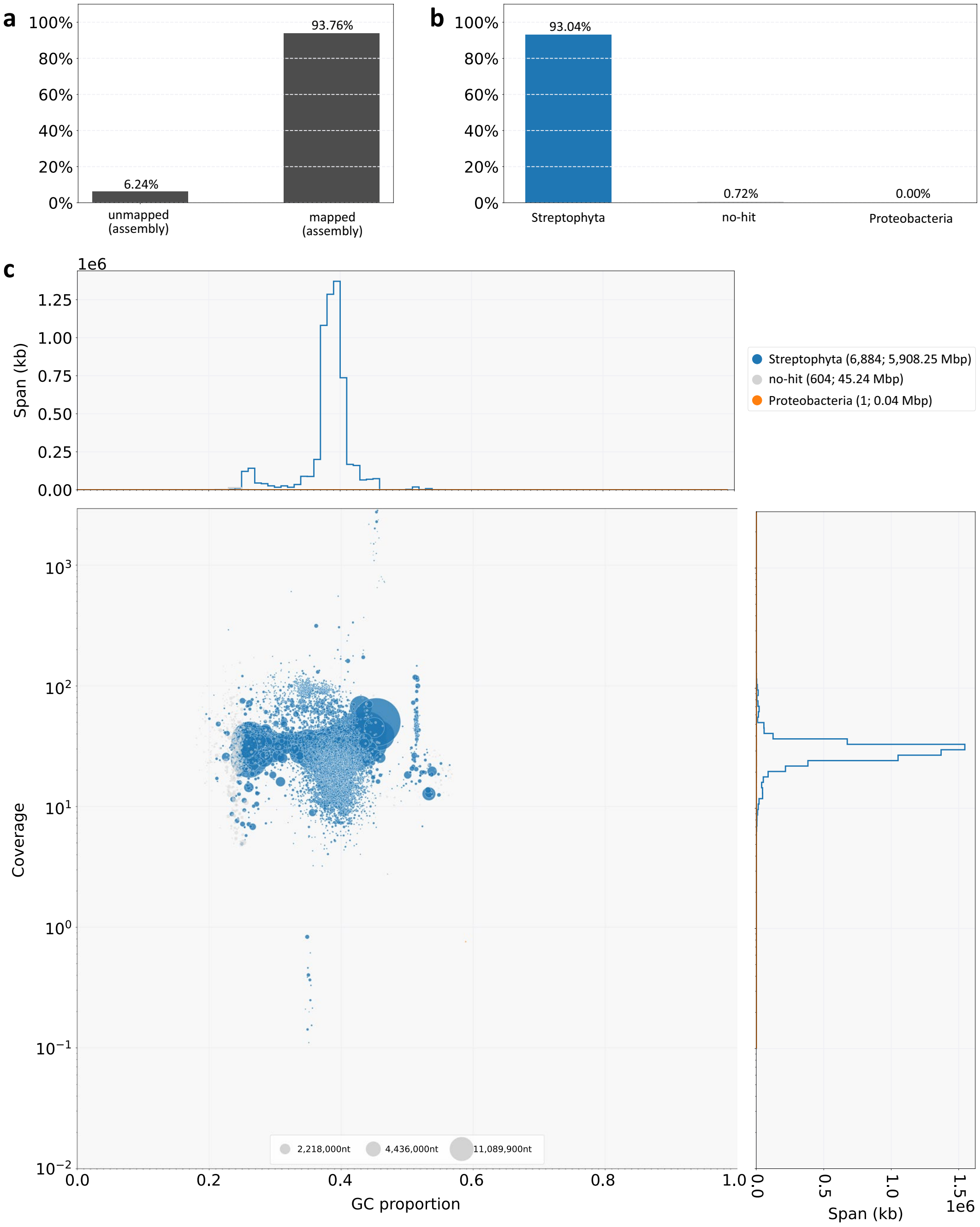

Fig. 3 BlobTools results for the un-scaffolded assembly. (**a**) percentage of PacBio HiFi reads mapped to the assembly contigs (**b**) breakdown of contigs (weighted by mapped reads) by taxonomic class, showing Streptophyta, as well as possible contaminants (Proteobacteria), and contigs of unknown taxonomic class (no-hit). (**c**) contigs plotted according to their average sequence coverage and GC content. The size of each bubble represents the length of each contig, with colours assigned by taxonomic class. Histograms show GC content and average sequence coverage of contigs, weighted by length.

Data Records

The datasets presented in this study comprise

- raw Pacific Biosciences HiFi long reads of LS007 genomic DNA, available at
- EBI ENA https://identifiers.org/ena.embl:ERP155791 (2024), FASTQ[53]
- scaffolded assembly of LS007 along with annotations as an EMBL format file on NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_963859935.3 (2024)[54]

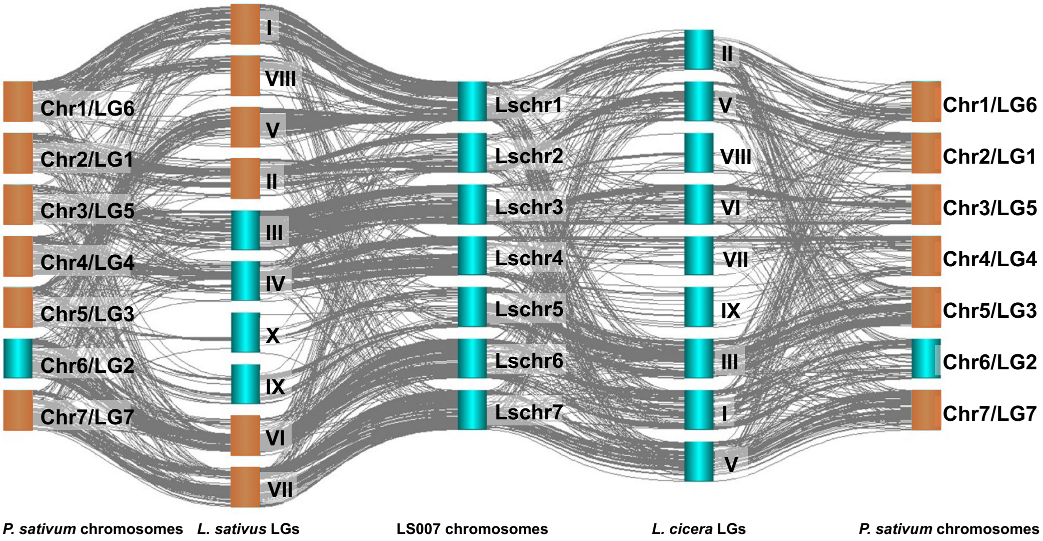

Fig. 4 Syntenic relationships plot between LS007 pseudomolecules (center), *Lathyrus sativus* RAIPUR-4 x LS87-124-4-1 LGs (center left) and *L. cicera* BGE023542 x BGE008277 LGs (center right) and assembly *Pisum sativum* cv. Caméor chromosomes v1a[49] (outer left and outer right). Brown boxes indicate chromosomes/LGs with inverted orientation from the original publications.

- CENH3 ChIP-seq sequencing, Illumina paired end, available at EBI ENA https://identifiers.org/ena.embl:ERP139716 (run accessions ERR12509730-ERR12509733), FASTQ[30]
- Functional annotation and repeat annotation are available on Zenodo, https://doi.org/10.5281/zenodo.10671532[55]

Technical Validation

Sequence quality. We used BlobTools to assess the quality of the raw assembly prior to scaffolding (Fig. 3). Of the Illumina paired-end reads 93.76% (Fig. 3a) mapped to the raw assembly. This includes 93.04% of reads mapping to contigs that are classified as belonging to Streptophyta (Fig. 3b), with minimal potential contamination of Proteobacteria (one contig, <0.00001% of assembly) or unknown classification (604 contigs, 0.75% of assembly). Each point in Fig. 3c corresponds to a contig, with coordinates determined by read coverage and GC content; size by contig size and color by taxonomic affiliation.

Assembly benchmarking. The chromosome-scale scaffolded assembly presented here represents a significant improvement on previous grasspea assemblies, as shown in Table 1. Due to its high level of fragmentation, the previous LS007 draft assembly "Rbp"[7] was only partially scaffolded, with 42.7% of the assembly assigned to 7 chromosome-scale and 2 sub-chromosome-scale scaffolds. The Pusa-24 assembly developed by Rajarammohan *et al.* was scaffolded by aligning Pusa-24 contigs to the *Pisum sativum* cv. Caméor v1a assembly[6,49]. This approach means regions of the grasspea genome that are not sufficiently similar to the pea genome (e.g. regions lost in the pea genome or expanded in the grasspea genome) could not be scaffolded. In addition, any differences in chromosome structure between pea and grasspea will be missed, as exemplified by a translocation from *P. sativum* chr1 to chr5, compared to the ancestral Galegoid karyotype[49]. By using the Caméor genome as a scaffold, this structure is carried over into the Pusa-24 assembly. We used Merqury to obtain consensus quality values (QV) for the present assembly as well as the previously published assembly. While the Rbp draft assembly has a QV of 15.69, our new assembly has a QV of 42.37, based on WGS Illumina data.

Comparative mapping. Comparative mapping between the LS007 scaffolded assembly and the previously published *L. sativus*[47] and *L. cicera*[48] genetic linkage maps confirmed a high degree of synteny. Out of the 2149 molecular markers mapped on the *L. sativus* RAIPUR-4 x LS87-124-4-1 genetic map, BLAST hits were obtained for 2060 (95.9%) markers. Using the *L. cicera* BGE023542 x BGE008277 genetic map, out of 1468 molecular markers, BLAST hits in the LS007 scaffolded assembly were obtained for 1278 (87.0%) markers. About 86.1% (1775 markers) and 87.2% (1115 markers) of the molecular markers with BLAST hits from the *L. sativus* and *L. cicera* linkage maps respectively, were mapped without redundancy to the LS007 scaffolds. A total of 1735 (80.7% of the total mapped markers in *L. sativus*) and 1115 (76.0% of the total mapped markers in *L. cicera*) molecular markers were assigned to the 7 chromosome-scale scaffolds. After rearranging the orientation of *L. sativus* and *L. cicera* LGs according to the LS007 assembly, a clear macrosynteny was observed, mainly between the LS007 assembly and the *L. sativus* linkage map (Fig. 4).

Comparing the homologous regions of major and minor LGs of the *L. sativus* RAIPUR-4 x LS87-124-4-1 linkage map with the LS007 assembly clearly indicates that LGII, III, IV, VI and VII correspond to HiC scaffolds 12, 2, 4, 10 and 6 respectively (Table 2). Similarly, *L. cicera* BGE023542 x BGE008277 LGI, II, III, IV and V correspond to HiC scaffolds 10, 13, 8, 6 and 12 respectively (Table 2). The syntenic relationship between *L. sativus* and

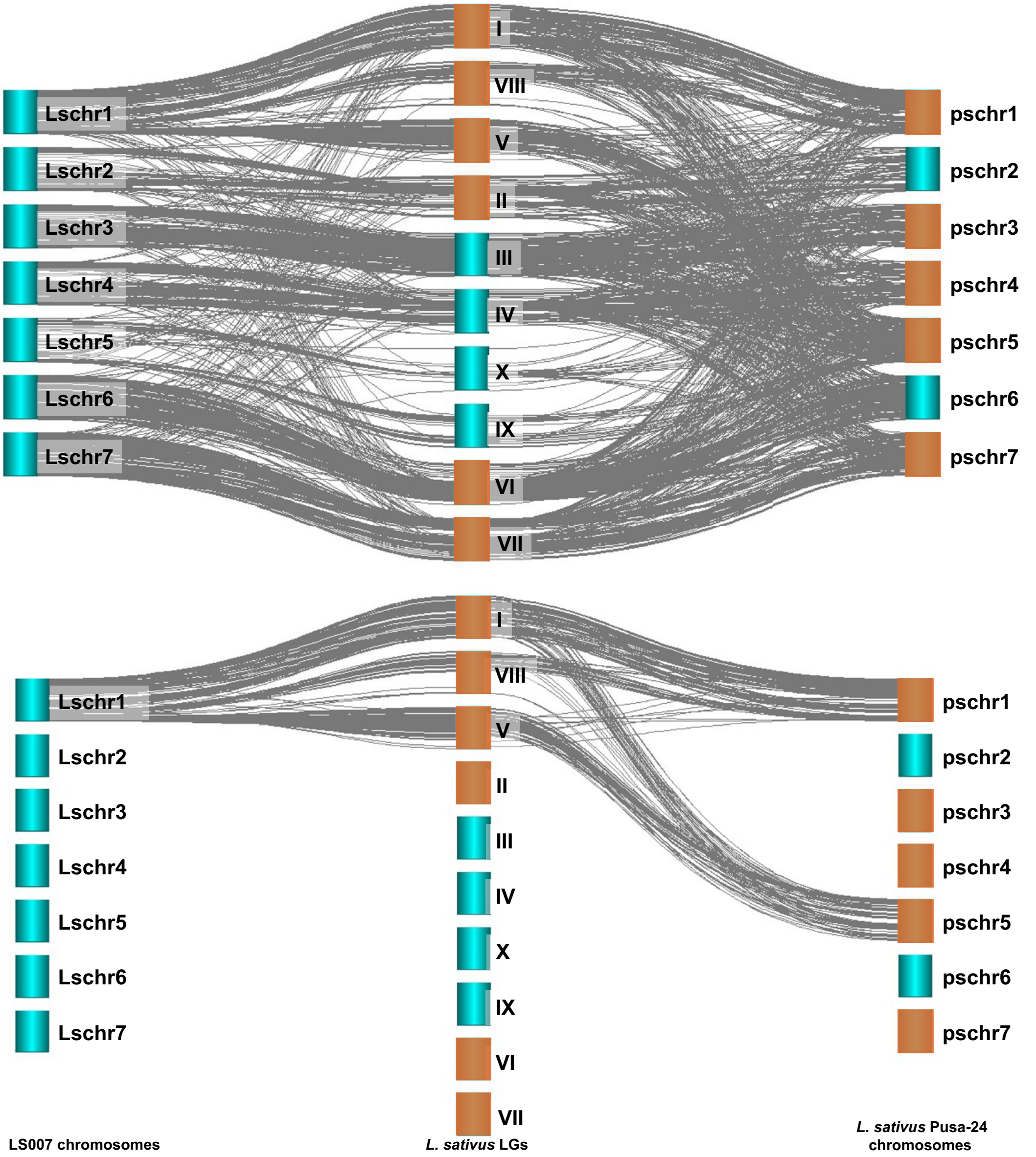

Fig. 5 Syntenic relationships plot among *L. sativus* RAIPUR-4 x LS87-124-4-1 linkage groups, LS007 HiC_Scaffolds and Pusa-24 chromosomes. On the top: All syntenic relationships. At the bottom: Detail on comparison between *L. sativus* LG V, VIII and I and LS007 Lschr1 (HiC_Scaffold_13) and Pusa-24 Ls_pschr1 and Ls_pschr5. Brown boxes indicate chromosomes/LGs with inverted orientation from the original publications.

L. cicera LGs and *P. sativum* Caméor reference genome[47] (Fig. 4) was used to assign chromosome designations to the chromosome-scale LS007 scaffolds (Table 2).

L. sativus RAIPUR-4 x LS87-124-4-1 LG I and the minor LG VIII map to about 2/3 of HiC_scaffold_13 and to pea chr1LG6 (Fig. 5). The whole of *L. sativus* LG V was also mapped to the end of HiC_scaffold_13 (and to the whole length of pea chr5LG3). This suggests a translocation between the *L. sativus* and *P. sativum* Caméor genome[47]. Likewise, *L. cicera* BGE023542 x BGE008277 LGII also mapped to LS007 HiC_scaffold_13 and to pea chr1LG6 and chr5LG3 (Fig. 5), supporting the chromosomal rearrangement between *P. sativum* and these two *Lathyrus* species. This matches the previously reported translocation between chr1 and chr5 of the *P. fulvum* and the *P. sativum sativum* lineages (which is not shared with *P. sativum elatius*)[49]. Indeed, this chromosomal rearrangement was also apparent when comparing the *L. sativus* RAIPUR-4 x LS87-124-4-1 linkage map with the *L. sativus* Pusa-24 genome assembly scaffolded based on the *P. sativum* Caméor genome (Fig. 5).

Finally, minor *L. sativus* RAIPUR-4 x LS87-124-4-1 LGs IX and X mostly map to the HiC_scaffold_8 and to the *P. sativum* chr5LG3 (Table 2). Both these LGs have fewer markers mapped, despite HiC_scaffold_8 being 700 Mbp long. Since the whole *L. cicera* BGE023542 x BGE008277 LG III mapped mainly to the HiC_scaffold_8 and the *P. sativum* chr5LG3, we assigned Lschr5 as the chromosomal designation to the LS007 HiC_scaffold_8.

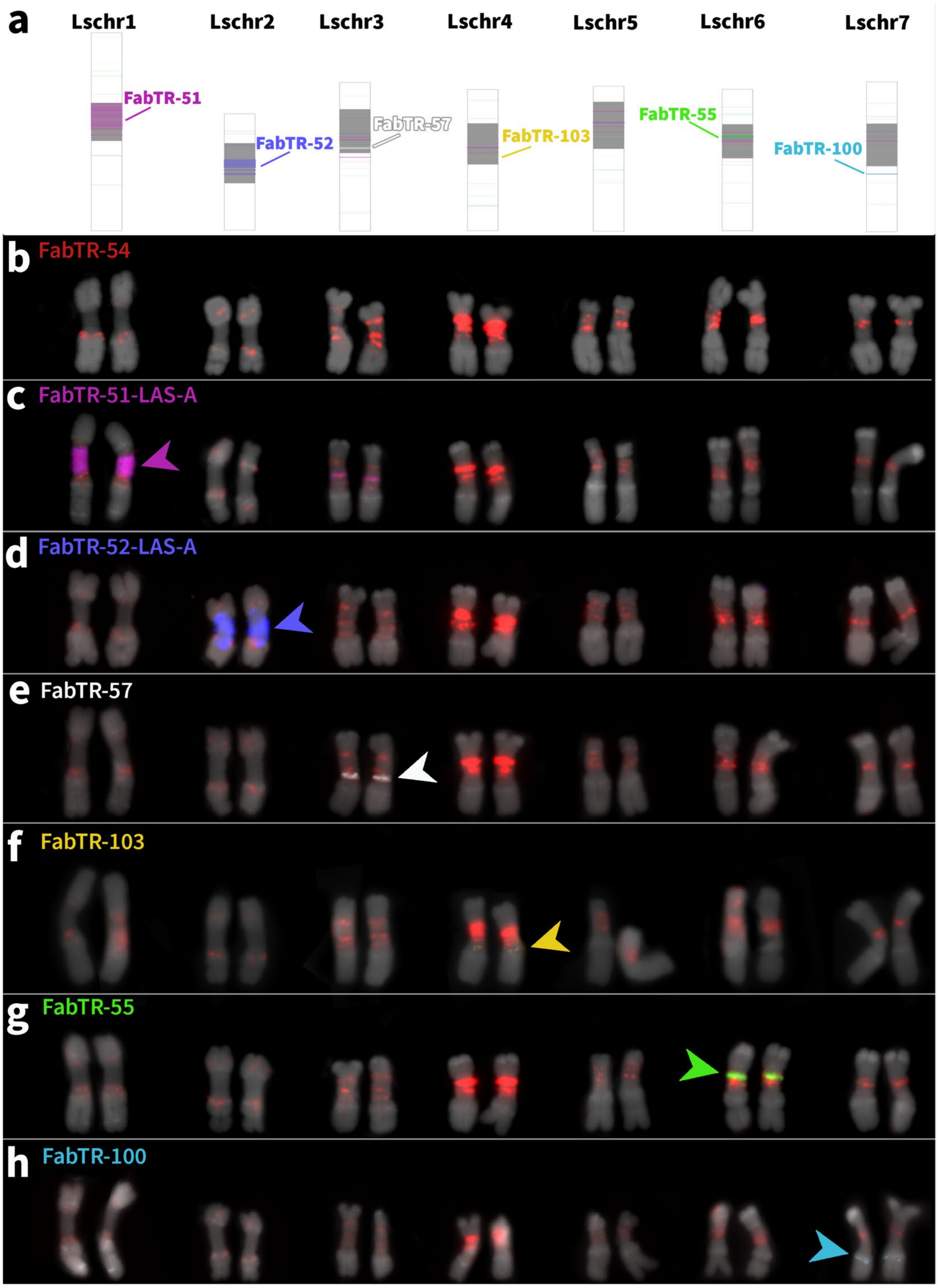

Fig. 6 Assignment of pseudomolecules to chromosomes using fluorescence *in situ* hybridization. (**a**) predicted locations of chromosome-specific FabTR satellites according to the genome assembly (centromere positions estimated from ChIP-seq are shown in grey), (**b**) FabTR-54 repeat (red) allowing chromosome discrimination according to hybridisation patterns, (**c**–**h**) hybridization signals of chromosome-specific satellites. Chromosomes were counterstained with DAPI (grey) in b–h.

Fluorescence *in situ* hybridization. As an independent validation of the assignment of pseudomolecules to chromosomes within the karyotype of *L. sativus*, we compared the distribution patterns of selected families of FabTR satellites (FabTR = Fabeae Tandem Repeats)[23,25] in the assembly (Fig. 6a) with those detected by FISH on metaphase chromosomes (Fig. 6b–h). Since these satellites are arranged into a small number of long arrays in the genome, they provide easily recognizable landmarks for distinguishing chromosomes. The probe for the FabTR-54 satellite provides hybridization signals on all chromosomes, which, together with the morphology

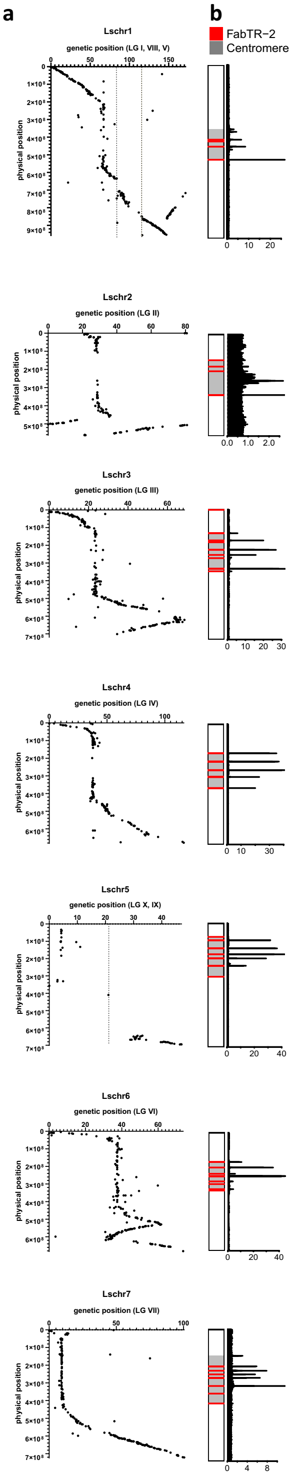

Fig. 7 Comparison of genetic and physical maps and centromere positions. (**a**) positions of genetic markers derived from the *L. sativus* RAIPUR-4 x LS87-124-4-1 in their linkage groups vs. their anchoring positions on the chromosome-scale scaffolds. Linkage groups mapping to the same scaffold (Lschr1 and Lschr5) are shown as concatenated (separated by dotted lines). (**b**) Positions of the centromeres in the assembly, determined by ChIP-Seq with the CENH3-2 antibody. The plots show the mean ChIP/input ratios calculated for 100 kb windows and reveal the positions of CENH3 domains as peaks in the graph (note that some smaller CENH3 domains are not visible at this magnification). The bars on the left represent pseudomolecules with highlighted locations of FabTR-2 satellites (red) and the extent of centromeric regions (grey) defined by the positions of the outermost CENH3 domains.

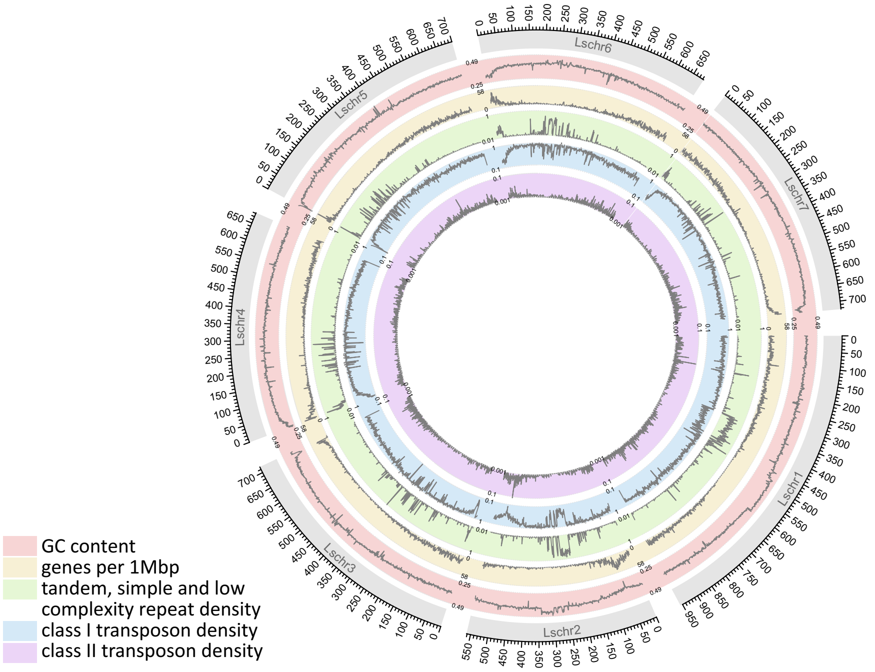

Fig. 8 Circos plot showing seven chromosome-scale scaffolds with tracks (from outermost to innermost) indicating GC content, number of genes per 1 Mbp, density of simple, tandem and low-complexity repeats (including satellite repeats), class I transposons and class II transposons. Densities of all features are shown averaged over 1Mbp windows.

of the chromosomes, allow all chromosomes within the karyotype to be distinguished[23] (Fig. 6b). In addition, we identified a set of six chromosome-specific satellites that were also present at the corresponding loci in the assembled pseudomolecules, allowing their unambiguous assignment to physical chromosomes (Fig. 6c–h). No chromosome-specific satellite was available for Lschr5.

Centromeres and validation of chromosome structure. The position of the centromeres in the assembled pseudomolecules was analyzed using ChIP-seq with CENH3 antibodies. Since *L. sativus* has two copies of the CENH3 gene (CENH3-1 and CENH3-2), both of which are expressed and corresponding proteins are localized in the centromeric chromatin[29], the experiments were performed in parallel with antibodies that distinguish the two protein variants. *L. sativus* has meta-polycentric chromosomes characterized by centromeres consisting of multiple CENH3 domains separated by long regions of CENH3-free chromatin. The positions of the outermost CENH3 domains define the extent of the primary constrictions (see supplementary table 2), which comprise up to one third of the chromosome length. In agreement with the previous cytogenetic studies of CENH3 distribution[28,29,31], the ChIP-seq enrichment signals of both CENH3 variants overlapped and were mostly localized on the arrays of the satellite repeat FabTR-2 (Fig. 7b). The observed distribution of FabTR-2/CENH3 positions in the pseudomolecules did not accurately reflect the corresponding FISH and CENH3 immunostaining patterns observed on some metaphase chromosomes, most likely due to the fact that some of the FabTR-2 arrays were missing or truncated in the assembly. Nevertheless, the multi-domain organization of the centromeres was evident on all pseudomolecules and their positions were generally consistent with the chromosome morphology observed in cytogenetic experiments (Fig. 7b). For comparison, the physical anchoring positions and positions in their linkage groups of markers derived from the *L. sativus* RAIPUR-4 x LS87-124-4-1 are shown in Fig. 7a,with good agreement between areas of low recombination and the positions of centromeric repeats. Linkage groups mapping to the same scaffold are shown as concatenated.

As shown in Fig. 7, in Lschr1, Lschr2, Lschr3 and Lschr6, the genetic positions of the markers are not a monotonously rising function of their physical positions along their entire length (while the linkage groups corresponding to Lschr5 do not contain enough markers for this analysis). Several factors could contribute to these discrepancies. Firstly, the *L. sativus* map[47] was developed from two genotypes (Raipur-4 from India and LS87-124-4-1 from Canada), which both differ from the line used in genome sequencing, LS007 from the UK. Hence genotypic differences in chromosome organization between LS007 and the parents of the mapping cross could result in some of these discrepancies. Secondly, there may be errors in the original maps that did not become apparent before due to the lack of a reference genome. This may be the case for the "V" shapes seen in the plots for Lschr3 and Lschr6, which imply markers that are in sequence in the genetic maps are split across stretches of sequence running in opposite directions. These features of the genetic map are similar to what might be expected of an inversion, causing it to appear shortened compared to its true length. Thirdly, errors during genome scaffolding could have resulted in contigs being placed in the wrong positions in the chromosomes. To check this possibility, we have performed collinearity analysis between the presented *L. sativus* LS007 assembly and the recent *P. sativum* cv. ZW6 assembly[11], shown in Supplementary Figure S1. Clear collinearity with ZW6, especially towards the telomeres, supports the overall correctness of the LS007 assembly. Full resolution of discrepancies due to genotypic differences, errors in the map or any residual errors in the assembly would likely require additional scaffolding data. This may also allow the placement of many of the remaining non-chromosome-scale scaffolds to generate a future telomere-to-telomere assembly, and the analysis of chromosome structure variants among grasspea diversity collections.

Gene annotation. To assess the quality of the generated data, both the gff3 file and the corresponding protein sequences were evaluated for BUSCO score[56] (Simao *et al.*[56]). The BUSCO pipeline (version 5.5.0) was executed with the following parameters: -f -c 16 -l viridiplantae_odb10 -m genome –offline. The genome and corresponding protein sequences were queried against the plants lineages embryophyta (embryophyta_odb10, n = 1614, v.2024-01-08, 1614 BUSCOs) and viridiplantae (viridiplantae_odb10, n = 425, v.2024-01-08) reference databases. The final assembly and the corresponding annotation were formatted using gff3toolkit (version 2.1.0) and converted to EMBL using EMBLmyGFF3 and accessioned as ERZ22626074.

In total, 31,719 protein-coding genes were annotated, with a mean gene length of 2620 bp, and an average number of 5.064 exons per gene (average exon length 252 bp). These encode 34,800 predicted proteins (1.097 transcripts per gene), of a mean length of 387 amino acids.

The submitted genome assembly achieved a genomic BUSCO score of 99.3% against both the lineages. Annotation completeness was evaluated using the protein output, resulting in 95.8% completeness for embryophyta and 96.9% for viridiplantae.

InterProscan[57] was used for a functional protein analysis. Genes found in the gene annotation were classified in protein families and structural domains and important sites were predicted. InterPro annotations were predicted using InterProScan v 5.53-87.0, with the parameters "-t p -dp -pa -appl Pfam,ProDom-2006.1,SuperFamily-1.75 --goterms –iprlookup". eggNOG-mapper v 2.1.12 was used for an orthology-based functional annotation. The orthology-based functional annotation circumvents collapsing annotations from close paralogs or duplicate genes with a higher chance of being involved in functional divergence. EggNOG uses precomputed Orthologous Groups (OGs) and phylogenies from the EggNOG database[58,59].

The distribution of genes and repeats across the seven chromosome-scale scaffolds is shown in Fig. 8.

Code availability

Source code for the gene annotation is available on github (https://github.com/gitbackspacer/grasspea_annotation).

Received: 19 February 2024; Accepted: 5 September 2024;
Published online: 27 September 2024

References

1. Dixit, G. P., Parihar, A. K., Bohra, A. & Singh, N. P. Achievements and prospects of grass pea (*Lathyrus sativus* L.) improvement for sustainable food production. *The Crop Journal* **4**, 407–416 (2016).
2. Kislev, M. E. Origins of the cultivation of *Lathyrus sativus* and L. cicera (Fabaceae). *Economic Botany* **43**, 262–270 (1989).
3. Coward, F., Shennan, S., Colledge, S., Conolly, J. & Collard, M. The spread of Neolithic plant economies from the Near East to northwest Europe: a phylogenetic analysis. *Journal of Archaeological Science* **35**, 42–56 (2008).
4. Lambein, F., Travella, S., Kuo, Y.-H., Van Montagu, M. & Heijde, M. Grass pea (*Lathyrus sativus* L.): orphan crop, nutraceutical or just plain food? *Planta* https://doi.org/10.1007/s00425-018-03084-0 (2019).
5. Campbell, C. G. *Grass Pea: Lathyrus Sativus L. Promoting the conservation and use of underutilized and neglected crops* vol. 18 (International Plant Genetic Resources Institute, 1997).
6. Rajarammohan, S. *et al.* Genome sequencing and assembly of *Lathyrus sativus* - a nutrient-rich hardy legume crop. *Sci Data* **10**, 32 (2023).
7. Edwards, A. *et al.* Genomics and biochemical analyses reveal a metabolon key to β-L-ODAP biosynthesis in Lathyrus sativus. *Nat Commun* **14**, 876 (2023).
8. Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob DNA* **10**, 1 (2019).
9. Macas, J., Koblížková, A. & Neumann, P. Characterization of Stowaway MITEs in pea (Pisum sativum L.) and identification of their potential master elements. *Genome* **48**, 831–839 (2005).
10. Macas, J., Neumann, P. & Pozárková, D. Zaba: a novel miniature transposable element present in genomes of legume plants. *Mol Genet Genomics* **269**, 624–631 (2003).
11. Yang, T. *et al.* Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat Genet* **54**, 1553–1563 (2022).
12. Sanches, M. *et al.* Grass pea (*Lathyrus sativus*) interesting panoply of mechanisms to cope with contrasting water stress conditions – a controlled study of sub populational differences in a worldwide collection of accessions. *Agricultural Water Management* **292**, 108664 (2024).
13. Jones, A. *et al.* High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. *PLOS ONE* **16**, e0253830 (2021).
14. Schalamun, M. *et al.* Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. *Molecular Ecology Resources* **19**, 77–89 (2019).
15. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
16. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. Preprint at https://doi.org/10.12688/f1000research.12232.1 (2017).
17. Laetsch, D. R., Koutsovoulos, G., Booth, T., Stajich, J. & Kumar, S. DRL/blobtools: BlobTools v1.0.1. *Zenodo* https://doi.org/10.5281/zenodo.845347 (2017).
18. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013).
19. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

20. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
21. Dudchenko, O. *et al. De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
22. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101 (2016).
23. Vondrak, T. *et al.* Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *The Plant Journal* **101**, 484–500 (2020).
24. Aliyeva-Schnorr, L., Ma, L. & Houben, A. A Fast Air-dry Dropping Chromosome Preparation Method Suitable for FISH in Plants. *J Vis Exp* e53470 https://doi.org/10.3791/53470 (2015).
25. Macas, J. *et al.* In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. *PLoS ONE* **10**, e0143424 (2015).
26. Macas, J., Neumann, P. & Navrátilová, A. Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. *BMC Genomics* **8**, 427 (2007).
27. Macas, J. *et al.* Assembly of the 81.6 Mb centromere of pea chromosome 6 elucidates the structure and evolution of metapolycentric chromosomes. *PLOS Genetics* **19**, e1010633 (2023).
28. Neumann, P. *et al.* Centromeres Off the Hook: Massive Changes in Centromere Size and Structure Following Duplication of CenH3 Gene in Fabeae Species. *Molecular Biology and Evolution* **32**, 1862–1879 (2015).
29. Neumann, P. *et al.* Epigenetic Histone Marks of Extended Meta-Polycentric Centromeres of Lathyrus and Pisum Chromosomes. *Frontiers in Plant Science* **7** (2016).
30. Macas, J. *et al.* Long read sequencing and centromere characterization of Fabeae species (2022).
31. Ávila Robledillo, L. *et al.* Extraordinary Sequence Diversity and Promiscuity of Centromeric Satellites in the Legume Tribe Fabeae. *Molecular Biology and Evolution* **37**, 2341–2356 (2020).
32. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
33. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
34. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
35. Stovner, E. B. & Sætrom, P. epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* **35**, 4392–4393 (2019).
36. Novak, P. kavonrtep/TideCluster: 0.0.8. *Zenodo* https://doi.org/10.5281/zenodo.7885626 (2023).
37. Gao, Y., Liu, B., Wang, Y. & Xing, Y. TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* **35**, i200–i207 (2019).
38. Novak, P. Domain based annotation of transposable elements - DANTE (2023).
39. Novak, P. kavonrtep/dante_ltr: 0.2.3.2. *Zenodo* https://doi.org/10.5281/zenodo.8183566 (2023).
40. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
41. Novák, P., Neumann, P. & Macas, J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat Protoc* **15**, 3745–3776 (2020).
42. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker (2013).
43. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
44. Novak, P. Various bioinformatics utilities (2023).
45. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
46. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**, lqaa108 (2021).
47. Santos, C., Polanco, C., Rubiales, D. & Vaz Patto, M. C. The MLO1 powdery mildew susceptibility gene in Lathyrus species: The power of high-density linkage maps in comparative mapping and synteny analysis. *The Plant Genome* **14**, 1–15 (2021).
48. Santos, C., Martins, D., Rubiales, D. & Vaz Patto, M. C. Partial Resistance Against Erysiphe pisi and E. trifolii Under Different Genetic Control in Lathyrus cicera: Outcomes from a Linkage Mapping Approach. *Plant Disease* **104**, 2875–2884 (2020).
49. Kreplak, J. *et al.* A reference genome for pea provides insight into legume genome evolution. *Nature Genetics* **51**, 1411–1422 (2019).
50. BioBam Bioinformatics. OmicsBox – Bioinformatics Made Easy (2019).
51. Bayer, M. *et al.* Comparative visualization of genetic and physical maps with Strudel. *Bioinformatics* **27**, 1307–1308 (2011).
52. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res* **51**, D445–D451 (2023).
53. Vigouroux, M. *et al.* PRJEB70892 - *Lathyrus sativus* LS007 HiFi genome sequencing, PacBio raw data. *European Nucleotide Archive* https://www.ebi.ac.uk/ena/browser/view/PRJEB70892, https://identifiers.org/ena.embl:ERP155791 (2024).
54. Vigouroux, M. *et al.* PRJEB70892 - *Lathyrus sativus* LS007 HiFi genome sequencing, scaffolded genome assembly. *NCBI* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_963859935.3/, https://identifiers.org/ncbi/insdc.gca:GCA_963859935.3 (2024).
55. Vigouroux, M. *et al.* Supporting files for research paper 'A chromosome-scale reference genome of *Lathyrus sativus*' https://doi.org/10.5281/zenodo.10671532 (2024).
56. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
57. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **49**, D344–D354 (2021).
58. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825–5829 (2021).
59. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).
60. Emmrich, P. M. F. *et al.* A draft genome of grass pea (*Lathyrus sativus*), a resilient diploid legume. 2020.04.24.058164 Preprint at https://doi.org/10.1101/2020.04.24.058164 (2020).

Acknowledgements

We would like to thank Laura Avila Robledillo for her advice on chromosome preparation and FISH experiments. We would also like to thank Noel Ellis for his advice and insight on validating the genome scaffolding through SNP mapping. J.M., P.N., A.K. and L.O. were supported by the Czech Science Foundation grant 24-10036S. CS was supported by Fundação para a Ciência e Tecnologia (FCT, Portugal) through 2017.00198.CEECIND, research contract by stimulus of scientific employment. PE is supported by the John Innes Foundation. MVig, RHMW, JC, MVic, BS and CM were supported by the Biotechnology and Biological Sciences Research Council Institute Strategic Programmes, 'Molecules from Nature' (BB/P012523/1) and 'Harnessing Biosynthesis for Sustainable Food and Health' (BB/X01097X/1), and a grant from the Templeton World Charity Foundation, Inc. (TWCF0400). Computational resources and data storage facilities were in part provided by the ELIXIR-CZ

Research Infrastructure Project (LM2023055). This research was also funded by Fundação para a Ciência e Tecnologia (FCT, Portugal) through the R&D Research Unit GREEN-IT—Bioresources for Sustainability (UIDB/04551/2020 DOI10.54499/UIDB/04551/2020 and UIDP/04551/2020 DOI 10. 54499/UIDP/04551/2020) and the LS4FUTURE Associated Laboratory (LA/P/0087/2020).

Author contributions

M.Vig. performed the assembly and HiC scaffolding, wrote the relevant methods sections, prepared the HiC contact map and wrote the relevant text. R.H.M.W. performed DNA extraction for HiFi and HiC sequencing, and wrote the relevant text, supported gene annotation and data uploads. JC - performed the gene annotation and data uploads, wrote relevant methods and technical validation sections. P.N. performed the repeat annotation and wrote the relevant methods section, contributed to Fig. 1. A.K. - performed ChIP-seq. L.C.O. - performed FISH and prepared relevant text and figure. P.P. - performed Blobtools analysis and prepared relevant figure and methods. M.Vic. - contributed to assembly and annotation strategy. C.S. - performed comparative mapping analysis, wrote relevant text and prepared relevant figures. M.C.V.P. - supervised comparative mapping, critically revised manuscript. J.M. - supervised repeat annotation, performed centromere mapping and prepared relevant text and figure, critically revised manuscript. B.S. - supervised assembly, scaffolding and gene annotation, critically revised manuscript. C.M. - initiated and designed the project, supported project management, critically revised manuscript. P.M.F.E. - general project supervision, wrote abstract, background and summary, prepared Fig. 1, coordinated the work and manuscript preparation. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03868-y.

Correspondence and requests for materials should be addressed to P.M.F.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.