Chromosome-scale assembly of the African yam bean genome

Waweru, Bernice

Cold Spring Harbor Laboratory

https://doi.org/10.1101/2023.10.31.564964

Chromosome-scale assembly of the African yam bean genome

Bernice Waweru[1,2*], Isaac Njaci[1,3*], Edwin Murungi[4], Rajneesh Paliwal[5], Collins Mulli[1], Mary Maranga[6,7], Davies Kaimenyi[8,9], Beatus Lyimo[10], Helen Nigussie[11], Bwihangane Birindwa Ahadi[12,13], Ermias Assefa[14], Hassan Ishag[15], Oluwaseyi Olomitutu[3], Michael Abberton[3], Christopher Darby[2], Cristobal Uauy[2], Nasser Yao[1], Daniel Adewale[16§], Peter Emmrich[2], Jean-Baka Domelevo Entfellner[1§], Oluwaseyi Shorinola[1,2,17§]

[1]International Livestock Research Institute, P.O. Box 30709, Nairobi 00100, Kenya.
[2]The John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK.
[3]School of Agriculture and Food Sustainability, The University of Queensland, Brisbane, QLD, Australia.
[4]Department of Medical Biochemistry, Kisii University, P.O. Box 408-40200, Kisii, Kenya.
[5]Genetic Resources Center, International Institute of Tropical Agriculture, Oyo Road, Ibadan 200001, Nigeria.
[6]Department of Biochemistry, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000, Nairobi 00200, Kenya.
[7]Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland.
[8]Bioscience Research Centre (PUBReC), Pwani University, P.O Box 195-80108, Kilifi, Kenya.
[9]Institut für Mikrobiologie und Biochemie, Hochschule Geisenheim University, Von-Lade-Str. 1, 65366 Geisenheim, Germany.
[10]Nelson Mandela African Institute of Science and Technology, Arusha, Tanzania.
[11]Department of Microbial Cellular and Molecular Biology, Addis Ababa University, Ethiopia.
[12]Université Evangélique en Afrique, UEA, Faculty of Agriculture and Environment sciences, Bukavu, Democratic Republic of the Congo.
[13]Université Officielle de Bukavu, UOB, Faculty of Sciences, Bukavu, Democratic Republic of the Congo.
[14]Genomics and Bioinformatics Research Directorate, Bio and Emerging Technology Institute, Addis Ababa, Ethiopia.
[15]College of Veterinary Sciences, University of Nyala, Nyala, Sudan.
[16]Department of Crop Science and Horticulture, Federal University Oye-Ekiti, Ikole-Ekiti Campus, Nigeria.
[17]School of Bioscience, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK.

[*]Joint first authors
[§]Joint corresponding authors.

Corresponding authors: Oluwaseyi Shorinola (o.shorinola@bham.ac.uk); Jean-Baka Domelevo Entfellner (J.DomelevoEntfellner@cgiar.org); Daniel Adewale(d.adewale@gmail.com)

Abstract

Genomics-informed breeding of locally adapted, nutritious, albeit underutilised African crops can help mitigate food and nutrition insecurity challenges in Africa, particularly against the backdrop of climate change. However, utilisation of modern crop improvement tools including genomic selection and genome editing for many African indigenous crops is hampered by the scarcity of genetic and genomic resources. Here we report on the assembly of the genome of African yam bean (*Sphenostylis stenocarpa)*, a tuberous legume crop that is indigenous to Africa. By combining long and short read sequencing with Hi-C scaffolding, we produced a chromosome-scale assembly with an N50 of 69.5 Mbp and totalling 649 Mbp in length (77 - 81% of the estimated genome size based on flow cytometry). Using transcriptome evidence from Nanopore RNA-Seq and homology evidence from related crops, we annotated 31,614 putative protein coding genes. We further show how this resource improves anchoring of markers, genome-wide association analysis and candidate gene analyses in Africa yam bean. This genome assembly provides a valuable resource for genetic research in Africa yam bean.

Background and Summary

African yam bean (*Sphenostylis stenocarpa* (Hochst. Ex. A. Rich) Harms) is an underutilised tuberous legume which produces edible protein-rich seeds and starch-rich tubers (Fig. 1). It is a tropical African crop[1] that originated from Ethiopia from where its distribution extended to West and Central Africa[2]. African yam bean (hereafter referred to as AYB) is important for food and nutritional security in local communities in sub-saharan Africa. AYB is a rich source of dietary protein with up to 30% and 10% protein content in the seeds and tubers, respectively[3,4]. Its seeds and tubers are also low in fat and rich in carbohydrates, minerals and vitamins[3,4]. In addition, AYB exhibits high nitrogen-fixing ability[5] and is drought tolerant. These attributes may have allowed it to thrive in marginal soils under low-input farming systems and intercropping, especially in Ghana and Nigeria[6,7].

AYB, however, is largely underutilised due to the hardness of the seed coat, leading to long cooking time and the presence of anti-nutritional factors which reduce protein digestibility[3]. Also, the need for staking of plants has greatly hampered its cultivation on a commercial scale. Its production has been sustained indigenously through intercropping with major crops, especially yam - *Dioscorea spp*. To date, minimal genomic information is available to assist breeding efforts aimed at unlocking the full potential of AYB, thereby limiting its contribution to food and nutritional security in Africa.

Here, we present the first chromosome-scale assembly of the AYB genome using Illumina short-read and Oxford Nanopore long-read sequencing platforms (Fig. 2). Using homology and transcript evidence, we performed a gene annotation of the

AYB's genome (Fig. 2). We further demonstrate the usefulness of this genome resource for genetics analyses and trait mapping in AYB.

Methods

Size estimation of AYB genome by flow cytometry

Fresh 10 mg leaf samples of AYB and soybean (*Glycine max*, used as standard) were immersed in 1 mL of ice-chilled Galbraith buffer (45 mM $MgCl_2$, 30 mM sodium citrate, 20 mM 3-(N-morpholino) propanesulfonic acid, 0.1% w/v Triton X-100, pH 7) and sliced using a scalpel. The supernatant was filtered through one layer of Miracloth (pore size 22 - 25 µm). An aliquot of 600 µL of filtrate were mixed with propidium iodide to a concentration of 50 µM and RNAse A to 20 µg/mL and incubated for 1.5 h on ice. A FACSCantoII flow cytometer (Becton Dickinson) was used to measure nuclei, with flow rate adjusted to 20 - 50 events/s and results were analysed using FCSalyser (v. 0.9.18 alpha). The genome size of *Sphenostylis stenocarpa* was estimated following the method described by Dolezel et al (2007)[8] by dividing the mean position of its fluorescence peak by the mean position of the corresponding soybean peak and multiplying by the estimated soybean genome size of 1.10 - 1.15 Gbp[9] (Fig. 3, Supplementary Table 1). Based on this range we estimate the size of the AYB genome as 804 - 841 Mbp.

Sample selection, library preparation and sequencing

Illumina DNA sequencing: Seeds of AYB accession TSs11[10] were germinated in a Petri dish on filter paper moistened with tap water. The sprouted seedlings were transferred to soil and allowed to grow in the greenhouse facility at the International Livestock Research Institute (ILRI, Kenya) for a month. DNA was extracted from young leaves using a DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol, recovering a total of 10 µg. DNA was quantified using a Qubit 2.0 Fluorometer and dsDNA BR Assay (Invitrogen, Paisley, United Kingdom) and integrity was confirmed by gel electrophoresis on a 0.8% agarose gel.

Three samples of 50 ng each of genomic DNA were sheared and processed using the Nextera DNA Library Prep Kit (Illumina, USA) according to the manufacturer's instructions. Three runs of paired-end (2 x 150 bp) sequencing were performed on an Illumina MiSeq (Illumina) to generate 25 Gbp of raw data, representing ~30x of the AYB genome.

Nanopore DNA sequencing: Two grams of young leaves of AYB accession TSs11 were harvested, frozen in liquid nitrogen and stored at -80 ^{0}C. The leaves were then ground in liquid nitrogen using a pestle and mortar. High Molecular Weight (HMW) DNA was extracted from the ground sample with Carlson lysis buffer (100 mM Tris-HCl, pH 9.5, 2% CTAB, 1.4 M NaCl, 1% PEG 8000, 20 mM EDTA) followed by purification using the Qiagen Genomic-tip 100/G as described on the Oxford Nanopore Technologies (ONT, UK) HMW plant DNA extraction protocol. The ONT SQK-LSK109 ligation sequencing kit protocol was used to prepare sequencing libraries from the HMW DNA. This involved repairing and 3' adenylation of 1 µg of

HMW genomic DNA with the NEBNext FFPE DNA Repair Mix and the NEBNext® Ultra™ II End Repair/dA-Tailing Modules (New Englang Biolab, NEB). Sequencing adapters were then ligated using the NEBNext Quick Ligation Module (NEB). After library purification with AMPure XP beads (Beckman Coulter), sequencing was conducted at ILRI using R9.4.1 flow cells on an ONT MinION sequencer. High-accuracy base calling was done using Guppy basecaller[11] (v4.1.1) generating 9.5 million reads totalling 42.4 Gbp of sequence that represents 50 - 53x of the estimated genome size (Fig. 2).

Nanopore RNA sequencing: Two grams of young and disease-free leaves, stem and root tissues of AYB accession TSs11 were harvested and ground with mortar and pestle in liquid nitrogen. HMW RNA was extracted with the following extraction buffer [100 mM Tris–HCl (pH 8.0), 25 mM EDTA, 2 M NaCl, 2% CTAB (w/v), 2% PVP (w/v) and 2% β-mercaptoethanol (v/v)], followed by removal of residual DNA using DNASE I (RNase-free) kit (Thermo Fisher Scientific). The library was prepared following the Oxford Nanopore SQK-PCS109 PCR-cDNA sequencing kit. A total of 50 ng total RNA was transcribed using Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific). Full length transcripts were selected by PCR amplification using the LongAmp Taq Master Mix and the product was purified with AMPure XP beads (Beckman Coulter). An aliquot of 1 µL of Rapid Adapter was added to the amplified cDNA library. The libraries were sequenced at ILRI using R9.4.1 flowcells on the ONT MinION sequencers. Real-time data acquisition and high accuracy base-calling were conducted using the MinKNOW software with the Guppy basecaller generating 7.1 Gbp of sequence from 17.1 million reads (Fig. 2).

De novo assembly

Genome assembly was done primarily with the ONT long reads generated above. Briefly, the reads were assembled using Flye *de novo* long read assembler (v2.9)[12] with default parameters generating 10,329 contigs with total assembly length of 701.6 Mbp (Fig. 2). The draft assembly was further polished for error correction with Illumina short reads generated above from the same AYB accession TSs11 using HyPo hybrid polisher[13] (v1.0.3) with parameters -s 700m -c 30 -p 96 and -t 64. The polished draft assembly had an N50 of 781,337 bp, and a total assembly of 701.3 Mbp (Table 1). The draft assembly was further scaffolded using Chromatin conformation capture data as described below.

Hi-C scaffolding

Chromatin conformation capture (Hi-C) scaffolding was performed by Phase Genomics (Seattle, USA) using the Proximo Hi-C 2.0 Kit. For this, fresh leaves from young AYB accession TSs11 plants were frozen in liquid nitrogen, ground to powder and cross-linked using formaldehyde solution before being sent to Phase Genomics for library preparation following the manufacturer's protocol. Sequencing of the Hi-C library was performed at Phase Genomics using Illumina HiSeq platform generating a total of 275,166,448 paired-end reads. Reads were aligned against the polished assembly using BWA-MEM[14] with options -5SP and -t 8 specified and the other parameters set to the defaults. PCR duplicates were then mapped using SAMBLASTER[15], which were later excluded from the analysis. Non-primary and secondary alignments were flagged and filtered with Samtools[16] using the -F 2304 filtering flag. Putative misjoined contigs were broken using Juicebox[17] based on the

Hi-C alignments, and the same alignment procedure was repeated from the beginning on the corrected assembly. Phase Genomics Proximo Hi-C genome scaffolding platform was used to create chromosome-scale scaffolds from the corrected assembly following the method similar to that described by Bickhart et al (2017)[18]. Ordering of the scaffolds into pseudomolecules was done by LACHESIS[19].

The scaffolded assembly contains 11 pseudomolecules (representing the 11 AYB chromosomes) with 649.8 Mbp of sequence, and N50 of 69.5 Mbp (Table 1). There were also 8,422 short contigs with total and average length of 51.8 Mbp and 6.1 Kbp, respectively, that were not anchored into chromosomes. Summary statistics and evaluation of the completeness of the chromosome-scale genome assembly was evaluated using QUality ASsessment Tool (QUAST)[20] (ver 5.0.2) and Benchmarking Universal Single-Copy Orthologs (BUSCO)[21] (v5.2.2), respectively (see Technical Validation section).

Table 1: AYB assembly statistics before and after scaffolding.

Assembly Metric	Polished Assembly	Scaffolded (Hi-C) Assembly
Number of contigs/scaffolds (> 500 bp)	10,329	11
Total assembly length (bp)	701,349,621	649,801,261
N50 (bp)	781,088	69,519,929
L50(bp)	194	4
N75 (bp)	292,358	47,252,539
L75(bp)	567	7
Longest contig/scaffold (bp)	9,386,731	107,191,003
Mapping back rate of ONT reads		97.82%

Synteny with closely related species genomes

We examined the syntenic relationship between the HiC-scaffolded genome of AYB and those of closely related species including common bean (*Phaseolus vulgaris)*[22] and lablab (*Lablab purpureus)*[23]. For this, long-read based genome assemblies and annotation dataset for *P. vulgaris* and *L. purpureus* were obtained from Ensembl plant[24] and e!DAL[25] respectively. The blastp option from BLAST v2.7.1 was used to compare the AYB protein to *P. vulgaris* and *L. purpureus* with parameters: -max_target_seqs 1 -evalue 1e-10 -qcov_hsp_perc 70. MCscanX[26] algorithm was subsequently used to identify collinear blocks between the AYB-phaseolus and AYB-lablab genome pairs with parameters: -s 20 and -m 10. Visualisation of synteny linkages was made by R[27] (v3.3.1) and circos[28] (v0.69-4). Six of AYB chromosomes show direct one-to-one syntenic relationships with lablab and common bean chromosomes, while the other five AYB chromosomes show syntenic relationship across two or more chromosomes in either lablab or common bean. Based on these syntenic relationships, we assigned chromosome names to the Hi-C-scaffolded AYB pseudomolecules.

Repeat annotation

The Extensive de novo TE Annotator[29] (EDTA v1.9.7) pipeline was used to annotate the transposable elements (TE) in the AYB genome. The pipeline incorporates different tools to annotate predominant TE classes found in plant genomes using structure and homology-based detection methods. The tools include LTRharvest[30], LTR_FINDER[31], LTR_retriever[32], TIR-Learner[33], HelitronScanner[34], RepeatModeler2[35] and RepeatMasker[36]. The outputs of each tool are combined and filtered into a comprehensive non-redundant TE library. The inbuilt genome annotation function in EDTA was then used to produce a final non-overlapping repeat annotation for the AYB genome. Data visualisation and summary were carried out in R[37] using the Tidyverse suite[38]. In total 624,517 TEs and 78,100 unclassified repeats accounting for 74.08% of the total assembly were identified (Table 2, Fig. 5 and Fig. 6).

Table 2: The number of TEs, TE families and the proportion of occupied assembly length by different classes of repeats identified and annotated in AYB.

Class	**Order**	**Superfamily**	**Number of TEs**	**Number of Families**	**% of Assembly**
Class I	LTR-RT	Copia	172471	3149	22.01
		Gypsy	183229	1488	24.13
		unknown	187360	1379	17.73
	LINE	unknown	351	6	0.04
Class II	TIR	CACTA	28514	494	2.4
		MUDR-Mutator	20256	264	1.23
		PIF-Harbinger	800	23	0.08
		Tc1-Mariner	675	25	0.03
		hAT	10466	116	0.59
	MITE	CACTA	39	32	0
		MUDR-Mutator	1255	174	0.03
		PIF-Harbinger	45	3	0
		Tc1-Mariner	2	2	0

		hAT	4496	22	0.21
	Helitron	Helitron	14242	27	1.44
Other	Pararetrovirus		316	1	0.01
	Unclassified repeat		78100	702	4.15
Total			**702617**	**7907**	**74.08**

Gene prediction and functional annotation of genome

We combined transcript and homology evidence to annotate the gene content of the AYB genome. The transcript evidence was generated from 17,117,377 ONT long-read RNA reads totalling 7.1 Gbp of sequencing data used for *de novo* assembly of 60,249 transcripts. Briefly, Minimap2[39] (v2.22) was used to index the AYB genome assembly and the RNA reads were mapped to the indexed assembly. Samtools[16] (v1.9) was used to sort mapped reads by coordinates that were used to assemble transcripts with Stringtie2[40] (v2.0.1). Transdecoder[41] (v2.0.1) was then used to identify candidate CDS regions and select transcripts with a minimum protein length of 100 amino acids.

We combined the *de novo* transcripts with protein homology evidence from four well-annotated plant genomes (*Arabidopsis thaliana* TAIR10, *Phaseolus vulgaris* v1.0, *Glycine max* v2.1, *Vigna_angularis* v1.1) together with a soft-masked (for repeats) AYB genome as inputs into Funannotate[42] (v1.8.11) to identify protein coding genes. Funannotate *'predict'* uses *ab initio* gene predictors Augustus[43], PASA[44], SNAP[45] and GlimmerHMM[46] together with protein sequences as evidence to predict genes. Gene predictions from all four *ab initio* predictors are passed to EVidenceModeler[47] with various weights for integration. This resulted in 30,840 coding gene models totalling 31,614 transcripts with a median exon length of 231 bp and a median of three exons per transcript. Additionally, we detected 774 non-overlapping tRNA-encoding genes using tRNAscan-SE[48] for tRNA prediction. The gene and transposable element distribution across the genome are inversely correlated (Fig. 5).

Protein domains were annotated using InterProScan-5.25-64.0[49] based on InterPro protein databases, including TIGRFAM, SUPERFAMILY, PANTHER, Pfam, PRINTS and ProDom. We also used eggNOG-mapper[50] (v2.1.7) to annotate predicted gene models. Funannotate *'annotate'* uses results of InterProScan and eggNOG-mapper to annotate putative functions of protein sequences using PFAM[51], UniProtKB[52] and Gene Ontologies[53] databases. In total, functional descriptions were assigned to 25,241 (81.85%) of the genes.

Gene family analysis

To delineate gene families, AYB was compared to other legumes *L. purpureus*, *P. vulgaris, Vigna unguiculata,* and *Mycrotyloma uniflorum* (with *Solanum tuberosum* as an outgroup) using OrthoFinder[54] (v2.5.4). This analysis placed 26,038 (84.4%) of the 30,840 AYB proteins into orthogroups. Clustering using Venn Diagrams[55] revealed 1,296 AYB proteins (4.2%) segregated in 384 species-specific orthogroups (Fig. 7).

Technical Validation

Genome and annotation completeness

We mapped the ONT long-reads data back to the genome assembly and analysed the alignment with Qualimap[56] (v.2.2.2). The alignment mapping rate was 97.8% (Table 1) with an average read coverage of 39x. We also evaluated the completeness of the genome assembly and annotation using BUSCO[21] (v5.2.2). A highly conserved set of single-copy orthologs from embryophta_odb10 and fabales lineages were used as references. For the genome assembly, we obtained complete matches to 98.0% and 98.5% of the conserved single-copy orthologs in the fabales and embryophyta lineages, respectively (Fig. 8). Similarly, 90.4% and 91.4% of the conserved single copy orthologs showed complete matches to the gene annotation of AYB (Fig. 8). These high percentages suggest a high degree of accuracy and completeness of the genome assembly and gene annotation.

Marker mapping and association

We also examined the usefulness of the AYB genome for positionally anchoring markers for genetic analyses. Previous efforts to anchor a set of Genotyping-By-Sequencing (GBS) markers generated from a collection of AYB accessions using the common bean genome as reference only mapped 15.48% of the markers to a unique syntenic position, thus limiting the number of markers used for genome-wide association analyses (GWAS)[57]. Using the chromosome-scale assembly of AYB as reference, we could anchor 92% of the total 5,142 DArTseq-SNPs markers to unique positions in the AYB genome. The distribution of the markers across the genome tallies with the gene distribution highlighting the gene-centric nature of the GBS pipeline (Fig. 5). Furthermore, we used a subset of 1,460 quality-filtered SNPs (Call Rate > 0.70, Marker repeatability > 0.95, MAF > 0.05, missing < 0.05) for evaluating how the chromosome-scale genome assembly support GWAS and candidate gene analyses. For this, we used Best Linear Unbiased Estimates (BLUE, combined years and locations) of seed yield traits (hundred seed weight - HSW; seed length -SL;

seed width - SW; and seed thickness - ST) from a landrace population of 195 AYB accessions which were phenotyped in 2018 and 2019 under optimal field condition in three different locations of IITA research farms in Nigeria (Ibadan, Kano and Ubiaja). More information about field trait analysis can be obtained from Olomitutu et al (2022)[57].

The Generalised Linear Model (GLM) and Mixed Linear Model (MLM) were used in TASSEL[58] (v5.2.87) for identifying marker-trait association for seed yield-traits. Significant SNP association with the traits were determined by adjusting p-value threshold (a = 0.10) for FDR procedure proposed by Benjamini and Hochberg (BH)[59]. Significant marker-trait associations were found for SW and ST traits, two highly correlated traits, on Ss07 and on unanchored contigs. SNP 29420736-57-G/T was associated with both SW and ST traits on Ss07 (4.78 Mbp) of the AYB genome, suggesting a possible pleiotropic effect. The unanchored SNP 29420736-57-G/T was associated with SW traits. The contribution of these associated SNPs to the phenotypic variation ranged between 8.38% to 11.19%.

Candidate genes were searched within 1 Mbp interval around the position of SNP 29420736-57-G/T on Ss07 (± 500 Mb, 4283198 bp to 5283198 bp) in the AYB genome. In total sixteen genes were identified (Supplementary Table 2). Out of 16 candidate genes, nine genes were involving in grain development process in which seven genes were related to seed development role (Spste.TSs11.07G209790.1[60], Spste.TSs11.07G209800.1[61], Spste.TSs11.07G209840.1[62], Spste.TSs11.07G209850.1[62], Spste.TSs11.07G209860.1[62], Spste.TSs11.07G209910.1[63], and Spste.TSs11.07G209920.1[63]), one gene for seed shape (Spste.TSs11.07G209820.1[64,65]), and two genes for seed size (Spste.TSs11.07G209790.1[60] and Spste.TSs11.07G209920.1[66]) in plants.

Similarly, Olomitutu et al (2022)[57] reported nine candidate genes in common bean by blasting SNPtag of SNP 29420736-57-G/T in legume information system database[67]. The encoding product of these common bean candidate genes were similarly involved in regulation of seed development[68], seed/fruit size[69], seed size[70–73], grain shape[64,65], and grain size[74,75]. The mechanism regulating seed traits in AYB needs further exploration. The SNP position and candidate genes information in the AYB genome provided in this study might help to improve AYB yield. These results also indicate that AYB genome will play a central role in precise mapping of SNP markers and genome-wide allele mining for agronomical, biotic, abiotic and nutrition value traits in future AYB crop breeding.

Data Records

The genome assembly and annotation have been deposited in the following repositories:

1. ENA/NCBI/DDBJ accessions: BioProject PRJEB57813 on ENA and on NCBI, chromosomes have accessions OY731398 to OY731408 and the whole **genome assembly is GCA_963425845.**
2. ENA only: ENA project ERP142818 is already published on INSDC as EDTAPRJEB57813, sample is ERS16321187 (completed) and annotated assembly is ERZ21776326 (completed).

Code Availability

Open-source software were used for the analyses reported. The software versions and custom parameters used (if different from default) are indicated in the Methods.

Acknowledgements

This work was supported by the UK Biotechnology and Biological Sciences Research Council through the GCRF Strategic Training Awards for Research Skills (STARS) program grants BB/R020272/1 (ABCF Bioinformatics Community of Practice (BiX COP)) and BB/T017422/1 (Building capacity in third-generation genomics and bioinformatics for agricultural biosciences in Africa). OS was also supported by the Royal Society FLAIR Fellowship (FLR\R1\191850).

Authors Contribution

OS, JDE, BW, IN, DA, PE, EM, CU, CD, NY and MA designed the experiments. DA, OS, PE, IN, CM, MA, and OO performed the plant sampling. PE, OS, JDE and CM performed genome size estimation, CM, OS, PE, HN, BW, EA, JDE extracted DNA and RNA for sequencing. CM, OS, PE, BW, IN conducted the DNA and RNA sequencing (Illumina and Nanopore), IN, BW, MM, DK, OS, JBE, PE, EM, BBA, HN, BL, HI and EA produced the genome and transcript assemblies. IN and PE performed the repeat annotation. BW, IN, OS, EM and JDE performed the gene annotation. EM, BL, BW, OS, EA did the gene family analyses. OS, PE and HN conducted the synteny analyses. RP, OS, MA, OO and DA did the marker and

GWAS analyses. BW, IN, RP, EM, MM, DK, BL, HN, BBA, EA, HI, CU, CD, DA, PE, JDE and OS wrote the manuscript.

Competing Interests

The authors declare that there is no conflict of interest.

References

1. Potter, D. & Doyle, J. J. Origins of the African Yam bean (Sphenostylis stenocarpa, leguminosae): evidence from morphology, isozymes, chloroplast DNA, and linguistics. *Econ. Bot.* **46**, 276–292 (1992).
2. Adesoye, A. I. & Nnadi, N. C. Mitotic chromosome studies of some accessions of African yam bean Sphenostylis stenocarpa (Hochst. Ex. A. Rich.) Harm. *Afr. J. Plant Sci.* **5**, 835–841 (2011).
3. Adewale, B. D. & Nnamani, C. V. Introduction to food, feed, and health wealth in African yam bean, a locked-in African indigenous tuberous legume. *Front. Sustain. Food Syst.* **6**, (2022).
4. George, T. T., Obilana, A. O. & Oyeyinka, S. A. The prospects of African yam bean: past and future importance. *Heliyon* **6**, (2020).
5. Assefa, F. & Kleiner, D. Nodulation of African yam bean (Sphenostylis stenocarpa) by Bradyrhizobium sp. isolated from Erythrina brucei. *Biol. Fertil. Soils* **25**, 209–210 (1997).
6. Klu, G. Y. P., Amoatey, H. M., Bansa, D. & Kumaga, F. K. Cultivation and use of African yam bean Sphenostylis stenocarpa in the Volta Region of Ghana. *J. Food Technol. Afr.* **6**, 74–77 (2001).
7. Daniel, A. B. Chapter 18 - African yam bean (Sphenostylis stenocarpa hochst ex. A. Rich) Harms). in *Neglected and Underutilized Crops* (eds. Farooq, M. & Siddique, K. H. M.) 487–514 (Academic Press, 2023). doi:10.1016/B978-0-323-90537-4.00030-2.
8. Doležel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using

flow cytometry. *Nat. Protoc.* **2**, 2233–2244 (2007).

9. Arumuganathan, K. & Earle, E. D. Estimation of nuclear DNA content of plants by flow cytometry. *Plant Mol. Biol. Report.* **9**, 229–241 (1991).

10. Africa Yam Bean TSs-11 Accession Passport data. https://my.iita.org/accession2/accession/TSs-11.

11. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).

12. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

13. Kundu, R., Casey, J. & Sung, W.-K. HyPo: Super Fast &amp; Accurate Polisher for Long Read Genome Assemblies. *bioRxiv* 2019.12.19.882506 (2019) doi:10.1101/2019.12.19.882506.

14. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

15. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

16. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

17. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).

18. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).

19. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).

20. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

21. 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).

22. Schmutz, J. *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014).

23. Njaci, I. *et al.* Chromosome-level genome assembly and population genomic resource to accelerate orphan crop lablab breeding. *Nat. Commun.* **14**, 1915 (2023).

24. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

25. Arend, D. *et al.* e!DAL - a framework to store, share and publish research data. *BMC Bioinformatics* **15**, 214 (2014).

26. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).

27. The R Project for Statistical Computing. https://www.r-project.org/.

28. Krzywinski, M. I. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* (2009) doi:10.1101/gr.092759.109.

29. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18 (2019).

30. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 1–14 (2008).

31. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 1–3 (2019).

32. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

33. Su, W., Gu, X. & Peterson, T. TIR-learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol. Plant* **12**, 447–460 (2019).

34. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci.*

111, 10263–10268 (2014).

35. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).

36. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **25**, 4.10.1-4.10.14 (2009).

37. R: The R Project for Statistical Computing. https://www.r-project.org/.

38. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

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

40. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).

41. Haas, BJ. TransDecoder: Finding coding regions within transcripts. https://github.com/TransDecoder/TransDecoder/wiki#transdecoder-find-coding-regions-within-transcripts.

42. Palmer, J. M. & Stajich, J. Funannotate v1. 8.1: Eukaryotic genome annotation. *Httpsdoi Org105281zenodo* **4054262**, (2020).

43. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).

44. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

45. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 1–9 (2004).

46. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

47. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).

48. Chan, P. P. & Lowe, T. M. tRNAscan-SE: searching for tRNA genes in genomic

sequences. in *Gene prediction* 1–14 (2019).

49. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

50. 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).

51. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

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

53. Consortium, G. O. The Gene Ontology (GO) database and informatics resource. in (2003).

54. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

55. Calculate and draw custom Venn diagrams. *Van de Peer Lab* http://bioinformatics.psb.ugent.be/webtools/Venn/.

56. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

57. Olomitutu, O. E. *et al.* Genome-Wide Association Study Revealed SNP Alleles Associated with Seed Size Traits in African Yam Bean (Sphenostylis stenocarpa (Hochst ex. A. Rich.) Harms). *Genes* **13**, (2022).

58. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).

59. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

60. Guo, C., Zhou, J. & Li, D. New Insights Into Functions of IQ67-Domain Proteins.

Front. Plant Sci. **11**, (2021).

61. Puentes-Romero, A. C., González, S. A., González-Villanueva, E., Figueroa, C. R. & Ruiz-Lara, S. AtZAT4, a C2H2-Type Zinc Finger Transcription Factor from Arabidopsis thaliana, Is Involved in Pollen and Seed Development. *Plants* **11**, (2022).

62. Vigeolas, H. *et al.* Combined Metabolomic and Genetic Approaches Reveal a Link between the Polyamine Pathway and Albumin 2 in Developing Pea Seeds. *Plant Physiol.* **146**, 74–82 (2008).

63. Tian, S. *et al.* Genome wide analysis of kinesin gene family in Citrullus lanatus reveals an essential role in early fruit development. *BMC Plant Biol.* **21**, 210 (2021).

64. Hu, Z. *et al.* A Kelch Motif-Containing Serine/Threonine Protein Phosphatase Determines the Large Grain QTL Trait in Rice. *J. Integr. Plant Biol.* **54**, 979–990 (2012).

65. Wang, Q. *et al.* Dissecting the Genetic Basis of Grain Size and Weight in Barley (Hordeum vulgare L.) by QTL and Comparative Genetic Analyses. *Front. Plant Sci.* **10**, (2019).

66. Siou-Luan He & Shin-Lon Ho. Functional Characterization of a WD-Repeat Protein Gene (OsWD1) in Rice. *Am. J. Agric. For.* **6**, 18–27 (2018).

67. Priyam, A. *et al.* Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol. Biol. Evol.* **36**, 2922–2924 (2019).

68. Kang, Y.-J. *et al.* Fine mapping and candidate gene analysis of the quantitative trait locus gw8.1 associated with grain length in rice. *Genes Genomics* **40**, 389–397 (2018).

69. Ma, Y., Yang, C., He, Y., Tian, Z. & Li, J. Rice OVATE family protein 6 regulates plant development and confers resistance to drought and cold stresses. *J. Exp. Bot.* **68**, 4885–4898 (2017).

70. Li, N. & Li, Y. Ubiquitin-mediated control of seed size in plants. *Front. Plant Sci.* **5**, (2014).

71. Li, N., Xu, R. & Li, Y. Molecular Networks of Seed Size Control in Plants. *Annu. Rev. Plant Biol.* **70**, 435–463 (2019).

72. Guo, Y. *et al.* Quantitative Trait Loci for Seed Size Variation in Cucurbits – A Review.

Front. Plant Sci. **11**, (2020).

73. Zhong, J. *et al.* A putative AGO protein, OsAGO17, positively regulates grain size and grain weight through OsmiR397b in rice. *Plant Biotechnol. J.* **18**, 916–928 (2020).

74. Zhang, W. *et al.* Fine mapping of GS2, a dominant gene for big grain rice. *Crop J.* **1**, 160–165 (2013).

75. Watt, C., Zhou, G. & Li, C. Harnessing Transcription Factors as Potential Tools to Enhance Grain Size Under Stressful Abiotic Conditions in Cereal Crops. *Front. Plant Sci.* **11**, (2020).

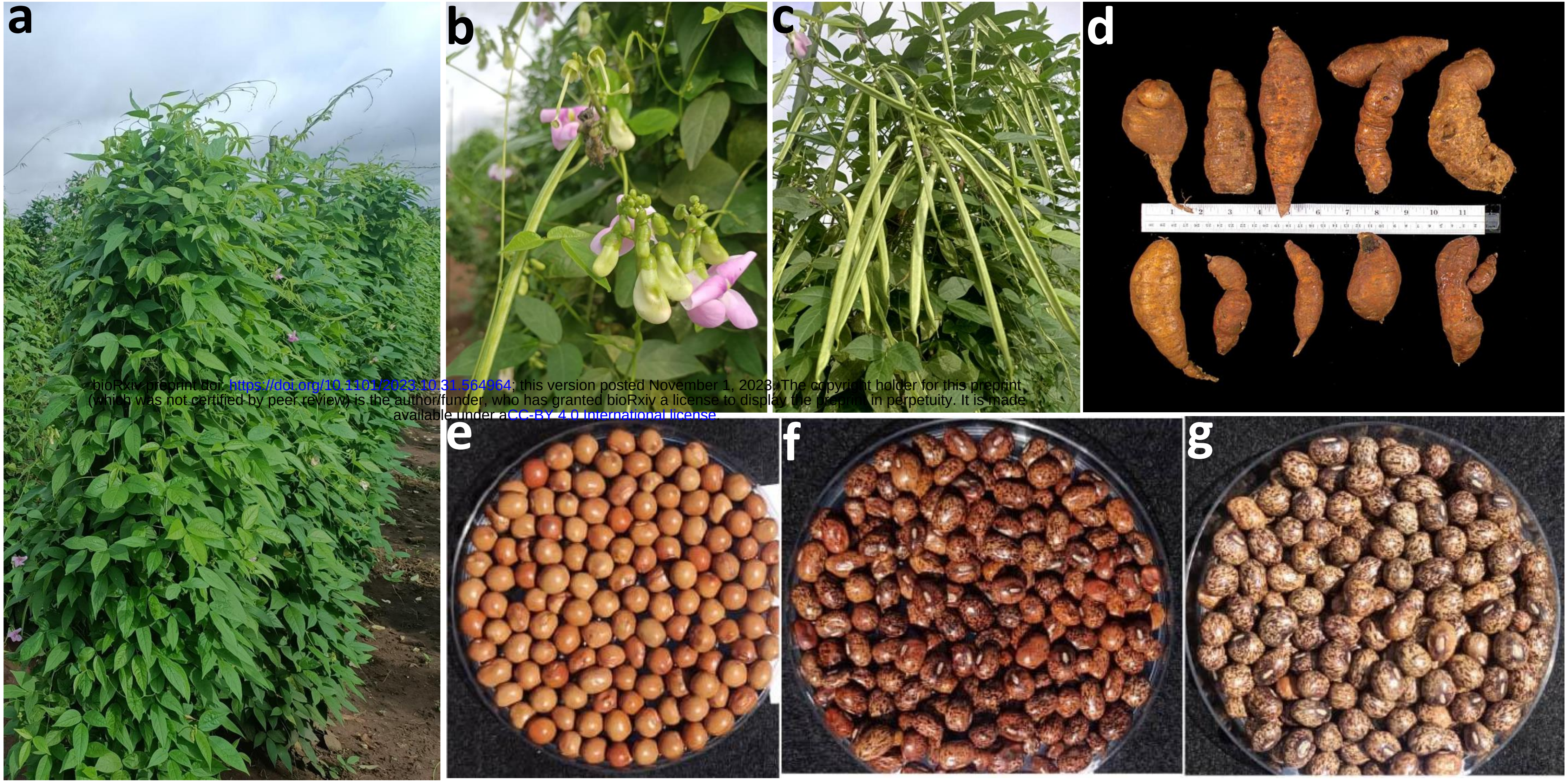

Fig. 1: Africa yam bean - an African indigenous tuberous legume. Figure shows (a) full grown plants in the field (b) flowers (c) pods (d) root tubers of different shapes and sizes (e - g) different coloured seeds.

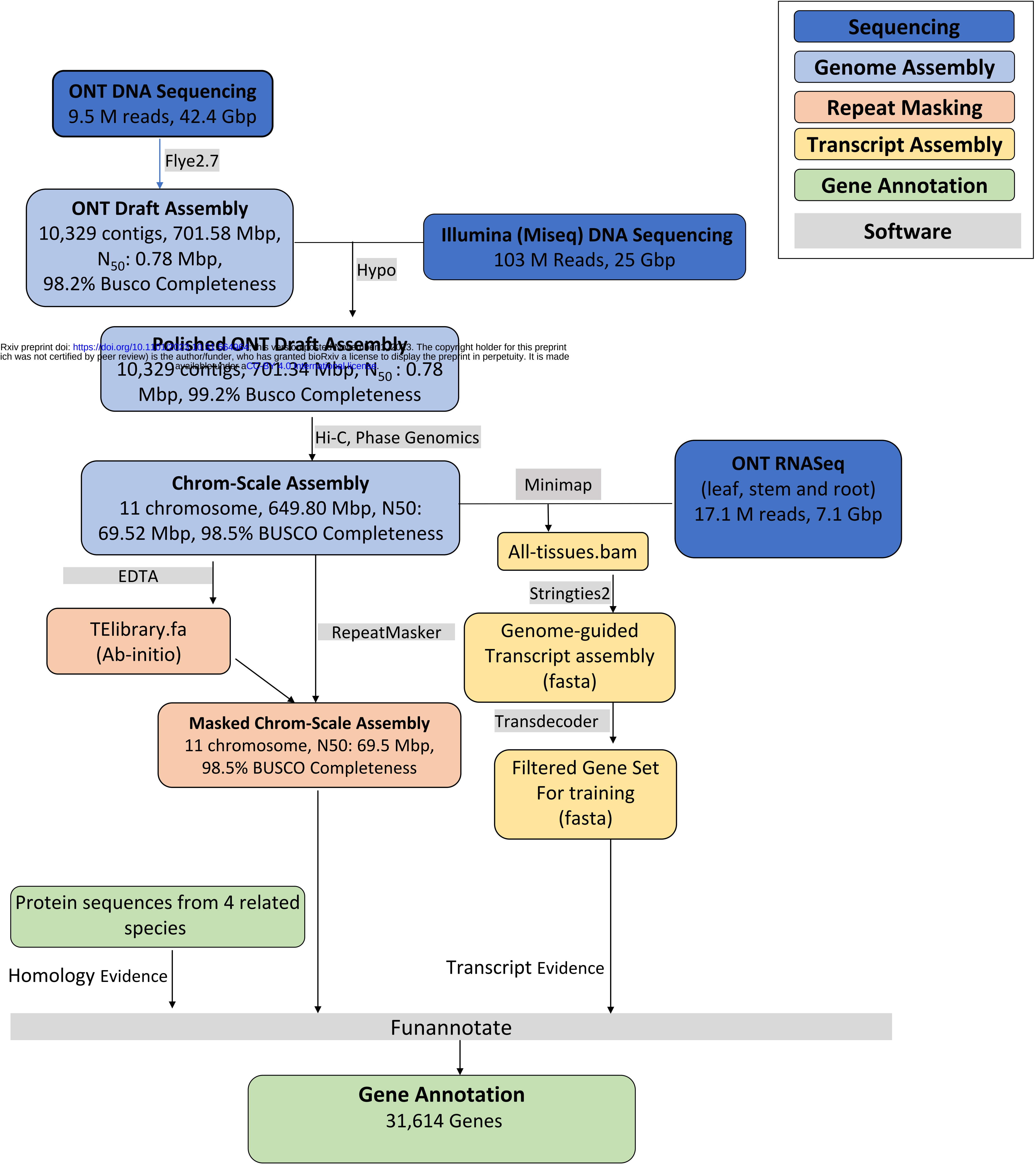

Fig. 2: AYB genome sequencing and annotation workflow. Overview of the workflow used for the sequencing, assembly, masking and annotation of the African Yam Bean genome. The boxes are color-coded by the different stages involved in the workflow (top right box) . Software used for each step are indicated in grey boxes..

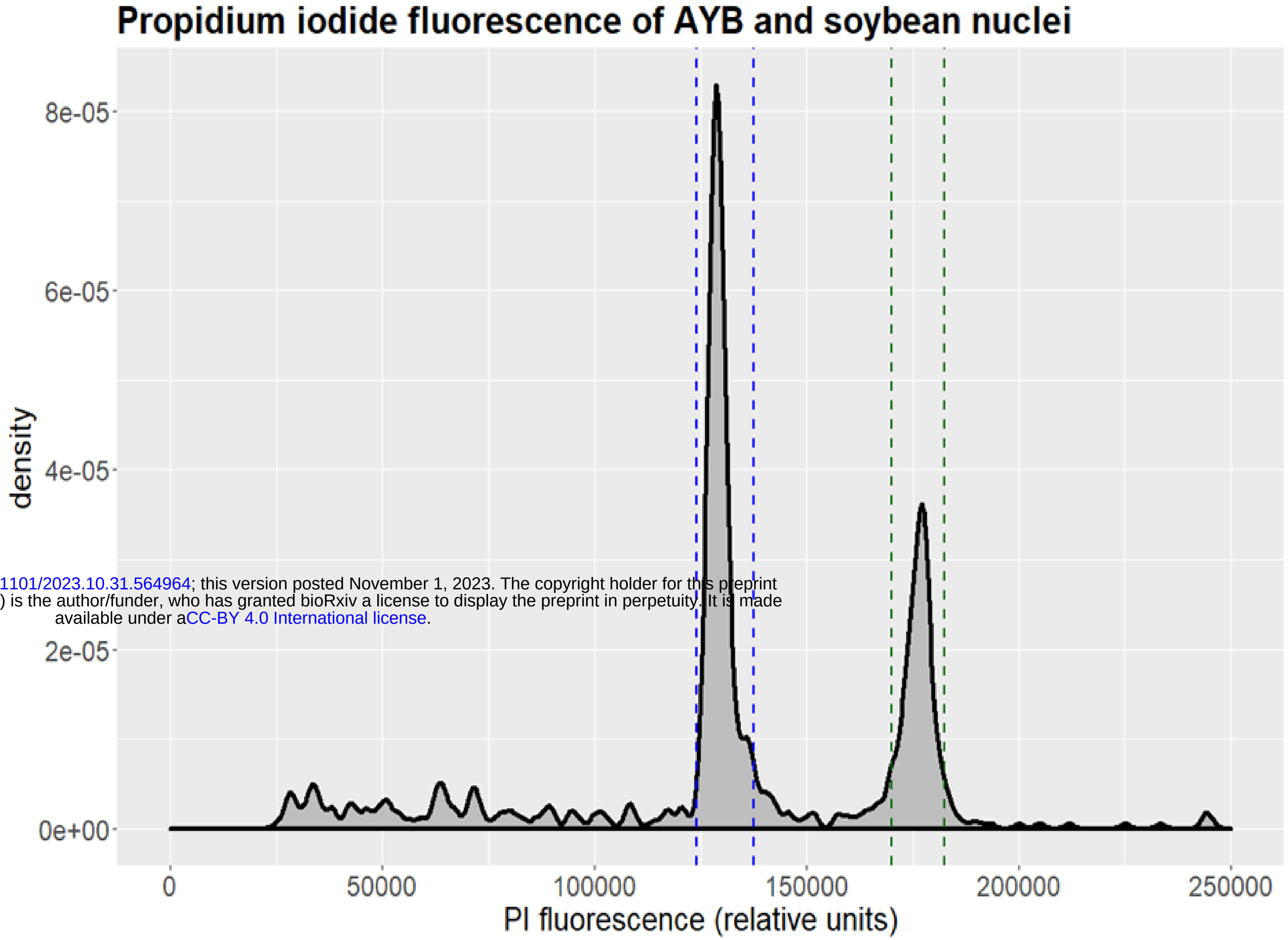

Fig. 3. Genome size estimation of AYB. Density plot showing results of a representative flow cytometry run, excluding events caused by cell debris. Propidium iodide fluorescence amplitude (in relative units) is plotted against event density. The interval assumed to be AYB nuclei is delimited by blue dashed lines, the interval assumed to be soybean nuclei is delimited by green dashed lines. Three biological replicates were performed.

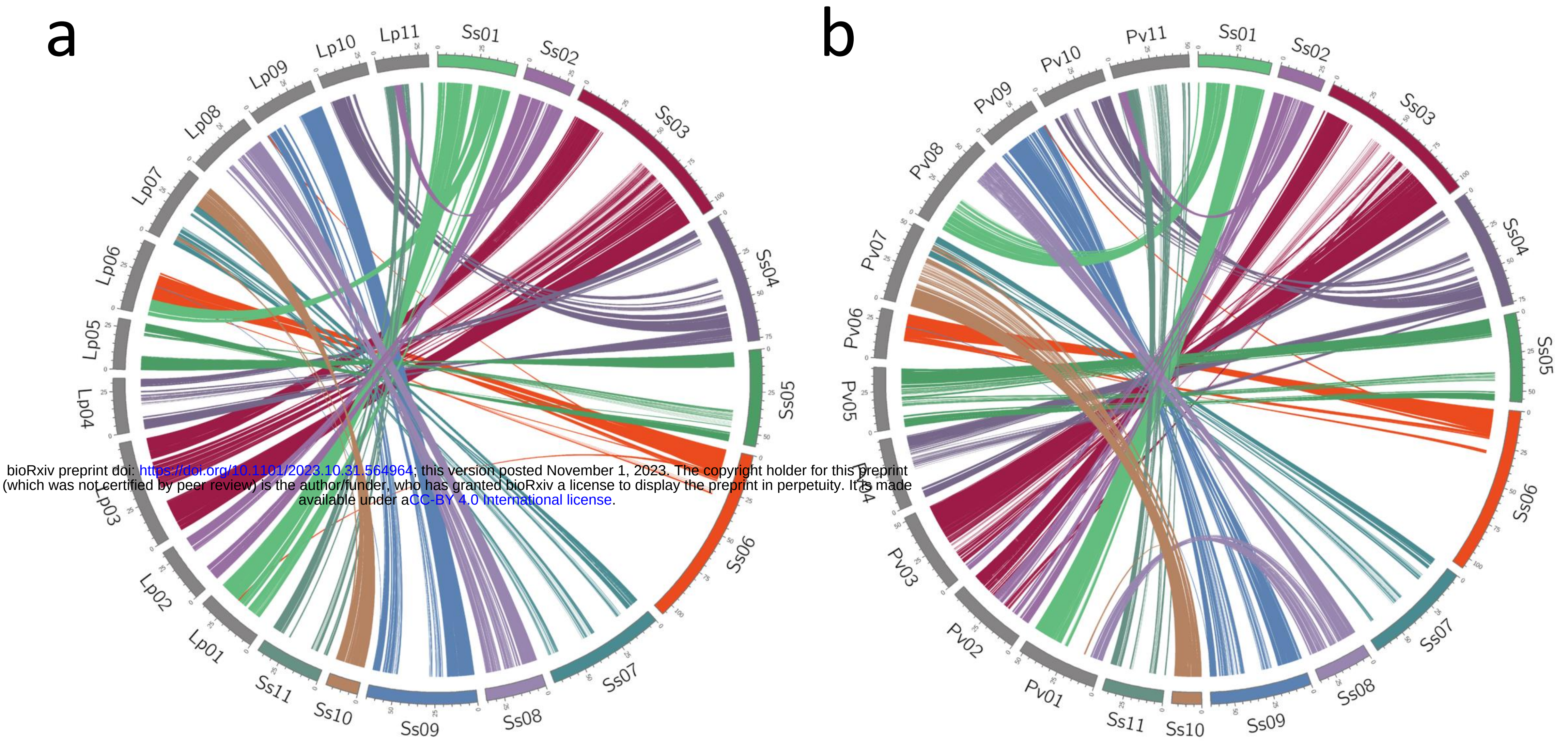

Fig. 4: AYB synteny to related legumes. Syntenic relationships between AYB chromosomes to the genomes of (a) lablab (*Lablab purpureus*) and (b) common bean (*Phaseolus vulgaris*). Where possible, AYB chromosome were renamed to reflect these syntenic relationships.

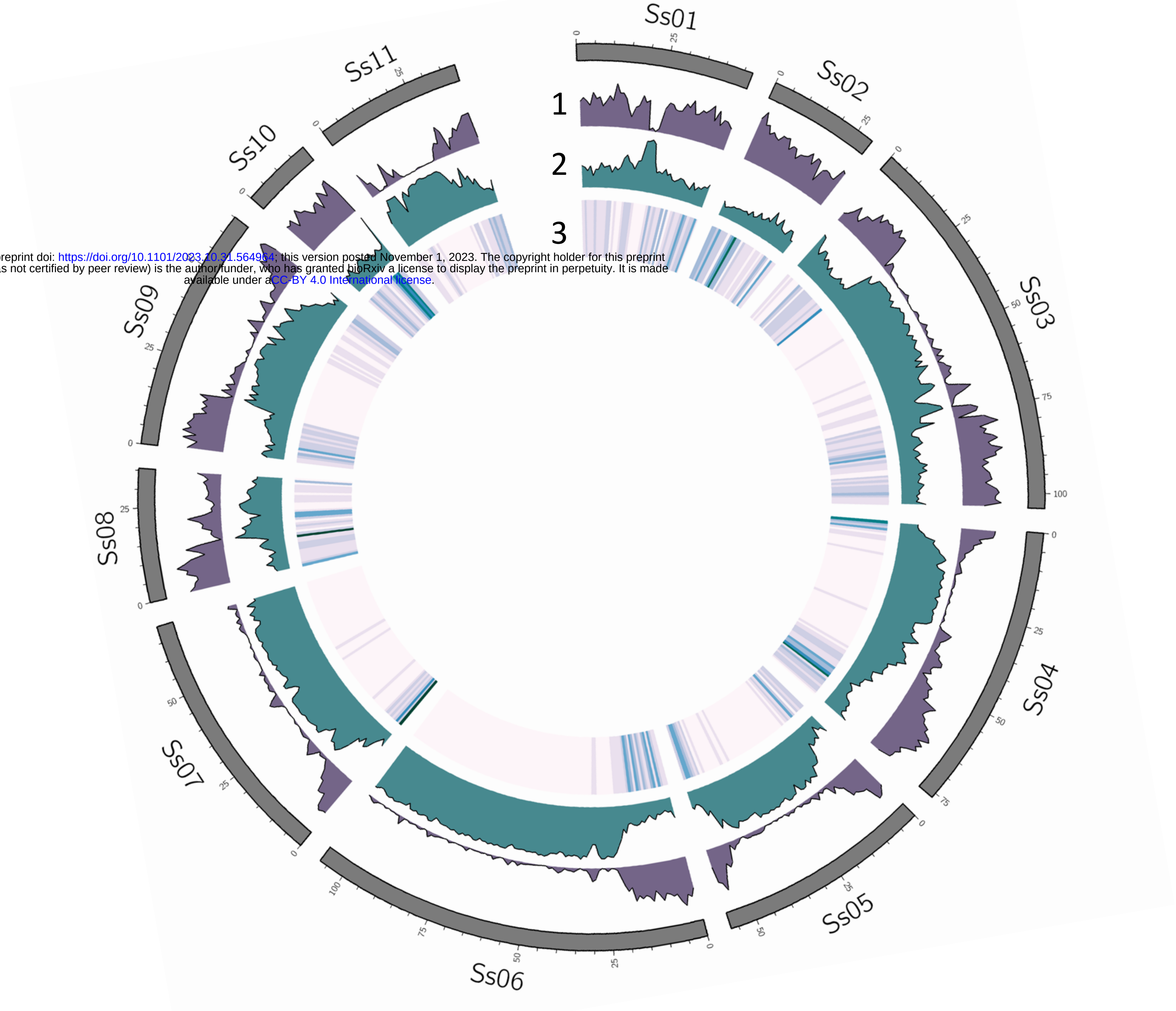

Fig. 5: Gene, repeat and marker distribution in the AYB genome. The outer to the inner track show 1) gene density, 2) repeat density, 3) heatmap of marker distribution.

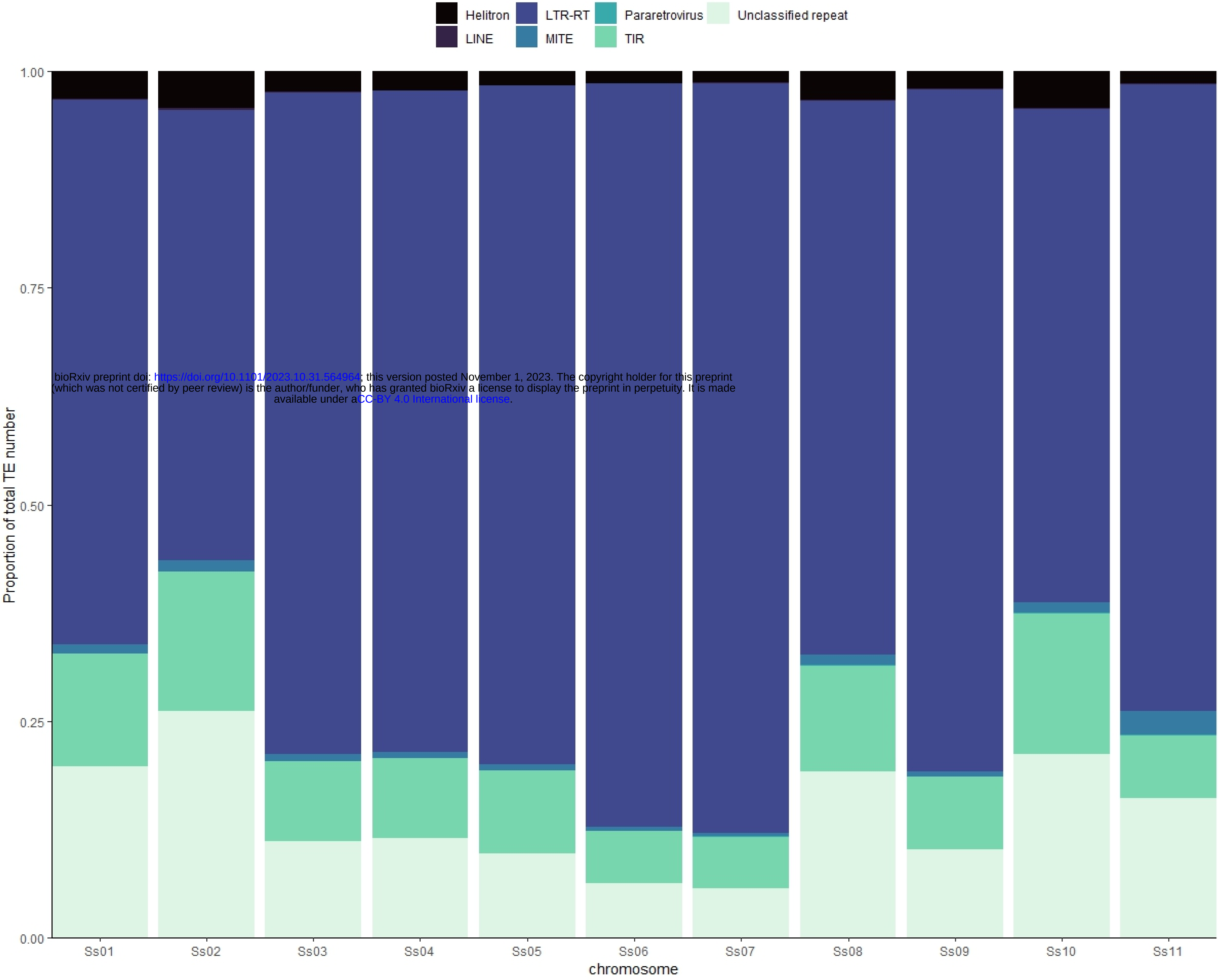

Fig. 6: Distribution of transposable elements across the AYB genome. Chromosomal repeats content in the AYB genome showing proportional abundance of identified transposable element Orders on each chromosome.

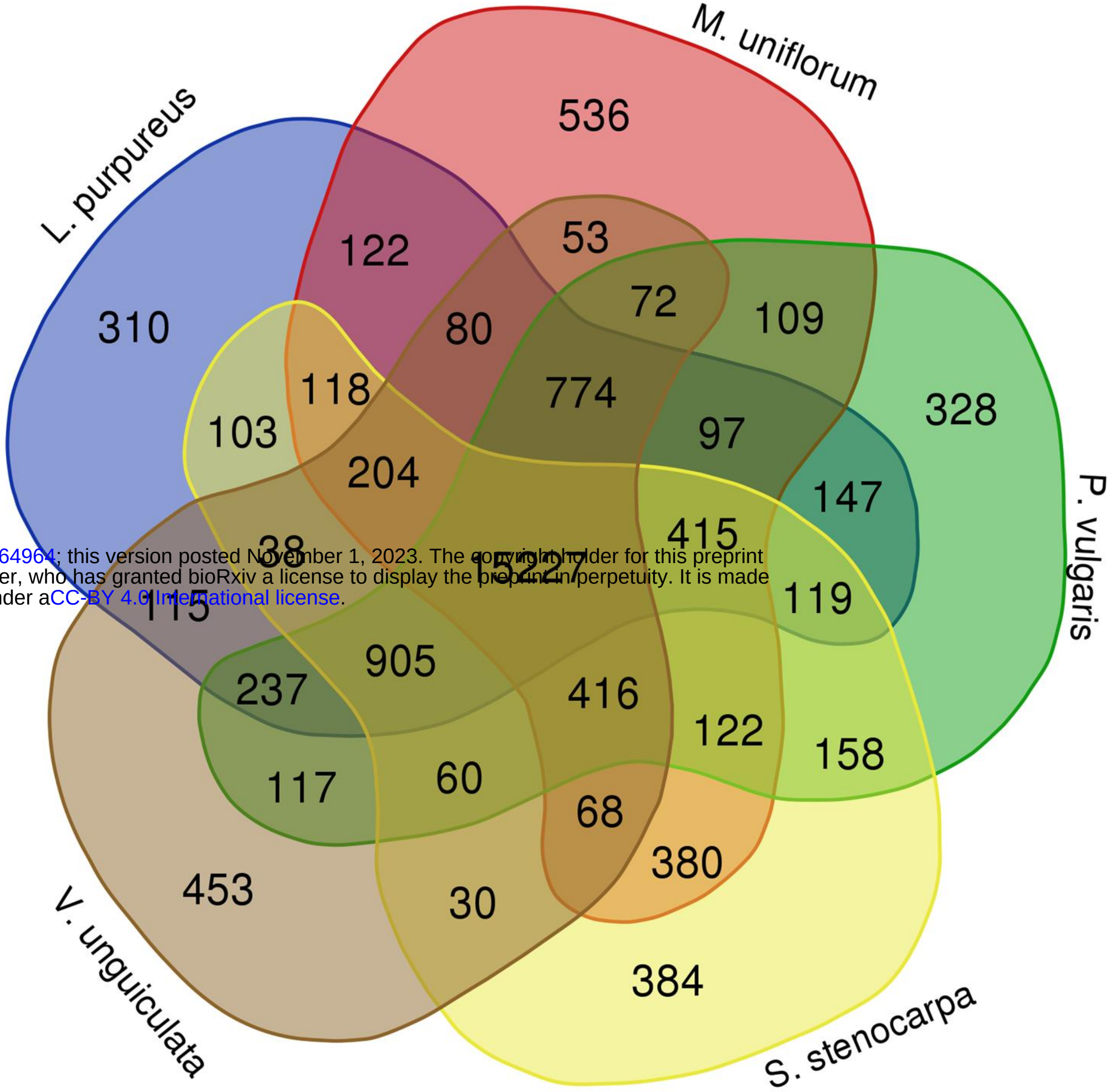

Fig. 7: Gene families in *AYB.* Venn diagram of the number of gene families common among and unique to AYB (S. stenocarpa), *Lablab purpureus*, *Phaseolus vulgaris*, *Vigna unguiculata* and *Macrotyloma uniflorum*.

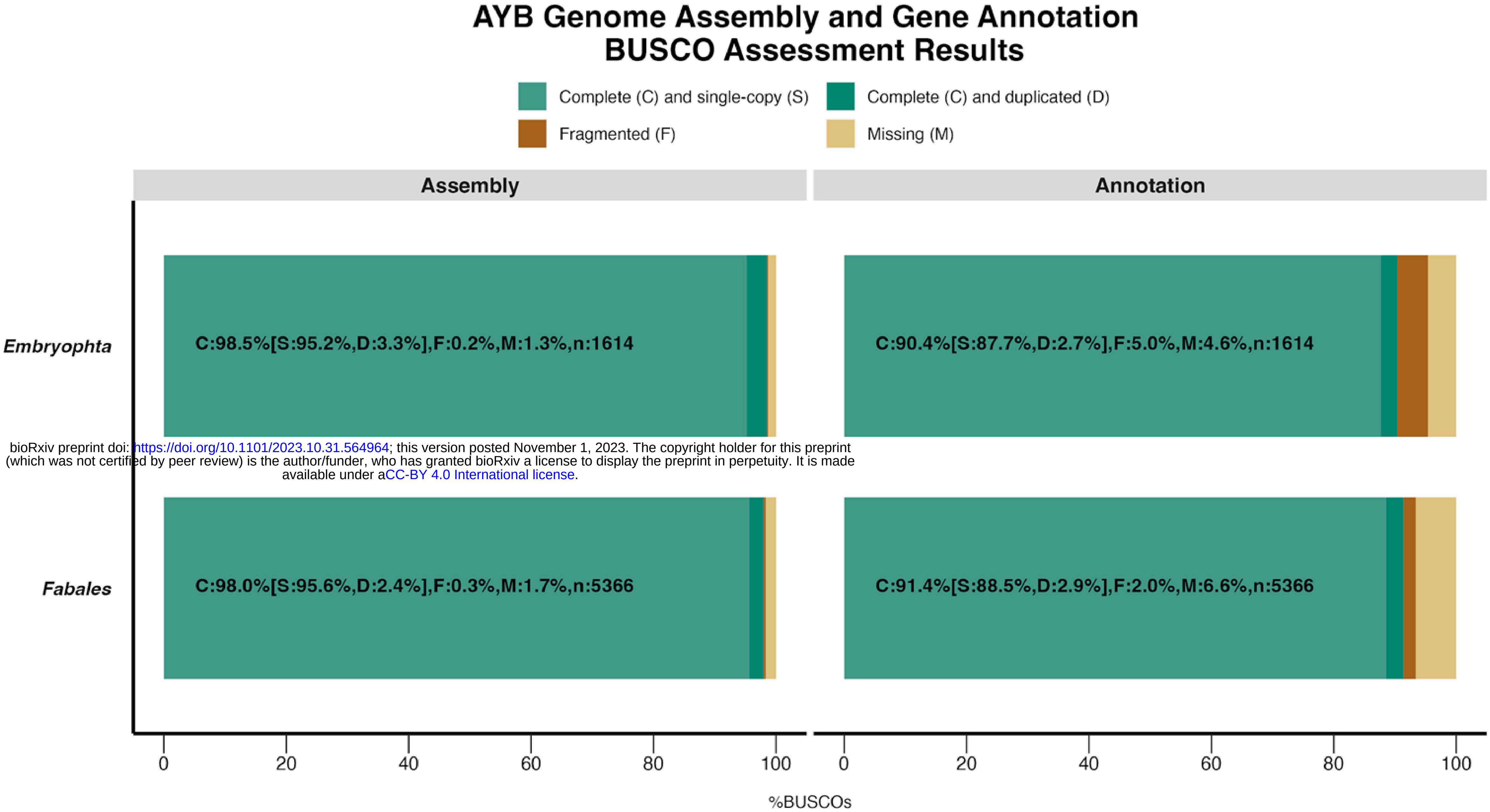

Fig. 8: AYB Genome assembly completeness: BUSCO scores of the AYB genome and gene annotation using the embryophyta and fabales reference lineages.

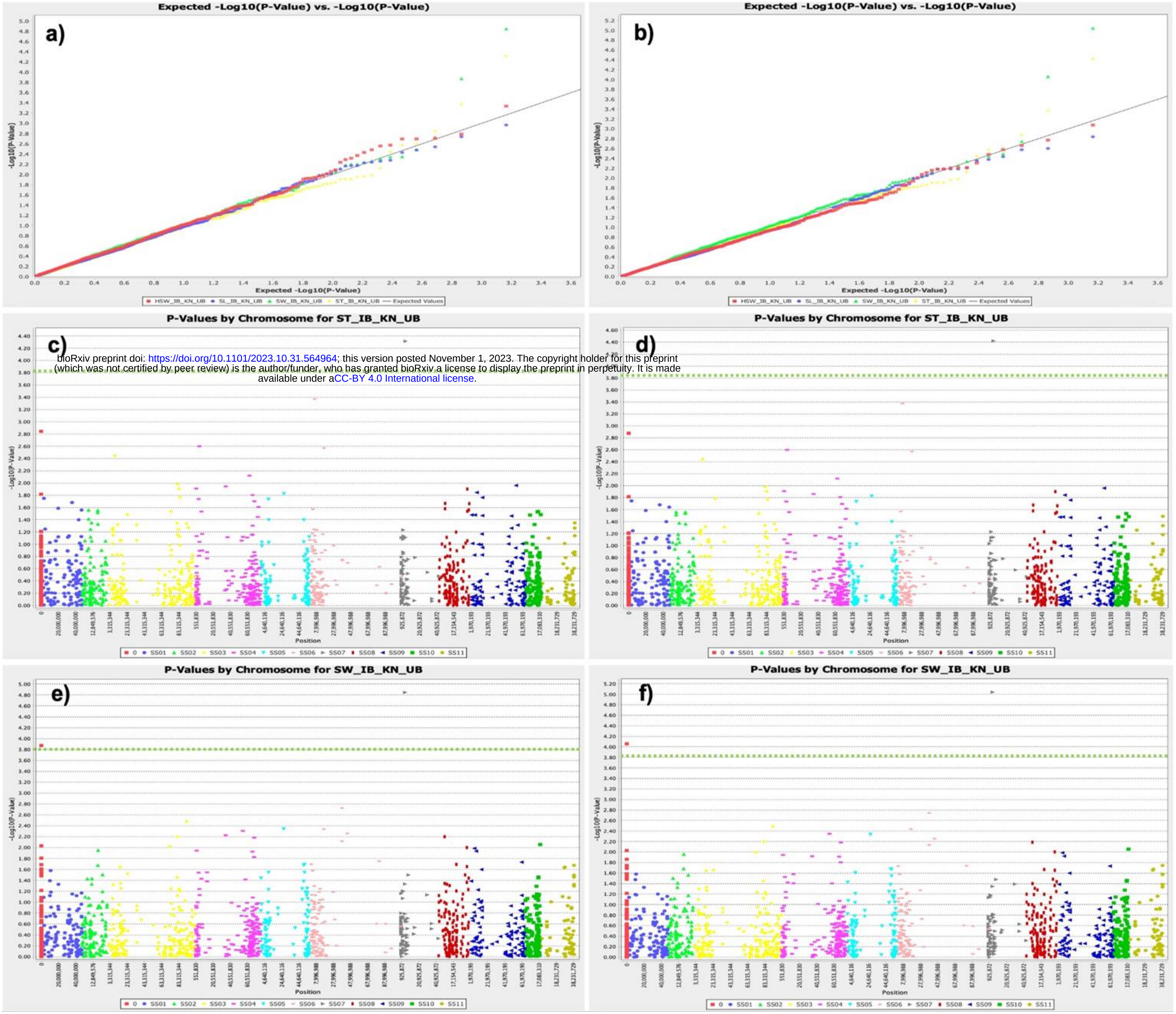

Fig. 9: GWAS in AYB. Manhattan and quantile-quantile plot of GWAS analysis for seed thickness (ST) and seed width (SW) in African yam bean collection. (a & b) Q-Q plot for GWAS results for ST and SW traits using GLM and MLM statistical models respectively. (c & d) Manhattan plot for GWAS results of ST trait using GLM and MLM statistical models respectively. (e & f) Manhattan plot for GWAS results of SW trait using GLM and MLM statistical models respectively. The green horizontal dotted lines indicate the significant threshold for associated SNPs in GWAS analysis.