The autotetraploid potato genome provides insights into highly heterozygous species

Fang Wang 1,2,3,4,*, Zhiqiang Xia 5,6,†, Meiling Zou 5,6,†, Long Zhao 1,2,†, Sirong Jiang 5,6, and Jian Wang 1,2,3,4,*

1Academy of Agriculture and Forestry Sciences, Qinghai University, Xining, China
2National Key Laboratory of Sanjiangyuan Ecology and Plateau Agriculture and Animal Husbandry, Qinghai University, Xining, China
3Key Laboratory of Qinghai-Tibet Plateau Biotechnology Ministry of Education, Xining, China
4Qinghai Provincial Key Laboratory of Potato Breeding, Xining, China
5Sanya Nanfan Research Institute of Hainan University, Hainan Yazhou Bay Seed Laboratory, Sanya, China
6College of Tropical Crops Hainan University, Hainan University, Haikou, China

Received 17 December 2021; revised 24 May 2022; accepted 21 June 2022.
*Correspondence (Tel +8613709738081; fax +869715311193; email jianwang2197@163.com (J.W.); Tel +8613637691088; email zqxia@hainanu.edu.cn (Z.X.); Tel +8698966279257; email wangwenquan@itbb.org.cn (W.W.))
†The authors contributed equally to this work.

Summary

Potato (Solanum tuberosum L.) originated in the Andes and evolved its vegetative propagation strategy through short day-dependent tuber development. Herein, we present a high-quality, chromosome-scale reference genome sequence of a tetraploid potato cultivar. The total length of this genome assembly was 2.67 Gb, with scaffold N50 and contig N50 sizes of 46.24 and 2.19 Mb, respectively. In total, 1.69 Gb repetitive sequences were obtained through de novo annotation, and long terminal repeats were the main transposable elements. A total of 126 070 protein-coding genes were annotated, of which 125 077 (99.21%) were located on chromosomes. The 48 chromosomes were classified into four haplotypes. We annotated 31 506 homologous genes, including 5913 (18.77%) genes with four homologues, 11 103 (35.24%) with three homologues, 12 177 (38.65%) with two homologues and 2313 (7.34%) with one homologue. Mismatch repair, genome-wide association analysis revealed that cytochrome P450, flavonoid synthesis, chalcone enzyme, glycosyl hydrolase and glycosyl transferase genes were significantly correlated with the flesh colours of potato tuber in 150 tetraploid potatoes. This study provides valuable insights into the highly heterozygous autotetraploid potato genome and may facilitate the development of tools for potato cultivar breeding and further studies on autotetraploid crops.

Introduction

Potato (Solanum tuberosum L.) is amongst the four major food crops grown worldwide (Li et al., 2016; Wang et al., 2020a). Diploid to hexaploid potato varieties (2n = 24, 36, 48, 60 and 72) are found in nature, with each species being geography specific. S. tuberosum ssp. tuberosum was introduced to Europe in the 15th century; however, its cultivation began only in the 17th century. Potato originated in the Andes and evolved into a short day-dependent tuber-forming crop, which characterizes its vegetative propagation strategy (Kloosterman et al., 2013). The flesh colour of potato tuber is usually white or yellow, whereas potatoes with purple, red, blue or black tubers are named ‘coloured potato’. Coloured potatoes contain an anti-ageing organic component, anthocyanin, in addition to nutrients (Kyoungwon et al., 2016; Liu et al., 2015). Being an essential food resource, they play a vital socioeconomic role (Stokstad, 2019).

Potato belongs to the Solanaceae family, and they have been classified into different groups according to different taxonomic opinions (Spooner et al., 2014; Spooner and Heterscheid, 2005). Based on the International Code of Botanical Nomenclature, Bukasov (1971) and Lechnovich (1971) classified potatoes into 21 species, including S. tuberosum subsp. Andigenum and S. tuberosum subsp. Tuberosum. Hawkes (1990) recognized seven species of potato (Hawkes, 1990). Ochoa and Center (1990), Ochoa (1999) identified nine species of potato. Dodds (1962) classified potatoes into five species, including Stenotonum, Phureja, Chaucha, Andigena and Tuberosum, based on the International Code of Nomenclature of Cultivated Plants. Recently, potatoes were considered to have four species, namely S. tuberosum (2n = 48), S. ajanhuiri (2n = 24), S. juzepczukii (2n = 36) and S. curtisii (2n = 60). S. tuberosum can be further divided into two subspecies, S. tuberosum ssp. andigena (2n = 48) and S. tuberosum ssp. tuberosum (2n = 48; Spooner et al., 2007; Spooner et al., 2014). Potato polyploids are allopolyploids originated from diploid hybridizations, and a diploid cultivar S. stenotonum (2n = 24) is the ancestor of all potato cultivars. S. tuberosum ssp. tuberosum, a long-day autotetraploid, is a variant of the more primitive S. tuberosum ssp. andigena, with the variation and selection caused by a change from a short-day to a long-day environment (Hawkes, 1990). S. stenotonum and S. sparsipilum were hybridized to generate a double diploid, S.
tuberosum ssp. anigena, whereas S. tuberosum ssp. tuberosum is a homologous heterozygous tetraploid and the most widely cultivated potato species worldwide. Moreover, local landraces of native potatoes from the South American highlands hold great importance, particularly for small stakeholders.

Several genome sequences of diploid potato cultivars identified are of great significance in potato research (Aversano et al., 2015; Leisner et al., 2018). The Potato Genome Sequencing Consortium (2011; Zhou et al., 2020). A homozygous double-haploid material, DM1-3516 R44 (DM), was obtained from a haploid material, which was produced using the anther culture of S. phureja through chromosome doubling (Paz and Veilleux, 1999). The bacterial artificial chromosome (BAC) sequences of a diploid hybrid material, RH89-039-16 (RH), containing the blood of common cultivated species have been generated to perform the anchor and comparison of the genome sequence. Finally, a 727-Mb genome was assembled (Os, 2006; The Potato Genome Sequencing Consortium, 2011). The genome sequence of a wild potato species S. commersonii with outstanding resistance to diseases and adverse environments, especially cold stress, is also reported (Aversano et al., 2015). The genome of a self-compatible diploid inbred line M6, which exhibits self-compatibility, good tuber market quality and disease resistance, has also been completed sequenced (Leisner et al., 2018). Using the third-generation sequencing technology, the genome of the diploid hybrid RH was completely sequenced (Zhou et al., 2020). Moreover, considerable progress has been made in sequencing the genome of tetraploid potato cultivars. The disclosure of the genome of tetraploid potato cultivars such as Atlantic and Otava has laid a foundation for further tetraploid potato research (Hoopes et al., 2022; Sun et al., 2022).

Diploid-scale potato breeding effectively combines resistance genes and traits for diploid germplasm; however, the yield of diploid potatoes remains low. The yield of tetraploid potato cultivars is significantly higher than that of diploid potato cultivars. The interaction amongst genes in potato heterosis determines productivity and adaptability; according to this principle, each locus in tetraploid potato has four homologues and the interactions observed within and between loci are strong. Because of the high heterozygosity in genomes, tetraploid potatoes are highly adaptable and can stabilize and exhibit high yields under diverse environmental conditions. Moreover, the reference genome of diploid potato exhibits limitations in solving problems related to large structural variations or complex genome rearrangements in tetraploid cultivated potato genomes. Thus, additional reference genomes of tetraploid potatoes are urgently required. Although tetraploid heterozygous potatoes have advanced characteristics and high yield potential, the genome sequencing and assembly of these potatoes are more challenging than those of diploid potatoes. The advancement of third-generation sequencing technology has led to improvements in reading length and sequencing accuracy, thereby making the analysis of complex gene compositions possible. Sequencing of a series of highly complex genomes, such as octoploid strawberry (Edger et al., 2019), rye (Li et al., 2021), water lily (Zhang et al., 2020) and tea (Wang et al., 2020b), has been completed. “Qingshu No.9” (Q9, 2n = 4x = 48) is amongst the selective breeding potato varieties in China. This variety offers various crucial social and economic benefits and the highest yield, and it is grown in the largest planting area in China. We, herein, sequenced and assembled the genome of Q9, a homotetraploid heterozygote potato species. Then, we compared diploid and tetraploid potato genomes to investigate the evolution of haplotypes in polyploid potatoes. Based on different dicotyledon families, we performed the evolutionary analysis of potato genomes at a large scale. Furthermore, we performed a genome-wide association analysis on the tuber flesh colour of 150 tetraploid heterozygous potatoes to identify candidate genes. The study findings highlight the contribution of a high-quality reference genome to the analysis of complex genomes and the development of the potato industry.

Results

Genome sequencing, assembly and annotation

PacBio single-molecule real-time (SMRT) sequencing and Hi-C sequencing were combined to generate a de novo genome assembly of the tetraploid potato Q9 (2n = 4x = 48). In total, 509.82 Gb raw reads equivalent to 190X genome coverage for Q9 were generated. After quality control, the clean reads were assembled, yielding a 2.67-Gb genome assembly with a contiguous N50 size of 2.19 Mb (Table S1). The initial sequence assembly was polished using NextPolish, and the splicing errors were corrected. The Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão et al., 2015) analysis conducted to evaluate the quality and completeness of the genome assembly revealed an assembly quality of 99.3% (Figure S1). The genome assembly was corrected in synteny analysis by using the Hi-C technology, and a high-quality, chromosome-scale reference genome was constructed. Finally, an ultrahigh-density genome map (Figure 1a) of Q9 (scaffold N50: 46.24 Mb) containing 48 chromosomes included in four haplotypes was constructed.

The repetitive sequences (1.69 Gb), accounting for 63.44% of the total genome size, were identified using RepeatMasker, with long terminal repeats being the main transposable elements (Figure 1b, Table S2). By performing de novo gene prediction and homologous gene prediction with Maker (Holt and Yan-dell, 2011) along with the AUGUSTUS pipeline (Mario et al., 2006), genes and their functions were annotated for the Q9 genome. The average coding sequence length was approximately 1042 bp, and each gene contained an average of 4.78 exons. Through filtering, 126 070 protein-coding genes were finally obtained, of which 125 077 were distributed on 48 chromosomes (Figure 1c; Table S3). The transcriptome data were mapped to the four Q9 genome haplotypes to verify the expression patterns of different haplotypes in different potato tissues. The gene expression at both distal regions of the chromosome was abundant in different parts, whereas that near the centromeric region was minimal or low (Figure 1d–i).

Synteny and whole-genome duplication analyses

In this study, based on gene families of Solanaceae plants including potato, tomato, pepper and tobacco, as well as 15 other species, a phylogenetic tree was constructed. The Solanaceae species probably diverged from Theaceae 85 million years ago (MYA). Gene family expansion and contraction are crucial features of species selective evolution (Figure S2a). During evolution, Q9 acquired more new genes and gene families than other potato varieties. However, during individual species evolution, gene families were acquired and lost in varying degrees. The synteny analysis was conducted amongst the haplotypes A1, A2,
A3 and A4 in the highly heterozygous autotetraploid Q9 genome, and a certain synteny was observed amongst these four haplotypes (Figure 2a). Approximately 13,328 genes exerted synteny between any two haplotypes. Of them, 42.69% exerted synteny between the haplotypes A1 and A2, 43.87% exerted synteny between A2 and A3 and 41.07% exerted synteny between A3 and A4 (Figure 2b, Figure S3). The collinearity between Q9 and Otava was consistent with that between Q9 and Atlantic (Figure S4). The heterozygosity of Q9 (1.91%) was significantly higher than those of Otava (1.53%) and Atlantic (1.78%), as determined using genomescope2 (Table S4, Figure S5). Single nucleotide polymorphism (SNP) distribution in the Q9 genome was found to be as uniform as that in Otava, and more long-distance SNPs were found in Atlantic (Figures S6–S8).

We analysed the expression of tetraploid potato genomic homologous genes. The expressions of most homologous genes were consistent between each of the two subgenomes (Figure S9). In total, 31,506 differentially expressed homologous genes were identified. Of them, 5913 (18.77%), 11,103 (35.24%) and 12,177 (38.65%) exhibited four, three and two differentially expressed homologous genes, respectively (Table 1). Obvious differences were noted in the expression of homologous genes from different tissues (Figure S10). This result might be attributed to the high heterozygosity of Q9, confirming that
chromosomal fusion or divergence occurred in the Q9 genome during evolution and environmental adaptation.

The synonymous substitution rate (Ks) of Q9 was calculated to determine the whole-genome duplication (WGD) events that occurred in the Q9 genome. The peak value of the homologous gene pairs was 1, indicating that a WGD event occurred 1 MYA (Figure S2 b). The divergence time of Q9 was closest to that of tomato, followed by pepper and tobacco, whereas it was far from that of the Theaceae family plants. Q9 diverged before the divergence of other Solanaceae species, which is consistent with the results of the phylogenetic tree. The Ks of four homologous chromosomes were generally consistent with each other in the four haplotypes (Figure S2c).

The four genomes, namely Q9, DM, RH and M6, were compared and analysed to determine differences between these genomes. In total, 16 988 gene families were shared by the four genomes, with Q9 possessing the most gene families (Figure S11). Of them, 7723, 598, 2277 and 858 gene families were unique to Q9, DM, RH and M6, respectively. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed to explore differences in functional genes between autotetraploid and diploid potatoes. Substantial differences were noted in mismatch repair (MMR), homologous recombination, DNA replication and nucleotide excision repair (Figure 3a).

Further analysis conducted to explore the mechanism of MMR regulation revealed that the key genes of mismatch recognition were **MutS2, PMS1, MLH1/3, MSH2/3/6/7, RFC1/2/3** and PCNA and that the related excision gene was **ExoI** (Figure 3b). The comparison of three diploid genomes and a tetraploid genome showed that **MLH3, MSH6/7** and **RFC3** were significantly expanded (Figure 3c). Statistical analysis revealed that the 70 **MLHs** identified in the tetraploid potato genome were concentrated in the distal regions of chromosomes 1, 2 and 4 (Figure 3d); the 32 identified **MSHs** were distributed on chromosomes 1, 3, 6, 7, 9 and 11 (Figure 3e) and the 32 identified **RFCs** were located on chromosomes 1, 2, 6, 7, 8 and 11, of which chromosomes 1, 8 and 11 had the highest number of genes (Figure 3f). Many DNA MMR-related genes were enriched in the autotetraploid potato genome. In addition, MMR-related differentially expressed genes were identified in various potato tissues.

---

Table 1 The number of genes and homologous genes of four haplotypes in Q9

|     | No. of genes with A1 | No. of genes with A2 | No. of genes with A3 | No. of genes with A4 | No. of genes with 4 homologues | No. of genes with 3 homologues | No. of genes with 2 homologues | No. of genes with 1 homologue |
|-----|---------------------|---------------------|---------------------|---------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| Chr1| 4373                | 4069                | 4100                | 4709                | 907                           | 1797                          | 1199                          | 146                           |
| Chr2| 2223                | 3402                | 2839                | 2817                | 1266                          | 888                           | 1050                          | 89                            |
| Chr3| 3007                | 2610                | 2370                | 3620                | 237                           | 1102                          | 1384                          | 278                           |
| Chr4| 2828                | 2568                | 3601                | 2137                | 293                           | 846                           | 1371                          | 310                           |
| Chr5| 2156                | 2227                | 2669                | 2283                | 573                           | 916                           | 593                           | 141                           |
| Chr6| 2644                | 2356                | 3197                | 1782                | 244                           | 854                           | 1217                          | 263                           |
| Chr7| 1781                | 1897                | 2837                | 2857                | 335                           | 834                           | 968                           | 225                           |
| Chr8| 2636                | 2950                | 1742                | 2219                | 648                           | 891                           | 646                           | 114                           |
| Chr9| 2457                | 2304                | 2658                | 2078                | 583                           | 809                           | 830                           | 89                            |
| Chr10|1878               | 2237               | 1750                | 2484                | 262                           | 722                           | 958                           | 228                           |
| Chr11|2340               | 2223               | 1786                | 1758                | 331                           | 663                           | 838                           | 149                           |
| Chr12|2521               | 1708               | 3279                | 2110                | 234                           | 781                           | 1123                          | 281                           |
| Total| –                   | –                  | –                   | –                   | 5913                          | 11 103                        | 12 177                        | 2313                          |
The DNA repair system plays a role in the face of cumulative effects of replication errors, environmental damages and ageing to maintain genome integrity and stability.

**Genome-wide association analysis of tetraploid cultivated potatoes**

A total of 150 tetraploid cultivated species from China and The International Potato Center (CIP) were selected for simplified sequencing, and 10,364 high-quality variation sites were detected. These high-quality variation sites were then clustered with ADMIXTURE, and the largest subgroups \( K \) were inferred as 1–12. The cross-validation (CV) error of each \( K \) value was calculated. At the \( K \) value ranging from 1 to 3, the CV error gradually increased; at \( K > 3 \), the CV error decreased rapidly; at \( K = 5 \), the CV error was the smallest (0.22) and at \( K > 5 \), the CV error gradually increased (Figure S12). Therefore, \( K = 5 \) was deemed as the most appropriate value; accordingly, the entire potato population was divided into five subgroups.

Differences in the distribution pattern amongst the five subgroups could be observed on the principal component 1 (PC1) axis, and the clustering results were consistent with the group structure division (Figure 4a). In accordance with the \( Q \) value of each material in these five subgroups, each material was classified into the subgroup with the largest \( Q \) value. In total, 86, 11, 3, 26 and 24 genetic resources were found from subgroups 1, 2, 3, 4 and 5, respectively (Figure 4b). The clustering results were consistent with the group structure division, and the five subgroups could be roughly clustered together (Figure 4c).

The association analysis was performed on the tuber flesh colour of 150 potato varieties, and the Manhattan plot was constructed (Figure 4d). A total of 108 significant SNP sites were obtained, with \( P < 0.01 \) as the threshold (Table S5). Fifty-eight genes, including 13 cytochrome P450 genes, were identified, of which 11 were distributed on chromosomes 2, 4, 5 and 6. In addition, three flavonoid synthesis-related genes and a chalcone enzyme gene were identified on chromosome 5A4. Additionally,
22 glycosyl hydrolase and glycosyl transferase genes, 18 TFs and a methyltransferase gene were identified (Table S6).

**Discussion**

Currently, diploid hybrid potato genomes, namely DM, M6 and RH, have been sequenced. The genome sizes of DM and M6 haplotypes have been reported to be 727 Mb (The Potato Genome Sequencing Consortium, 2011) and 882 Mb (Leisner et al., 2018), respectively, with the genome size of the diploid heterozygous genome RH being 1.67 Gb (Zhou et al., 2020). These genomes are important in potato research; specifically, the RH genome is close to those of commercially cultivated tetraploid potatoes. Owing to the limitations of sequencing technology, tetraploid genomes were previously difficult to be sequenced. The Otava and Atlantic reference genomes of tetraploid potatoes recently provided a basis for the research of tetraploid cultivated potatoes. However, obtaining completed assemblies and homologous chromosomes from highly heterozygous and complex autotetraploid cultivated potato genomes was still challenging. These challenges could be overcome with the development of sequencing technology, particularly third-generation sequencing. In this study, the high-quality genome of tetraploid potato Q9 (2.67 Gb) was obtained. This genome is larger than the Atlantic genome and slightly smaller than the Otava genome, and its heterozygosity is significantly higher than that of Atlantic and Otava. The contig N50 size of the Q9 genome (2.19 Mb) was larger than that of Otava (2.1 Mb). The tetraploid potato reference genome will provide numerous vital resources for autotetraploid homologous genes to potato breeders and researchers.

Ancient WGD evolutionary events (also known as ancient polyploidization events) occur commonly in plants, representing a powerful evolutionary force for the emergence of new gene functions and new species (Otto, 2007). Solanaceae and Theaceae species probably diverged 85 MYA. A WGD event occurred in tetraploid cultivated potatoes at approximately 1 MYA. The comparison of the tetraploid potato with three diploid potatoes indicated that the gene families derived from the tetraploid potatoes were involved in MMR. MMR is a crucial and conserved keeper of genetic information (Elez, 2021). The MMR protein plays a key role in DNA damage processing and signal processing. MMR can increase the fidelity of DNA replication by several orders of magnitude, whereas the lack of MMR can lead to the emergence of a mutant phenotype. The state of MMR can also affect interstitial and mitochondrial reorganization, DNA damage signals, apoptosis and cell type-specific processes; however, the role of MMR in these processes remains unclear (Jiricny, 2006; Li, 2008; Marti et al., 2002). Compared with those of diploid potatoes, MSHs and MLHs in MMR proteins of tetraploid potatoes exhibited significant expansion. MSHs have a vital role in maintaining genome integrity and stability, and the lack of
MSHs remarkably increases the DNA recombination rate (Emmanuel et al., 2006; Hoffman, 2004). The key proteins MLHs in MMR can also affect the recombination rate (Dion et al., 2010). Therefore, these genes were speculated to be critical for maintaining the stability of the autotetraploid potato genome.

Furthermore, the genome-wide association analysis of tetraploid cultivated potato species by using the tetraploid potato genome could associate candidate genes with specific chromosomes. In this study, many significant tuber colour-associated SNP sites were found through the genome-wide association analysis of potato flesh colour. Amongst them, 11 cytochrome P450 genes were distributed on chromosomes 2, 4, 5 and 6. F3’H and F3’S/H are cytochrome P450 monooxygenases; F3’H could catalyse various oxidation reactions that depend on NADPH or NADPH, whereas F3’S/H could catalyse hydroxylation at the 5’- or 3’-position, 5’-position of the flavonoid B ring in the presence of NADPH, and O2 and is the only enzyme system known to catalyse the 5’-hydroxylation reaction of the B ring. Studies have indicated that F3’S/H abundance in plant species at the transcription level can strongly affect anthocyanin synthesis (Groteveld, 2006; Petroni and Tonelli, 2011). One candidate gene, chalcone synthase (CHS), was identified on chromosome 5A4. CHS is the first key enzyme gene for flavonoid synthesis in plants. This enzyme catalyses the condensation of 4-coumarin coenzyme A (CoA) and malonic acid CoA to form chalcone and provides a basic carbon framework for flavonoids (Gu et al., 2019; Li et al., 2020; Xie et al., 2016; Yang et al., 2018). In addition, three candidate genes related to flavonoid synthesis and directly related to colour were found in the A2 haplotype. Moreover, genes related to glycosyl hydrolase, glycosyl transferase and the molecular modifications of anthocyanins were identified in this study. Anthocyanins are unstable under physiological pH conditions because of the presence of exposed hydroxy groups. Flavonoid 3-O-carbonyl transferase transfers glucose transferase on UDP-glucose to the C3 hydroxyl of the anthocyanin molecule to form coloured anthocyanins (Albert et al., 2014; Holton and Cornish, 1995; Zhang et al., 2019a). In addition, a methyl transferase gene was found on chromosome 8A4. Methylolation modification mostly occurred on the C3 and C5 hydroxyl groups of the anthocyanidin molecule and sometimes on the C5 and C7 hydroxyl groups. Methylation modification stabilizes the B ring of anthocyanins, thereby reducing the chemical activity of the entire molecule and increasing their water solubility (Bernini et al., 2011; Wen, 2006). In the present study, 18 TFs, which were distributed in the four haplotypes (A1, A2, A3 and A4), were also identified. TFs play a vital role in the regulation of colour formation (Guo et al., 2014; Hao et al., 2020). This study provides autotetraploid cultivated potato assembly strategies and the basic data for promoting the breeding of different potato varieties.

Materials and methods

Materials and sequencing

‘Qingshu No. 9’ (Q9, S. tuberosum ssp. tuberosum) is a late-maturing potato variety that was obtained from selective breeding through hybridization between 3 875 213 and APHRODITE at the Institute of Biotechnology, Qinghai Academy of Agriculture and Forestry Sciences (Liao et al., 2012). Q9 offers the advantages of high yield, high-quality, drought tolerance, cold tolerance, late blight resistance and ring rot resistance and is one of the most crucial autotetraploid cultivated potatoes in the main potato-producing regions of China. We used the SMRT technology (Pacific Biosciences, Menlo Park, CA, USA) to sequence the tetraploid Q9 genome and reads with a total length equivalent to ~190X genome coverage were obtained. Transcriptomes of different Q9 tissues were sequenced and used for the genome assembly. In addition, 150 tetraploid potato accessions, including 44 from China, 102 from CIP and 4 from other countries, were collected (Table S7). The CTAB method was used to extract DNA, and the AFSM technology was used for simplified sequencing (Xia et al., 2014).

Genome assembly and chromosome construction

The extracted DNA was sequenced on a PacBio Sequel 2 platform (Pacific Biosciences, Menlo Park, CA, USA) with the CCS model and a total of three SMRT cells. The resulting BAM files were subjected to the pbccs program (https://ccs.how/) with default parameters, resulting in 509.82-Gb HiFi reads. We further used the hisfams (Cheng et al., 2012) assembler ⟨v. 0.15.1-c329⟩ to assemble these HiFi reads with default parameters. The resulting ‘p_utg.gfa’ file contained the haplotype-resolved unigene graph without small bubbles and was used to extract phased contigs.

A total of 312.78-Gb Hi-C reads were mapped to the initially phased contigs by using BWA-MEM (Li and Durbin, 2009), and uniquely mapped reads were extracted using SAMtools with the command ‘samtools view -bq 40’. We then used the ALLHIC – corrector program implemented in the ALLHIC (Zhang et al., 2019b) package to correct the misjoined or chimeric contigs based on the contact pattern of Hi-C signals. Additionally, an allelic contig table was prepared by aligning coding sequences from a diploid potato assembly, downloaded from Phytozome (v. 12.0) and the phased contigs by using a GMAP-based approach (Wu and Watanabe, 2005), guided by an ALLHIC GitHub document. The second round of Hi-C read mapping was performed on the corrected contigs. The resulting BAM file was filtered and subjected to the standard ALLHIC phasing pipeline, including pruning, partition, optimization and construction, along with the corrected contigs. The scaffolding results were further assessed using the Hi-C heat map, plotted using the ALLHIC_plot function and evaluated using dot plot analysis; dot plots were generated from the jcvi MCscan package (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)). Based on the single-copy homologous plant-specific database, the completeness of the genome assembly was assessed using BUSCO (Simão et al., 2015).

Gene prediction and functional annotation

Genome-wide repetitive sequences were annotated de novo with RepeatMasker (http://www.repeatmasker.org). Based on the repeat-masked genome, the protein-coding genes were predicted using ab initio calculations, a combination of conservation of protein homologues and transcript assembly. The homology-based prediction was performed using the protein sequences of potato genomes (DM, M6 and RH) and the genome of Solanaceae crops such as tobacco, red pepper and tomato. The transcriptomes of six different Q9 tissues, namely roots, stems, leaves, stolons, flowers and pedicels, were used for auxiliary annotation. MAKER was used to integrate these results into final gene models (Holt and Yandell, 2011). AUGUSTUS software was used for gene prediction integration (Mario et al., 2006).
Construction of phylogenetic tree and estimation of evolutionary time

Single-copy orthologous genes of Solanaceae plants, including potato (one tetraploid and three known diploids), tomato, pepper and tobacco, as well as 20 other species, including 8 monocotyledonous and 18 dicotyledonous plants, were identified using the OrthoMCL (Li et al., 2003) program. FastTree (v. 2.1.9) (Price et al., 2010) was used to construct a maximum likelihood (ML) tree based on these single-copy orthologous genes. This ML tree was converted into a super-time-scale phylogenetic tree by r8s using the calibration time on the TimeTree website (Kumar et al., 2010) was used to construct a maximum likelihood (ML) tree based on these single-copy orthologous genes. This ML tree was converted into a super-time-scale phylogenetic tree by r8s

Synteny and Ks analyses

MCScanX with default parameters was used to identify collinear blocks. Protein sequences were used as queries to search against the genomes of other plant species and find the best matching pair (Wang et al., 2012). Each aligned block represented homologous pairs derived from a common ancestor. The Nei–Gojobori method implemented in PAML was adopted to calculate the Ks value (the number of synonymous substitutions for each synonymous site) of the homologous gene in the collinear region, and the median Ks was considered to represent the collinear region (Yang, 2007). The values of all gene pairs were plotted to determine the assumed WGD events of tetraploid potatoes. The formula \( t = \frac{Ks}{2r} \) represents the neutral substitution rate and was used to estimate the duplication time of tetraploid potatoes and the divergence time of tetraploid potatoes from other tree species.

The SNP invocation and comment

The Perl script (http://afsmseq.sourceforge.net/) was used to filter the original sequencing data, and the total number of reads obtained from sequencing was obtained simultaneously. Reads were allocated to each individual by using barcodes designed in accordance with the AFSM technology, and the number of reads of each individual was calculated. Bowtie2 software (Langmead and Salzberg, 2012) was used to map the corrected sequencing reads with the tetraploid genome, and SAMtools (Li et al., 2009) and VCFtools (http://vcftools.sourceforge.net/) were used to detect the SNP and indel sites.

Population structure and genetic diversity

The genetic distance matrix of samples was constructed using PHYLIP (http://evolution.genetics.washington.edu/phylip.html), and Notepad++ software was used to adapt the genetic distance matrix file into the appropriate format. The NJ method was used to construct the phylogenetic tree, whereas iTOL (https://itol.embl.de/) was used to construct the phylogenetic tree. Principal component analysis (PCA) was conducted using GCTA software for the potato population materials based on the selected SNPs (Yang et al., 2011). R software was used to calculate the vector of each PC and draw the PCA scatter plot. In addition, ADMIXTURE software (Alexander et al., 2009) was used to analyse the population structure and estimate the optimal number of subgroups. PLINK software (Purcell et al., 2007) was used to adjust the input file format of ADMIXTURE and input the file. The K value representing the number of subgroups ranged from 1 to 12, and the most appropriate value was selected in accordance with the obtained CV error. The genetic component coefficient (Q) of each material in each subgroup was used to construct the population genetic structure matrix.

Genome-wide association analysis of fresh potato tuber colour

High-quality SNPs and indels were used to perform a genome-wide association analysis on the potato flesh colour of this population (Tables S7 and S8). The compressed mixed linear model of TASSEL 5.0 software (Bradbury et al., 2007) was used for the association analysis, and the key SNP sites and candidate genes were obtained.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NO. 31660417), the Science and Technology Department of Qinghai Province (2019-NK-A1 and 2020-HZ-806), the Startup Fund of Hainan University (KQYQD(ZR)-20101) and China Agriculture Research System of MOF and MARA (NO. CARS-9).

Conflicts of interest

The authors declare no competing interests.

Authors’ contributions

Z.X. and W.J. conceived the study. Z.X., F.W., M.Z., L.Z., Y. Z., Y. M., H. S., S.J., C.Z. and Y.B. participated in various aspects including biological sample collection, sample preparation, quality control and conducted the experimental work. Z.X. created partial bioinformatics scripts and analysed the data. W.J. provided potato accessions for sequencing and performed other experiments. Z.X., L.Z., S.J., C.Z. and Y.B. performed genome sequencing, assembly, annotation and analyses. Z.X., L.Z. and S.J. performed transcriptome and GWAS analyses. M.Z., Z.X., L.Z., S.J., C.Z. and Y.B wrote and improved the manuscript. W.W. reviewed and contributed to improving the manuscript. Z.X. revised the last version of the manuscript. Z.X. and W.J. supervised the whole study. All authors read and approved the final manuscript.

Data availability statement

The datasets generated and analysed during this study are available from the corresponding author upon request. The autotetraploid potato genome sequence and resequencing data have been deposited in the Genome Sequence Archive at the National Genomics Data Center and are accessible at http://bigd.big.ac.cn/ under BioProject numbers PRJCA006096 and PRJCA006099, respectively. The gene expression matrices have also been deposited in the Genome Sequence Archive and are accessible at http://bigd.big.ac.cn/ under the BioProject number PRJCA006877.

References

Albert, N.W., Davies, K.M., Lewis, D.H., Zhang, H., Montefiori, M., Brendolise, C., Bose, M.R. et al. (2014) A conserved network of transcriptional activators and repressors regulates anthocyanin pigmentation in eudicots. Plant Cell, 26, 962–980.
2004 Fang Wang et al.

Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664.

Aversano, R., Contaldi, F., Ercolano, M.R., Grosso, V. and Iorizzo, M. (2015) The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *Plant Cell* **27**, 954–968.

Bernini, R., Crisante, F. and Ginnasi, M.C. (2011) A Convenient and safe O-Methylation of flavonoids with dimethyl carbonate (DMC). *Molecules* **16**, 1418–1425.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635.

Bukasov, S.M. (1971) Flora of cultivated plants in the USSR. volume 9. potatoes. In *Kul'turn-aya flora SSSR. 9. Kartofel'*.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H. (2012) Haplotype-resolution de novo assembly using phased assembly graphs with hisfim. Nat. Methods **18**, 170–175.

Dion, E., Li, L., Jean, M. and Belzile, F. (2010) An Arabidopsis MLH1 mutant exhibits reproductive defects and reveals a dual role for this gene in mitotic recombination. *Plant J.* **51**, 431–440.

Dodds, K.S. (1962) The potato and its wild relatives. *Bot. Stud.* **4**, 517–539.

Edger, P.P., Poorten, T.J., Vanburen, R., Hardigan, M.A., Colle, M., McKain, M.R. et al. (2019) Origin and evolution of the octoploid strawberry genome. Nat. Genet. **51**, 541–547.

Elez, M. (2021) Mismatch repair: from preserving genome stability to enabling mutation studies in Real-Time single cell. *Cell* **10**, 1535.

Emmanuel, E., Yehuda, E., Melamed-Bessudo, C., Avivi-Ragolsky, N. and Levy, E. (2011) A Convenient and safe O-Methylation of flavonoids with dimethyl carbonate (DMC). *Molecules* **16**, 1418–1425.

Fang Wang et al. (2022) The potatoes of south America: Bolivia. Cambridge: Cambridge University Press.

Holton, T.A. and Cornish, E.C. (1995) Genetics and biochemistry of anthocyanin biosynthesis in *Liriodendron tulipifera*. *Plant Biotechnology Journal* **18**, 435–439.

Hoopes, G., Meng, X., Hamilton, J.P., Achakkagari, S.R., de Alves, F.G.F., Dion, E., Li, L., Jean, M. and Belzile, F. (2010) An Arabidopsis MLH1 mutant exhibits reproductive defects and reveals a dual role for this gene in mitotic recombination. *Plant J.* **51**, 431–440.

Kloosterman, B., Abelenda, J.A., Gomez, M.M., Oortwijn, M., de Boer, J.M., Edger, P.P., Poorten, T.J., Vanburen, R., Hardigan, M.A., Colle, M., McKain, M.R. et al. (2015) Origin and evolution of the octoploid strawberry genome. Nat. Genet. **51**, 541–547.

Li, W., Wang, Z., Yang, J., He, H., Jin, H., Li, X., Ren, T. et al. (2021) A high-quality genome assembly highlights yve genomic characteristics and agronomically important genes. Nat. Genet. **53**, 574–584.

Liao, W., Zhuo, G., Ni, M., Wang, J., Zhang, Y., Qu, J. et al. (2012) The introduction, evaluation and utilization of full powder processed potato new variety “Qingshu No9”. *Tibet J. Agric. Sci.* **34**, 36–41.

Liu, Y., Lin-Wang, K., Deng, C., Warran, B., Wang, Y., Wu, B., Yang, H. et al. (2015) Comparative transcriptome analysis of white and purple potato to identify genes involved in anthocyanin biosynthesis. *PloS One* **10**, e0121948.

MARIO, S., Oliver, K., Ifran, G., ALEC, H., STEPHAN, W. and BURKhard, M. (2006) *Augustus*: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435–439.

Martí, T.M., Kunz, C. and Fleck, O. (2002) DNA mismatch repair and mutation avoidance pathways. *J. Cell. Physiol.* **191**, 28–41.

Ochoa, C.M. (1999) Las papas de sudamerica. Peru: Centro Internacional de la Papa.

Ochoa, C.M. and Center, J.P. (1990) *The potatoes of south America: Bolivia*. Cambridge: Cambridge University Press.

Os, V.H. (2006) Construction of a 10,000-marker ultradense genetic recombination map of potato: A framework for accelerating gene isolation and a genomewide physical map. *Genetics*, **173**, 1075–1087.

Otto, S.P. (2007) The evolutionary consequences of polyploidy. *Cell* **131**, 452–462.

Paz, M.M. and Veilleux, R.E. (1999) Influence of culture medium and in vitro conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids. *Plant Breed.* **118**, 53–57.

Petroni, K. and Tonelli, C. (2011) Recent advances in the regulation of anthocyanin synthesis in reproductive organs. *Plant Sci.* **181**, 219–229.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS ONE* **5**, e9490.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D. et al. (2007) PLINK: a tool set for Whole-Genome association and Population-based linkage analyses. *Am. J. Human Genet.* **81**, 559–575.

Siimao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 2310–2312.

Spooner, D.M., Ghislain, M., Simon, R., Jarsky, S.H. and Gaviléno, T. (2014) Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. *Botanical Rev.* **80**, 283–383.

Spooner, D.M. and Hettiserscheidt, W.L.A. (2005) Origins, evolution, and group classification of cultivated potatoes. In *Darwin's Harvest: New Approaches to the Origins, Evolution, and Conservation of Crops* (Motley, T.J., Zerega, N. and Cross, H., eds), pp. 285–307. New York: Columbia University Press.

Spooner, D.M., Nunez, J., Trujillo, G., Herrera, M.D.R., Guzman, F. and Ghislain, M. (2007) Extensive simple sequence repeat genotyping of potato landraces.
supports a major reevaluation of their gene pool structure and classification. Proc. Natl Acad. Sci. USA, 104, 19398–19403.

Stokstad, E. (2019) The new potato. Science, 363, 574–577.

Sun, H., Jiao, W.B., Krause, K., Campoy, J.A. and Schneeberger, K. (2022) Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat. Genet. 54, 342–348.

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature, 475, 189–195.

Wang, X., Feng, H., Chang, Y., Ma, C., Wang, L., Hao, X., Li, A. et al. (2020b) Population sequencing enhances understanding of tea plant evolution. Nat. Commun. 11, 4447.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X. et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, e49.

Wang, R., Wang, F. and Wang, J. (2020a) Study on changes of genomic DNA methylation level in potato under drought stress. Mol. Plant Breed. 18, 2250–2257.

Wen, X. (2006) Methylated flavonoids have greatly improved intestinal absorption and metabolic stability. Drug Metab. Dispos. 34, 1786–1792.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics, 21, 1895–1875.

Xia, Z., Zou, M., Zhang, S., Feng, B. and Wang, W. (2014) AFSM sequencing approach: a simple and rapid method for genome-wide SNP and methylation site discovery and genetic mapping. Sci. Rep. 4, 7300.

Xie, L., Li, F., Zhang, S., Zhang, H., Qian, W., Li, P., Zhang, S. et al. (2016) Mining for candidate genes in an introgression line by using RNA sequencing: the anthocyanin overaccumulation phenotype in Brassica. Front. Plant Sci. 7, 1245.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Yang, J., Lee, S.H., Goddard, M.E. and Vischer, P.M. (2011) GCTA: a tool for Genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82.

Yang, C.Q., Yang, W.Y. and Liu, J. (2018) Advances on chemical ecology of plant flavonoids. Nat. Product Res. Develop. 30, 2009–2016.

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., Chang, X. et al. (2020) The water lily genome and the early evolution of flowering plants. Nature, 577, 1–6.

Zhang, Q., Wang, L., Liu, Z., Zhao, Z., Zhao, J., Wang, Z. et al. (2019a) Transcriptome and metabolome profiling unveil the mechanisms of Ziziphus jujuba Mill. peel coloration. Food Chem. 312, 125903.

Zhang, X., Zhang, S., Zhao, Q., Ming, R. and Tang, H. (2019b) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat. Plants, 5, 833–845.

Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J.P. et al. (2020) Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat. Genet. 52, 1–6.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Assembly quality assessment by BUSCO.

Figure S2 Genome evolution and gene family analysis.

Figure S3 Collinearity analysis for the four haplotypes of Solanum tuberosum Q9.

Figure S4 Collinearity analysis of Q9 genome, Otava genome and Atlantic genome (a) Comparison of collinearity between Q9 and Otava; (b) Comparison of collinearity between Q9 and Atlantic.

Figure S5 Comparing the heterozygosity of different cultivated species, the order from left to right is Q9, Otava and Atlantic.

Figure S6 SNP distribution map of 10KB window in Q9, Otava and Atlantic.

Figure S7 SNPs density map with spacing less than 150bp identified by three cultivated potatoes.

Figure S8 Box diagram of the proportion of SNPs in long-distance SNPs (spacing >1000 bp) (long-distance SNP/all SNPs).

Figure S9 Expression of homologs in every two haplotypes.

Figure S10 Expression analysis of homologous genes in all four haplotypes in 12 groups of chromosomes. From left to right are flower, leaf, pedicel, root, stem and stolon.

Figure S11 Shared and unique gene families among DM, M6, RH and Q9 potato species.

Figure S12 Cross-validation (CV) errors.

Table S1 Statistics of the sequenced and assembled genome of potato Q9.

Table S2 The type, length and percentage of repeat sequences in potato Q9.

Table S3 Gene annotation statistics of Q9 genome.

Table S4 Heterozygosity statistics of tetraploid potato Q9, Otava and Atlantic.

Table S5 Significant SNP sites of flesh color.

Table S6 The candidate gene of flesh color.

Table S7 Source and flesh color of 150 tetraploid cultivated potatoes.

Table S8 Identification standard of potato flesh color.