scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Hippophae rhamnoides* variety

Xingyu Yang[1,2,8], Shujie Luo[1,3,8], Shihai Yang[4,8], Ciren Duoji[5], Qianwen Wang[1], Zhiyu Chen[2,6,7], Danni Yang[1], Tianyu Yang[2,6,7], Xi Wan[1,2], Yunqiang Yang[1,6,7], Tianmeng Liu[3] ✉ & Yongping Yang[6,7] ✉

***Fructus hippophae (Hippophae rhamnoides spp. mongolica×Hippophae rhamnoides sinensis)*, a hybrid variety of sea buckthorn that *Hippophae rhamnoides* spp. *mongolica* serves as the female parent and *Hippophae rhamnoides sinensis* serves as the male parent, is a traditional plant with great potentials of economic and medical values. Herein, we gained a chromosome-level genome of *Fructus hippophae* about 918.59 Mb, with the scaffolds N50 reaching 83.65 Mb. Then, we anchored 440 contigs with 97.17% of the total genome sequences onto 12 pseudochromosomes. Next, *de-novo*, homology and transcriptome assembly strategies were adopted for gene structure prediction. This predicted 36475 protein-coding genes, of which 36226 genes could be functionally annotated. Simultaneously, various strategies were used for quality assessment, both the complete BUSCO value (98.80%) and the mapping rate indicated the high assembly quality. Repetitive elements, which occupied 63.68% of the genome, and 1483600 bp of non-coding RNA were annotated. Here, we provide genomic information on female plants of a popular variety, which can provide data for pan-genomic construction of sea buckthorn and for the resolution of the mechanism of sex differentiation.**

Background & Summary

Sea buckthorn (*Hippophae*), belonging to the Elaeagnaceae family, is a diploid (2n = 2x = 24) deciduous plant with high exploitation values[1,2]. Most sea buckthorn is cultivated in cold zones of Europe and Asia[3,4]. *Hippophae* is rich in ascorbic acid, carotenoids, healthy fatty acids, and other secondary metabolites[5–7]. Previous studies have primarily focused on its medicinal value. Extracts from the leaves and orange-yellow fruit have immunomodulatory potential and antioxidant, anti-viral, and wound-healing properties[8–11]. Sea buckthorn is also used in traditional medicine for the treatment of pulmonary, cardiac, gastrointestinal, blood, or metabolic disorders[12–16]. It is therefore crucial to decode the genomic information of Sea buckthorn. Three genome of *Hippophae* were published last year, including *Hippophae rhamnoides ssp. sinensis*, *Hippophae tibetana*, and *Hippophae gyantsensis* which revealing differences in their biological data, such as the genome size and percentage of repeated sequences[17–19]. The decoding of further genomic information from other *Hippophae* subspecies and popular varieties is therefore of importance.

Rapid advances in sequencing technology have made it possible to obtain accurate and high-throughput data at a very low cost[20,21]. However, there is currently no research on *Fructus hippophae* genomic information. Studies on *Fructus hippophae* are currently limited to compounds and their related protein targets of Hippophae Fructus oil (HFO), relying on the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP: https://old.tcmsp-e.com/tcmsp.php)[22]. Other studies have focused on methods of extracting and purifying flavonoids, tannins, and other novel nutritional supplements from *Fructus hippophae*, which depend on spectrophotometry, chromatography and other chemical methods[23–25]. Herein, we integrated three different

[1]Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, kunming, 650000, China. [2]University of Chinese Academy of Sciences, Beijing, 100049, China. [3]Dali University, Dali, 671000, China. [4]Yunwang Industrial Corporation, Ltd, Tibet, 850000, China. [5]Service Center for Forestry and Grassland Bureau of Sangzhuzi District in Xizang, Xizang, 850000, China. [6]Plant Germplasm and Genomics Center, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. [7]Institute of Tibetan Plateau Research at Kunming, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. [8]These authors contributed equally: Xingyu Yang, Shujie Luo, Shihai Yang. ✉e-mail: tianmeng_liu@163.com; yangyp@mail.kib.ac.cn

Sequence	Platform	Total bases	GC content (%)	Q30 (%)	Sequence depth (×)
DNA reads	MGISEQ-T7	56050695900	30.30	90.32	61
RNA reads	MGISEQ-T7	85726685700	42.16	93.08	93
Hi-C reads	MGISEQ-T7	79086390300	33.59	94.28	86

Table 1. Characteristics of NGS data for genome assembly.

Sequence	Platform	Total bases	GC content (%)	N50 (bp)	Sequence depth (×)
ONT reads	Nanopore	98574842275	29.55	36264	107

Table 2. Characteristics of ONT data for genome assembly.

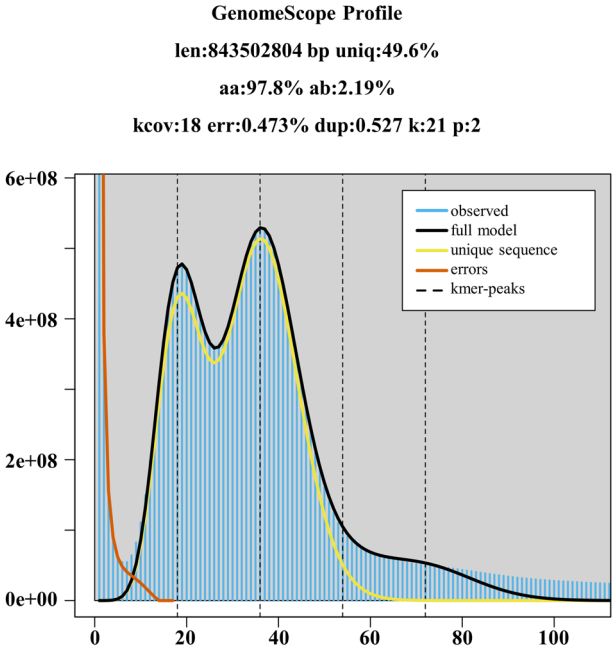

Fig. 1 K-mer distribution (K = 21) of *Fructus hippophae* genome using GenomeScope 2.

sequencing datasets for genome assembly, including short reads based on next generation sequencing (NGS) on the MGI platform, Oxford Nanopore Technologies (ONT) long reads, and high-throughput chromatin conformation capture (Hi-C) reads. Structural annotation of protein-coding genes was then carried out by *de novo*, homology and transcriptome assembly strategies. Next, gene functional annotation was performed by alignment with public databases. These genome-related data will provide a valuable resource for the study of sea buckthorn.

Methods

Plant materials and genome sequencing. To study the genome of *Fructus hippophae*, fresh young leaves were collected from the same wild *Fructus hippophae* tree which planted in Shigatse, Tibet, China. Total genomic DNA and RNA were extracted using the modified cetyl trimethylammonium bromide (CTAB) method and E.Z.N.A. Total RNA Kit I (Omega Bio-Tek, Norcross, GA, USA), respectively[26]. Then, 150 bp paired-end libraries with an insert size of 250 bp were constructed and sequenced at the MGISEQ-T7 platform. The Hi-C sequencing library was constructed according to the published protocol, and then the crosslinked chromatin was digested with DpnII and ligated after biotinylation. DNA fragments were enriched via the interaction between biotin and blunt-end ligation, and then the enriched library was sequenced on the MGISEQ-T7 platform. DNA Long reads were generated by the Nanopore platform and processed by IsoSeq technology using the SMRT method. A totals of 56 Gb of raw MGI short-read data (61× coverage) with a Q30 exceeding 90% (Table 1), 98 Gb of passed Nanopore long-reads data (93× coverage, the N50 length reaching 36264 bp, the average length reaching 25977 bp) (Table 2), and 79.09 Gb of Hi-C data (86× coverage) with the Q30 reaching 94.23% (Table 1) finally were obtained from the whole-genome sequencing.

Estimatie of genome size. Sequence adaptors, duplications, and low-quality reads from the original paired-end short DNA reads were filtered by Fastp[27] with the parameters -n 0 -l 140. Then, 55 Gb of the clean reads from the MGI library were used to estimate the size, heterozygosity, and repeat content of the genome using

Features	Statistics
Sequenced genome size (Mb)	921.69
Number of contigs	723
Contig N50 (bp)	14835755
Contig N90 (bp)	420505
Max contig size (bp)	45242531

Table 3. Characteristics of the *Fructus hippophae* genome at contig level.

Features	Statistics
Number of Chromosomes	12
Scaffold N50 (bp)	83648241
Scaffold N90 (bp)	55834877
GC content (%)	30
Max scoffold size (bp)	105172879
Total Size (Mb)	918.59

Table 4. Characteristics of the *Fructus hippophae* genome at scaffold level.

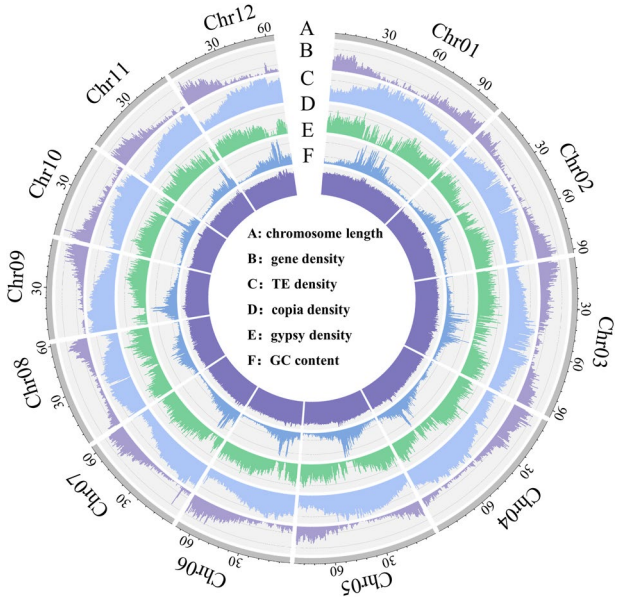

Fig. 2 Circos plot of distribution of the *Fructus hippophae* genomic elements. The tracks indicate (**A**) length of chromosomes, (**B**) distribution of genes on different chromosomes, (**C**) distribution of transposable elements on different chromosomes, (**D**) distribution of copia elements on different chromosomes, (**E**) distribution of gypsy elements on different chromosomes, (**F**) GC content of different chromosomes. The densities of genes, TEs, copia elements, gypsy elements and GC were calculated in 500 kb windows.

Jellyfish[28], with a 21-mer frequency and the parameter set as reads_cutoff = 1k, obetaining 47716688158 k-mer. Next, the Genomescope v2.0[29] was used to analyze the K-mer frequency distribution. Ultimately, the genome size was estimated to be 843 Mb with 2.19% heterozygosity and 49.6% repetitive sequences (Fig. 1).

Genome *de novo* assembly. The genome was assembled by integrating the clean Nanorpore long reads, MGI short reads and Hi-C reads. First, *de novo* genome assembly was performed by NextDenovo (v2.5.0) (https://github.com/Nextomics/NextDenovo) with the high-quality ONT reads. Then, the clean NGS reads were used for four-rounds of self-correction and three-rounds of consensus correction by Nextpolish (v1.4.1)[30] with the task parameter = best. Next, the redundant sequences resulting from heterozygosity were removed with the purge-dups (v1.2.5)[31] pipeline. After assembly, a 921.69 Mb draft genome, including 723 contigs and the N50 reaching 14.8 Mb, was obtained (Table 3). Additionally, Hicpro (v3.1.0)[32] was used to further validate the Hi-C reads, and 3D-DNA[33] was then used to organize and anchor the contigs into draft chromosomes. Manual check and refinement to the cluster, order, and orientation of the draft assembly were carried out using Juicebox assembly tools[34]. Ultimately, the final genome was 918.59 Mb in size and consisted of 253 scaffolds with an N50

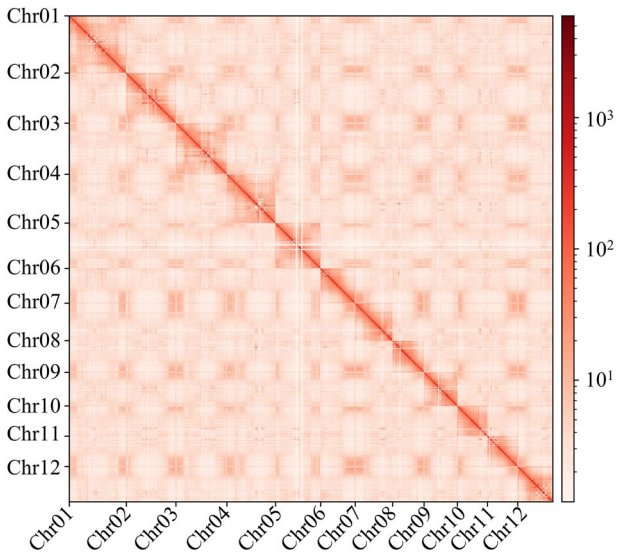

Fig. 3 Heatmap of genome-wide Hi-C data of *Fructus hippophae* chromosomes. The frequency of Hi-C interaction links is represented by colors, ranges from orange (low) to dark red (high).

Type		count	masked (bp)	masked (%)
LTR	unknown	129541	66497775	7.24
	Gypsy	236098	159424221	17.36
	Copia	215332	155535260	16.94
TIR	DTA	85840	25710480	2.80
	DTH	28546	7676719	0.84
	DTM	78246	49437657	5.38
	DTC	88257	38449970	4.19
	DTT	38064	8859044	0.96
non-LTR	LINE_element	7529	3229551	0.35
	tRNA_SINE	1123	169924	0.02
	Penelope	1854	744765	0.08
	unknown	25098	23311951	2.54
non-TIR	helitron	63779	25541983	2.78
others	DNA_transposon	45599	11813872	1.29
	low_complexity	78	467391	0.05
	repeat_region	25227	7930950	0.86
Total		1070211	584801513	63.68

Table 5. Summary of transposable elements in *Fructus hippophae* genome.

length up to 83.64 Mb, including 12 pseudochromosomes that accounted for 97.14% of the total genome length (Table 4). Circos plot of the distribution of the genomic elements (Fig. 2) was generated by shinyCircos v2.0 (https://venyao.xyz/shinyCircos/) and the heatmap of genome-wide Hi-C data (Fig. 3) of the *Fructus hippophae* genome chromosomes was drawn by hicexplorer.

Repetitive elements identification. The transposable elements (TEs) in the genome were identified and annotated by Extensive de-novo TE Annotator (EDTA) v2.1.2[35] and classified by TEsorter (v1.3)[36], DeepTE[37], and LTR_FINDER[38]. Finally, 913550 bp of repeat elements were predicted, occupying 61.02% of the total genome length. The TEs could be classified into five categories after annotation, including long terminal repeats (LTR), tandem inverted repeats (TIR), non-LTR, non-TIR, and others. Of these, *Gypsy* occupied the highest proportion (35.65%) and was evenly distributed on 12 pseudochromosomes in the genome, followed by *Copia* with 19.81% occupation and high abundance in the central region of the genome (Table 5, Fig. 2).

Protein-coding genes prediction. Simultaneously, Repeatmasker (v4.1.2-p1)[39] software was used for repeat masking. The masked genome was then subjected to gene prediction. First, structure annotation of the protein-coding genes was predicted using braker[40] and tsebra[41] software by integrating evidence from homology-, *de nove*- and transcriptome-based annotations. Maker (v3.01.04)[42] and EVidenceModeler (v1.1.1) pipelines[43]

Database	anno_num	ratio(%)
COG	13800	37.83
GO	18384	50.40
KEGG	16111	44.17
KOG	19494	53.44
Swissprot	25632	70.27
TrEMBL	36226	99.31
NR	36004	98.71
Total_annotated	36226	99.31

Table 6. Statistical analysis of the functional gene annotations of the *Fructus hippophae* genome.

	Counts	Masked (bp)
miRNA	185	23364
tRNA	1041	77105
snRNA	245	23439
rRNA	750	489313
spliceosomal_RNA	84	11878
orthers	9626	858501
Total	10376	1483600

Table 7. Classification of non-coding RNA in the *Fructus hippophae* genome.

BUSCO	%
Genome Complete Buscos	98.8
Complete and aingle-copy Buscos	87.2
Complete and duplicated Buscos	11.6
Fragemented Buscos	0.2
Missing Buscos	1.0

Table 8. Statistics for genome assessment using BUSCO.

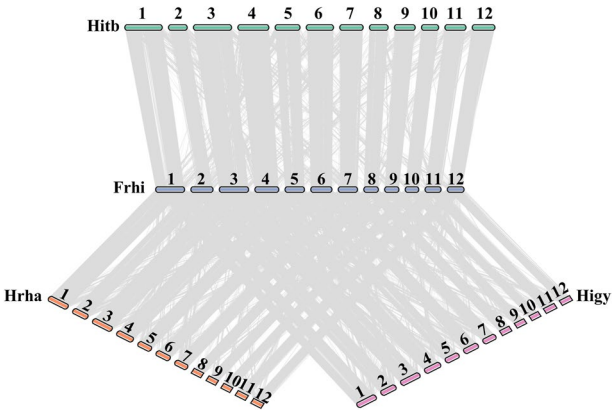

Fig. 4 Genome synteny is observed among *F. hippophae* and three other Sea buckthorn species: Hrha for *Hippophae rhamnoides* ssp. *sinensis*, Frhi for Fructus hippophae, Hitb for *Hippophae tibetana*, and Higy for *Hippophae gyantsensis* genomes. Chromosome numbers 1–12 represent the chromosomes 1 through 12 of the four Sea buckthorn species.

were used to integrate the evidence for non-redundant gene models, and the GFF3 file locating the gene, coding sequence, protein, and mRNA positions was obtained. Finally, a total of 36475 protein-coding genes were predicted, with gene lengths of 158 to 127368 bp. Additionally, 35943 (98.54%) of the predicted genes were allocated to the 12 chromosomes, and the gene distribution showed a higher density at the ends of the chromosomes.

Genes function and non-coding RNA annotation. The functional annotations of the predicted genes were further annotated by homologous searches against public databases using BLASTP[44] with the e-value cut-off = 1e-10, including NR, Swissprot[45], Translated European Molecular Biology Laboratory (TrEMBL)[46], KOG, GO[47], KEGG[48] and COG. Overall, 99.31% of the genes were functionally annotated. Among them 98.71%, 99.31%, 70.27%, 53.44%, 44.17%, 50.4%, and 37.83% gene were annotated in NR, TrEMBL[46], Swissprot[45], KOG, KEGG[48], GO[47] and COG[49] databases, respectively (Table 6). Non-conding RNAs were identified using cmscan[50] search against the RNA families database (Rfam)[51] with default parameters. Finally, 10376 non-coding RNAs(1483600 bp), including 1041 transfer RNA (77105 bp), 750 ribosomal RNA (489313 bp), 84 spliceosomal nuclear RNA (11878 bp), 185 microRNA (23364 bp) and 9626 other types of RNA (858501 bp) were identified in *Fructus hippophae* (Table 7).

Data Records

The genomic WGS sequencing data were deposited in the NCBI Sequence Read Archive (SRA) database under the BioProject PRJNA1003561.

The genomic NGS data were deposited in the SRA at NCBI SRR25591597[52].

The genomic ONT data were deposited in the SRA at NCBI SRR25591606[53] and SRR25591605[54].

The RNA short reads of leaves and stems with 3 dupication were deposited in the SRA at NCBI SRR25591604[55], SRR25591603[56], SRR25591602[57], SRR25591601[58], SRR25591600[59], SRR25591599[60], SRR25591596[61], SRR25591595[62], SRR25591594[63], SRR25591593[64], SRR25591592[65], and SRR25591591[66].

Hi-C data were deposited in the SRA at NCBI SRR25591598[67].

The final chromodome assembly and genome annotation files are available in GenBank[67].

Technical Validation

Here, several strategies were taken to assess the genome quality. The completeness of the non-redundant draft genome was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO)[68] with the embryophyta odb10 dataset, which consists of 1614 single copy genes with the default parameters. Revealing that 98.8% of these genes exhibited complete coverage. Among them, 87.2% were complete and only 1% were missing (Table 8). Additionally, coverage was also estimated by mapping the NGS reads and ONT reads to the assembled genome with BWA-mem2 (v2.2) (https://github.com/bwa-mem2/bwa-mem2) and minimap2, respectively. The coverage was calculated by SAMtools[69], indicating that 93.9% of the DNA short reads mapped to the assembled genome. Furthermore, the clean RNA reads were aligned back to the draft genome using HISAT2,with 99.96% of the uniquely mapped transcriptome reads suggesting comprehensive genome coverage. Given the existence of published sea buckthorn genomes, we also compared the gene structure between *F. hippophae* and other three sea buckthorn species using JCVI. Blocks with a span lower than 10 were filtered out, revealing a strong collinearity relationship (Fig. 4). In summary, the combined results from BUSCO, mapping coverage, and collinearity analysis demonstrate the high quality of our *F. hippophae* genome.

Code availability

No specific code was developed in this work.

Received: 2 January 2024; Accepted: 19 June 2024;
Published online: 13 July 2024

References

1. Zhao, J., Zhang, Z. H., Zhou, H. D., Bai, Z. F. & Sun, K. The study on Sea Buckthorn (Genus *Hippophae* L.) fruit reveals cell division and cell expansion to promote morphogenesis. *Plants.* **12**, 1005, https://doi.org/10.3390/plants12051005 (2023).
2. Andersson, S. C., Olsson, M. E. & Johansson, E. Carotenoids in sea buckthorn (*Hippophae rhamnoides* L.) berries during ripening and use of pheophytin a as a maturity marker. *J.Agric.Food Chem.* **57**, 250–258, https://doi.org/10.1021/jf802599f (2009).
3. Ruan, C. J., Rumpunen, K. & Nybom, H. Advances in improvement of quality and resistance in a multipurpose crop: sea buckthorn. *Crit Rev Biotechnol.* **33**, 126–144, https://doi.org/10.3109/07388551.2012.676024 (2013).
4. Mishra, K. P. *et al.* A comparative analysis of immunomodulatory potential of Seabuckthorn leaf extract in young and old mice. *Biomed. Aging Pathol.* **1**, 61–64, https://doi.org/10.1016/j.biomag.2011.02.001 (2011).
5. Andersson, S. C., Rumpunen, K., Johansson, E. & Olsson, M. E. Tocopherols and tocotrienols in sea buckthorn (*Hippophae rhamnoides* L.) berries during ripening. *J Agric Food Chem.* **56**, 6701–6706, https://doi.org/10.1021/jf800734v (2008).
6. Suomela, J. P., Ahotupa, M., Yang, B., Vasankari, T. & Kallio, H. Absorption of flavonols derived from sea buckthorn (*Hippophaë rhamnoides* L.) and their effect on emerging risk factors for cardiovascular disease in humans. *J Agric Food Chem.* **54**, 7364–7369, https://doi.org/10.1021/jf061889r (2006).
7. Zadernowski, R., Naczk, M., Czaplicki, S., Rubinskiene, M. & Szalkiewicz, M. Composition of phenolic acids in sea buckthorn (*Hippophae rhamnoides* L.) berries. *J Am Oil Chem Soc.* **82**, 175–179, https://doi.org/10.1007/s11746-005-5169-1 (2005).
8. Tanwar, H., Shweta, S. D., Singh, S. B. & Ganju, L. Anti-inflammatory activity of the functional groups present in *Hippophae rhamnoides* (Seabuckthorn) leaf extract. *Inflammopharmacology.* **26**, 291–301, https://doi.org/10.1007/s10787-017-0345-0 (2018).
9. Jiang, F. *et al.* Flavonoids from sea buckthorn inhibit the lipopolysaccharide-induced inflammatory response in RAW264.7 macrophages through the MAPK and NF-κB pathways. *Food Funct.* **8**, 1313–1322, https://doi.org/10.1039/c6fo01873d (2017).
10. Mishra, K. P., Chanda, S., Karan, D., Ganju, L. & Sawhney, R. C. Effect of Seabuckthorn (*Hippophae rhamnoides*) flavone on immune system: an *in-vitro* approach. *Phytother Res.* **22**, 1490–1495, https://doi.org/10.1002/ptr.2518 (2008).
11. Padwad, Y. *et al.* Effects of leaf extract of Seabuckthorn on lipopolysacchride induced inflammatory response in murine macrophages. *Int. Immunopharmacol.* **6**, 46–52, https://doi.org/10.1016/j.intimp.2005.07.015 (2006).
12. Zhou, J. Y., Zhou, S. W., Du, X. H., Zeng, S. Y. Protective effect of total flavonoids of seabuckthorn (*Hippophae rhamnoides*) in simulated high-altitude polycythemia in rats. *Molecules.* **17**, https://doi.org/10.3390/molecules171011585 (2012).
13. Maheshwari, D. T., Yogendra, K. M. S., Verma, S. K., Singh, V. K. & Singh, S. N. Antioxidant and hepatoprotective activities of phenolic rich fraction of Seabuckthorn (*Hippophae rhamnoides* L.) leaves. *Food Chem Toxicol.* **49**, 2422–2428, https://doi.org/10.1016/j.fct.2011.06.061 (2011).

14. Basu, M. *et al*. Anti-atherogenic effects of seabuckthorn (*Hippophaea rhamnoides*) seed oil. *Phytomedicine.* **14**, 770–777, https://doi.org/10.1016/j.phymed.2007.03.018 (2007).
15. Upadhyay, N. K., Kumar, R., Siddiqui, M. S. & Gupta, A. Mechanism of wound-healing activity of *Hippophae rhamnoides* L. leaf extract in experimental burns. *Evid Based Complement Alternat Med.* **2011**, 659705, https://doi.org/10.1093/ecam/nep189 (2009).
16. Zhuang, X. Y., Zhang, W., Pang, X. F. & Wu, W. B. Combined effect of total flavonoids from seed residues of *Hippophae rhamnoides* L. and zinc on advanced glycation end products-induced endothelial cell dysfunction. *Food Chem.* **133**, 905–911, https://doi.org/10.1016/j.foodchem.2012.02.001 (2012).
17. Wu, Z. *et al*. Genome of *Hippophae rhamnoides* provides insights into a conserved molecular mechanism in actinorhizal and rhizobial symbioses. *New Phytol.* **235**, 276–291, https://doi.org/10.1111/nph.18017 (2022).
18. Chen, M. *et al*. Chromosome-level genome assembly of Hippophae gyantsensis. *Sci Data.* **11**, 126, https://doi.org/10.1038/s41597-024-02909-w (2024).
19. Wang, R. *et al*. How to survive in the world's third poplar: Insights from the genome of the highest altitude woody plant, *Hippophae tibetana* (Elaeagnaceae). *Front Plant Sci.* **13**, 1051587, https://doi.org/10.3389/fpls.2022.1051587 (2022).
20. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* **17**, 333–351, https://doi.org/10.1038/nrg.2016.49 (2016).
21. Huddleston, J. *et al*. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685, https://doi.org/10.1101/gr.214007.116 (2017).
22. Huang, J. *et al*. The possible mechanism of Hippophae fructus oil applied in tympanic membrane repair identified based on network pharmacology and molecular docking. *J Clin Lab Anal.* **36**, e24157, https://doi.org/10.1002/jcla.24157 (2022).
23. Tunde, J., Vicas, L. G., Marian, E. & Vicas, S. L. A new natural antioxidant supplement-design and development. *Farmacia.* **64**, 135–142 (2016).
24. Costel, S. & Anamaria, S. Evaluation of polyphenolic fingerprints and antioxidant profiles of wild fruits. *J. Food Sci. Technol.* **51**, 1442–1440, https://doi.org/10.1111/ijfs.13111 (2016).
25. Jia, Q. *et al*. Rapid qualitative and quantitative analyses of anthocyanin composition in berries from the Tibetan Plateau with UPLC-quadruple-Orbitrap MS and their antioxidant activities. *Eur J Mass Spectrom (Chichester).* **26**, 301–308, https://doi.org/10.1177/1469066720926435 (2020).
26. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol Biol Rep.* **15**, 8–15, https://doi.org/10.1007/BF02772108 (1997).
27. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
28. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
29. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *nature communications* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
30. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics.* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
31. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
32. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259, https://doi.org/10.1186/s13059-015-0831-x (2015).
33. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scafolds. *Science.* **356**, 92–95, https://doi.org/10.1126/science.aal332 (2017).
34. Durand, N. C. *et al*. 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 (2016).
35. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
36. Zhang, R. G. *et al*. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res.* **9**, uhac017, https://doi.org/10.1093/hr/uhac017 (2022).
37. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics.* **36**, 4269–4275, https://doi.org/10.1093/bioinformatics/btaa519 (2020).
38. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, w265–w268, https://doi.org/10.1093/nar/gkm286 (2007).
39. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics.* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
40. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automic eukaryotic genome annotation with GeneMARK-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics.* **3**, Iqaa108, https://doi.org/10.1093/nargab/lqaa108 (2021).
41. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics.* **22**, 566, https://doi.org/10.1186/s12859-021-04482-0 (2021).
42. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinformatics.* **48**, 4.11.1–14.11.39, https://doi.org/10.1002/0471250953.bi0411s48 (2014).
43. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
44. Kent, W. K. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664, https://doi.org/10.1101/gr.229202 (2002).
45. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
46. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48, https://doi.org/10.1093/nar/28.1.45 (2000).
47. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
48. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462, https://doi.org/10.1093/nar/gkv1070 (2016).
49. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science.* **278**, 631–637, https://doi.org/10.1126/science.278.5338.631 (1997).
50. Madeira, F. *et al*. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279, https://doi.org/10.1093/nar/gkac240 (2022).
51. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **1**, D121–D124, https://doi.org/10.1093/nar/gki081 (2005).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591597 (2024).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591606 (2024).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591605 (2024).

55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591604 (2024).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591602 (2024).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591601 (2024).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591600 (2024).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591599 (2024).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591596 (2024).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591595 (2024).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591594 (2024).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591593 (2024).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591592 (2024).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591591 (2024).
66. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25591598 (2024).
67. *NCBIGenBank* https://identifiers.org/ncbi/insdc.gca:GCA_033030585.1 (2024).
68. Simão, F. A., Waterhouse, R. M., Loannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
69. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics*. **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).

Acknowledgements

This research was supported by Regional Science and Technology Collaborative Innovation Project of Shigatse Bureau of Science and Technology(QYXTZX-RKZ2021-07), the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502), Yunling Scholar Project to Yang Yongping, Regional Science and Technology Collaborative Innovation Project of Shigatse Bureau of Science and Technology(QYXTZX-RKZ2022-01), the Major Program of National Natural Science Foundation of China (31590820, 31590823), the National Natural Science Foundation of China (31601999 and 41771123), and the 13th Five-year Informatization Plan of Chinese Academy of Sciences,Grant No. XXH13506.

Author contributions

Tianmeng Liu and Yongping Yang conceived the study and supervised the project. Xingyu Yang Shujie Luo and Shihai Yang wrote the manuscript and participated in the data analysis. Ciren Duoji, Qianwen Wang, Zhiyu Chen, Danni Yang, Tianyu Yang, Xi Wan, Yunqiang Yang collected the samples, performed the figures drawing and upload the data. All authors have read, revised, and approved the final manuscript for submission.

Competing interests

The authors declear no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.L. or Y.Y.

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

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