Genome Survey of Sphallerocarpus gracilis Based on High-throughput Sequencing

Shiming Qi
College of Agriculture and Ecological Engineering
Chunmei Zhang
zhangcm@hxu.edu.cn
College of Agriculture and Ecological Engineering
Fang Yan
Key Laboratory of Hexi Corridor Resources Utilization of GanSu
Xifeng Zhang
College of Agriculture and Ecological Engineering
Gang Zhao
College of Agriculture and Ecological Engineering
Hai Song
Key Laboratory of Hexi Corridor Resources Utilization of GanSu
Ye Chen
College of Agriculture and Ecological Engineering
Zhenrong Liu
College of Agriculture and Ecological Engineering

Article

Keywords: Sphallerocarpus gracilis, High-thoughput Squencing, Genome size, K-mer analysis, Ploidy, Heterozygosity

Posted Date: January 24th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5782050/v1

Additional Declarations: No competing interests reported.

Abstract

Sphallerocarpus gracilis is a high-value medicinal and green health food product. The analysis of the genomic characteristic information of *S. gracilis* can lay a theoretical foundation for whole genome sequencing and molecular mechanism research of the biosynthesis of bioactive active ingredients. In this study, survey genome sequencing technology was employed to evaluate the genomic characteristics of *S. gracilis* using K-mer analysis, and smudgeplot analysis was used to evaluate its chromosome ploidy. The K-mer analysis results showed that the genome size of the sample was approximately 1,071 Mb, and the corrected genome size was 1,063 Mb. The heterozygosity rate, the proportion of repeat sequences, and GC content were determined 1.22%, 76.33%, and 35.70%, respectively. Based on the smudgeplot analysis, the maximum possible ploidy of the analyzed species was AB type, corresponding to a diploid plant. Blast analysis revealed *S. gracilis* to have a close relative relationship with Daucus carota (4.78%). In summary, the results indicate that the genome of *S.gracilis* is a complex and large genome with high heterozygosity and repetition and a large genome. This study provides a theoretical basis for future whole genome sequencing and related research.

Introduction

Sphallerocarpus gracilis, commonly known as Xiaoyeshan Red Radish and Huangfeng, is a single plant of *Sphallerocarpus*, belonging to the family *Apiaceae* [1]. It is scattered in the northwest, northeast, and north of China, particularly in Qilian Mountain and Yanzhi Mountain (Zhangye City, Gansu Province). It is rare to form a large regional population advantage in the growth of wild *S. gracilis* in Shandan County, Zhangye City [2]. The fleshy roots of *S. gracilis* are large and conical, and are used to cook porridge and various other dishes. *S. gracilis* grows in alpine humid grasslands and semi-steppe and mountain beaches at an altitude of 1700–3000 m. It is a non-polluting and non-toxic green food. *S. gracilis* is rich in more than 10 types of beneficial amino acids, vitamins A and vitamin C, Ca, P, Fe, Zn, and other minerals and trace elements [3]. It is known as “small ginseng” and is a natural and important medicinal and edible resource.

S. gracilis and its products are favored by consumers, yet its market supply is limited. This has led to excessive excavation by local farmers in spring and autumn, causing serious damage to wild resources. These actions, combined with the deterioration of the ecological environment and the predatory management of the local people, have resulted in a decline in the number of wild *S. gracilis* populations. At present, the functional genome, gene sequence information, and genetic background of endangered wild *S. gracilis* germplasm are scarce. Thus, research that adopts genomic resources is required for the protection of *S. gracilis* and to determine its genetic structure.

For non-model organisms, genome investigation and analysis are essential for functional gene mining and molecular mechanism research in the absence of genome data [4]. With the continuous development of sequencing technology, the K-monomeric unit (K-mer) method has become important in studying characteristics such as genome size, repeatability and heterozygosity [5-7], and has been applied to

species such as *Euphrasia* [8], *Aspalathus linearis* [9], *Reseda pentagyna* and *Reseda lutea* [10], and *Platostoma palustre* A.J.Paton [11]. The current research on *S. gracilis* focuses on polysaccharides function [12-14], biological activity composition [15], hepatoprotective effects [16,17], transcriptome sequencing analysis [18,19], etc., while studies on the genome of *S. gracilis* are lacking.

In this study, K-mer analysis was performed to evaluate the genome size, heterozygosity, repeatability, GC content, chromosome ploidy, and related species relationship with wild *S. gracilis* by genome survey sequencing technology. The results provide theoretical reference for genome assembly, molecular mechanism research of pharmacodynamic component synthesis, functional gene resource mining, and new drug development and innovation of *S. gracilis*.

Results

Morphological characteristics of S. gracilis

To investigate the genomic characteristics of wild *S. gracilis*, we collected plant samples from the nature reserve near Shandan Horse Farm three, Shandan County, Zhangye City, Gansu Province, China (Fig. 1A). Wild *S. gracilis* is a perennial herb with a plant height of 50–120 cm (Fig. 1B). The root is large and conical, the stem is round and multi-branched, the cauline leaves are 2–3 pinnately divided, and the umbellate inflorescence is small (Fig. 1B).

Genome sequencing and quality control of S. gracilis

Based on the second-generation Illumina NovaSeq sequencing platform, the genomic DNA of *S. gracilis* was sequenced to obtain 109.64 Gb raw sequencing data, resulting in 96.16 Gb quality sequencing data after filtering (Table 1). The distribution of each read sequence length over all sequences was 150 bp (Fig. 2), indicating that the sequencing quality was stable and accurate. An evaluation of the sequencing quality showed that the Q20 and Q30 values of *S. gracilis* were 98.09% and 94.74% (Table 1), respectively, revealing that the genomic data were reliable and could be used for the subsequent analysis. Most of the quality values of the sequencing data in the *S. gracilis* genome were greater than 35 (Fig. 3), indicating that read quality of the genome sequencing and the reliability of the sequencing results were high. The complementary bases of A and T, and C and G in the genome sequencing data *S. gracilis* were essentially the same, and the position base N was 0. The GC content of the genome was approximately 35.70% (Table 1). The GC distribution map showed that the GC content of the sequencing results followed a normal distribution (Fig. 4), indicating that there was no bias and that the data was not contaminated by exogenous species. In addition, the proportion of A, G, C, and T in each base position in the sequencing results was balanced (Fig. 4), further confirming the high reliability of the sequencing results. The GC content distribution exhibited a single peak, demonstrating normally (Fig. 5). The sequencing results were consistent with the GC content of all the genes expressed in the species. However, due to the low sequencing quality of the first few bases and the deviation of the DNA template

amplification, these bases exhibited large fluctuations, which is expected. In summary, the results prove that the genome survey of *S. gracilis* has obtained high-quality sequencing data.

Table 1
Survey sequencing data of the *Sphallerocarpus_gracilis* genome

Type	ReadNum	BaseCount (Gb)	ReadLength (bp)	Q20 (%)	Q30 (%)	GC Content (%)
raw	730,929,594	109.64	150	97.11	93.44	36.18
dedup	651,373,830	96.16	147	98.09	94.74	35.70

Evaluation of the genome size and heterozygosity of S. gracilis

K-mer analysis was employed to estimate the genome size, heterozygosity, and repeatability of the species, with a Kvalue of 17. The results of the K-mer analysis showed that the size of the *S. gracilis* genome was about 1,071 Mb, with a corrected genome size of 1,063 Mb (Table 2, Fig. 5). The heterozygosity rate was 1.22%, and the proportion of repeat sequences was 76.33% (Table 2, Fig. 6). The K-mer depth distribution revealed the presence of a peak at 1/2 of the main peak position (Fig. 6). This indicates that the heterozygosity of *S. gracilis* genome was high. Moreover, the K-mer curve showed obvious tailing (Fig. 6), suggesting that the content of repetitive sequences in *S. gracilis* genome was high. These results indicate that the *S. gracilis* genome belongs to a species with a large genome, high heterozygosity, and numerous repeat sequences.

Table 2
Genomic characteristics of *Sphallerocarpus_gracilis* (K = 17)

K-mer number	K-mer Depth	Genome Size (bp)	Revised Genome Size (bp)	Heterozygous Ratio (%)	Repeat (%)
85,719,933,597	80	1,071,499,170	1,063,284,596	1.22	76.33

Determination of the S. gracilis ploidy using smudgeplots

The genomic structure and ploidy of *S. gracilis* were analyzed using smudgeplots. The peak value of the AB ploidy of the *S. gracilis* genome was determined as 0.63 (Table 3) and followed a single peak curve (Fig. 7), it to be a heterozygous diploid. This is consistent with the reported genome ploidy of carrot (*Daucus carota*).

Table 3
Smudgeplot analysis results of the *Sphallerocarpus_gracilis* genome

peak	kmers	kmers (%)	summit B / (A + B)	summit A + B
AB	16,443,246	0.63	0.48	78.73
AABB	4,361,639	0.17	0.49	162.2
AAB	3,437,341	0.13	0.32	114.51
AAAABB	980,872	0.04	0.32	233.75
AAAAB	706,114	0.03	0.19	209.9

Relative species of S. gracilis

The NCBI database was used to compare with the nucleic acid sequence of *S. gracilis*. The read matching rates of *Daucus carota* (4.78%), *Anthriscus Hoffm* (1.16%), *Hedera helix* (1.10%), *Ostericum sieboldii* (1.00%), and *Apium graveolens* (0.97%) were relatively high (Table 4). This suggests that *Daucus carota* is a relative species of *S. gracilis*. No abnormal results were observed with other species such as animals in the comparison data.

Table 4
Comparison of high-quality data NT library of *Sphallerocarpus gracilis* plants

Genus	Kingdom	Blast number	Hit number	Percentage of hits (%)	Percentage of extraction (%)	Median identity (%)
Daucus	Viridiplantae	956	3,659	26.13	4.78	91.25
Anthriscus	Viridiplantae	232	3,659	6.34	1.16	99.31
Hedera	Viridiplantae	220	3,659	6.01	1.10	84.91
Ostericum	Viridiplantae	199	3,659	5.44	1.00	99.33
Apium	Viridiplantae	194	3,659	5.30	0.97	100.00
Ferula	Viridiplantae	152	3,659	4.15	0.76	99.16
Pulicaria	Viridiplantae	80	3,659	2.19	0.40	83.99
Heracleum	Viridiplantae	69	3,659	1.89	0.34	100.00
Torilis	Viridiplantae	63	3,659	1.72	0.32	98.67
Pternopetalum	Viridiplantae	52	3,659	1.42	0.26	98.67
Solanum	Viridiplantae	46	3,659	1.26	0.23	84.62
Sphallerocarpus	Viridiplantae	41	3,659	1.12	0.21	99.33
Zizia	Viridiplantae	40	3,659	1.09	0.20	100.00
Saposhnikovia	Viridiplantae	38	3,659	1.04	0.19	99.33
Panax	Viridiplantae	32	3,659	0.87	0.16	100.00
Gossypium	Viridiplantae	32	3,659	0.87	0.16	86.07
Chenopodium	Viridiplantae	29	3,659	0.79	0.14	88.36
Angelica	Viridiplantae	28	3,659	0.77	0.14	99.33
Vigna	Viridiplantae	27	3,659	0.74	0.14	83.19
Ipomoea	Viridiplantae	27	3,659	0.74	0.14	84.21
Medicago	Viridiplantae	26	3,659	0.71	0.13	85.83
Cuminum	Viridiplantae	26	3,659	0.71	0.13	98.67
Impatiens	Viridiplantae	24	3,659	0.66	0.12	84.92
Geum	Viridiplantae	21	3,659	0.57	0.10	90.20
Scutellaria	Viridiplantae	21	3,659	0.57	0.10	84.93
Peucedanum	Viridiplantae	20	3,659	0.55	0.10	99.67

Genus	Kingdom	Blast number	Hit number	Percentage of hits (%)	Percentage of extraction (%)	Median identity (%)
Musa	Viridiplantae	18	3,659	0.49	0.09	90.34
Hansenia	Viridiplantae	18	3,659	0.49	0.09	98.67
Ballota	Viridiplantae	18	3,659	0.49	0.09	83.21
Osmorhiza	Viridiplantae	17	3,659	0.46	0.08	98.67
Clematis	Viridiplantae	17	3,659	0.46	0.08	100.00
Hymenidium	Viridiplantae	16	3,659	0.44	0.08	99.33
Cicer	Viridiplantae	16	3,659	0.44	0.08	88.83
Fraxinus	Viridiplantae	16	3,659	0.44	0.08	83.72
Physospermopsis	Viridiplantae	16	3,659	0.44	0.08	99.67
Bupleurum	Viridiplantae	16	3,659	0.44	0.08	99.33
Ligusticum	Viridiplantae	15	3,659	0.41	0.07	98.01
Sinocarum	Viridiplantae	15	3,659	0.41	0.07	98.66
Arachis	Viridiplantae	15	3,659	0.41	0.07	96.00
Meeboldia	Viridiplantae	14	3,659	0.38	0.07	98.01
Theobroma	Viridiplantae	14	3,659	0.38	0.07	86.15

Discussion

The analysis of whole genome information based on sequencing technology lays a foundation for the study of plant origin, evolution, reproduction, development, resistance and sex regulation. Considering the large differences in the heterozygosity and repeat content of the genomes of different species, it is important to determine the genome characteristics before whole genome sequencing. A genome survey is a low-depth sequencing method based on small fragment libraries that can quickly obtain the genome size, heterozygosity, and weight by K-mer analysis [7,20]. The analysis of filtered high-quality data revealed that the heterozygosity of the *S. gracilis* genome was 1.22% and the proportion of repetitive sequences was 76.33% (Table 2). This indicates that the *S. gracilis* genome is complex with high repetition and high heterozygosity. A heterozygosity exceeding 0.8% is typically considered to be high [21]. A heterozygosity increases the difficulty of genome-wide assembly and interferes with the estimation progress of K-mer, making the estimation result deviate from the actual size [22]. The GC content of the *S. gracilis* genome was 35.70% (Table 1), which is within the acceptable range of 25–65%, indicating the feasibility of the genome assembly [23]. Based on the K-mer and smudgeplot analysis, the genome size of *S. gracilis* was

estimated to be 1,063 Mb, indicating that it is an AB-type diploid plant (Tables 2 and 3, Fig. 7). This is consistent with previous research that reported the karyotype of *S. gracilis* to be 2n = 20 [24]. Moreover, most *Apiaceae* plants are diploid [25,26], such as *Daucus carota* (2n = 18) [27,28], *Coriandrum sativum* (2n = 22)[29], and *Apium graveolens* (2n = 22) [30], with genome sizes of 421 Mb, 2130 Mb, and 2.21 Gb, respectively. Our results suggest that *S. gracilis* is a species with high repetition, high heterozygosity, and a large genome. The genomic characteristic data of *S. gracilis* obtained in this study lay a foundation for subsequent genome sequencing.

The study of genomics can reveal the genetic diversity, genome evolution, and gene function of species. The phylogenetic tree can directly show the genetic relationship and evolution process [31]. We used high-quality reads to compare NCBI nucleic acid data. The similarity of plants included in the NT library did not exceed 10%, only *Daucus carota* was 4.78% (Table 4). This may be attributed to the limited sequence information of *S. gracilis* and its approximate species included in the NT library. The morphological characteristics of *S. gracilis* are similar to those of carrot (*Daucus carota*) plants (Fig. 1B), and *S. gracilis* is also known as the “small red carrot”. Moreover, the genomes of *S. gracilis* and carrot are different by a factor of just 2.5. Thus, the advanced research results of carrot [28,32,33] can provide a reference for further research on *S. gracilis*.

With the rapid development of sequencing and analysis techniques, the genomes of large, highly repeated, and highly heterozygous species have been sequenced at the fine chromosome level [34–38]. The publication of these high-quality reference genomes provides a basis for the study of the origin and evolution of important economic plants, the protection and utilization of germplasm resources, the molecular mechanism of important component anabolism, and the breeding of new varieties. It also provides a reference for whole genome sequencing and assembly strategies of complex genome species. In this study, the genome size, chromosome ploidy and related species of *S. gracilis* were estimated by K-mer analysis. This provides a basis for the subsequent development of the fine mapping of the whole genome of *S. gracilis*.

Materials and methods

Plant materials and genome DNA extraction

In January 2024, wild *S. gracilis* was collected with soil from three fields of Junmachang, Shandan County, Zhangye City, Gansu Province (101.05° N, 38.32° W) and brought back to the laboratory of Hexi University (Fig. 1A). According to the conventional cultivation method, the root segment of *S. gracilis* was planted in a flowerpot with a diameter of the bottom of 18 cm, a diameter of the upper of 30 cm, and a depth of 38 cm. The potted soil comprised equal amounts of sterilized nutrient soil with a volume of 2/3 of the flowerpot. After one month of plant growth, young leaves were selected, frozen in liquid nitrogen, and stored in a refrigerator at − 80°C. According to the manufacturer’s instructions, total genomic DNA was extracted from the young leaf tissues using the SteadyPure Plant Genomic DNA

Extraction Kit (Accurate Biotechnology, Co., Ltd). The quality, purity and concentration of DNA samples were detected by 1% agarose electrophoresis and a Nanodrop2000 Spectrophotometer (Thermo Scientific, USA).

Library preparation and sequencing

The genome survey was completed by Wuhan OneMore Technology Co., Ltd. A library of 300–400 bp fragments was constructed from the DNA sample. The DNA fragments were subjected to end repair, 3' A tailing and ligation with adaptors [39]. Double-end (PE, paired-end, 150) sequencing was then performed on the constructed library based on the DNBseq sequencing platform.

Quality control of sequencing data

The raw data obtained by sequencing were filtered by FASTQC v0.12.0 [40] to obtain high-quality data (clean data) for the analysis of the GC content, heterozygosity, and genome size. For the analysis, the adaptor sequences of reads were removed [41], the inaccurate bases at both ends of the reads were cut off, and five bases at the left and right ends were cut off. Moreover, reads containing more than 10% N were removed and read pairs with more than 20% of the base mass fraction less than 20 in a read were discarded.

Estimation of the S. gracilis genome size using K-mer analysis

Based on the clean data, the K-mer method [20] was employed to estimate the genome size of *S. gracilis*. The K value was set as 17 and the K-mer of the four bases in A, T, C, and G was counted. The Lander–Waterman algorithm was used to calculate the K-mer total and depth. The K-mer curve frequency distribution was drawn based on the calculated K-mer. The K-mer depth C value was obtained using the curve and the genome size was estimated.

Ploidy estimation of S. gracilis

Smudgeplot v0.4.0 [6] was used to estimate ploidy levels of *S. gracilis* from modified reads generated by the default settings of MECAT v2.0 [42]. Smudgeplot extracts heterozygous K-mer pairs from the K-mer database of sequencing data and trains heterozygous K-mer pairs. By comparing the total number and relative coverage of K-mer pairs, the number of heterozygous K-mers pairs was counted to analyze the genome structure.

Comparative analysis of near-source species of S. gracilis

To study the similarity between *S. gracilis* and its related species, we randomly selected 10,000 single-end reads data from the filtered high-quality data, and compared them with the NCBI nucleotide database (NT library, July 4, 2024) using Blast.

Declarations

Competing interests

The authors have no conflict of interest declaring.

Permissions Statement

We have obtained permission or authority for the collection, sequencing and related research work of *S. gracilis* plant materials. The wild *S. gracilis* has been deposited in the Herbarium of Agricultural and Ecological Engineering College of Hexi University. The wild *S. gracilis* materials were identified by Dr. C.Z. and Dr. S.Q.

Author Contribution

S.Q. and C.Z. planned and designed the research. Y.C. and R.Z. collected plant materials. S.Q., C.Z., F.Y., Z.X., G.Z., and H.S. performed the experiments. S.Q. and C.Z. drafed and revised the manuscript. All the authors reviewed and approved the manuscript.

Acknowledgements

We are grateful to the professional editors of Charlesworth Author Services for critical reading and revision of the manuscript. This work was supported by National Natural Science Foundation of China (No. 32160745); Natural Science Foundation of Gansu Province (No.22JR5RG566).

Data Availability

Sequence data that support the findings of this study have been deposited in the NCBI Sequence Read Archive (SRA) on January 17,2025 with the primary accession code PRJNA1211825, entitled Sphallerocarpus gracilis Genome sequencing and assembly.

References

1. ZY, W., AM, L., YC, T., ZD, C. & DZ, L. The families and genera of Angiosperms in China a comprehensive analysis. (Beijing: Science Press, 2003).
2. Huixian, J., Qing, Z., Xiangqing, Y., Zhenxia, Z. & Mingyan, Z. Studies on distribution and content of trace elements of Shandan Huangshen. *Acta Botanica Boreali-Occidentalia Sinica***21**, 188–190, doi:https://doi.org/1000-4025-(2001)01-0188-03 (2001).

3. Ye, C., Tianren, C. & Guanghong, L. Study on technology of *Sphallerocerpus gracilis* series food. *Food Science and Technology***11**, 96–97 + 100, doi:https://doi.org/10.13684/j.cnki.spkj.2003.11.033 (2003).
4. Hong, L. *et al.* Construction and analysis of telomere-to-telomere genomes for 2 sweet oranges: Longhuihong and Newhall (*Citrus sinensis*). *GigaScience***13**, doi:https://doi.org/10.1093/gigascience/giae084 (2024).
5. Kim, J. H. *et al.* Estimation of the genome sizes of the chigger mites *Leptotrombidium pallidum* and *Leptotrombidium scutellare* based on quantitative PCR and k-mer analysis. *Parasites & Vectors***7**, 279, doi:https://doi.org/10.1186/1756-3305-7-279 (2014).
6. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.***11**, 1432, doi:https://doi.org/10.1038/s41467-020-14998-3 (2020).
7. Karikari, B., Lemay, M.-A. & Belzile, F. *k*-mer-based genome-wide association studies in plants: advances, challenges, and perspectives. *Genes***14**, 1439, doi:https://doi.org/10.3390/genes14071439 (2023).
8. Becher, H., Sampson, J. & Twyford, A. D. Measuring the invisible: the sequences causal of genome size differences in Eyebrights (*Euphrasia*) revealed by k-mers. *Front. Plant Sci.***13**, 818410, doi:https://doi.org/10.3389/fpls.2022.818410 (2022).
9. Mgwatyu, Y., Stander, A. A., Ferreira, S., Williams, W. & Hesse, U. Rooibos (*Aspalathus linearis*) genome size estimation using flow cytometry and k-mer analyses. *Plants***9**, 270, doi:https://doi.org/10.3390/plants9020270 (2020).
10. Al-Qurainy, F. *et al.* Estimation of genome size in the endemic species *Reseda pentagyna* and the locally rare species *Reseda lutea* using comparative analyses of flow cytometry and K-mer approaches. *Plants (Basel, Switzerland)***10**, doi:https://doi.org/10.3390/plants10071362 (2021).
11. Zhao, Z. *et al.* Genome survey sequencing and characterization of simple sequence repeat (SSR) markers in *Platostoma palustre* (Blume) A.J.Paton (Chinese mesona). *Sci. Rep.***12**, doi:https://doi.org/10.1038/s41598-021-04264-x (2022).
12. Guo, J., Liu, Q., Wang, C., Shi, J. & Zhang, J. A polysaccharide isolated from *Sphallerocarpus gracilis* protects PC12 cells against hydrogen peroxide-induced injury. *Int. J. Biol. Macromol.***129**, 1133–1139, doi:https://doi.org/10.1016/j.ijbiomac.2018.11.208 (2019).
13. Li, Y. *et al.* Extraction, purification, characterization, and bioactive properties of polysaccharides from *Sphallerocarpus gracilis*. *Starch-Starke***73**, doi:https://doi.org/10.1002/star.202100082 (2021).
14. Xu, Y. *et al.* Sulfated modification of the polysaccharide from *Sphallerocarpus gracilis* and its antioxidant activities. *Int. J. Biol. Macromol.***87**, 180–190, doi:https://doi.org/10.1016/j.ijbiomac.2016.02.037 (2016).
15. Shihan, B. *et al.* Variation in the chemical composition and biological activities of *Sphallerocarpus gracilis* stems and leaves among different harvesting time. *Science and Technology of Food Industry***42**, 27–42, doi:https://doi.org/10.13386/j.issn1002-0306.2020030315 (2021).

16. Lu, Y. *et al.* Protective effect of free phenolics from *Lycopus lucidus* Turcz. root on carbon tetrachloride-induced liver injury in vivo and in vitro. *Food & Nutrition Research***62**, doi:https://doi.org/10.29219/fnr.v62.1398 (2018).
17. Guo, J. *et al. Sphallerocarpus gracilis* polysaccharide protects pancreatic β-cells via regulation of the bax/bcl-2, caspase-3, pdx-1 and insulin signalling pathways. *Int. J. Biol. Macromol.***93**, 829–836, doi:https://doi.org/10.1016/j.ijbiomac.2016.08.083 (2016).
18. Chunmei, Z. *et al.* Sequencing and analysis of transcriptome on regulating gene expression of *Sphallerocarpus gracilis* in Qilian Mountains under drought stress. *Agricultural Research in the Arid Areas***42**, 68–79, doi:https://doi.org/10.7606/j.issn.1000-7601.2024.03.08 (2024).
19. Chunmei, Z., Fang, Y., Hai, S., Xifeng, Z. & Ye, C. Sequencing and bioinformatic analysis for transcriptome of Shandan *Sphallerocarpus gracills* leaf. *Journal of Guangxi Normal University (Natural Science Edition)***40**, 247–256, doi:https://doi.org/10.16088/j.issn.1001-6600.2021080902 (2022).
20. Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology***35**, 62–67, doi:https://doi.org/10.48550/arXiv.1308.2012 (2013).
21. Li, G. Q. *et al.* Genome survey and SSR analysis of *Apocynum venetum*. *Biosci. Rep.***39**, doi:https://doi.org/10.1042/bsr20190146 (2019).
22. Zhou, P. *et al.* A first insight into the genomic background of *Ilex pubescens* (Aquifoliaceae) by flow cytometry and genome survey sequencing. *BMC Genomics***24**, 270, doi:https://doi.org/10.1186/s12864-023-09359-5 (2023).
23. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.***12**, R18, doi:https://doi.org/10.1186/gb-2011-12-2-r18 (2011).
24. Dongli, Z., Zhong, H., Wei, C., Xiaowen, J. & Lunshan, W. Karyotype analysis of *Sphallerocarpus gracilis* and *Trif olium pratense* from China. *Acta Botanica Boreali-Occidentalia Sinica***21**, 1026–1030, doi:https://doi.org/1000-4025-(2001)05-1026-05 (2001).
25. Kumar, G. & Pandey, A. Selfish genetic drive of b chromosomes in diploid and autotetraploid coriander (*Coriandrum sativum* L.). *Cytology and Genetics***56**, 466–474, doi:https://doi.org/10.3103/S0095452722050073 (2022).
26. O'LEARY, N., I. CALVIÑO, C., GREIZERSTEIN, E., MARTÍNEZ, S. & POGGIO, L. Further cytogenetical studies on diploid and polyploid species of *Eryngium* L. (Saniculoideae, Apiaceae) from Argentina. *Hereditas***140**, 129–133, doi:https://doi.org/10.1111/j.1601-5223.2004.01795.x (2004).
27. Iorizzo, M. *et al.* A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.***48**, 657–666, doi:https://doi.org/10.1038/ng.3565 (2016).
28. Wang, Y.-H. *et al.* Telomere-to-telomere carrot (*Daucus carota*) genome assembly reveals carotenoid characteristics. *Hortic. Res.***10**, doi:https://doi.org/10.1093/hr/uhad103 (2023).
29. Song, X. *et al.* Deciphering the high-quality genome sequence of coriander that causes controversial feelings. *Plant Biotechnol. J.***18**, 1444–1456, doi:https://doi.org/10.1111/pbi.13310 (2020).

30. Li, M.-Y. *et al.* The genome sequence of celery (*Apium graveolens* L.), an important leaf vegetable crop rich in apigenin in the *Apiaceae* family. *Hortic. Res.***7**, 9, doi:https://doi.org/10.1038/s41438-019-0235-2 (2020).
31. Liu, Y. *et al.* Chromosome-level genome of putative autohexaploid *Actinidia deliciosa* provides insights into polyploidisation and evolution. *Plant J.***118**, 73–89, doi:https://doi.org/10.1111/tpj.16592 (2024).
32. Coe, K. *et al.* Population genomics identifies genetic signatures of carrot domestication and improvement and uncovers the origin of high-carotenoid orange carrots. *Nat. Plants***9**, 1643–1658, doi:https://doi.org/10.1038/s41477-023-01526-6 (2023).
33. Duan, A.-Q., Deng, Y.-J., Liu, H., Xu, Z.-S. & Xiong, A.-S. An anthocyanin activation gene underlies the purple central flower pigmentation in wild carrot. *Plant Physiol.***196**, 1147–1162, doi:https://doi.org/10.1093/plphys/kiae391 (2024).
34. Song, X. *et al.* Coriander genomics database: a genomic, transcriptomic, and metabolic database for coriander. *Hortic. Res.***7**, 55, doi:https://doi.org/10.1038/s41438-020-0261-0 (2020).
35. Yang, Z. *et al.* The chromosome-scale high-quality genome assembly of *Panax notoginseng* provides insight into dencichine biosynthesis. *Plant Biotechnol. J.***19**, 869–871, doi:https://doi.org/10.1111/pbi.13558 (2021).
36. Han, X. *et al.* The chromosome-level genome of female ginseng (*Angelica sinensis*) provides insights into molecular mechanisms and evolution of coumarin biosynthesis. *Plant J.***112**, 1224–1237, doi:https://doi.org/10.1111/tpj.16007 (2022).
37. Niu, S. *et al.* The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell***185**, 204–217.e214, doi:https://doi.org/10.1016/j.cell.2021.12.006 (2022).
38. Wang, Z. *et al.* Chromosome-level genome assembly of *Cnidium monnieri*, a highly demanded traditional Chinese medicine. *Sci. Data.***11**, 667, doi:https://doi.org/10.1038/s41597-024-03523-6 (2024).
39. Khan, M. W., Habibi, N., Shaheed, F. & Mustafa, A. S. Draft genome sequences of five clinical strains of *Brucella melitensis* isolated from patients residing in Kuwait. *Genome Announc.***4**, e01144–e01116, doi:https://doi.org/10.1128/genomea.01144-16 (2016).
40. Andrews, S. Babraham Bioinformatics -FastQC A quality control tool for high throughput sequence data. *Available at www.bioinformatics.babraham.ac.uk/projects/fastqc/.*, Accessed June 27, 2015 (2013).
41. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics***34**, i884-i890, doi:https://doi.org/10.1093/bioinformatics/bty560 (2018).
42. Xiao, C.-L. *et al.* MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods***14**, 1072–1074, doi:https://doi.org/10.1038/nmeth.4432 (2017).

Figures

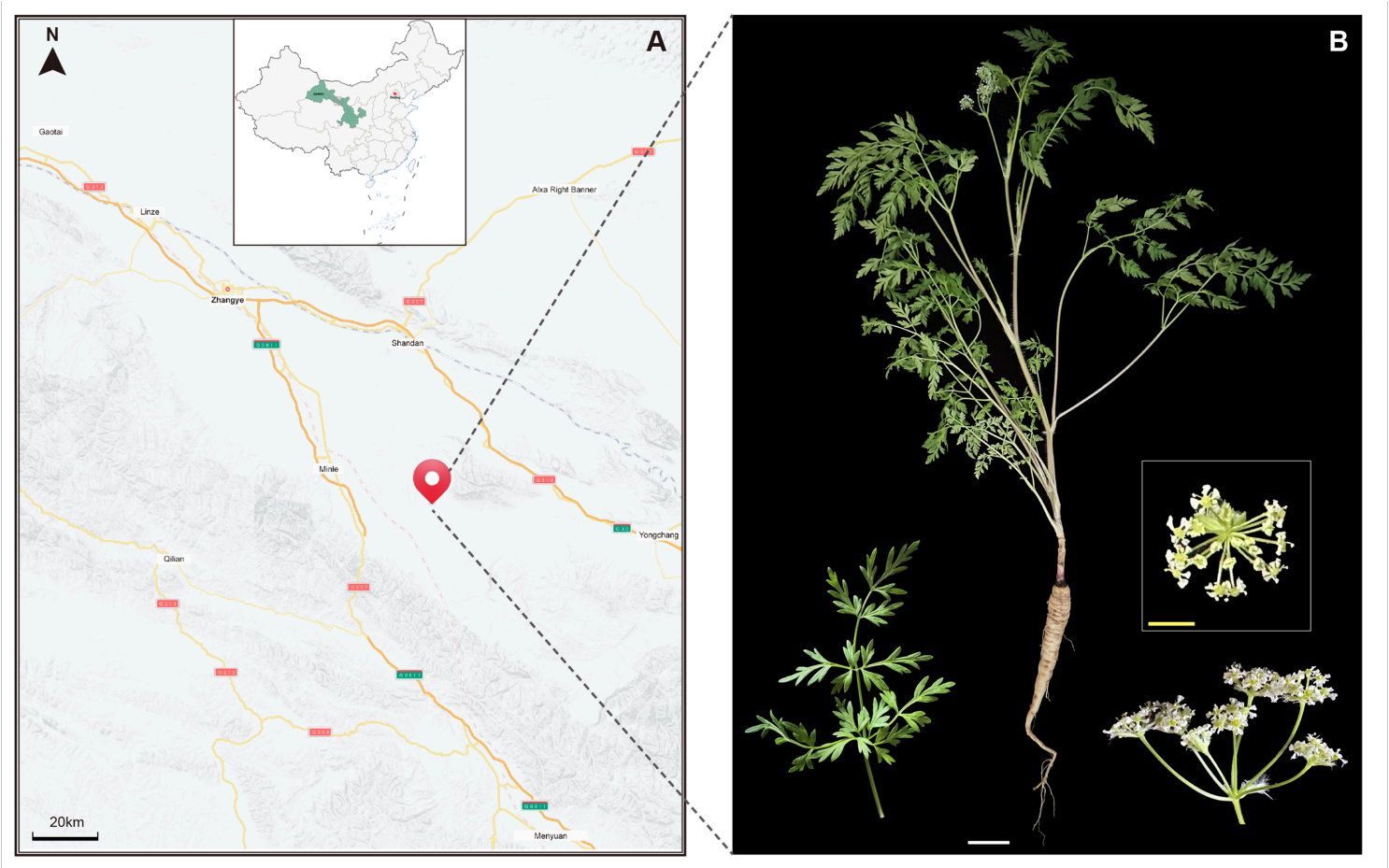

Figure 1

Location and plant morphological characteristics of *Sphallerocarpus gracilis*. (A) Location of *S. gracilis* in the nature reserve of Shandan County, Zhangye City. (B) Morphological characteristics of a single *S. gracilis* plant growing in Shandan County Nature Reserve. The white and yellow scales denote 20 mm and 5 mm, respectively.

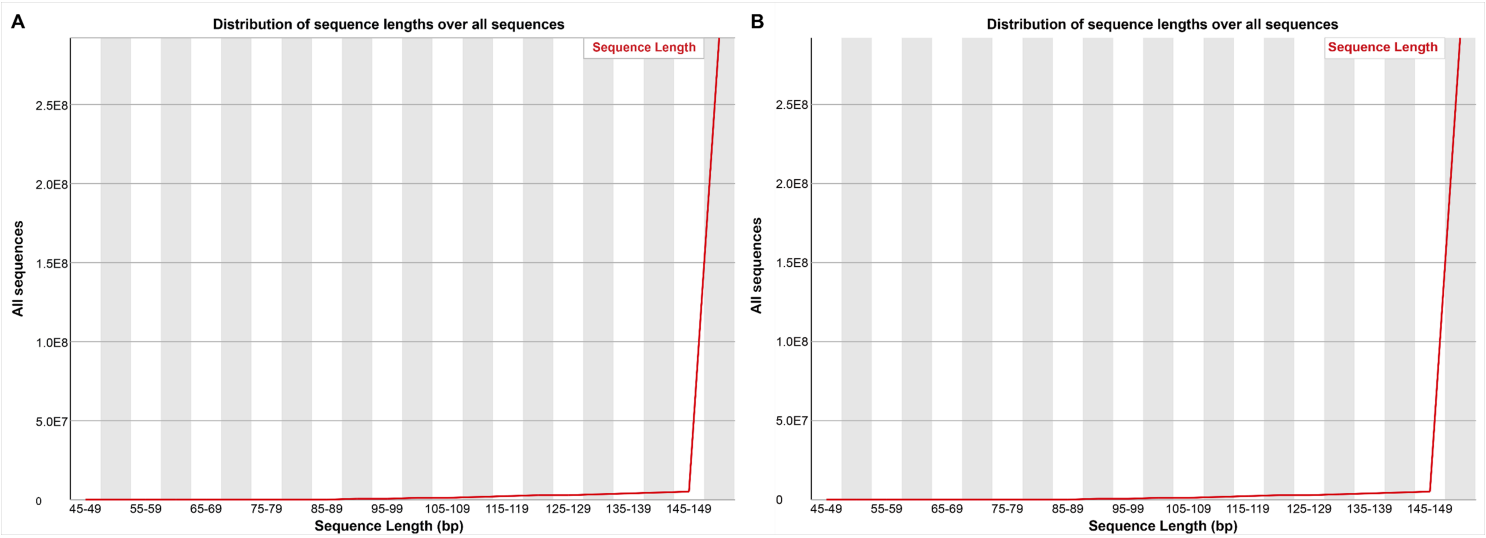

Figure 2

Sequence length distribution of all *Sphallerocarpus gracilis* sequences. Sequence length distribution of the first-end (A) and other-end (B) sequencing reads of the double-end sequencing sequence.

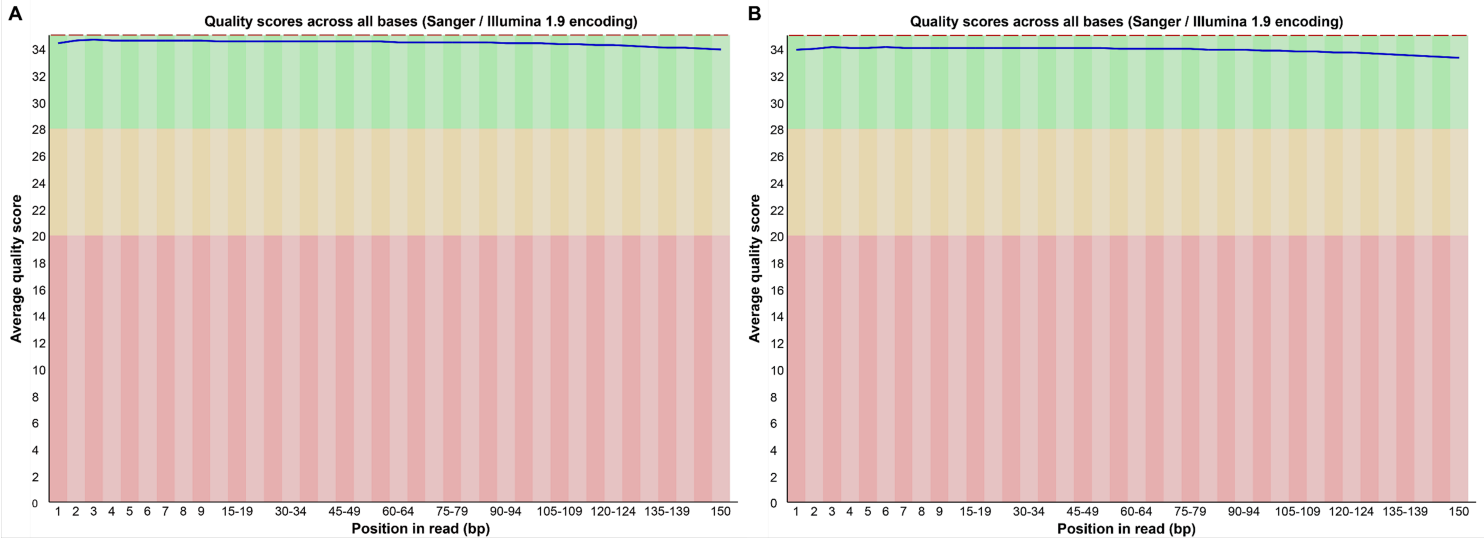

Figure 3

Proportion of bases T, C, A, and G in the *Sphallerocarpus gracilis* genome. A. Proportion of bases T, C, A, and G at each base position in the first read; B: Proportion of bases T, C, A, and G at each base position in the second read.

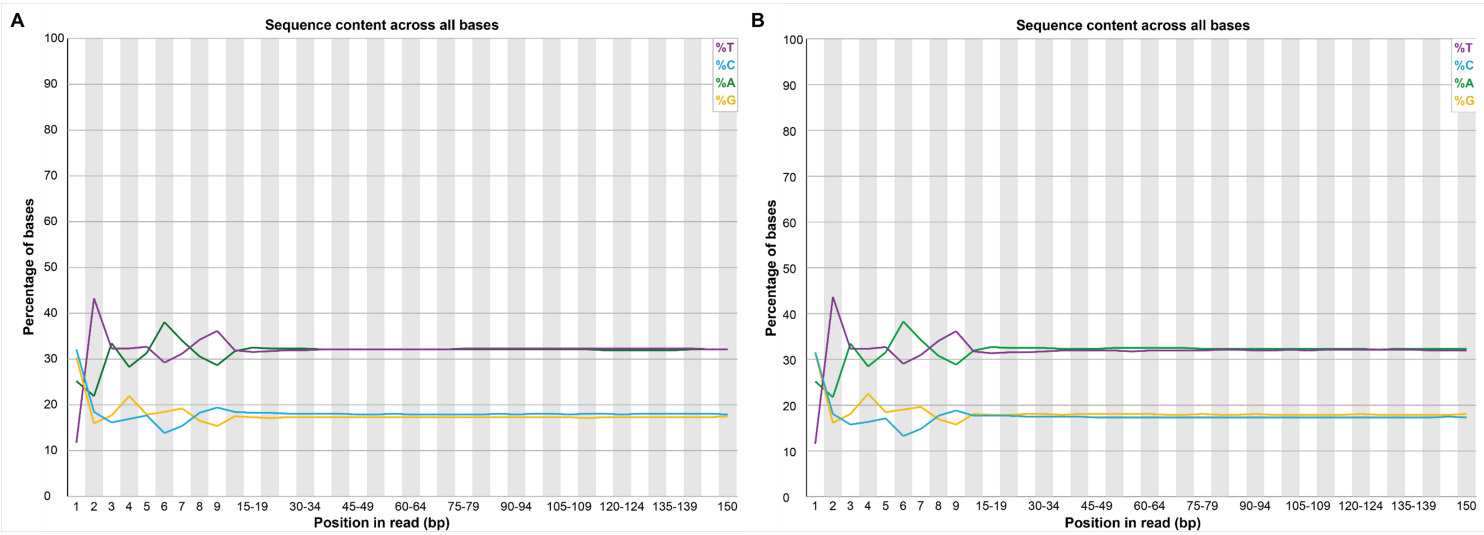

Figure 4

Proportion of bases T, C, A, and G in the *Sphallerocarpus gracilis* genome. A. Proportion of bases T, C, A, and G at each base position in the first read; B: Proportion of bases T, C, A, and G at each base position in the second read.

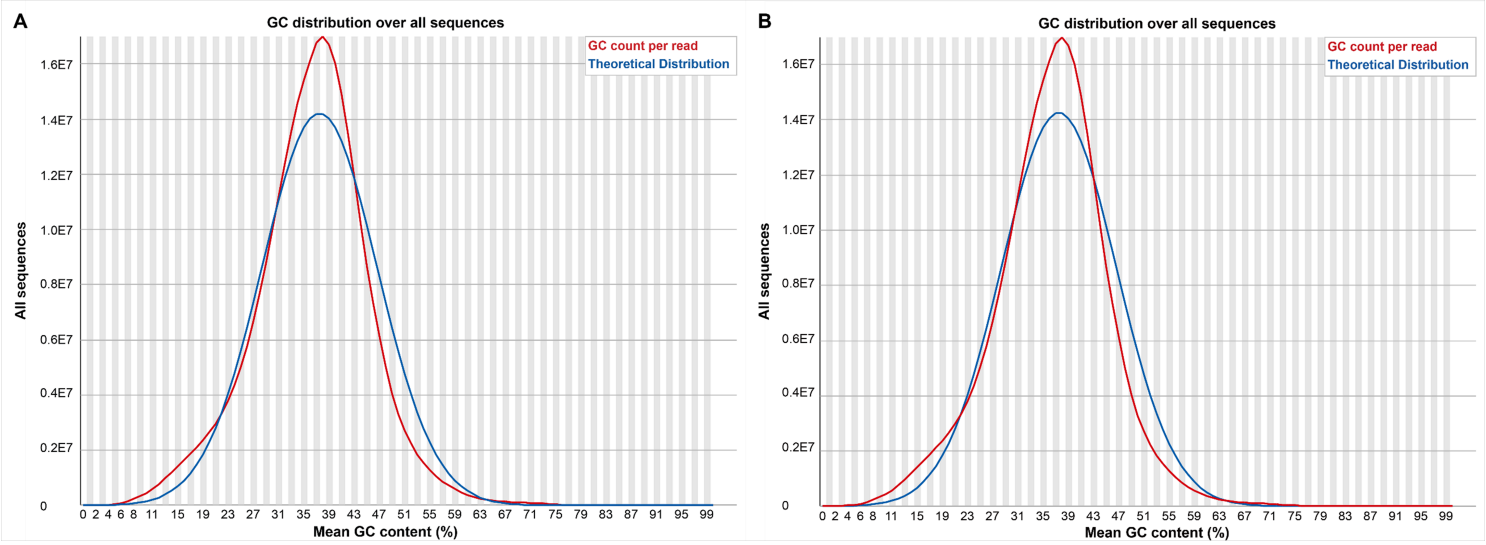

Figure 5

Distribution structure of the GC content in the *Sphallerocarpus gracilis* genome. Distribution structure of the GC content of the first-end (A) and other-end (B) sequencing reads of the double-end sequencing sequence.

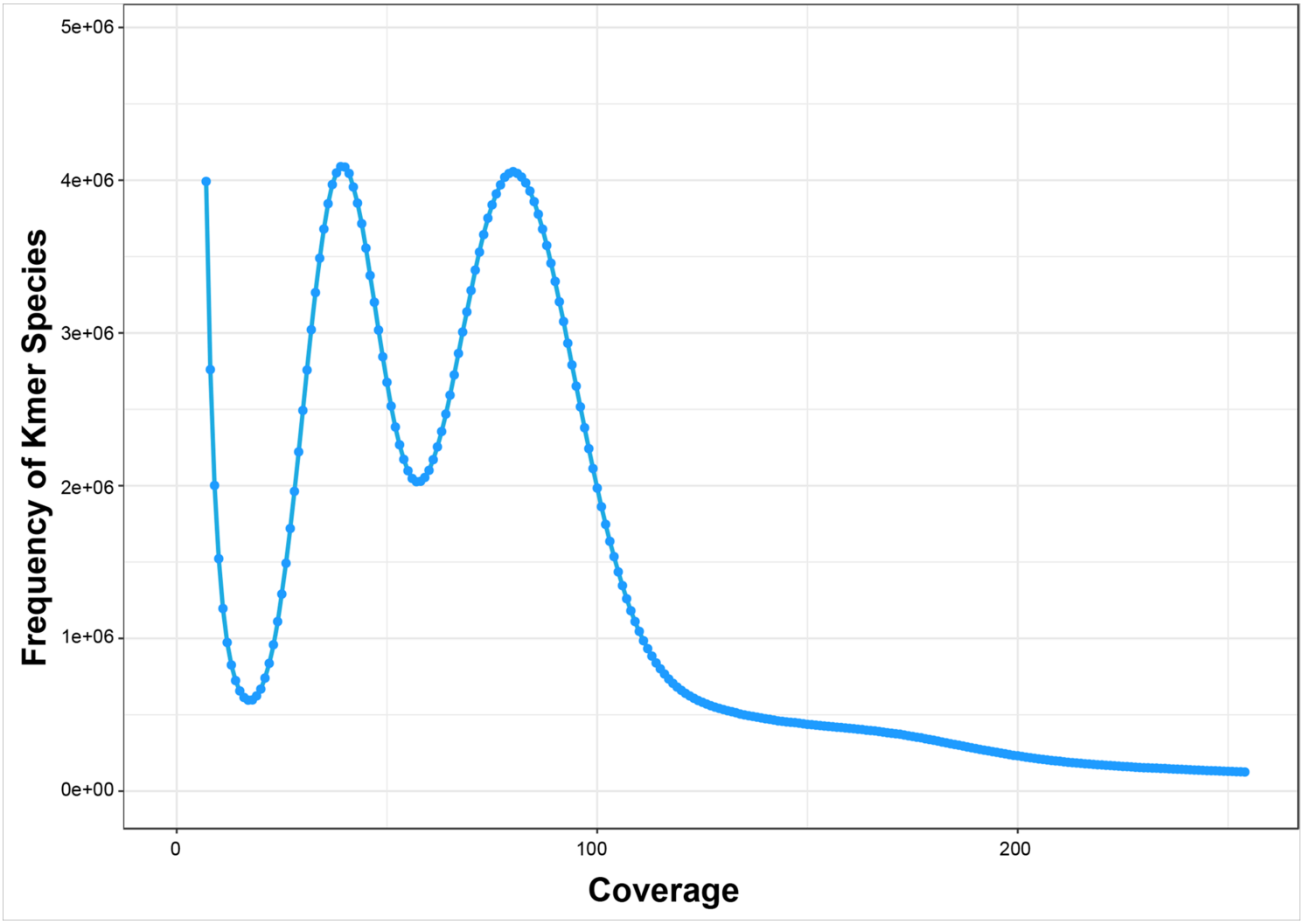

Figure 6

K-mer species frequency and depth distribution of *Sphallerocarpus gracilis*. x- and y-axis represent the coverage and frequency of the K-mer species, respectively.

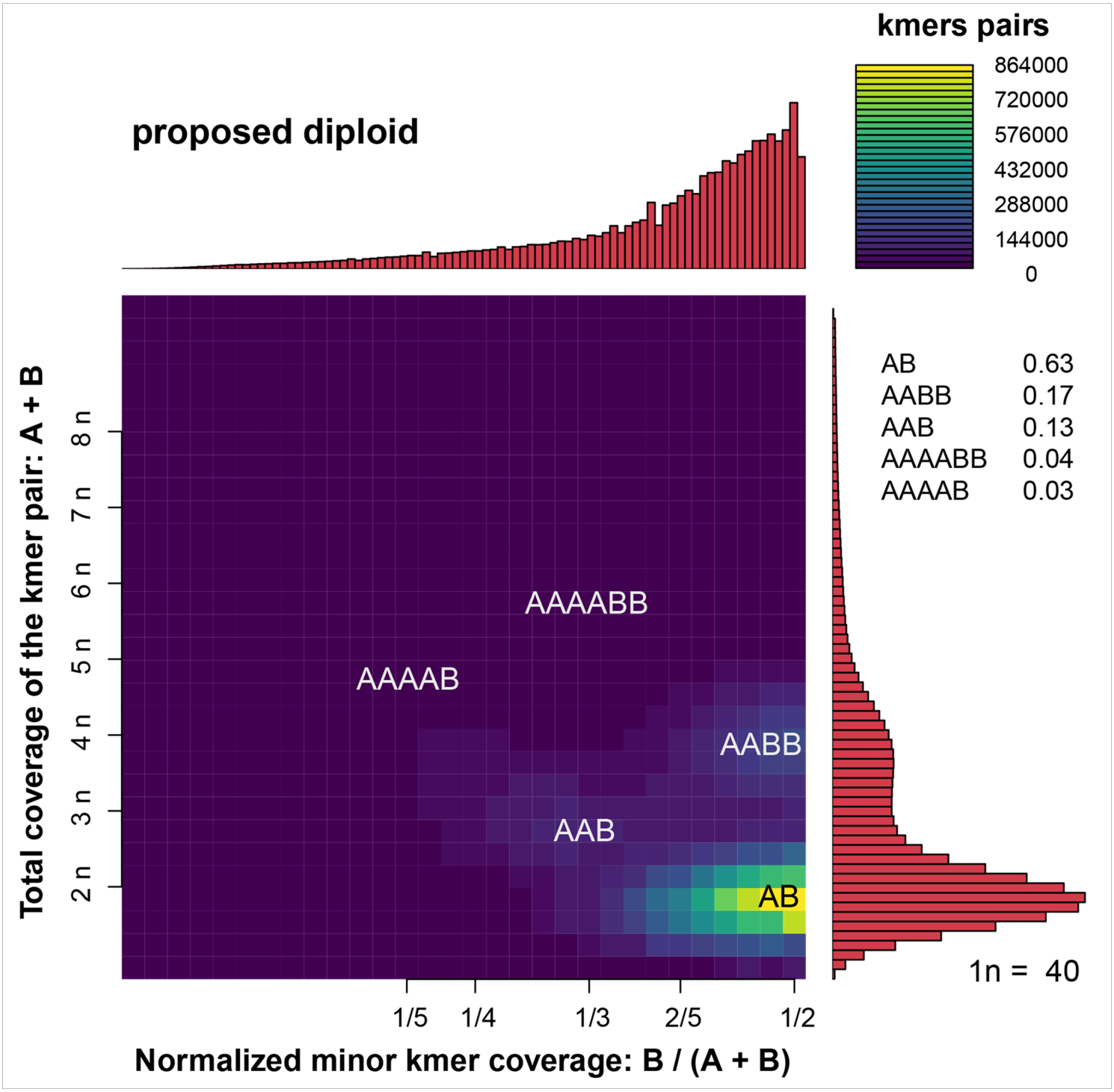

Figure 7

Smudgeplots for the diploid *Sphallerocarpus gracilis* based on real datasets.