Planta (2024) 260:124
https://doi.org/10.1007/s00425-024-04551-7

ORIGINAL ARTICLE

Chromosome-level genome assembly and functional annotation of *Citrullus colocynthis*: unlocking genetic resources for drought-resilient crop development

Anestis Gkanogiannis[1] · Hifzur Rahman[1] · Rakesh Kumar Singh[1] · Augusto Becerra Lopez-Lavalle[1]

Received: 1 April 2024 / Accepted: 11 October 2024 / Published online: 23 October 2024

Abstract

Main conclusion **The chromosome-level genome assembly of *Citrullus colocynthis* reveals its genetic potential for enhancing drought tolerance, paving the way for innovative crop improvement strategies.**

Abstract This study presents the first comprehensive genome assembly and annotation of *Citrullus colocynthis*, a drought-tolerant wild close relative of cultivated watermelon, highlighting its potential for enhancing agricultural resilience to climate change. The study achieved a chromosome-level assembly using advanced sequencing technologies, including PacBio HiFi and Hi-C, revealing a genome size of approximately 366 Mb with low heterozygosity and substantial repetitive content. Our analysis identified 23,327 gene models, that could encode stress response mechanisms for species' adaptation to arid environments. Comparative genomics with closely related species illuminated the evolutionary dynamics within the Cucurbitaceae family. In addition, resequencing of 27 accessions from the United Arab Emirates (UAE) identified genetic diversity, suggesting a foundation for future breeding programs. This genomic resource opens new avenues for the de novo domestication of *C. colocynthis*, offering a blueprint for developing crops with enhanced drought tolerance, disease resistance, and nutritional profiles, crucial for sustaining future food security in the face of escalating climate challenges.

Keywords Desert plant · Gene annotation · Wild crop relatives · Drought tolerance · Comparative genomics · Genetic variation

Introduction

Climate change, coupled with burgeoning human demands, has prompted scientists to intensify efforts in harnessing genetic diversity and conserving Crop Wild Relatives (CWRs) (Bohra et al. 2022). The use of CWR to develop resilience in domesticated/related crops is seen as part of a viable solution assuming that lacking-genetic-traits can be transferred to domesticated crops resulting in better crop varieties. They have been exploited as a valuable source of 'game-changing' alleles related to adaptive traits to counter abiotic and biotic stresses resulting from climate change, and to improve yield and other desired traits (Bohra et al. 2022; Renzi et al. 2022). The economic value of CWRs in augmenting agricultural productivity is substantial, contributing an estimated $186.3 billion to the global economy in 2020 (Pimentel et al. 1997; Tyack et al. 2020; Bohra et al. 2022).

These wild relatives offer novel genetic diversity that is not available in cultivated varieties and breeders often dream for such traits in the cultivated species. The major question is how to mine alleles underpinning the dream traits from gene pool. Advances in genomics and genome assisted breeding have accelerated the identification of valuable alleles for use in crop improvement. However, such information could be utilized in breeding domain only if CWR's genomic constellation is unveiled using advanced sequencing approaches.

Genome sequencing of major crop species is already completed; hence, major attention has been shifted to

Communicated by Stefan de Folter.

Anestis Gkanogiannis and Hifzur Rahman have contributed equally to this work.

✉ Anestis Gkanogiannis
a.gkanogiannis@biosaline.org.ae

[1] International Center for Biosaline Agriculture, ICBA, P.O. Box 14660, Dubai, United Arab Emirates

pan-genome and wild relatives' genomes to widen the gene pool, and to expand the breeding opportunities. In view of this, *Citrullus colocynthis* (L.) Schrad ($2n = 2x = 22$), a close relative of domesticated watermelon, which has not been sequenced so far, has been targeted as candidate for genome sequencing to unveil its genetic secrets for its possible use in breeding programs. The Cucurbitaceae family, encompassing economically significant species such as watermelon (*Citrullus lanatus*), melon (*Cucumis melo*), cucumber (*Cucumis sativus*), and various Cucurbita species exemplifies the diversity and potential of CWRs in agriculture. Among these, *Citrullus colocynthis* (L.) Schrad stands out as particularly noteworthy (Chomicki and Renner 2015). Native to the arid regions of the Sahara and Arabian deserts in Africa, Asia and the Mediterranean, *C. colocynthis* has evolved sophisticated mechanisms to conserve water under extreme drought conditions, attributable in part to its deep-root system (Si et al. 2010; Wang et al. 2014). Being resilient, *Citrullus colocynthis* has been explored as rootstock for cultivated watermelon to improve the stress tolerance capability of cultivated watermelon. *Citrullus colocynthis* has been used as rootstock for controlling diseases caused by *Fusarium* species viz., wilt, fusarium crown and root rot in cultivated watermelons (Borgi et al. 2009). Furthermore, drought tolerance and fruit quality of watermelon have been improved by grafting them onto the rootstock of colocynth (Bigdelo et al. 2017; Bikdeloo et al. 2021).

Beyond its environmental resilience, *C. colocynthis* is valued for its medicinal properties and as a source of nutritional oil, boasting bioactive compounds (polyphenols, glycosides, triterpenes, and cucurbitacin) beneficial for a range of ailments, including diabetes and cancer (Dane et al. 2007; Hussain et al. 2014). The seeds are particularly rich in linoleic and oleic acids, highlighting their potential for nutritional and therapeutic applications (Sawaya et al. 1983). *Citrullus colocynthis* seed contains high protein content with higher concentrations of essential amino acids like tryptophan, arginine, methionine, etc. (Sawaya et al. 1986; Council 2006; Ogundele et al. 2012; Al-Snafi 2016; Mariod et al. 2022). The seeds contain up to 25% oil, comprising 70% unsaturated fatty acids and 51% polysaturated fatty acids, and are considered a good source of essential fatty acids (Berwal et al. 2022). By leveraging the genetic material of CWRs like *C. colocynthis*, there exists a pathway to characterize, understand, and enhance the genetic basis of stress tolerance, disease resistance, nutritional and medicinal factors, and other agronomically important traits in cultivated species.

Despite its significant potential, genomic research on *C. colocynthis* has not kept pace. Challenges in assembling its genome arise from its monoecious nature and high percentage of repeated sequences. However, recent advancements in long-read sequencing technologies, such as PacBio Single-Molecule Real-Time (SMRT) sequencing and High-throughput Chromosome Conformation Capture (Hi-C), offer promising avenues for overcoming these obstacles.

To improve the agricultural resilience and sustainability of watermelon along with other cucurbits, it is imperative to sequence the genome of *Citrullus colocynthis*. Through the process of deciphering its genome, scientists can identify the genetic pathways that underpin the winning characteristics to enhance water use efficiency of watermelon. Therefore, in this study, we present a comprehensive chromosome-level whole-genome and transcriptome sequencing of *Citrullus colocynthis*, aiming to elucidate its genome structure and contribute valuable insights into plant adaptation, evolution, and the genetic basis of its remarkable drought tolerance and medicinal properties. Through this endeavor, we seek to unlock the genetic secrets of *C. colocynthis*, paving the way for its utilization in developing the next generation of climate-resilient crops.

Materials and methods

Plant material

Pure seeds of *Citrullus colocynthis* (accession RMS257) were obtained from the International Center for Biosaline Agriculture gene bank. The seeds were grown under greenhouse conditions at the International Center for Biosaline Agriculture. For PacBio long reads, DNBSEQ short reads and Hi-C sequencing, newly emerged fresh leaves were collected from single plants in 50 ml tubes and flash-frozen in liquid nitrogen. For RNA sequencing, the leaf, flower, young fruit, and root tissues were separately collected, rinsed in deionized water followed by pat drying using paper towel and then flash-frozen in liquid nitrogen. All tissues were stored in − 80 ℃ until further use.

High-molecular-weight DNA extraction and PacBio sequencing

The frozen leaves were grinded using liquid nitrogen in pestle and mortar for high-molecular-weight genomic DNA extraction. About 250–300 mg ground powder was used for DNA isolation using Genomic-tips Kit 100/G (QIAGEN, Germantown, MD, USA) as per manufacturer's protocol. After extraction, the concentration, integrity, and purity of the DNA were determined using Qubit® fluorometer (Thermo Fisher Scientific, Wilmington, DE, USA), agarose gel (0.8%), and NanoDrop (Thermo Fisher Scientific, Wilmington, DE, USA) respectively.

The PacBio SMRTbell libraries were constructed and sequenced by Nucleome (Hyderabad, India) according to the manufacturer's instructions (Pacific Biosciences, Melon Park, CA). In brief, PacBio HiFi Sequencing libraries were constructed following the manufacturer's protocol using a SMRTbell Express Template Prep Kit 2.0. Five micrograms of high-molecular-weight genomic DNA was fragmented to ≈ 20 kb targeted size using Diagenode's Megaruptures 3 system (Diagenode, Denville, USA). The fragmented DNA was end repaired and then ligated to hairpin adapters. Following digestion of incompletely formed SMRTbell templates were removed with Exonuclease III and VII, and the prepared libraries were purified using AMPure PB beads (Agencourt Bioscience, Beverly, MA). DNA molecules between 17 and 90 kb were selected by BluePippin electrophoresis (Sage Science, Beverly, MA). A Sequel II Binding Kit 2.0 (Pacific Biosciences, Melon Park, CA) was used to anneal the sequencing primer and bind the polymerase to the SMRTbell templates. About 80 pM of the library was loaded onto one SMRTcell and sequenced in PacBio Sequel II system in CCS/HiFi mode.

Hi-C library preparation

Chromatin conformation capture data were generated by Nucleome (Hyderabad, India) using the Proximo Hi-C kit. Following the manufacturer's instructions, fresh leaves were frozen in liquid nitrogen, ground, crosslinked using a formaldehyde solution. After the completion of the sample of crosslinking and cell lysis, the samples were digested with *DpnII*, *DdeI*, *HinfI*, and *MseI*. A Hi-C library was further constructed as described by Lieberman-Aiden et al. (2009), using a Proximo Hi-C Plant Kit (Phase Genomics, Seattle, WA). The resulting library was sequenced by Illumina Novaseq 6000 system (Illumina, San Diego, CA).

Transcriptome sequencing

Total RNA for leaf, flower, fruit, and root samples were isolated using Trizol reagent (Ambion, Carlsbad, CA, USA) following manufacturer's recommendations and isolated total RNA was purified using Nucleospin RNA cleanup kit (MACHEREY–NAGEL, Germany). Purified RNA quantity was measured using Qubit 3.0 fluorometer (Thermofisher Scientific, Massachusetts, USA) using RNA HS assay kit (Thermofisher Scientific, Massachusetts, USA). RNA purity was checked using NanoDrop 2000 (Thermo Fisher Scientific, Massachusetts, USA). The integrity of RNA was evaluated on 1% agarose gel and an Agilent 2100 Bioanalyzer (Agilent Technologies, California, USA). The RNA with RIN ≥ 7 was subjected to cDNA synthesis and amplification using the NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module (New England Biolabs Inc., Massachusetts, USA) in conjunction with Iso-Seq Express Oligo Kit (Pacific Biosciences, Melon Park, CA). The Pronex beads (Promega, Wisconsin, USA) were used to purify the cDNA before amplification and later for size selection of the amplified product. The library was constructed using the SMRTbell Express template Preparation Kit 2.0 (Pacific Biosciences, Melon Park, CA) per manufacturer's protocol. The library was purified using Pronex beads (Promega, Wisconsin, USA) and the library size was assessed using Bioanalyzer (Agilent Technologies, California, USA). About 80 pM of the library was loaded onto One SMRTcell containing 8 M ZMW and sequenced in PacBio Sequel II system in CCS/HiFi mode.

Short read sequencing

DNA was extracted from flash-frozen leaf samples using a modified CTAB method (Porebski et al. 1997). The purity and quality of the isolated DNA were verified by agarose gel electrophoresis on 0.8% agarose gels and the concentration was determined on a Qubit 4 fluorometer using Qubit Broad range Assay Kit (Thermo Fisher Scientific Inc., Waltham, MA, United States). One microgram of genomic DNA was mechanically fragmented to an average size of 250 bp using the Covaris® M220 Focused-ultrasonicator™ (Covaris, Woburn, Massachusetts), and the size selection of fragmented DNA was done using MGIEasy DNAClean beads (MGI Tech; Shenzhen, China). A single-stranded circular DNA library was prepared using MGIEasy Universal DNA library Prep Set Ver.1.0 following the manufacturer's standard protocol for a 250 bp insert size, followed by DNA nano ball (DNB) formation based on the rolling circle amplification. The DNB was loaded into the flow cell (DNBSEQ-G400RS Sequencing Flow Cell Ver.3.0, MGI Tech; Shenzhen, China), and cPAS-based 100-bp paired-end sequencing was performed with DNBSEQ-G400RS High-throughput Sequencing Set Ver.3.1 (MGI Tech; Shenzhen, China).

Sequencing data pre-processing

There were four types of sequencing data: DNBSEQ paired short reads (DNA), PacBio HiFi long reads (DNA), Hi-C paired short reads (DNA), and PacBio IsoSeq long reads (RNA). Raw sequencing data were initially pre-processed to clean from adapter contamination and low-quality reads. After cleaning and pre-processing, around 90 billion bases (Gb) of DNBSEQ, around 30 billion bases (Gb) of PacBio HiFi, around 43 billion bases (Gb) of Illumina Hi-C and around 6 billion bases (Gb) of PacBio IsoSeq were available for further analysis. Table 1 summarizes the sequencing data available. Software used for data pre-processing was fastp ver. 0.23.2 (Chen et al. 2018), LongQC ver. 1.2.0c (Fukasawa et al. 2020), and qc3C ver. 0.5 (DeMaere and Darling 2021). All raw genomic and transcriptomic sequencing data have been deposited in the European Nucleotide Archive (ENA). Accession numbers and links, accessible through browsers of any member of the International Nucleotide

Table 1 Summary of the sequencing data used in the study, including the platform, read lengths, number of reads, total bases, and genome coverage

Technology	Mean read length	Reads	Bases	Coverage
DNBSEQ	2×100 bp	897.8 M	89.8 G	257×
PacBio HiFi	15263 bp (N50 19464 bp)	1.9 M	29.7 G	85×
Illumina Hi-C	2×150 bp	291.3 M	43.2 G	123×
PacBio IsoSeq	1819 bp (N50 1931 bp)	3.5 M	6.4 G	18×

The data were generated using PacBio HiFi for long-read sequencing, DNBSeq for short-read sequencing, and Illumina Hi-C for chromatin conformation capture

Sequence Database Collaboration (INSDC), are shown in Table 2.

Genome size and ploidy

K-mer ($k = 21$) spectra from DNBSEQ short reads (Fig. S1) and PacBio HiFi long reads (Fig. S2) were used to estimate genome size and ploidy. For ploidy estimation, smudgeplot ver. 0.2.5 (Ranallo-Benavidez et al. 2020) was used, whereas for k-mer counting KMC ver. 3.1.1 (Kokot et al. 2017) and for genome size estimation GenomeScope ver. 2.0 (Ranallo-Benavidez et al. 2020).

GenomeScope analyzes the k-mer frequency distribution of sequencing reads to estimate genome size and heterozygosity. It works by analyzing the number of unique k-mers (short sequences of length k) in the sequencing data and grouping them into distinct frequency bins. The frequency distribution of these k-mers can then be used to estimate the genome size and the level of heterozygosity in the sample. Smudgeplot, on the other hand, extracts heterozygous k-mer pairs from k-mer count databases and performs estimations based on the patterns in the k-mer histogram.

Contigs level assembly

For the assembling of raw cleaned sequencing data into contigs, hifiasm ver. 0.19.0-r534 (Cheng et al. 2021) was used, with default parameters and with the mode that uses both HiFi long reads for contiging and Hi-C short reads for haplotype phasing. Hifiasm assembles long-read sequencing data (such as PacBio HiFi) into high-quality genome sequences. It uses an overlap-layout-consensus approach to build a graph representation of the input reads, which is

Table 2 Accession numbers and related project identifiers associated with the sequencing and annotation data of *C. colocynthis*

Project identifiers	
Species	*Citrullus colocynthis*
NCBI taxonomy ID	252529
NCBI locus tag	CITCOLO1
tolID	ddCitColo1
Project accessions	PRJEB78362 (umbrella project) PRJEB78297 (genomic and transcriptomic data) PRJEB66071 (genome assembly and annotation)
BioSample accessions	SAMEA114392512 (sample for genome assembly) SAMEA115375937—SAMEA115375963 (for diversity)
Assembly/annotation accession	GCA_963978565.1 (assembly) ERZ23582334 (analysis)
Chromosomes accessions	OZ021735—OZ021745
Unplaced contigs accessions	CAXBTG010000001—CAXBTG010000074
Raw data runs accessions	
PacBio HiFi	ERR12083385
DNBSEQ	ERR12083395 (for genome assembly) ERR12726176—ERR12726202 (for diversity)
Illumina Hi-C	ERR12083391
PacBio IsoSeq	ERR12098375 (flower) ERR12098376 (fruit) ERR12098377 (leaf) ERR12098378 (root) ERR12083390 (all tissues)

These include the NCBI taxonomy ID, BioProject accessions, BioSample accessions, accession numbers for raw data, assembled and annotated genome. The data have been deposited in the European Nucleotide Archive (ENA) and is accessible through any member of the International Nucleotide Sequence Database Collaboration (INSDC) under the specified accession numbers. Hyperlinks are provided for the main accession numbers then used to generate a consensus sequence. It first detects overlaps between the long-read sequences to construct a graph representation of the reads. Then it simplifies the graph by removing spurious edges and nodes, which can reduce the complexity of the assembly graph and improve its accuracy. Finally, it generates contigs, or contiguous sequences, from the graph by traversing paths that represent potential sequences in the final assembly and using Hi-C data produces two fully phased haplotype assemblies.

Since it was found a small degree of heterozygosity during genome characteristics estimation, default mode of purging with hifiasm was used, that performs three rounds of dups purging and correction. Contig assemblies were evaluated with QUAST ver. 5.2.0 in long mode (Gurevich et al. 2013) that calculates various assembly statistics, assembly completeness, and quality measures with BUSCO ver. 3.0.2 (Seppey et al. 2019), Merqury ver. 1.3 (Rhie et al. 2020), and Meryl ver. 1.4. Lineage dataset eudicotyledons_odb10 (Creation date: 2020-09-10, number of species: 31, number of BUSCOs: 2326) was used for BUSCO search.

BUSCO (Benchmarking Universal Single-Copy Orthologs) assesses the completeness of genome assemblies or gene sets by searching for evolutionarily conserved genes expected to be present in a wide range of eukaryotic genomes. It uses a set of evolutionarily conserved genes, called orthologs, as a benchmarking dataset. These orthologs are present as single copies in most eukaryotic genomes and are highly conserved across a wide range of taxa. It then searches for these orthologs in the genome assembly or gene set being tested using a set of HMM profiles corresponding to each ortholog and scores the genome assembly or gene set based on the number of complete, fragmented, duplicated and missing orthologs. A high percentage of complete and single-copy orthologs indicates a high level of completeness in the assembly or gene set.

Merqury evaluates the completeness of assembly from available whole-genome short and long sequencing reads without needing a reference. It generates k-mer frequency spectra for the reads and compares them with spectra from the assemblies.

Contigs contamination

BlobToolKit ver. 4.1.2 (Challis et al. 2020) was used to check for contaminants in the contigs assembly. PacBio HiFi reads are aligned to contigs assembly with minimap2 ver. 2.28 (Li 2018, 2021). Then the NCBI nt database (Coordinators 2014) is used to search for contigs hits with blastn ver. 2.2.31 (Altschul et al. 1990). Contig coverage and blast hits are combined with BUSCO stats to generate plots of taxa abundances in the contigs assemblies.

Scaffolds level assembly

Contig level assemblies can further be improved at the scaffold or chromosome level. With proximity information from chromosome conformation capture technologies, contigs can be arranged and stacked into scaffolds, improving the contiguity and quality of assembly. For this, the project's Hi-C data was used, that can capture the 3D organization of the genome. In practice, Hi-C sequencing data can reveal proximity information of various genome parts. Using this information and the assumption that parts of genome's chromosomes tend to be organized inside the chromatin in close proximity, various software can infer contigs coming from the same chromosome, order them and/or stitch them together. This process is known as "scaffolding".

YaHS (Zhou et al. 2023) was used for this scaffolding process. Hi-C paired reads were aligned to contigs assemblies with BWA (Li and Durbin 2009) and in this way, proximity information is revealed, as reads from a pair are coming from parts of contigs that are close together inside the chromatin organization. YaHS then uses the proximity information to organize contigs into scaffolds, by stitching and ordering contigs to form larger and more contiguous blocks. Eventually, the goal is that these scaffolds are as complete as possible, stretching to full chromosomes.

Manual curation and chromosome-level assembly

Using JuicerTools (Durand et al. 2016) and JuiceBox Assembly Tools (Robinson et al. 2018), the Hi-C contact maps were visually inspected for potential misassemblies, misoriented contigs, or incorrectly placed contigs. Identified misassemblies were corrected by reordering, reorienting, or breaking contigs as necessary. Contigs that exhibited poor alignment or inconsistent proximity relationships were manually adjusted to ensure correct assembly.

Centromeric regions were identified based on reduced Hi-C contact frequencies and decreased HiFi coverage, which are characteristic of these repetitive regions. Telomeric sequences (TTAGG repeats) were located at the ends of the largest scaffolds, supporting their classification as full-length chromosomes. GRIT Rapid Curation workflow (Howe et al. 2021) was used for gaps, telomeres, and coverage identification.

The curated scaffolds were re-evaluated using QUAST (Quality Assessment Tool for Genome Assemblies) and BUSCO (Benchmarking Universal Single-Copy Orthologs) to confirm the improvements in assembly accuracy and completeness. Merqury plots were also generated to assess the k-mer completeness of the final chromosome assemblies.

Repeat identification and annotation of chromosome assembly

RepeatModeler v2.0.4 (Flynn et al. 2020) and RepeatMasker v4.1.5 (Smit et al. 2013) were used to identify and mask repetitive DNA sequences in the chromosome assembly.

RepeatModeler operates de novo, meaning it identifies repeat families directly from genomic sequences without requiring prior knowledge. It integrates various computational methods and tools, such as RECON v1.08 (Bao and Church 2002) and Tandem Repeats Finder v4.09.1 (TRF) (Benson 1999), to scan a genome for regions that appear repeatedly. Once identified, these regions are clustered based on sequence similarity to form repeat families. The tool then generates consensus sequences for each family, representing the "typical" sequence for each identified repeat class. Essentially, RepeatModeler helps to discover and categorize new and previously unidentified repeat families in a genome.

RepeatMasker, on the other hand, uses a library of known repeat sequences (which can be outputs from RepeatModeler, curated databases, or others) to scan and annotate genomic sequences for the presence of these repeats. It goes through the genome sequence and matches sections of it against its library of known repetitive elements. When a match is found, that genome section is masked, marked or replaced to indicate a repetitive region. By doing so, RepeatMasker helps annotate and sometimes obscure these repetitive regions to simplify further analyses, ensuring they do not interfere with processes like gene prediction.

Structural and functional annotation of chromosome assembly

Funannotate v.1.8.15 pipeline (Palmer 2020) was used to identify genic regions and assign functions to them. Funannotate is a comprehensive genome annotation pipeline primarily designed for fungal genomes but adaptable to other eukaryotes. At its core, Funannotate integrates several bioinformatics tools into a cohesive workflow to streamline the annotation process. The pipeline starts with a genome assembly and first employs ab initio gene prediction tools like AUGUSTUS v3.5.0 (Stanke et al. 2006) and GeneMark v3.68 (Besemer and Borodovsky 2005) to predict gene models. These predictions are refined using protein evidence from public databases like UniProt DB version 2023_03 (Consortium 2021) as well as transcriptomic evidence from various tissues of the organism under consideration. Concurrently, Funannotate identifies other genomic features, such as tRNAs and ncRNAs. After gene prediction, the pipeline annotates the predicted proteins' functions based on homology searches against known protein databases, InterProScan v5.63 (Jones et al. 2014), Gene Ontology release 2023-07-27 (Ashburner et al. 2000), Pfam v35.0 (El-Gebali et al. 2019) and eggNOG v5.0 (Huerta-Cepas et al. 2019; Cantalapiedra et al. 2021). The result is a fully annotated genome with predicted gene models and their inferred functions. HMMER v3.3.2 (Finn et al. 2011), DIAMOND v2.1.6 (Buchfink et al. 2021), MMseqs2 v14.7 (Steinegger and Söding 2017), and Prodigal v2.6.3 (Hyatt et al. 2010) are used in various steps for the identification of the functional annotation.

To explore the *C. colocynthis* transcriptome for different parts of the plant, PacBio IsoSeq reads from different plant tissues were aligned to the genome assembly with minimap2 ver. 2.28 (Li 2018, 2021). Then alignments were quantified with salmon ver. 1.10.3 (Patro 2017) to produce per-tissue gene expression matrices. Raw gene expression counts were analyzed with custom scripts in R ver. 4.3.3. PCA was calculated with the "prcomp()" function of the "stats" package. Gene and tissue distances were calculated with the "Euclidean distance" mode of the "dist()" function of the "stats" package and were hierarchically clustered with the "complete" agglomerative method of the "hclust()" function of the same package. Gene expression heatmap was calculated and plotted with the "pheatmap" package. Finally, the "VennDiagram" package was used for graphically presenting the intersections of gene sets between tissues.

Ortholog analysis

OrthoFinder v2.5.5 (Emms and Kelly 2015, 2019) was used for inferring gene orthogroups, gene trees and species phylogenetic trees between *Citrullus colocynthis* and seven species from the same order (Cucurbitales), plus one species (*Cicer arietinum*), from the same clade (Rosids/Fabids), as outgroup. OrthoFinder determines orthologs from orthogroup gene trees, starting with the proteomes of each species provided in FASTA format, which contain the amino acid sequences of the proteins. Orthogroups are identified using the OrthoFinder algorithm, and an unrooted gene tree for each orthogroup is constructed using DendroBLAST (Kelly and Maini 2013). From this collection of unrooted orthogroup trees, OrthoFinder infers an unrooted species tree using the Species Tree from All Genes (STAG) algorithm (Emms and Kelly 2018). STAG works by identifying the most closely related genes within both single-copy and multi-copy orthogroups. The resulting STAG species tree is subsequently rooted using the Species Tree Root Inference from Duplication Events (STRIDE) algorithm (Emms and Kelly 2017), which roots the tree by identifying and utilizing well-supported gene duplication events found across the orthogroup trees to estimate the most probable root location. STRIDE was specifically designed to root species trees using only the information available in gene trees, making it a powerful tool for rooting species trees in phylogenomic analyses. Finally, iTOL (Letunic and Bork 2007) v6.8.1 was used to plot the phylogenetic tree. For another view of the relationship between Cucurbitales, the intersection of orthogroups among them was defined and plotted in R with the UpSetR package (Conway 2017).

Diversity of 27 samples

Raw reads of 27-sample population were initially cleaned and pre-processed with fastp ver. 0.23.2 (Chen et al. 2018). Cleaned reads were aligned to the reference assembly sequence with bwa ver. 0.7.17 (Li and Durbin 2009) and variants were called from alignments with GATK ver. 4.4.0.0 (Van der Auwera and O'Connor 2020). Variants were filtered to keep only high-quality biallelic and polymorphic SNPs with GATK and parameters QD > 2.0, MQ > 40, FS < 60, SOR < 3.0, MQR > -12.5, RPR > -8.0, DP > 3, DP < 100, GQ > 20, missingness < 5% and MAF > 5%. A neighbor-joining (Saitou and Nei 1987) phylogenetic tree was constructed from the filtered set of SNPs with fastreeR (Gkanogiannis 2023) Bioconductor package in R that was finally plotted with iTOL (Letunic and Bork 2007).

Variant calling on assembly

Single-nucleotide variants (SNVs) and short insertion-deletion variants (INDELs) were annotated in the protein-coding regions of the *Citrullus colocynthis* assembly. For this, the DNBSEQ paired short reads, that were also been used to estimate genome length and heterozygosity, were used with the Sarek v.3.4.0 (Hanssen et al. 2024) pipeline to perform germline variant calling with GATK (Van der Auwera and O'Connor 2020). Finally, SNVs and INDELs were annotated using snpEff v5.1.0 (Cingolani et al. 2012).

Results

Citrullus colocynthis genome assembly

The genome of *C. colocynthis* was assembled into contigs from 1.9 million PacBio High Fidelity (HiFi) long reads (Rhoads and Au 2015; Wenger et al. 2019), that were scaffolded into chromosome-level scaffolds with the aid of 291 million paired-end High-throughput Chromosome Conformation Capture (Hi-C) short reads (Lieberman-Aiden et al. 2009). Gene models were extracted from 3.5 million PacBio Isoform Sequencing (IsoSeq) long RNA reads (Gonzalez-Garay 2016). Genome characteristics (length and ploidy) were estimated from the HiFi reads and from around 900 million paired-end DNA Nanoball Sequencing (DNBSEQ) short reads (Meslier et al. 2022). Table 1 summarizes the sequencing data used in this work. Raw sequencing data were deposited to the International Nucleotide Sequence Database Collaboration (INSDC), through the European Nucleotide Archive (ENA). Table 2 lists accession numbers and links to the data.

C. colocynthis genome was estimated by k-mer analysis ($k = 21$) of around 90 billion bases (Gb) of DNBSEQ paired short reads to be around 359 million bases long (Mb), diploid, and of 0.16% heterozygosity. Figure. S1 presents the k-mer spectra plot. By k-mer analysis ($k = 21$) of around 30 Gb of PacBio HiFi long reads, the genome was estimated to be around 347 Mb long, diploid, and 0.13% heterozygous. The k-mer spectra plot is shown in Fig. S2. It is although unlikely that the heterozygosity rate estimation is accurate. Looking at the k-mer spectra plots, no heterozygous k-mer peak is visible, and therefore real heterozygosity must be much less. The number of chromosomes has been previously defined as $2n = 22$ (Li et al. 2016), which agrees with all the similar species of the *Citrullus* genus.

HiFi reads were assembled into contigs and phased into two haplotypes using around 43 Gb of Illumina chromosome conformation capture paired short reads (Hi-C). Contig assembly for the Haplotype I contains 88 contigs measuring around 366 Mb long, with N90 (Lander et al. 2001) value around 26 Mb, N50 value around 31 Mb, L90 value 11, covering around 99% of the read k-mer spectrum, with around 67 assembly quality value QV and containing almost 97% complete (around 96% single and around 1% duplicated) copies of the 2326 Benchmarking Universal Single-Copy Orthologs (BUSCOs) (Seppey et al. 2019) of the Eudicotyledons plants (odb10). The length of the 95 contigs of the Haplotype II contig assembly was around 357 Mb, N90 around 6 Mb, N50 around 29 Mb, L90 14, covered by around 98% of the reads of 67 QV and containing around 96% (around 95% single and around 1% duplicated) copies of Eudicotyledons BUSCOs. Table 3 and Table S1 summarize quality measures of contigs assembly, whereas Fig. S3 and Fig. S4 present their Merqury and Nx plots, respectively. Only a single contig in Haplotype I assembly (length around 45 thousand bp) was found to be contaminant, belonging to a plant-feeding mite (*Tetranychus urticae*), and therefore was removed from the assembly. Figure. S5 presents contamination analysis plots for the two haplotypes of contigs assembly.

Using the Hi-C reads and manual curation, contig assemblies could be further improved by ordering and placing contigs into scaffolds. Hi-C data quality was problematic, more than 70% of Hi-C reads mapped on duplicate loci on the contig assembly; therefore, manual curation was necessary. Total scaffolds assembly size was not significantly changed for the two haplotypes, measuring again around 366 Mb and 357 Mb, respectively. Nevertheless, there was an improvement on the total scaffold number, reducing from 88 to 85 for the first haplotype and from 95 to 88 for the second haplotype. The N90 and N50 measures increased to almost 30 Mb and 28 Mb for each haplotype (N90) and

Table 3 Key statistics for the primary (Haplotype I) and secondary (Haplotype II) genome assemblies and annotation of *C. colocynthis*

Category	Type	Haplotype I	Haplotype II
Genome size Mb (k-mer analysis)		358.93 (short reads)	
Heterozygosity		0.159% (k-mer), 0.006% (actual)	
Contig	Number/size	88	95
	N50	30,853,444	29,357,918
	N90	25,551,156	5,828,716
	Longest	36,926,313	36,881,511
	Total contig length	366,111,899	356,939,597
	BUSCO (complete/single/duplicate)	96.7/95.7/1.0	95.7/94.5/1.2
Scaffold	Number/size	85	88
	N50	31,341,397	30,674,282
	N90	29,688,828	27,765,897
	Longest	36,926,313	36,881,511
	Total scaffold length	366,069,512	356,935,950
	BUSCO (complete/single/duplicate)	96.7/95.7/1.0	95.7/94.5/1.2
11 Chromosomes length		361,176,561	347,173,164
Repeat (% of genome)	Total	58.86	
	Retro/transposons/unclassified	31.44/3.95/19.19	
Predicted genes	Total number/mRNA/tRNA	23,327/21,969/1,358	
	Mean gene length	3531.84 bp	
Functional annotation	Number (% of coding genes)	19,157 (82.12)	
	GO/INTERPROSCAN/EGGNOG	14,573/17,347/19,157	

Metrics such as genome size, heterozygosity, contig and scaffold numbers, N50, and BUSCO completeness are included. The data highlights the quality and completeness of the assemblies. Primary assembly's length (366 Mbp) is very close to the predicted one (359 Mbp). Annotation of the primary assembly revealed that most of it (58.9%) is repeated. More than 95% of single and complete Eudicotyledons BUSCOs and less than 1% duplicates were found. Annotation found 23,327 gene models, of which 19,157 are coding ones and 14,573 were assigned with Gene Ontology terms

to more than 31 Mb and 30 Mb, respectively (N50). 90% of the total genome is covered by ten scaffolds for each haplotype assembly (L90). There was no change in the k-mer space and BUSCO gene space completeness. In Table 3 and Table S2, quality measures of the scaffolds assembly can be seen. Their Merqury and Nx plots are presented in Fig. S6 and Fig. S7, respectively, whereas the Hi-C contact maps of the final curated scaffolds assembly are shown in Fig. S8.

Eleven largest scaffolds of each haplotype's scaffolds assembly presented various evidence for almost completely representing the true chromosomes of *C. colocynthis*. Plant telomeric sequences (TTAGG) were identified on both arms of most of the 11 largest scaffolds for each haplotype assembly. HiFi reads coverage was uniform across their length, except areas around their middle, representing centromeres. Genome repetition was also increased in those centromeric areas. For this, those two haplotypes' assemblies are considered as the chromosome-level assemblies of the two haplotypes of the *C. colocynthis* sample. A graphical representation of the 11 chromosomes of the first haplotype (Haplotype I), with various genomic feature evidence across them, is shown in Fig. 1a. The 11 chromosome lengths can be found in Table S3. As the assembly of the first haplotype presented better qualitative characteristics (better QV score for example), it was selected as the base for further analyses.

Repeat and gene annotation

Repeat identification, on the Haplotype I assembly, revealed a high percentage of repetition in the *C. colocynthis* chromosome assembly. Around a third of the genome (31.44%) was identified to be retroelements, out of which 27.47% are Long Terminal Repeats (LTR retrotransposons), 3.48% are Long and 0.49% are Short Interspersed Nuclear Elements (LINE and SINE Non-LTR retrotransposons). An additional 3.95% of the genome consists of DNA transposons and 19.19% more is complex repeat elements of unknown type, raising the total percentage of interspersed repeats to 54.58%. Less complex and RNA type of repetition was identified at 4.32% of the genome. In total, 58.9% of the assembly was found to be repeat elements. Table 3 and Table S4 contain detailed statistics of repeat elements identification, whereas Fig. 1b presents their distribution along the 11 chromosomes.

Structural annotation yielded 23,327 gene models, of which 21,969 are mRNA and 1,358 are tRNA. The average length of genes was found to be about 3,531 bp. Functional

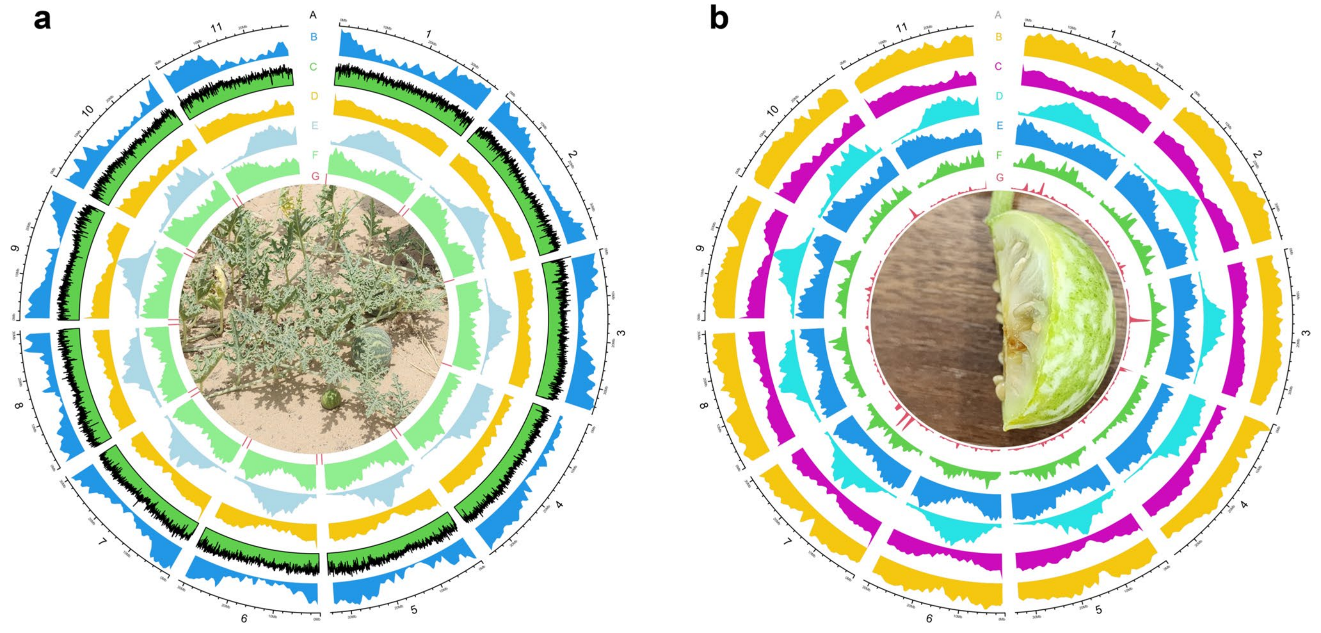


Fig. 1 Circos diagrams that provides a comprehensive visualization of the chromosomal and genomic features of *C. colocynthis* **a** The outermost ring (A) displays the chromosomal karyotype. The second ring (B) illustrates the distribution of genes across the chromosomes, providing insights into gene density and localization. The third ring (C) represents the HiFi coverage of sequencing data across the chromosomes, indicating the depth of coverage for genomic regions. The fourth ring (D) highlights simple repeats, while the fifth ring (E) maps the distribution of retrotransposons, emphasizing regions with high transposable element activity. The sixth ring (F) shows low-complexity repeat elements, which are typically associated with genomic regions prone to structural variation. The innermost ring (G) marks the telomeric regions of the chromosomes, indicating the ends of linear chromosomes that protect the genome from degradation. Gene features were annotated with Funannotate (Palmer 2020), whereas repeat regions were identified with RepeatModeler (Flynn 2020) and RepeatMasker (Smit et al. 2013). **b** Detailed representation of repeated elements across the genome. The first track (A) displays the chromosomal karyotype. The second track (B) represents the density of unknown type repeats. The third (C), fourth (D), and fifth (E) tracks show the density of simple, retro and low-complexity repeats, respectively. The sixth track (F) represents the density of DNA type, whereas the final track (G) the density of RNA type repeats along *C. colocynthis* chromosomes

annotation assigned Gene Ontology terms (Ashburner et al. 2000) to 14,573 genes, InterProScan (Jones et al. 2014) annotation to 17,347 genes, and eggNOG (Huerta-Cepas et al. 2019; Cantalapiedra et al. 2021) annotation to 19,157 genes. Table 3 and Table S5 present more detailed structural and functional annotation statistics. A graphical representation of the density of genes, repeats, HiFi data coverage and telomeric sequences can be seen in Fig. 1a.

Comparative transcriptome analysis

To investigate tissue-specific gene expression patterns, a comparative transcriptome analysis of four distinct tissues from the same sample (root, leaf, flower, and fruit) was performed. The analysis included hierarchical clustering, heatmap visualization, principal component analysis (PCA), and the identification of shared and unique gene sets among the tissues. The hierarchical clustering dendrogram (Fig. 2a) illustrates the relationships between the tissues based on their Euclidean distance calculated from raw gene expression counts. The clustering revealed that the leaf and flower tissues are more closely related in terms of their gene expression profiles compared to the root and fruit tissues, which clustered separately. This suggests that similar biological processes or functional genes are active in the leaf and flower tissues, distinguishing them from the root and fruit. A heatmap (Fig. 2b) was generated to further visualize the gene expression levels across the four tissues. This plot not only clusters the tissues (columns) but also the genes (rows), highlighting distinct patterns of expression. The heatmap confirms the dendrogram results and identifies specific gene groups that are expressed in each tissue, indicating potential tissue-specific functions. The PCA plot (Fig. 2c) provides a visual representation of the variance in gene expression data across the tissues. The first two principal components capture most of the variance (39.3% and 32%, respectively), with distinct separation among the tissues. The root and fruit tissues are clearly segregated along the principal components, while the leaf and flower tissues are positioned closer together, reinforcing the hierarchical clustering results. Finally, a Venn diagram (Fig. 2d) was used to compare the size of gene sets expressed in each tissue. The diagram shows significant overlap among the tissues, particularly among the leaf, flower, and root, with

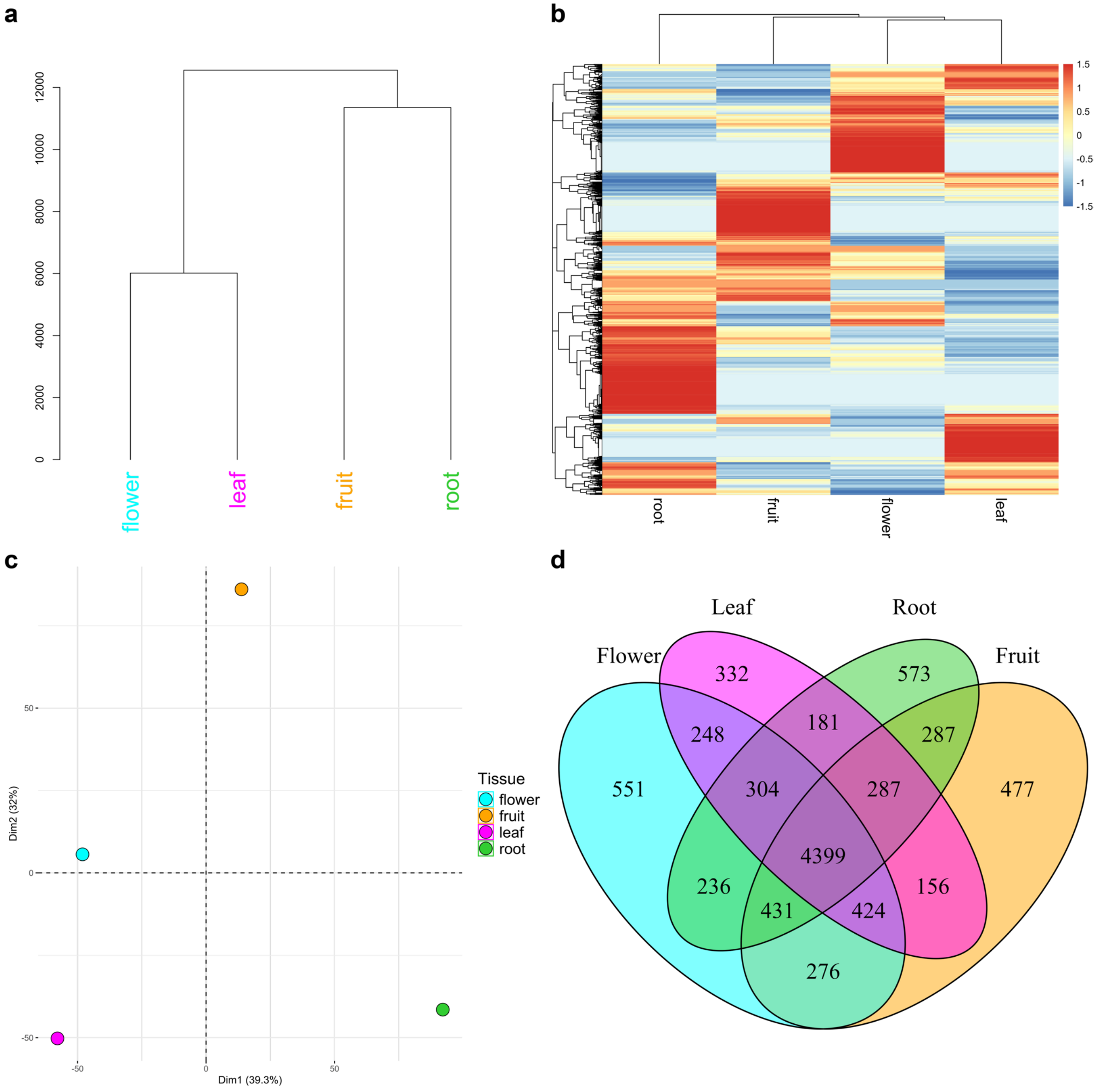


Fig. 2 Comparative transcriptome analysis for different tissues (root, leaf, flower, fruit) of the *C. colocynthis* sample. **a** Dendrogram showing the hierarchical clustering of the four tissues, based on Euclidean distances calculated from raw gene expression counts. Tissues that are closer on the dendrogram have more similar gene expression profiles. Leaf and flower tissues cluster together, indicating similar transcriptomic profiles, while root and fruit tissues form separate clusters. **b** Heatmap displaying gene expression levels across the four tissues. Rows represent genes, and columns represent tissues. The color gradient from blue (low) to red (high) indicates relative expression levels. Both genes and tissues were clustered hierarchically, revealing tissue-specific expression patterns and the overall similarities and differences among the tissues. **c** PCA plot that visualizes the variance in gene expression across the four plant tissues. Each point represents a tissue, plotted according to its principal component scores. The plot shows clear separation among the tissues, with root and fruit being distinct from the closely related leaf and flower tissues, highlighting their differing gene expression profiles. **d** Venn diagram that illustrates the overlap gene sets expressed in the tissues. Each oval represents the genes expressed in one tissue. The overlaps show shared genes, with 4399 genes expressed in all tissues. The diagram also highlights unique genes in each tissue, indicating specialized functions in each tissue type

4,399 genes shared across all tissues. In addition, the analysis revealed unique sets of genes specific to each tissue, with the flower tissue having 551, the leaf 332, the root 573, and the fruit 577 unique genes, respectively, suggesting specialized functions in each tissue.

Closely related species

Comparative genomics analysis was performed between *Citrullus colocynthis* and seven closely related species from the Cucurbitales order, plus one species from the Rosids/Fabids clade, as outgroup. These were *Cucumis sativus* (Cucumber), *Cucumis melo* (Muskmelon or Melon), *Momordica charantia* (Bitter gourd or Balsam pear), *Cucurbita maxima* (Pumpkin or Winter squash), *Cucumis melo var. makuwa* (Melon Oriental), *Cucurbita moschata* (Winter crookneck squash), *Citrullus lanatus* (Watermelon) and *Cicer arietinum* (Chickpea) as outgroup. 220,848 (95.3% of total sum) genes from all 9 species were assigned to 20,393 orthogroups. Out of them, 9,929 orthogroups have genes from all 9 species. Table S6 contains more detailed statistics about the inferred orthogroups and the intersection relationship of orthogroups between the nine species. A graphical representation of the intersecting orthogroups is shown in Fig. 3. A phylogenetic tree of the relationship between these species is presented in Fig. 4a, which is inferred from the 9929 orthogroups with genes from all 9 species and gene duplication event information.

Diversity of 27 samples

Twenty-seven colocynth samples were collected from various UAE regions and were genotyped with DNBSEQ paired short reads. Around 1,990 million high-quality short reads (100-bp long) were generated in total, yielding an average per sample coverage of 20x (for genome length of 366 Mbp). Read duplication was very low, less than 1%, and more than 99.5% of them were successfully mapped to the *C. colocynthis* Haplotype I chromosome-level assembly, a result that signifies the high quality of the assembly. Detailed mapping statistics can be found in Table S7. A total of 9,586,449 variants were called, out of which 8,239,764 were SNPs and 1387,141 were short INDELs. The ratio of transitions vs. transversions in raw SNPs (Ts/Tv) was 2.31. After filtering to remove variants of low quality, multiallelic and variants on repeated

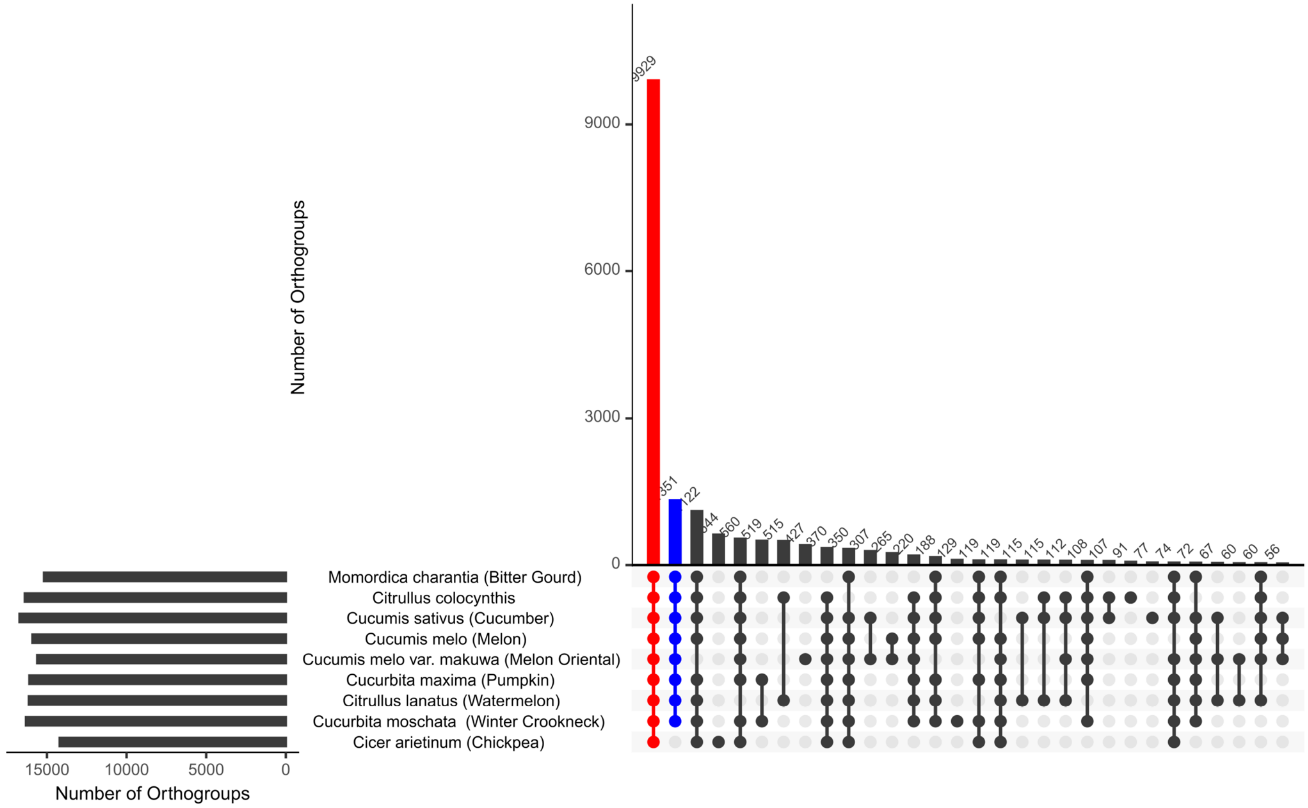


Fig. 3 Intersection relationships of orthogroups (groups of orthologous genes) among *C. colocynthis* and seven other species from the Cucurbitales order, plus one outgroup species (chickpea). Orthogroups are defined with OrthoFinder (Emms and Kelly 2015, 2019) and their intersection and plot with UpSetR (Conway 2017). The diagram illustrates the shared and unique orthogroups among the species, providing insights into their evolutionary relationships and functional conservation

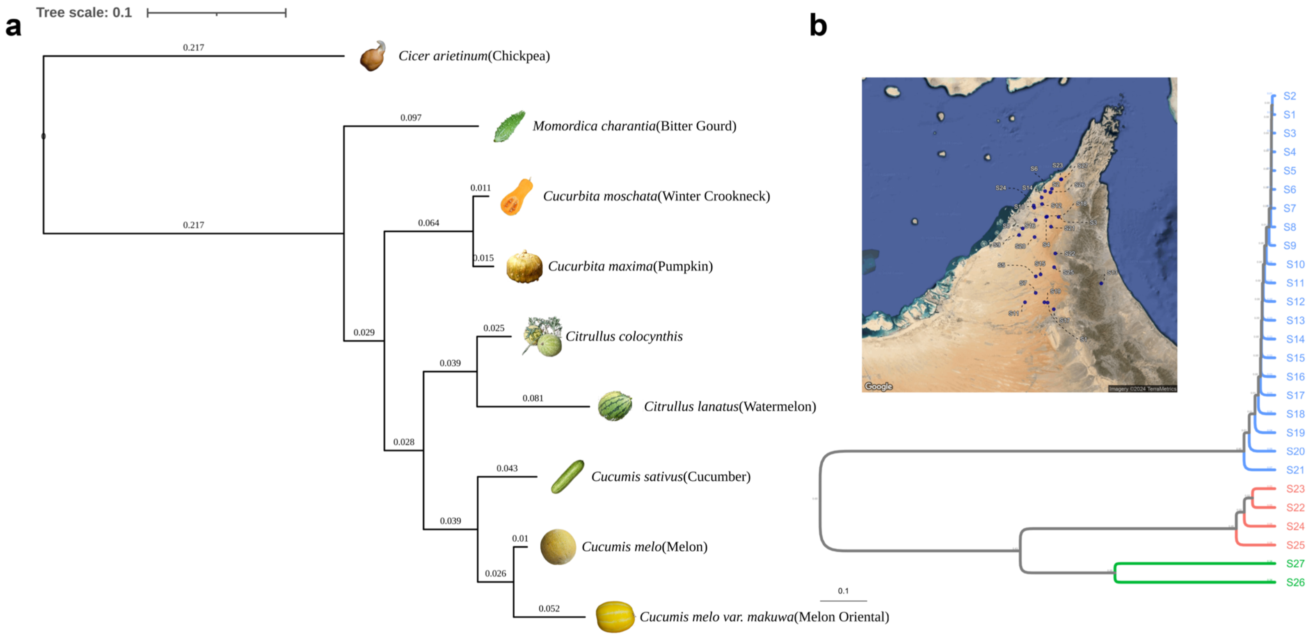


Fig. 4 Phylogenetic relationship and genetic diversity **a** Phylogenetic tree showing the relationship between *C. colocynthis* and closely related species within the Cucurbitaceae family. Chickpea is used as outgroup. The rooted tree is based on orthologous gene families that were identified by OrthoFinder (Emms and Kelly 2015, 2019). It is inferred with STAG (Emms and Kelly 2018) and is rooted with STRIDE (Emms and Kelly 2017). It is shown that *C. colocynthis* is a close relative to *C. lanatus* (watermelon). **b** Genetic diversity and clustering of 27 *C. colocynthis* samples collected from different regions of the UAE, showing their narrow genetic diversity and grouping into 3 distinct clusters. The tree was constructed with the Neighbor-Joining method (Saitou and Nei 1987) by first identifying Single Nucleotide Polymorphisms (SNPs) between them, that were then used to calculate their Euclidean distance matrix in the fastreeR R package (Gkanogiannis 2023). It was finally plotted with iTOL (Letunic and Bork 2007). This analysis underpins the evolutionary dynamics and potential founder effects within the *C. colocynthis* populations in the UAE

regions of the genome, 1,895,873 biallelic SNPs remained (Ts/Tv was 1.76) that were used to construct a phylogenetic tree of the relationship between the 27 samples, which is shown in Fig. 4b. The colocynth accessions collected from across UAE clustered into three groups, with narrow genetic diversity within the groups.

SNP and INDEL analysis

Single-nucleotide variants (SNVs) and short insertion-deletion variants (INDELs) were annotated in the protein-coding regions of the *Citrullus colocynthis* assembly to examine variant calling as a metric of genome annotation. The variant calling performed for Haplotype I assembly reported 21,249 heterozygous variants (275 SNPs and 20,974 INDELs). Given the fact that the short reads used for calling the variants were mappable at a rate of more than 99.5%, it translates to around 364 Mbp effective genome length and 1 heterozygous variant every 17 Kbp, or 0.0059% heterozygosity. Most variants were annotated to have extremely low or negligible impact, with the vast majority located at introns and upstream/downstream regions as shown in Fig. 5.

Discussion

Genome sequencing technologies have advanced dramatically in recent decades, enhancing accessibility of plant genomic data significantly. Since the advent of the first plant genome of *Arabidopsis thaliana*, achieved through whole-genome shotgun sequencing, the scientific community have seen the publication of over 200 plant genomes (https://www.plabipd.de/). Notably, the high-quality reference genome of *Citrullus lanatus* released by Guo et al. (2012) has become an indispensable resource for identifying candidate genes associated with critical traits. However, the genetic diversity unique to wild species, potentially lost through domestication or selective breeding, remains unexplored (Xie et al. 2019). The availability of multiple high-quality reference genomes from diverse genetic backgrounds is a prerequisite for effective mining of crop's genomic information (Yao et al. 2015), especially when incorporating wild germplasm into research.

To enhance watermelon genomic resources and facilitate its genetic improvement, this study embarked on whole-genome sequencing of *Citrullus colocynthis*. This species, a drought- and heat-tolerant wild relative of cultivated watermelon, offers invaluable insights into climate resilience. We present a diploid, chromosome-level genome assembly for *C. colocynthis*, integrating DNBSEQ short reads, PacBio HiFi long reads, and chromatin conformation capture Hi-C data.

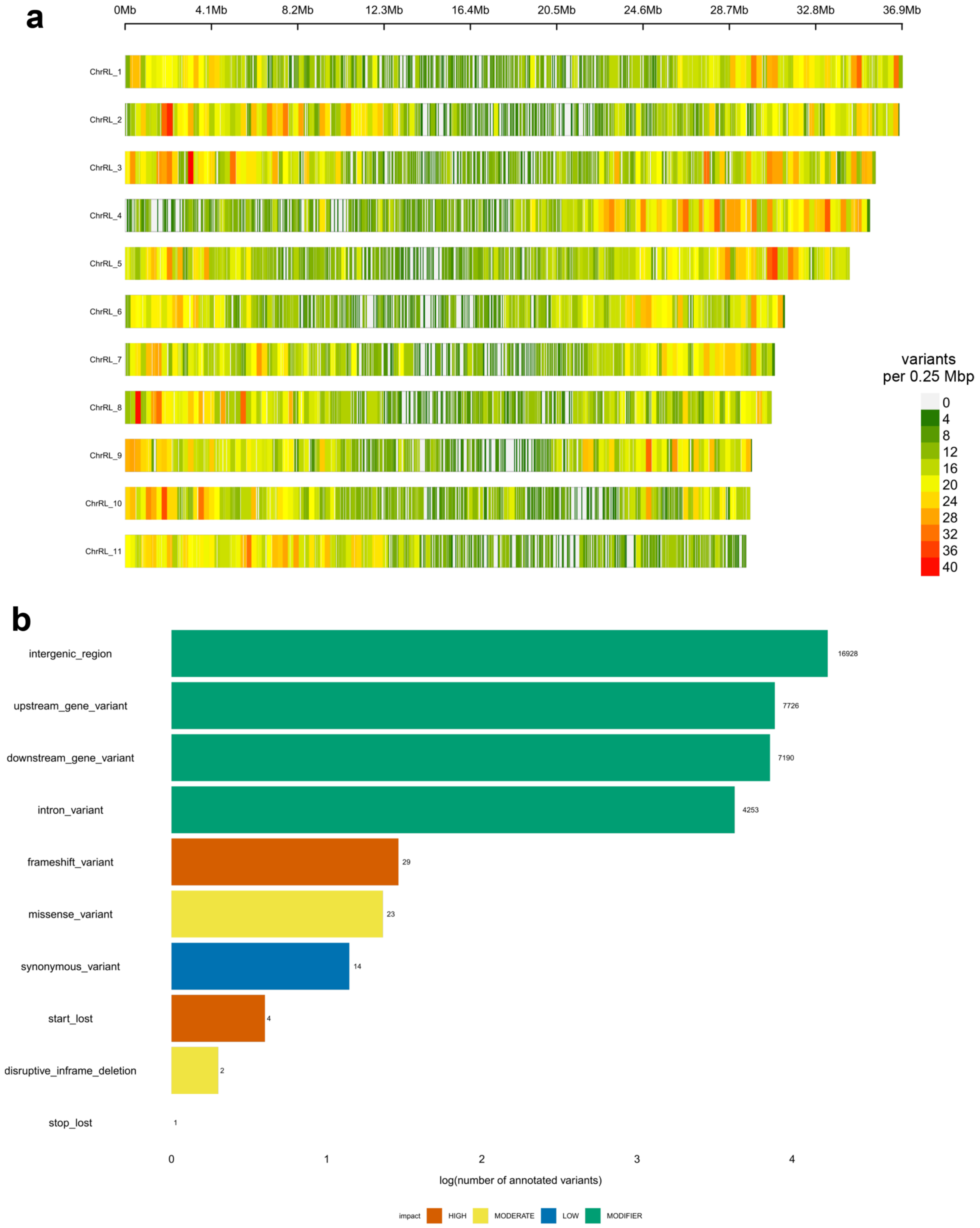
Our work resulted in two phased haplotype assemblies of 366.11 and 356.94 Mb, respectively, compared to the watermelon genome size of 353.5 Mb (Guo et al. 2012) and the Kordofan melon (*C. lanatus subsp. cordophanus*), which is approximately 367.9 Mb (Renner et al. 2021). Using Hi-C data and manual inspection of contact maps, we identified 11 large and 74 small contigs in the Haplotype I assembly, with the large contigs covering 98.66% of the colocynth genome and featuring telomeric sequences at both ends, suggesting their classification as pseudochromosomes. Chromosome counting confirmed the presence of $2n = 2x = 22$ chromosomes, consistent with other *Citrullus* species (Li et al. 2016; Guo et al. 2012; Renner et al. 2021). Interestingly, the heterozygosity rate measured by detecting SNPs between the two haplotype assemblies was extremely low at 0.006%, almost 30 times less than what k-mer analysis predicted.

The genome's completeness was affirmed by BUSCO analysis, indicating a 96.7% completeness with 95.7% single-copy genes and 1% duplicated genes. K-mer completeness analysis further supported a 99.7% completeness rate, underscoring the high quality of the *C. colocynthis* genome assembly.

In structural annotation, we identified and characterized protein-coding genes, non-coding RNA elements, and repetitive sequences, noting that over half of the genome (58.9%) consists of repetitive sequences, primarily LTR elements such as Copia and Gypsy. They were the most abundant repetitive elements, and long interspersed nuclear elements were the main proportion of non-LTR elements. This finding parallels observations in related species like watermelon and cucumber, where a significant portion of the genome is also composed of repetitive element (Guo et al. 2012). Meanwhile, 57.7% of the genome assembly in Kordofan melons was found to be repetitive sequences, predominantly LTRs. Identifying transposon hotspots and their distribution across chromosomes provides valuable insights into the genomic plasticity of this wild relative.

The results of the per-tissue transcriptomic analysis provide a comprehensive view of the gene expression profiles across four distinct plant tissues (root, leaf, flower, and fruit). However, it is crucial to consider the limitations of the current study, particularly the fact that RNA data were only available from the same single sample for all tissue types, without biological replicates or control/treatment conditions. This constraint inherently limits the generalizability of these findings and necessitates cautious interpretation. The hierarchical clustering and PCA results revealed distinct gene expression profiles among the tissues, with leaf and flower tissues showing closer transcriptomic relationships compared to root and fruit. While these findings are consistent with the expected biological functions of these tissues, the absence of replicates means that we cannot rule out the possibility that these observed differences may, in part, reflect individual sample variability rather than true biological differences. The lack of control conditions further complicates the interpretation, as it prevents from distinguishing between the effects of tissue type and other potential confounding factors such as environmental conditions or developmental stage. The heatmap and Venn diagram analyses provided additional insights into tissue-specific and shared gene expression patterns. However, without replicates, it is challenging to assess the statistical significance of these patterns or to determine whether the identified tissue-specific genes are consistently expressed across different individuals of the same species. The presence of unique gene sets in each tissue, particularly in the flower, may be indicative of specialized functions, but these findings should be validated with additional samples and experimental conditions to confirm their biological relevance. Moreover, the absence of control/treatment conditions means that it is difficult to evaluate how gene expression in these tissues might vary under different environmental or stress conditions, which would be critical for understanding the dynamic nature of transcriptome in response to external factors. Future studies should aim to include biological replicates and controls to allow for more robust statistical analyses and to capture the full range of gene expression variability within and between different plant tissues. Despite these limitations, the data presented here lay the groundwork for further exploration and underscores the potential of transcriptomic studies in uncovering tissue-specific functions and pathways in *C. colocynthis*.

Phylogenetic analysis based on orthologous gene families grouped the eight species of Cucurbitaceae into two subgroups where bitter gourd out-grouped the rest of the species. This analysis placed *C. colocynthis* closely with watermelon, affirming its role as a wild close relative (Assis et al. 2000). In addition, genome resequencing of 27 accessions from across UAE grouped the accessions in 3 clusters where one cluster consisted of 21 accessions and the other minor clusters consisted of 4 and 2 accessions, respectively. Several attempts have been carried out to understand genetic diversity in colocynth using random multilocus markers, allozymes, etc. Genetic diversity analysis on 29 *Citrullus colocynthis* belonging to MENA and Asia regions along with watermelon and citron melon was performed using high-frequency oligonucleotides-targeting active gene markers (Levi et al. 2017). In this study, the colocynth accessions were separated into two clusters and further separated themselves from watermelon and citron melon accessions with few admixtures. Furthermore, colocynth accessions have shown more than 70% genetic similarity across watermelon


a
0Mb
4.1Mb
8.2Mb
12.3Mb
16.4Mb
20.5Mb
24.6Mb
28.7Mb
32.8Mb
36.9Mb
ChrRL_1
ChrRL_2
ChrRL_3
ChrRL_4
ChrRL_5
ChrRL_6
ChrRL_7
ChrRL_8
ChrRL_9
ChrRL_10
ChrRL_11
variants per 0.25 Mbp
0
4
8
12
16
20
24
28
32
36
40
b
intergenic_region
16928
upstream_gene_variant
7726
downstream_gene_variant
7190
intron_variant
4253
frameshift_variant
29
missense_variant
23
synonymous_variant
14
start_lost
4
disruptive_inframe_deletion
2
stop_lost
1
0
1
2
3
4
log(number of annotated variants)
impact
HIGH
MODERATE
LOW
MODIFIER

◂**Fig. 5** Distribution of heterozygous Single Nucleotide Polymorphisms (SNPs) and insertion-deletion variants (INDELs) across the *C. colocynthis* genome. Variants were identified with GATK (Van der Auwera and O'Connor 2020) and annotated with snpEff (Cingolani et al. 2012). **a** Genome-wide distribution of heterozygous variants, identified in the Haplotype I assembly of *C. colocynthis*. The distribution highlights the density and location of these variants across the genome. **b** Functional impact of these variants, distinguishing between those with predicted low or negligible impact and those affecting coding sequences, introns, and regulatory regions. This analysis underscores the low overall heterozygosity of the genome, with most variants located in non-coding regions, reflecting the genetic stability of this species in its arid environment

accessions. Their genetic diversity across 38 colocynth accessions collected from different locations of Rajasthan, India has separated the accessions in two major clusters with more than 70% similarity when analyzed together with both random-amplified polymorphic DNA (RAPD) and inter simple sequence repeat (ISSR) markers (Verma et al. 2017). Genetic diversity analysis in 22 accessions collected from Egypt using ISSR markers and morphological variations divided the accessions in to 2 clusters (Badr et al. 2018). Badr et al. (2024) further used 10 Start Codon Targeted (SCoT) markers across 15 Egyptian colocynth accessions and found more than 77.5% similarity across accessions which separated in two clusters. Mazhar et al. (2024) conducted genetic variability analysis in 20 accessions of *Citrullus colocynthis* collected from Pakistan using 13 ISSR markers. In the present study, genetic diversity analysis in 27 accessions collected from UAE using SNP markers revealed a narrow genetic diversity within the *C. colocynthis*'s germplasm suggesting a potential founder effect in the region, whereby a small number of accessions are introduced to UAE and then might have been distributed throughout UAE through *zoochory*. This underscores the importance of broadening our collection and preservation efforts to bolster climate resilience in watermelon and other Cucurbitaceae crops.

In light of the genomic assembly and analysis of *C. colocynthis*, the potential for de novo domestication of this species presents a promising avenue for agricultural innovation. The comprehensive genomic insights provided by our study lay the groundwork for targeted genetic improvements through technologies such as CRISPR/Cas9, aiming to enhance drought tolerance, disease resistance, and nutritional profiles. De novo domestication, as argued by Fernie and Yan (2019), could significantly shorten the timeline for domesticating *C. colocynthis*, bypassing the generations of selective breeding required for traditional domestication processes. By directly editing the genome of *C. colocynthis* to introduce or enhance desirable traits, we can expedite the development of new crop varieties capable of thriving in changing environmental conditions and meeting the growing demands for food security (Fernie and Yan 2019; Gasparini et al. 2021). This strategy underscores the critical role of advanced genomic research and editing techniques in the future of crop development and agriculture.

In conclusion, the genomic assembly and analysis of *C. colocynthis* presented in this study provide a foundational step toward understanding the genetic basis of its remarkable traits, such as drought tolerance and medicinal properties. The *C. colocynthis* genome described here will be a valuable, if not indispensable, tool for Cucurbitaceae researchers conducting comparative genomic and evolutionary studies and watermelon breeders conducting crop improvement programs for climate resilience. By unraveling the complex genetic architecture of *C. colocynthis*, we have set the stage for future agricultural innovations through de novo domestication. Leveraging cutting-edge genome editing technologies, such as CRISPR/Cas9, offers a unique opportunity to harness the wild genetic diversity of *C. colocynthis* to develop new crop varieties tailored to the challenges of modern agriculture. This approach, as suggested by Fernie and Yan (2019) and Gasparini et al. (2021), could significantly accelerate the domestication process, bypassing traditional breeding limitations and directly introducing desirable traits into *C. colocynthis*. The potential for creating crops that are more resilient to environmental stressors and more nutritious and productive represents a pivotal advancement toward sustainable agriculture and food security. Our study contributes to the genomic resources available for *Citrullus* species and paves the way for innovative breeding strategies that could transform the agricultural landscape.

Conclusion

In conclusion, the chromosome-level genome for *C. colocynthis*, the wild close relative of watermelon, represents a significant advancement in the genomic study of the Cucurbitaceae family. Our high-quality reference genome and comprehensive gene annotation, encompassing 23,327 gene models, offer a robust platform for exploring evolutionary dynamics and functional genomics within this diverse plant family. This resource not only deepens our understanding of fruit crop evolution but also holds great potential for accelerating crop improvement initiatives. Given the cross-compatibility of colocynth with watermelon, our high-resolution, chromosome-level genome provides a substantial resource for genetic improvement of not only watermelon but also other Cucurbitaceae members either through breeding or genetic engineering.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00425-024-04551-7.

Acknowledgements We thank Dr. Mohammad Shahid and the genebank of the International Center for Biosaline Agriculture for providing the *C. colocynthis* seeds for this study. We also thank Dr. Srinivasan Samineni, ICBA's plant breeder, for providing the material for the diversity analysis of 27 colocynth samples. Authors would like to acknowledge Beijing Genomics Institute (BGI), China, for their support and guidance in establishing Desert Life Science Laboratory (DLSL) at International Center for Biosaline Agriculture. Authors would further like to acknowledge Dr. Tong Wei, BGI, for his guidance in carrying out whole-genome sequencing of colocynth. This work was supported by internal funding by the International Center for Biosaline Agriculture. The authors have no competing interests to declare that are relevant to the content of this article.

Author contributions AG and HR designed and conceived the research. HR collected the samples, performed the library preparation and sequencing of short reads, and supervised the sequencing of the PacBio, Hi-C, and IsoSeq libraries. AG assembled and annotated the genome, performed the ortholog, diversity of 27 samples, variants analysis and rest of bioinformatics analyses. RKS and ABLL supervised the study. AG and HR drafted the paper. All authors contributed to the final version of the manuscript.

Data availability Raw genome and transcriptome sequencing reads, the assembled genome, and its relevant annotations (repeats, genes, proteins, CDS and cDNA sequences) have been deposited into the European Nucleotide Archive (ENA) of the European Bioinformatics Institute (EBI), which is member of the International Nucleotide Sequence Database Collaboration (INSDC) along with NIG's DNA Data Bank of Japan and NCBI's GenBank, under the umbrella Bioproject id: PRJEB78362, that contains Bioproject PRJEB78297 (sequencing data) and Bioproject PRJEB66071 (assembly and annotation). Table 2 presents all the relevant accession numbers. Assembly and related data are accessible at https://www.ebi.ac.uk/ena/browser/view/GCA_963978565.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Altschul SF et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Al-Snafi AE (2016) Chemical constituents and pharmacological effects of *Citrullus colocynthis*—A review. IOSR J Pharm 6(3):57–67

Ashburner M et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29. https://doi.org/10.1038/75556

Assis JG et al (2000) Implications of the introgression between *Citrullus colocynthis* and *C. lanatus* characters in the taxonomy, evolutionary dynamics and breeding of watermelon. Pl Genet Resources Newslett. 121:15–19

Badr A, Zaki H (2024) Genetic diversity of *Citrullus colocynthis* populations using phytochemical analysis and SCoT marker variations. Genet Resour Crop Evol 71:2341–2353

Badr A et al (2018) Genetic diversity of colocynth (*Citrullus colocynthis* Schrader) populations in the eastern desert of egypt as revealed by morphological variation and ISSR polymorphism. Feddes Repertorium 129:173–184

Bao G, Church GM (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Berwal MK et al (2022) The bioactive compounds and fatty acid profile of bitter apple seed oil obtained in hot Arid Environments. Horticulturae. 8:259

Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451–W454

Bigdelo M et al (2017) Evaluation of bitter apple (*Citrullus colocynthis* (L.) Schrad) as potential rootstock for watermelon. Aust J Crop Sci 11:727–732

Bikdeloo M et al (2021) Morphological and physio-biochemical responses of watermelon grafted onto rootstocks of wild watermelon [*Citrullus colocynthis* (L.) Schrad] and commercial interspecific cucurbita hybrid to drought stress. Horticulturae. 7(10):359

Bohra A et al (2022) Reap the crop wild relatives for breeding future crops. Trends Biotechnol 40:412–431

Borgi Z, Hibar K, Boughalleb N, Jabari H (2009) Evaluation of four local colocynth accessions and four hybrids, used as watermelon rootstocks, for resistance to fusarium wilt and fusarium crown and root rot. Afr J Plant Sci Biotechnol 3:37–40

Buchfink B, Reuter K, Drost HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368

Cantalapiedra CP et al (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293

Challis R et al (2020) BlobToolKit—interactive quality assessment of genome assemblies. G3 Genes Genomes Genetics. 10:1361–1374

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890

Cheng H et al (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18:170–175. https://doi.org/10.1038/s41592-020-01056-5

Chomicki G, Renner SS (2015) Watermelon origin solved with molecular phylogenetics including Linnaean material: another example of museomics. New Phytol 205:526–532

Cingolani P et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6:80–92. https://doi.org/10.4161/fly.19695

Consortium T.U (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489

Conway J et al (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33(18):2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Coordinators NCBIR (2014) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 42:D7–D17

Council NR (2006) Lost Crops of Africa: Volume II: Vegetables. The National Academies Press, Washington.

Dane F, Liu J, Zhang C (2007) Phylogeography of the Bitter Apple, *Citrullus Colocynthis*. Genet Resour Crop Evol 54:327–336

DeMaere MZ, Darling AE (2021) qc3C: reference-free quality control for Hi-C sequencing data. PLoS Comput Biol 17:1–20

Durand NC et al (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 3:95–98. https://doi.org/10.1016/j.cels.2016.07.002

El-Gebali S et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238

Emms DM, Kelly S (2017) STRIDE: species tree root inference from gene duplication events. Mol Biol Evol 34(12):3267–3278

Emms DM, Kelly S (2018) STAG: Species Tree Inference from All Genes. bioRxiv. p. 267914.

Fernie AR, Yan J (2019) De novo domestication: an alternative route toward new crops for the future. Mol Plant 12:615–631

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37. https://doi.org/10.1093/nar/gkr367

Flynn JM et al (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci 117:9451–9457. https://doi.org/10.1073/pnas.1921046117

Fukasawa Y et al (2020) LongQC: a quality control tool for third generation sequencing long read data. G3 Genes Genomes Genetics. 10:1193–1196

Gasparini K, Moreira JDR, Peres LEP, Zsögön A (2021) De novo domestication of wild species to create crops with increased resilience and nutritional value. Curr Opin Plant Biol 60:102006

Gkanogiannis A (2023) fastreeR: phylogenetic, distance and other calculations on VCF and Fasta files. Bioconductor. https://doi.org/10.18129/B9.bioc.fastreeR

Gonzalez-Garay ML (2016) Introduction to isoform sequencing using pacific biosciences technology (Iso-Seq). In: Wu J (ed) Transcriptomics and gene regulation. Springer, Dordrecht, pp 141–160

Guo S et al (2012) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet. https://doi.org/10.1038/ng.2470

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hanssen F et al (2024) Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery. NAR Genom Bioinf 6(2):lqae031

Howe K et al (2021) Significantly improving the quality of genome assemblies through curation. Gigascience. 10:giaa153

Huerta-Cepas J et al (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314

Hussain AI et al (2014) *Citrullus colocynthis* (L.) Schrad (bitter apple fruit): a review of its phytochemistry, pharmacology, traditional uses and nutritional potential. J Ethnopharmacol 155:54–66

Hyatt D et al (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 11:119. https://doi.org/10.1186/1471-2105-11-119

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

Kelly S, Maini PK (2013) DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments. PLoS ONE 8:e58537

Kokot M, Długosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33:2759–2761

Lander ES et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128

Levi A et al (2017) Genetic diversity in the desert watermelon *Citrullus colocynthis* and its relationship with *Citrullus* species as determined by high-frequency oligonucleotides-targeting active gene markers. J. Am. Soc. Hort. Sci. 142(1):47–56

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

Li H (2021) New strategies to improve minimap2 alignment accuracy. Bioinformatics 37:4572–4574

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Lieberman-Aiden E et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293

Li KP et al (2016) Cytogenetic relationships among *Citrullus* species in comparison with some genera of the tribe Benincaseae (Cucurbitaceae) as inferred from rDNA distribution patterns. BMC Evol Biol 16:85

Mariod AA, Jarret RL (2022) Chapter 12—Antioxidant, antimicrobial, and antidiabetic activities of *Citrullus colocynthis* seed oil. Multiple biological activities of unconventional seed oils. Academic Press, New York, pp 139–146. https://doi.org/10.1016/b978-0-12-824135-6.00005-2

Mazher M et al (2024) Evaluation of genetic diversity and population structure of *Citrullus colocynthis* based on physiochemical and inter simple sequence repeat (ISSR) markers. Genet Resour Crop Evol. https://doi.org/10.1007/s10722-024-01913-8

Meslier V et al (2022) Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Scientific Data 9:694

Ogundele JO, Oshodi AA, Amoo IA (2012) Comparative Study of Amino Acid and Proximate Composition of *Citrullus colocynthis* and *Citrullus vulgaris* Seeds. Pak J Nutr 11:247–251

Palmer JM (2020) Funannotate v1.8.1: a fungal genome annotation and comparative genomics pipeline. Zenodo. https://doi.org/10.5281/zenodo.4054262. Accessed Aug 2023

Patro R et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. https://doi.org/10.1038/nmeth.4197

Pimentel D et al (1997) Economic and environmental benefits of biodiversity. Bioscience 47:747–757

Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Report 15:8–15

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

Renner SS et al (2021) A chromosome-level genome of a Kordofan melon illuminates the origin of domesticated watermelons. Proc Natl Acad Sci 118:e2101486118

Renzi JP et al (2022) How could the use of crop wild relatives in breeding increase the adaptation of crops to marginal environments? Front Plant Sci. https://doi.org/10.3389/fpls.2022.1101822

Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21:245

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genom Proteom Bioinform 13:278–289

Robinson JT et al (2018) Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell Syst 6:256-258.e1

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sawaya WN, Daghir NJ, Khalil JK (1986) *Citrullus colocynthis* seeds as a potential source of protein for food and feed. J Agric Food Chem 34:285–288

Sawaya WN, Daghir NJ, Khan P (1983) Chemical characterization and edibility of the oil extracted from *Citrullus colocynthis* seeds. J Food Sci 48:104–106

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M (ed) Gene prediction: methods and protocols. Springer, New York, pp 227–245

Si Y et al (2010) Cloning and expression analysis of the Ccrboh gene encoding respiratory burst oxidase in *Citrullus colocynthis* and grafting onto *Citrullus lanatus* (watermelon). J Exp Bot 61:1635–1642

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0. http://www.repeatmasker.org. Accessed Aug 2023

Stanke M et al (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439

Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028

Tyack N, Dempewolf H, Khoury CK (2020) The potential of payment for ecosystem services for crop wild relative conservation. Plants. 9(10):1305

Van der Auwera GA, O'Connor BD (2020) Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media.

Verma KS et al (2017) RAPD and ISSR marker assessment of genetic diversity in *Citrullus colocynthis* (L.) Schrad: a unique source of germplasm highly adapted to drought and high-temperature stress. 3 Biotech 7(5):288. https://doi.org/10.1007/s13205-017-0918-z

Wang Z et al (2014) Analysis of the *Citrullus colocynthis* transcriptome during water deficit stress. PLoS ONE 9:e104657

Wenger AM et al (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155–1162

Xie M et al (2019) A reference-grade wild soybean genome. Nat Commun 10:1216

Yao W et al (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16:187

Zhou C, McCarthy SA, Durbin R (2023) YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 39:btac808