DNA Research, 2022, **30**, 1–9
https://doi.org/10.1093/dnares/dsac052
Advance access publication 25 December 2022
Resource Article: Genomes Explored

OXFORD

Resource Article: Genomes Explored

Chromosome-scale genome assembly of a Japanese chili pepper landrace, *Capsicum annuum* ‘Takanotsume’

Kenta Shirasawa[1,*], Munetaka Hosokawa[2,3], Yasuo Yasui[4], Atsushi Toyoda[5], and Sachiko Isobe[1]

[1]Department of Frontier Research and Development, Kazusa DNA Research Institute, Kisarazu, Japan
[2]Department of Agriculture, Kindai University, Nara, Japan
[3]Agricultural Technology and Innovation Research Institute, Kindai University, Nara, Japan
[4]Graduate School of Agriculture, Kyoto University, Kyoto, Japan
[5]Advanced Genomics Center, National Institute of Genetics, Mishima, Japan

[*]To whom correspondence should be addressed. Tel.: +81-438-52-3935. Fax: +81-438-52-3934. Email: shirasaw@kazusa.or.jp

Abstract

Here, we report the genome sequence of a popular Japanese chili pepper landrace, *Capsicum annuum* ‘Takanotsume’. We used long-read sequencing and optical mapping, together with the genetic mapping technique, to obtain the chromosome-scale genome assembly of ‘Takanotsume’. The assembly consists of 12 pseudomolecules, which corresponds to the basic chromosome number of *C. annuum*, and is 3,058.5 Mb in size, spanning 97.0% of the estimated genome size. A total of 34,324 high-confidence genes were predicted in the genome, and 83.4% of the genome assembly was occupied by repetitive sequences. Comparative genomics of linked-read sequencing-derived *de novo* genome assemblies of two *Capsicum chinense* lines and whole-genome resequencing analysis of *Capsicum* species revealed not only nucleotide sequence variations but also genome structure variations (i.e. chromosomal rearrangements and transposon-insertion polymorphisms) between ‘Takanotsume’ and its relatives. Overall, the genome sequence data generated in this study will accelerate the pan-genomics and breeding of *Capsicum*, and facilitate the dissection of genetic mechanisms underlying the agronomically important traits of ‘Takanotsume’.

Key words: *Capsicum annuum*, chromosome-scale genome assembly, long-read technology, optical mapping, genetic mapping

1. Introduction

The genus *Capsicum* includes four major species, *C. annuum*, *C. baccatum*, *C. chinense*, and *C. frutescens*, all of which are used as vegetables and spices.[1] Because of partial cross-compatibility among *Capsicum* species, attractive cultivars have been bred worldwide through both inter- and intraspecific crossing.[1] Therefore, pedigrees of interspecific hybrids are complicated and error prone during the breeding process. The availability of interspecific hybrids depends on the combinations of parental lines used for their generation.[2] Some combinations generate morphologically abnormal F_1 hybrids, which fail to survive.[3] This phenomenon is caused by a negative interaction between two independent genetic loci, a hypothesis also known as the Bateson–Dobzhansky–Muller (BDM) model, which has been observed in wide interspecific crosses in animals and plants, including pepper.[4]

To the best of our knowledge, the genomes of four *C. annuum* lines, one *C. baccatum* line, and one *C. chinense* line have been sequenced to date.[5–10] These sequences were constructed using two next-generation sequencing technologies: short-read sequencing and error-prone long-read sequencing. Since the genomes of *Capsicum* species are larger and more complex than those of their relatives, for example, *Solanum* species,[11–13] complete and high-quality genome sequencing of *Capsicum* might be difficult with the existent technologies. Therefore, the available sequence data have gaps, even though the sequences are assembled at the chromosome level.[5–10] Recent advances in sequencing technologies enable the generation of high-quality long reads, also known as HiFi reads.[14] Furthermore, techniques such as chromosome conformation capture,[15] which generates chromatin contact maps, and optical mapping,[16] which outputs high-resolution genome-wide restriction maps, are also available. These technologies could be used to assemble the genomes of multiple lines of different *Capsicum* species, generating the *Capsicum* pan-genome,[17,18] which would enhance our understanding of its genetic mechanisms and provide insights into *Capsicum* evolution.

‘Takanotsume’ (which in Japanese literally translates to ‘The Claw of the Hawk’) is a Japanese *C. annuum* landrace named after the shape of its fruit, which is similar to that of the nails of hawks. ‘Takanotsume’ plants exhibit indeterminate growth, with a spread-out branching habit, and are cultivated for the thin-fleshed fruits,[19] which are used as a spice with a pungency level of approximately 11,900 on the Scoville scale.[20] Because of the rapid water loss from its fruits post-harvest, ‘Takanotsume’ has become a popular cultivar for spice purposes, and its derivative lines, such as ‘Hontaka’ and ‘Daruma’, have been distributed all over Japan.[19] However, the pedigree of ‘Takanotsume’ is unclear. ‘Takanotsume’ also possesses some unique characteristics,

Received 30 September 2022; Revised 7 December 2022; Accepted 23 December 2022

including two independent genes, which confer interspecific cross-compatibility explained by the BDM model,[3] and high ribonuclease activity in leaves, which could combat chrysanthemum stunt viroid *in vivo*.[21]

To reveal the genetic mechanisms underlying the attractive traits of 'Takanotsume', a high-quality genome assembly is required. In this study, we employed the HiFi sequencing technology, together with optical mapping and genetic mapping methods, to generate a chromosome-scale genome sequence assembly of 'Takanotsume'. Comparative genomics revealed nucleotide sequence variations, chromosome structural rearrangements, and transposon-insertion polymorphisms within the *Capsicum* species. The genome sequence and variant information obtained in this study would be helpful for elucidating the genetic mechanisms controlling the unique traits of 'Takanotsume'.

2. Materials and methods

2.1 Plant materials

Capsicum annuum landrace 'Takanotsume', which is maintained through self-pollination at Department of Agriculture, Kindai University, Nara, Japan, as well as 13 *Capsicum* lines, including 6 *C. annuum* lines ('106', '110', 'Sweet Banana', 'California Wonder', 'Murasaki', and 'Nikko'), 2 *C. baccatum* lines ('28' and 'Aji Rojo'), and 5 *C. chinense* lines ('3686', '3687', 'Charapita', 'pun1', and 'Sy-2'), were used in this study. *C. annuum* lines '106' and 'Nikko' were crossed to generate an F_1 mapping population. Then, the *C. chinense* line 'pun1' was crossed with a '106' × 'Nikko' F_1 plant to obtain a mapping population (n = 118).

2.2 Genome sequencing and data analysis

A short-read sequence library of 'Takanotsume' was prepared using the TruSeq DNA PCR-Free Sample Preparation Kit (Illumina) and sequenced on the NextSeq500 instrument (Illumina) in paired-end 151 bp mode. After removing low-quality bases (quality value of <10) with PRINSEQ[22] and adaptor sequences (AGATCGGAAGAGC) with fastx_clipper in the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit), the genome size of 'Takanotsume' was estimated using Jellyfish (k-mer size = 17).[23]

2.3 Linked-read sequencing and assembly

Genomic DNA was extracted from the young leaves of 'Takanotsume', '3686', and 'Sy-2' plants using Genomic Tip (Qiagen, Hilden, Germany), and high-molecular-weight DNA (fragment length >40 kb) was selected with BluePippin (Sage Science, Beverly, MA, USA). Genomic DNA library was prepared using the Chromium Genome Library Kit v2 (10X Genomics, Pleasanton, CA, USA) and sequenced on the NovaSeq 6000 platform (Illumina, San Diego, CA, USA) to generate paired-end 150 bp reads. The obtained sequence reads were assembled with Supernova (10X Genomics), in which 2 billion reads in maximum were subsampled (maxreads = 2,000,000,000).

2.4 Long-read sequencing and assembly

The genomic DNA of 'Takanotsume' used for linked-read sequencing was also used for long-read sequencing. Briefly, the genomic DNA of 'Takanotsume' was sheared in a DNA Shearing Tube g-TUBE (Covaris, Woburn, MA, USA) by centrifugation at 1,600 × g. The sheared DNA was used for HiFi SMRTbell library preparation with the SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA). The resultant library was separated on BluePippin (Sage Science) to remove short DNA fragments (<20 kb), and sequenced with SMRT cell 8 M on Sequel II and Sequel IIe systems (PacBio). The obtained HiFi reads were assembled with Hifiasm[24] (version 0.15.2) with the default parameters.

2.5 Optical mapping

Genomic DNA was extracted from young 'Takanotsume' leaves using the Plant DNA Isolation Kit (Bionano Genomics, San Diego, CA, USA), in accordance with Bionano Prep Plant Tissue DNA Isolation Base Protocol. The isolated genomic DNA was treated with DLE-1 nickase, and labelled with a florescent dye supplied in the DLS DNA Labeling Kit (Bionano Genomics). The labelled DNA was run on the Saphyr Optical Genome Mapping Instrument (Bionano Genomics). The output reads were assembled and then merged with the HiFi assembly to generate hybrid scaffold sequences using Bionano Solve (Bionano Genomics) with the default parameters.

2.6 Genetic mapping and chromosome-level assembly

Genomic DNA was extracted from all F_2 individuals (n = 118) and their parental lines using the DNeasy Plant Mini Kit (Qiagen). The obtained DNA samples were digested with *Pst*I and *Msp*I to construct a double-digest restriction-site-associated DNA sequencing (ddRAD-Seq) library,[25] which was sequenced on HiSeq 4000 (Illumina) in paired-end mode. The obtained sequence reads were subjected to quality control (as described above), and mapped onto the hybrid scaffold sequences with Bowtie 2.[26] High-confidence biallelic single nucleotide polymorphisms (SNPs) were identified using the mpileup option of SAMtools,[27] and filtered using VCFtools[28] using the following criteria: read depth ≥5; SNP quality = 10; proportion of missing data <50%. The identified SNPs were subjected to linkage analysis using Lep-Map3.[29] Contig sequences were anchored to the genetic map, and pseudomolecule sequences were established with ALLMAPS.[30] Using D-genies,[31] the genome structure of 'Takanotsume' was compared with those of four *C. annuum* lines ('CM334' [GenBank accession no.: AYRZ00000000], 'Zunla-1' [ASJU00000000], 'UCD-10X-F1' [NPHV00000000], and 'CA59' [JAJQWV000000000]), one *C. baccatum* line ('PBC81' [MLFT00000000]), and one *C. chinense* line ('PI159236' [MCIT00000000]).

2.7 Gene and repeat prediction

Gene prediction was performed with BRAKER2,[32] based on the peptide sequences of the predicted genes of *C. annuum* line 'CM334' and RNA-Seq reads obtained from the Sequence Read Archive (SRA) database of the National Center of Biotechnology Information (accession nos.: SRR17837286–SRR17837292 and SRR17837303–SRR17837315). Simultaneously, gene sequences reported in the genome assemblies of *C. annuum* lines 'CM334' (v.1.55: 30,242 genes) and 'Zunla-1' (v2.0: 35,336 genes) were mapped onto the 'Takanotsume' genome assembly with Liftoff.[33] Genome positions of the predicted and mapped genes were compared with the intersect command of BEDtools.[34] Functional annotation of the genes was performed with Hayai-Annotation Plants.[35]

Repetitive sequences in the genome assembly of 'Takanotsume' were identified with RepeatMasker (https://www.repeatmasker.org) using repeat sequences registered in Repbase and a *de novo* repeat library built with RepeatModeler (https://www.repeatmasker.org).

2.8 Genetic diversity analysis

Whole-genome shotgun libraries of 13 *Capsicum* lines were prepared with the TruSeq DNA PCR-Free Sample Prep Kit (Illumina), in accordance with the manufacturer's protocol. The resultant libraries were sequenced either on HiSeq 2500 (Illumina) to generate paired-end 250 bp reads or on NextSeq500 (Illumina) and NovaSeq 6000 (Illumina) platforms to generate paired-end 151 bp reads. The reads were subjected to quality control (as described above) and mapped onto the pseudomolecule sequences of 'Takanotsume' with Bowtie2.[26] Sequence variants were detected using the mpileup and call commands of BCFtools,[27] and high-confidence biallelic SNPs were identified with VCFtools[28] using the following parameters: minimum read depth ≥8 (--minDP 8); minimum variant quality = 20 (--minQ 20); maximum missing data <0.5 (--max-missing 0.5); and minor allele frequency ≥ 0.05 (--maf 0.05). Effects of SNPs on gene function were estimated with SnpEff.[36] The population structure of the 13 *Capsicum* lines and 'Takanotsume' were evaluated with maximum-likelihood estimation of individual ancestries with ADMIXTURE[37] and principal component analysis with TASSEL.[38] Genetic distances among the 13 *Capsicum* lines and 'Takanotsume' were calculated with the neighbour-joining method implemented in TASSEL[38] and a dendrogram was drawn with iTOL.[39] Insertion polymorphisms of *Tcc* transposons, which have been reported to affect pungency level in chili pepper,[40] were investigated across the 13 *Capsicum* lines with PTEMD.[41]

3. Results

3.1 Assembly of Capsicum genomes

Genome size estimation with 74.7 Gb short-read data indicated that 'Takanotsume' has a homozygous genome, with an estimated haploid genome size of 3,168.4 Mb (Supplementary Figure S1).

The linked reads of 'Takanotsume' (497.2 Gb) were assembled into contigs, resulting in a total of 30,425 sequences (total length = 3,072.8 Mb; contig N50 length = 9.4 Mb) (Table 1, Supplementary Table S1). The linked-read sequencing-based genome assembly of 'Takanotsume' was designated as CAN_r0.1. The complete BUSCO score was 96.9% (Supplementary Table S2); however, the contigs were fragmented and exhibited short sequence contiguity.

To improve the 'Takanotsume' genome assembly, we employed the HiFi long-read sequencing technology. Five SMRT cells were used, generating 3,127,118 HiFi reads (total length = 66.0 Gb; N50 length = 21.4 kb; genome coverage = 20.8X). The reads were assembled into 610 primary contigs (total length = 3,094.6 Mb; N50 = 99.0 Mb) (Table 1, Supplementary Table S1) with the GC content of 34.9%. The complete BUSCO score was 97.4%, of which 95.6% were single-copy BUSCOs (Supplementary Table S2). The long-read-based genome assembly of 'Takanotsume' was designated as CAN_r1.0.

To extend the sequence contiguity, optical mapping was performed. Data amounting to 1,201.0 Gb (read length ≥150 kb) were generated, a subset (600 Gb) of which was employed for further analysis. Of the 600 Gb data, 563.8 Gb data (number of reads = 1,355,894; N50 length = 407.5 kb) were used for *de novo* assembly, generating 40 molecule maps (total length = 3,078.9 Mb; N50 = 247.1 Mb). In the subsequent hybrid scaffolding process, 2 and 16 conflicts in the 40 molecule maps and CAN_r1.0, respectively, were resolved. Then, a hybrid scaffold comprising 23 sequences (total length = 3,074.2 Mb; N50 = 253.3 Mb) was obtained (Table 1), which was designated as CAN_r1.1.

To anchor the CAN_r1.1 scaffold sequences to the chromosome, genetic mapping was performed. The DNA of the mapping population and their parental lines was subjected to ddRAD-Seq analysis, which generated 1.0 M reads per sample. After quality filtering, high-quality reads were mapped onto the CAN_r1.1 assembly, with a mapping rate of 92.5%. This resulted in the detection of 1,836 high-confidence SNPs. A linkage analysis of these SNPs resulted in a genetic map, with a total of 12 linkage groups and 1,736 SNPs, and a total genetic distance of 748.8 cM (Table 2). Eighteen CAN_r1.1 scaffolds were anchored to the genetic map (Supplementary Figure S2). Nine of these scaffolds were anchored to nine chromosomes (ch01, ch02, ch03, ch04, ch05, ch06, ch08, ch09, and ch10; one per chromosome), while the remaining nine scaffolds were anchored to ch11 (two scaffolds), ch12 (three scaffolds), and ch07 (four scaffolds) (Table 2). Multiple

Table 1. Statistics of the genome assemblies of three *Capsicum* lines belonging to two species

	Capsicum annuum				Capsicum chinense	
	'Takanotsume'				'3686'	'Sy-2'
Sequencing technology	Linked-read	HiFi	HiFi + optical mapping	HiFi + optical mapping + genetic mapping	Linked read	Linked read
Total contig size (bp)	3,072,766,948	3,094,642,642	3,074,206,442	3,058,489,554	3,019,584,847	3,000,496,031
No. of contigs	30,425	610	23	12	31,863	30,812
Contig N50 length (bp)	9,406,601	99,049,140	253,267,520	262,665,162	8,977,441	12,719,824
Longest contig size (bp)	61,314,963	177,899,727	337,042,926	337,042,926	45,657,838	74,881,574
Gap (bp)	60,035,580	0	8,701,469	8,181,982	60,254,330	60,914,730

Table 2. Statistics of 'Takanotsume' pseudomolecule sequences

Chromosome	No. of SNP loci on the genetic map	Map distance (cM)	No. of contigs	Physical distance (bp)	No. of high-confidence genes
1	168	102.9	1	337,042,926	4,257
2	147	63.8	1	177,533,547	3,376
3	245	83.3	1	292,038,349	4,179
4	118	71.0	1	253,125,799	2,720
5	81	61.4	1	251,194,792	2,248
6	163	52.5	1	253,267,520	3,113
7	123	58.0	4	267,785,325	2,485
8	149	47.1	1	175,912,755	2,624
9	99	31.6	1	266,007,691	2,047
10	138	52.0	1	244,603,042	2,354
11	162	63.2	2	277,312,646	2,278
12	143	62.0	3	262,665,162	2,643
Total	1,736	748.8	18	3,058,489,554	34,324

scaffolds anchored to a particular chromosome were concatenated with 100 Ns. In total, 12 pseudomolecules spanning 3,058.5 Mb were established (Table 1, Figure 1). This final assembly was designated as CAN_r1.2.pmol.

3.2 Gene and repeat prediction

A total of 102,153 protein-coding genes were predicted in the CAN_r1.2.pmol assembly. Genes from the previously established genome assemblies of 'CM334' and 'Zunla-1' (30,242 and 35,336, respectively) were aligned against the CAN_r1.2.pmol to compare the genomic positions of predicted genes. Of the total of 29,899 'CM334' and 34,482 'Zunla-1' genes mapped onto the CAN_r1.2.pmol, 24,724 'CM334' and 31,206 'Zunla-1' genes coincided with the genomic positions of 34,324 of the total 102,153 predicted genes. These 34,324 genes and the remaining 67,829 genes were defined as high- and low-confidence genes, respectively (Table 2, Figure 1). The complete BUSCO score of the high-confidence genes was 95.0% (Supplementary Figure S2). Functional annotation analysis showed that of the 34,324 high-confidence genes, 7,609, 15,746, and 10,581 sequences were assigned to Gene Ontology slim terms in the biological process, molecular function, and cellular component categories, respectively, and 1,959 genes had enzyme commission numbers (Table S3).

Repetitive sequences occupied a total physical distance of 2,549.4 Mb (83.4%) in the CAN_r1.2.pmol genome assembly (3,058.5 Mb). Nine major types of repeats were identified in varying proportions (Table 3, Figure 1). The dominant repeat types in the chromosome sequences were long-terminal repeats (63.1%, 1,928.4 Mb) including *Gypsy*- (54.0%, 1,651.9 Mb) and *Copia*-type retroelements (5.9%, 178.9 Mb). Repeat sequences unavailable in public databases totalled 273.0 Mb.

3.3 Sequence and structural variations within the genus Capsicum

First, genome structure variants between *C. annuum* and *C. chinense* were investigated. Genome sequences of two *C. chinense* lines, '3686' and 'Sy-2', were constructed with the linked-read technology. The genome of '3686' was 3,211.8 Mb in size, according to the *k*-mer frequency analysis (Figure S1), and the resultant assembly was 3,019.6 Mb in size, with 31,863 sequences and a contig N50 length of 9.0 Mb (Table 1, Supplementary Table S1). On the other hand, the genome size of 'Sy-2' was estimated as 3,303.2 Mb (Supplementary Figure S1), and the assembly size was 3,000.5 Mb, including 30,812 sequences with a contig N50 length of 12.0 Mb (Table 1, Supplementary Table S1). The complete BUSCO scores of '3686' and 'Sy-2' genomes were 96.0% and 96.6%, respectively (Supplementary Table S2). Finally, alignment analysis revealed that the '3686', 'Sy-2', and CAN_r0.1 sequences covered 85.6%, 85.2%, and 96.3% of the CAN_r1.2.pmol reference sequence.

Next, sequence variants were detected in six *C. annuum*, two *C. baccatum*, and five *C. chinense* lines. On average, 84.5 Gb short-read data were obtained from the 13 lines, and mapped onto CAN_r1.2.pmol, with mapping rates of 96.4% for *C. annuum*, 80.2% for *C. baccatum*, and 87.3% for *C. chinense*. Totals of 5.2, 32.9, and 43.8 million high-confidence SNPs were found in *C. annuum*, *C. baccatum*, and *C. chinense*, respectively. In the *C. annuum* lines, the SNP distribution pattern was biased (Figure 1, Supplementary Figure S3), with a high density on ch09, ch10, and ch11 of '106', '110', 'California Wonder', and 'Sweet Banana'. According to SnpEff results, the most prominent SNP type was modifier impact (98.5%) in intergenic regions and introns, followed by moderate impact (0.9%; leading to missense mutations), low impact (0.5%; synonymous mutations), and high impact (0.1%; nonsense mutations and mutations at splice junctions) (Supplementary Table S4). The admixture analysis indicated the 13 lines in addition to 'Takanotsume' were grouped into three clusters (K) (Figure 2a): (a) seven *C. annuum* lines, (b) one *C. baccatum* line, and (c) the remaining five *C. chinense* lines in addition to 'Aji Roji' (*C. baccatum*) (Figure 2b). When the number of K was increased to four, the seven *C. annuum* lines were separated into two groups (Figure 2b): (1) 'Takanotsume', 'Nikko', and 'Murasaki' and (2) '106', '110', 'California Wonder', and 'Sweet Banana'. Genetic distances represented by a dendrogram (Figure 2c) and PCA analysis (Supplementary Figure S4) well supported the admixture result.

A total of 263 polymorphic sites of transposon insertions were found across the 13 *Capsicum* lines. Interestingly, the

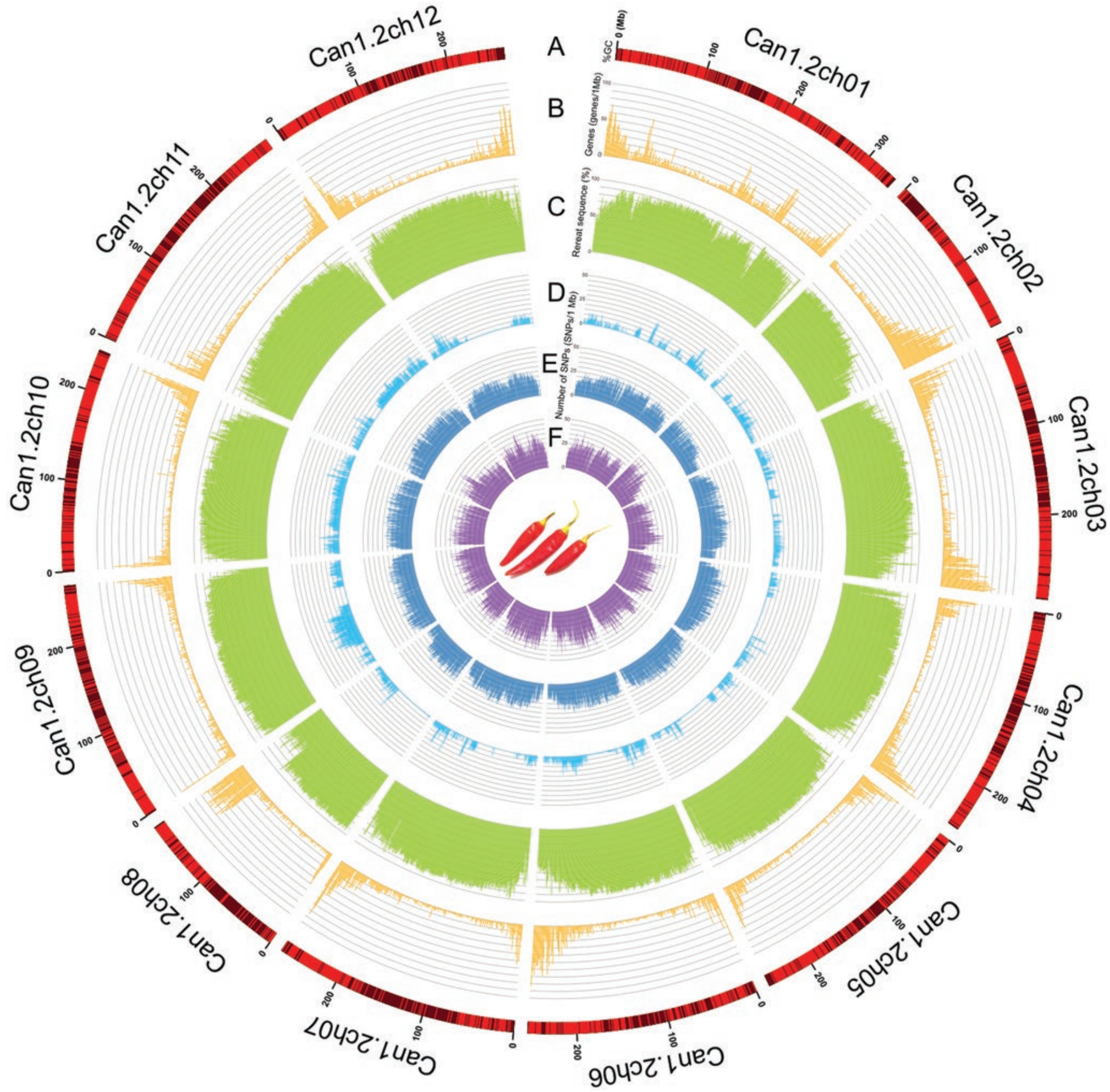

Figure 1. The *Capsicum annuum* 'Takanotsume' genome. (a) Length of the pseudomolecule sequences. Lines and boxes indicate high (≥35%) and low (<35%) GC content regions, respectively. (b) Number of high-confidence genes per a 1-Mb region. (c) Density of repetitive sequences per a 1-Mb region. (d–f) Number of SNPs per a 1-Mb region for *C. annuum* (d), *C. baccatum* (e), and *C. chinense* (f).

Table 3. Repetitive sequences in the 'Takanotsume' genome

Repeat type	Copy number	Length (bp)	Proportion of genome (%)
SINEs	69,118	11,253,375	0.4
LINEs	148,290	75,938,964	2.5
LTRs	1,381,959	1,928,422,101	63.1
DNA transposons	601,007	234,084,997	7.7
Small RNAs	49,406	9,285,310	0.3
Satellites	2,705	485,413	0.0
Simple repeats	279,862	13,411,099	0.4
Low complexity elements	48,352	2,600,900	0.1
Unclassified	983,937	272,965,980	8.9

SINEs, short interspersed nuclear elements; LINEs, long interspersed nuclear elements; LTRs: long-terminal repeats.

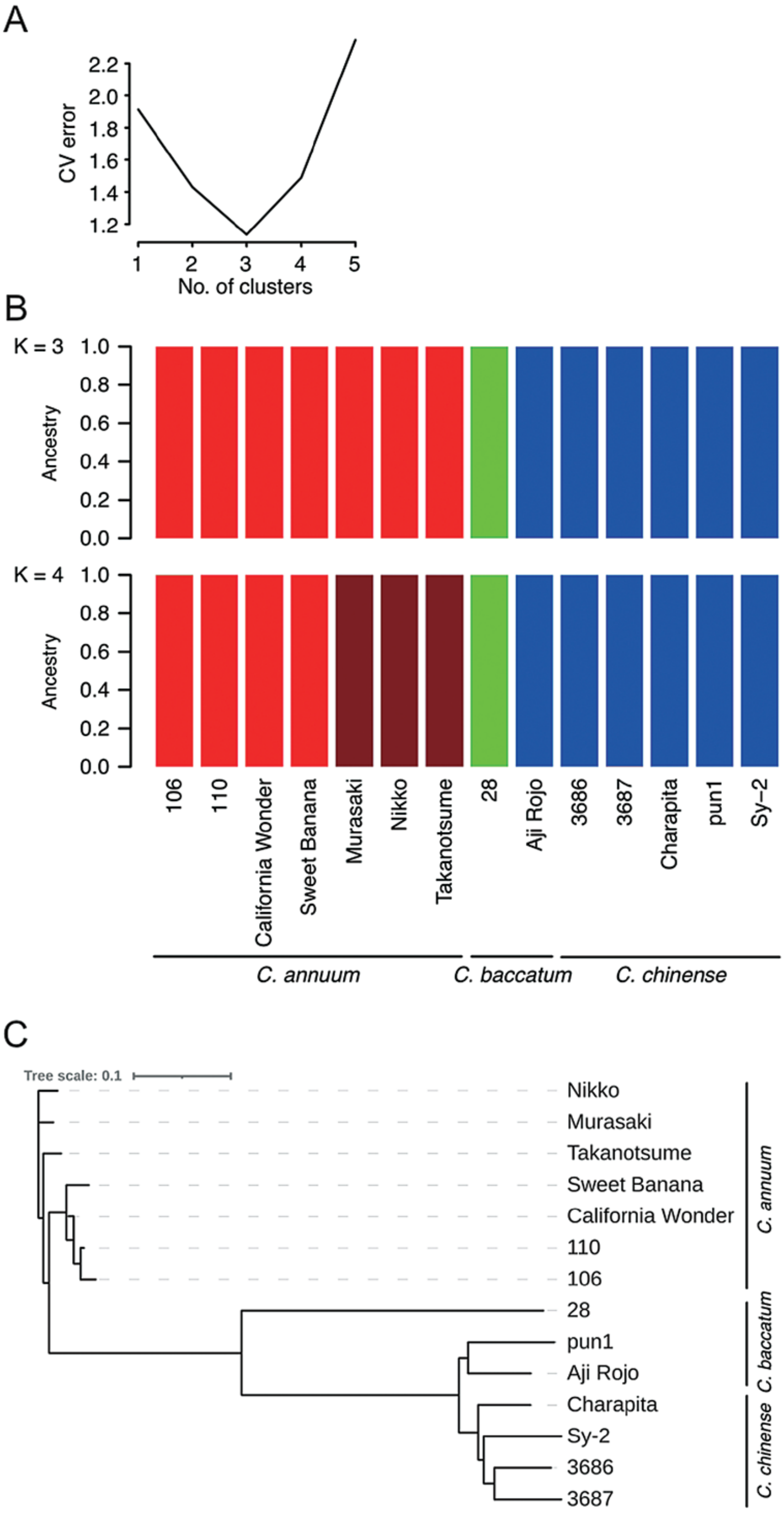

Figure 2. Genetic structure of the 14 *Capsicum* lines. (a) Cross-validation (CV) error plot for admixture analysis of K ranging from 1 to 5. (b) Population structure of the 14 capsicum lines. Each colour represents a distinct group. (c) A dendrogram based on genetic distances calculated with the neighbour-joining method.

number of polymorphic sites was biased in accordance with species. In the two *C. baccatum* and six *C. chinense* lines, numbers of polymorphism sites were ranged from 6 in pun1 to 83 in Sy-2 (Figure 3); however, no polymorphism sites was observed in any *C. annuum* lines investigated. Of the 263 sites, 20 transposons were detected in gene sequences while the remaining 243 insertions were found in intergenic regions (Supplementary Table S5).

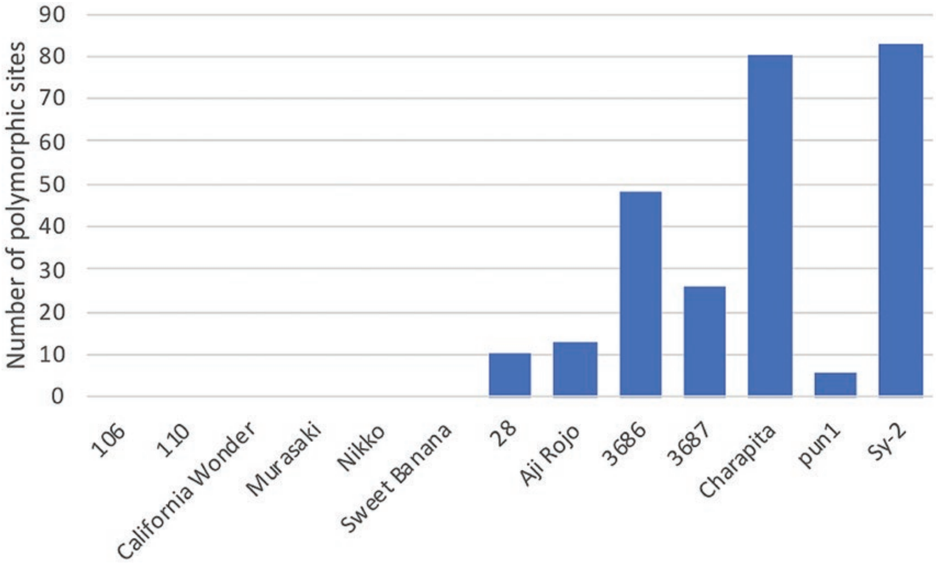

Figure 3. Number of polymorphic sites of transposon insertions.

Comparative genomics revealed that the genome structures of 'Takanotsume' and 'CA59' were well conserved; however, the chromosomes of five *Capsicum* lines ('CM334', 'Zunla-1', 'UCD-10X-F1', 'PBC81', and 'PI159236') were disrupted at the middle (Figure 4). Moreover, five potential translocations were detected in the 'Takanotsume' genome, including one on ch01 (compared with the ch08 of 'PBC81' and 'PI159236'), two on ch03 (one compared with the ch05 of 'PBC81' and another relative to the ch09 of 'PBC81'), one on ch05 (compared with the ch03 of 'PBC81'), and one on ch09 (compared with the ch03 of 'PBC81').

4. Discussion

Here, we present the chromosome-scale genome assembly (CAN_r1.2.pmol) of a popular Japanese chili pepper *C. annuum* landrace, 'Takanotsume' (Figure 1). The assembly spanned a total length of 3,058 Mb, which corresponded to 96.5% of the estimated genome size (Supplementary Figure S1, Table 1). Sequence gaps (total length = 8.2 Mb) were observed at 171 locations on 12 chromosomes (Table 1). The contiguity of this chromosome-level assembly was much greater than that obtained using the 10X Genomics Chromium technology (Table 1). The genome coverage of the 'Takanotsume' assembly was comparable with that of 'CA59' and higher than those of 'CM334', 'UCD-10X-F1', 'Zunla-1', and the relatives 'PBC81' and 'PI159236'. Moreover, sequence orders and orientations in the middle of the chromosomes were disrupted (Figure 4). This suggested that the genome structures varied within the *Capsicum* genus and/or there were misassembly points in the genomes of the five above-mentioned lines, probably because of the short-read sequencing technologies employed. To validate this assumption, further karyotyping studies with fluorescence *in situ* hybridization are required. In addition, genic regions in the CAN_r1.2.pmol assembly were also well annotated (Supplementary Tables S2 and S3). A total of 34,324 high-confidence genes in CAN_r1.2.pmol (Table 2) were supported by those in 'CM334' and/or 'Zunla-1'.

The population structure analyses indicated the three *Capsicum* species could be discriminated with the genetic variations of the genome (Figure 2, Supplementary Figure S4) except for 'Aji Rojo', which is a *C. baccatum* line but grouped in the *C. chinense* cluster. In our previous studies,[10, 42] 192 *Capsicum* lines mainly including four species, *C. annuum*, *C. baccatum*, *C. chinense*, and *C. frutescens*, were roughly classed into four groups representing the species; however, there were mismatch between the classifications based on morphological traits and those based on DNA sequence, probably due to misclassification of species based on morphological traits and/or genome introgression between different species.[42] These observations suggested that the concept of species might be reconsidered as discussed for a long time,[43,44] especially for crops including *Capsicum* because of the ease of the cross-compatibility between species. Indeed, in accordance with nuclear and plastid genotypes, 'Takanotsume' is suggested to be a derivative of the hybridization between *C. annuum* as a paternal parent and either *C. chinense* or *C. frutescens* as a maternal parent.[42] The 'Takanotsume' genome assembly from this study might contribute to clarify the mysterious pedigree.

'Takanotsume' exhibits attractive, agriculturally important phenotypes.[19] One of them is the restoration of hybrid breakdown in the progeny derived from crosses between *C. annuum* and either *C. baccatum* or *C. chinense*.[3] This phenomenon could be explained by the BDM model,[4] which was originally proposed >100 yrs ago; however, the molecular mechanisms still remain unclear. Owing to the high-quality genome assemblies and high coverage of the gene-rich regions, a map-based cloning strategy, together with gene editing and/or virus-induced gene silencing, would identify the genes capable of restoring hybrid breakdown in pepper. This would provide new insights into the molecular mechanisms responsible for the long-term unresolved BDM model. Another important characteristic of 'Takanotsume' is high ribonuclease activity in leaves.[21] This trait would be agronomically useful for the development of biopesticides to combat RNA viruses around the world. Identification of the genes responsible for the RNase activity in 'Takanotsume' would enable the regulation of enzyme activity and specificity.

In addition to nucleotide sequence polymorphisms, structural variations including copy–number variations (also known as presence–absence variations) and chromosomal rearrangements (such as translocations and inversions) can also

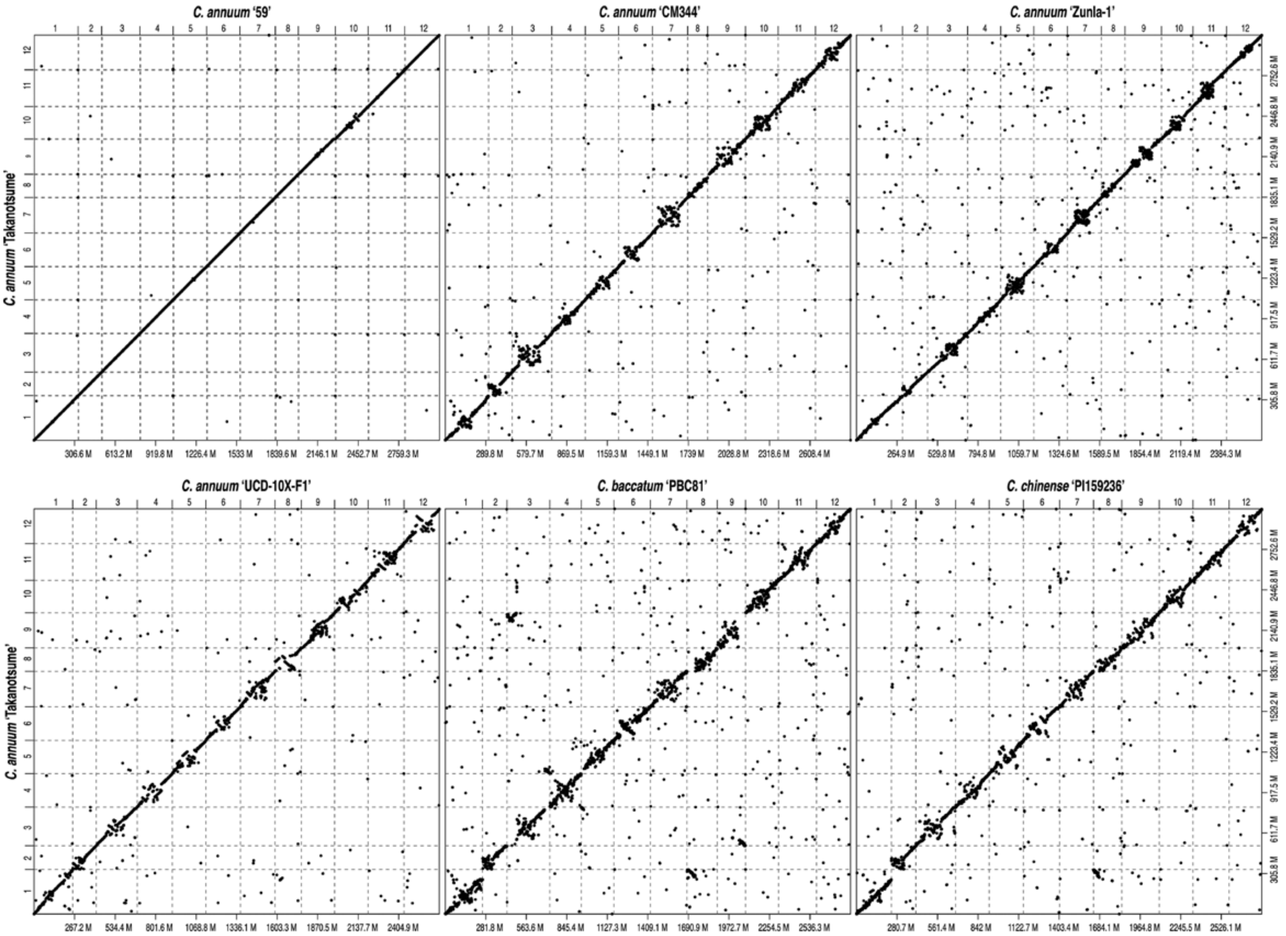

Figure 4. Comparative genomics of 'Takanotsume' and six divergent *Capsicum* lines belonging to three different species. Dots indicate structural similarities among the genomes of *Capsicum* species. Chromosome numbers are indicated above the *x*-axis and on the left-hand side of the *y*-axis, and genome sizes (Mb) are shown below the *x*-axis and on the right-hand side of the *y*-axis.

explain the within-species phenotypic variation. Transposon-insertion polymorphism (Figure 3, Supplementary Table S5) could also affect the phenotypic variations even within a species.[40] Therefore, a single reference genome sequence of a species is insufficient for gaining insights into its genomics and genetics.[45] A genome sequence established by sequencing the genomes of multiple lines of a species is called the pan-genome.[46] A pan-genome study of *Capsicum* recently conducted[17,18] will likely accelerate the pace of *Capsicum* genomics. The chromosome-level genome assembly of 'Takanotsume' constructed in this study is expected to contribute to the pan-genome study of *Capsicum*.

Acknowledgement

We thank Y. Kishida, C. Minami, K. Ozawa, H. Tsuruoka, and A. Watanabe (Kazusa DNA Research Institute) for technical assistance.

Funding

This study was supported in part by JSPS KAKENHI (16H02535, 16H06279 [PAGS], 20H02981, 22H05172, and 22H05181) and the Kazusa DNA Research Institute Foundation.

Data Availability

Raw sequence reads were deposited in the Sequence Read Archive (SRA) database of the DNA Data Bank of Japan (DDBJ) under the accession numbers DRA014624 and DRA014640–DRA014642. The assembled sequences are available at DDBJ (accession numbers AP026696–AP026707), Sol Genomics Network (https://solgenomics.net), and Plant GARDEN (https://plantgarden.jp).

Conflict of Interest

None declared.

Supplementary Data

Supplementary data are available at *DNARES* online.

References

1. Tripodi, P. and Kumar, S. 2019, The capsicum crop: an introduction. In: Ramchiary, N. and Kole C., eds., *The capsicum genome*. Springer International Publishing: Cham, pp. 1–8.
2. Walsh, B.M. and Hoot, S.B. 2001, Phylogenetic relationships of capsicum (Solanaceae) using DNA sequences from two noncoding regions: the chloroplast atpB-rbcL spacer region and nuclear waxy Introns, *Int. J. Plant Sci.*, **162**, 1409–18.

3. Yazawa, S., Sato, T., and Namiki, T. 1989, Interspecific hybrid dwarfism and geographical distribution of the dwarfness gene in *Capsicum*, *J. Jpn. Soc. Hotic. Sci.*, **58**, 609–18.
4. Orr, H.A. 1996, Dobzhansky, Bateson, and the genetics of speciation, *Genetics*, **144**, 1331–5.
5. Kim, S., Park, M., Yeom, S.-I., et al. 2014, Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species, *Nat. Genet.*, **46**, 270–8.
6. Qin, C., Yu, C., Shen, Y., et al. 2014, Whole-genome sequencing of cultivated and wild peppers provides insights into capsicum domestication and specialization, *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 5135–40.
7. Kim, S., Park, J., Yeom, S.-I., et al. 2017, New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication, *Genome Biol.*, **18**, 210.
8. Hulse-Kemp, A.M., Maheshwari, S., Stoffel, K., et al. 2018, Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library, *Hortic. Res.*, **5**, 4.
9. Liao, Y., Wang, J., Zhu, Z., et al. 2022, The 3D architecture of the pepper genome and its relationship to function and evolution, *Nat. Commun.*, **13**, 3479.
10. Shirasawa, K., Ban, T., Nagata, N., and Murakana, T. 2019, Impact of genomics on capsicum breeding. In: Ramchiary, N., and Kole, C., eds., *The capsicum genome*. Springer International Publishing: Cham, pp. 209–19.
11. Tomato Genome Consortium. 2012, The tomato genome sequence provides insights into fleshy fruit evolution, *Nature*, **485**, 635–41.
12. Xu, X., Pan, S., et al.; Potato Genome Sequencing Consortium. 2011, Genome sequence and analysis of the tuber crop potato, *Nature*, **475**, 189–95.
13. Hirakawa, H., Shirasawa, K., Miyatake, K., et al. 2014, Draft genome sequence of eggplant (*Solanum melongena* L.): the representative Solanum species indigenous to the old world, *DNA Res.*, **21**, 649–60.
14. Hon, T., Mars, K., Young, G., et al. 2020, Highly accurate long-read HiFi sequencing data for five complex genomes, *Sci. Data*, **7**, 1–11.
15. Dudchenko, O., Batra, S.S., Omer, A.D., et al. 2017, De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds, *Science*, **356**, 92–5.
16. Cao, H., Hastie, A.R., Cao, D., et al. 2014, Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology, *GigaScience*, **3**, 34.
17. Ou, L., Li, D., Lv, J., et al. 2018, Pan-genome of cultivated pepper (capsicum) and its use in gene presence-absence variation analyses, *New Phytol.*, **220**, 360–3.
18. Lee, J.-H., Venkatesh, J., Jo, J., et al. 2022, High-quality chromosome-scale genomes facilitate effective identification of large structural variations in hot and sweet peppers, *Hortic. Res.*, **9**, uhac210.
19. Kumazawa, S., Ohara, T., and Niiuchi, K. 1954, The differentiation of varieties of peppers in Japan, *J. Jpn. Soc. Hotic. Sci.*, **23**, 152–8.
20. Manikharda, Takahashi, M., Arakaki, M., et al. 2017, Physical properties, flavor characteristics and antioxidant capacity of shimatogarashi (*Capsicum frutescens*), *Food Sci. Technol. Res.*, **23**, 427–35.
21. Iraklis, B., Kanda, H., Nabeshima, T., et al. 2016, Digestion of chrysanthemum stunt viroid by leaf extracts of *Capsicum chinense* indicates strong RNA-digesting activity, *Plant Cell Rep.*, **35**, 1617–28.
22. Schmieder, R. and Edwards, R. 2011, Quality control and preprocessing of metagenomic datasets, *Bioinformatics*, **27**, 863–4.
23. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
24. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. 2021, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm, *Nat. Methods*, **18**, 170–5.
25. Shirasawa, K., Hirakawa, H., and Isobe, S. 2016, Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato, *DNA Res.*, **23**, 145–53.
26. Langmead, B. and Salzberg, S.L. 2012, Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, **9**, 357–9.
27. Li, H. 2011, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics*, **27**, 2987–93.
28. Danecek, P., Auton, A., Abecasis, G., et al. 2011, The variant call format and VCFtools, *Bioinformatics*, **27**, 2156–8.
29. Rastas, P. 2017, Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data, *Bioinformatics*, **33**, 3726–32.
30. Tang, H., Zhang, X., Miao, C., et al. 2015, ALLMAPS: robust scaffold ordering based on multiple maps, *Genome Biol.*, **16**, 3.
31. Cabanettes, F. and Klopp, C. 2018, D-GENIES: dot plot large genomes in an interactive, efficient and simple way, *PeerJ*, **6**, e4958.
32. Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. 2021, BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database, *NAR Genom. Bioinform.*, **3**, lqaa108.
33. Shumate, A. and Salzberg, S.L. 2020, Liftoff: accurate mapping of gene annotations, *Bioinformatics*, **37**, 1639–43.
34. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
35. Ghelfi, A., Shirasawa, K., Hirakawa, H., and Isobe, S. 2019, Hayai-annotation plants: an ultra-fast and comprehensive functional gene annotation system in plants, *Bioinformatics*, **35**, 4427–9.
36. Cingolani, P., Platts, A., Wang, L.L., et al. 2012, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3, *Fly*, **6**, 80–92.
37. Alexander, D.H., Novembre, J., and Lange, K. 2009, Fast model-based estimation of ancestry in unrelated individuals, *Genome Res.*, **19**, 1655–64.
38. Bradbury, P.J., Zhang, Z., Kroon, D.E., et al. 2007, TASSEL: software for association mapping of complex traits in diverse samples, *Bioinformatics*, **23**, 2633–5.
39. Letunic, I. and Bork, P. 2021, Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, *Nucleic Acids Res.*, **49**, W293–6.
40. Tanaka, Y., Asano, T., Kanemitsu, Y., et al. 2019, Positional differences of intronic transposons in pAMT affect the pungency level in chili pepper through altered splicing efficiency, *Plant J.*, **100**, 693–705.
41. Kang, H., Zhu, D., Lin, R., et al. 2016, A novel method for identifying polymorphic transposable elements via scanning of high-throughput short reads, *DNA Res.*, **23**, 241–51.
42. Shirasawa, K., Ishii, K., Kim, C., et al. 2013, Development of capsicum EST–SSR markers for species identification and in silico mapping onto the tomato genome sequence, *Mol. Breed.*, **31**, 101–10.
43. Wiley, E.O. 1978, The evolutionary species concept reconsidered, *Syst. Biol.*, **27**, 17–26.
44. Hausdorf, B. 2011, Progress toward a general species concept, *Evolution*, **65**, 923–31.
45. Yang, X., Lee, W.-P., Ye, K., and Lee, C. 2019, One reference genome is not enough, *Genome Biol.*, **20**, 104.
46. Morneau, D. 2021, Pan-genomes: moving beyond the reference, *Nat. Res.* https://www.nature.com/articles/d42859-020-00115-3