scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and improved annotation of onion genome (*Allium cepa* L.)

Heejung Cho[1,9]✉, Myunghee Jung[2,9], Seung Jae Lee[3,9], Ji Yeon Park[1], Yedomon Ange Bovys Zoclanclounon[1], Cheol-Woo Kim[4], JiWon Han[4], Jung Sun Kim[1], Do-Sun Kim[5], Younhee Shin[2], Yoon-Jung Hwang[6], Tae-Ho Lee[1], Si Myung Lee[1], Sang-Ho Kang[1], So Youn Won[1], Jin-Hyun Kim[1], Hye Yoon Jang[3], Hye-Eun Lee[5], Eun Su Lee[5], Sang-Choon Lee[7], Hyeonso Ji[8], Seong-Han Sohn[1] & Byoung Ohg Ahn[1]✉

Onion (*Allium cepa* L.) is an economically valuable crop, but its large, repeat-enriched genome makes genome assembly difficult and limits molecular breeding and biological studies. Herein, we present a chromosomal-level reference genome assembly of the double-haploid onion line DHW30006, constructed by combining PacBio, Illumina, and Hi-C sequencing approaches. The assembled genome totaled 12.77 Gb, with 65,730 gene models, and was anchored to eight pseudo-chromosomes covering 12.07 Gb (94.5%), with a scaffold N50 of 1.40 Gb. DHW30006 onion genome contained improved gene models covering approximately 580 Mb (4.54%) of the genic regions with an average gene length of 8,827 bp and 5.48 exons per gene. These gene models represented the most improved annotation among *Allium* genomes. This onion genome will serve as a valuable resource for breeding and biological research in *Allium* plants.

Background & Summary

Onion (*Allium cepa* L.) is one of the important vegetable worldwide. It has been cultivated and consumed for thousands of years, since ancient Egyptian times, as a rich source of health-beneficial nutrients and medicinal compounds with antioxidant, anti-inflammatory, antitumour, anticarcinogenic, cholesterol- and blood pressure-lowering, and antimicrobial activities[1].

The diploid bulb onion (2n = 2x = 16) has enormous genome, with a size of approximately 16 Gb/1C[2]. The average size of each onion chromosome is about 2 Gb, similar to the entire genome size of maize (2.2 Gb). In the onion genome, repetitive sequences occupy at least 95% of the genome[3]. The sequencing of onion BAC clones revealed that genes were sparsely dispersed across the genome, with no genes found in one BAC S1-D12 clone of 95 Kb and only one gene found in another BAC 1G-12–89 clone of 110 Kb[4]. Due to the high repeat content and high heterozygosity of the onion genome, its sequencing and assembly has been a challenging task.

The development of PacBio long-read sequencing and High-throughput Chromosome Conformation Capture (Hi-C) scaffolding techniques have facilitated the assembly of *Allium* plant genomes, which are usually over 10 Gb in size. Through the integration of these techniques, the first *Allium* genomic sequence was reported with garlic (*A. sativum*, 16.24 Gb)[5] in 2020, followed by onion (*A. cepa*, 14.94 Gb)[6] in 2021, bunching onion (*A. fistulosum*, 11.27 Gb)[7] in 2022 and improved assemblies of onion, garlic and bunching onion genomes are nearing completion in 2023[8]. However, while breakthroughs in sequencing technology have solved the problem of

[1]Genomics Division, National Institute of Agricultural Sciences, RDA, Jeonju, 54874, Republic of Korea. [2]Research and Development Center, Insilicogen, Inc., Yongin, 16954, Republic of Korea. [3]DNA Link Inc., Seoul, Republic of Korea. [4]Allium Vegetable Research Center, National Institute of Horticultural and Herbal Science, RDA, Muan, 58545, Republic of Korea. [5]Vegetable Division, National Institute of Horticultural and Herbal Science, RDA, Wanju, 55365, Republic of Korea. [6]Department of Chemistry Life Science, Sahmyook University, Seoul, 01795, Republic of Korea. [7]Phyzen Genomics Institute, Seongnam, 13488, Republic of Korea. [8]Gene Engineering Division, National Institute of Agricultural Sciences, RDA, Jeonju, 54874, Republic of Korea. [9]These authors contributed equally: Heejung Cho, Myunghee Jung, Seung Jae Lee. ✉e-mail: chohj78@korea.kr; boahn@korea.kr

	Allium cepa DHW30006
Sequencing data	
Pacbio Sequel II (Gb)	1006.5 (62.9 X)
Illumina Novaseq (Gb)	855.3 (53.5 X)
CHiCAGO (Gb)	483.2 (30.2 X)
Hi-C (Gb)	177.7 (11.1 X)
IsoSeq (Gb)	3.4
RNAseq (Gb)	46.5
(a) Assembly	
Total scaffolds length (Gb)	12.77
Anchored size on chromosomes (bp) (%)	12.07 (94.53%)
No. of scaffolds	5,357
No. of chromosomes	8
No. of unlocalized scaffolds	5,349
Longest scaffold size (Gb)	2.02
N50 scaffold length (Gb)	1.40
L50 scaffold count	4
Total contigs length (Gb)	12.77
No. of contigs	60,552
Longest contig size (Mb)	2.51
N50 contig length (Kb)	307.37
L50 contig count	12,550
GC (%)	33.26
Gap percent	0.04
BUSCO (Embryophyta) Complete (%)	91.38
Repeat (%)	76.91
(b) Annotation	
Number of gene models	65,730
Average gene length (bp)	8,827
Gene coverage (%)	4.54
Exons/Gene	5.48
Average exon length (bp)	206
Average intron length (bp)	1,720
BUSCO (Embryophyta) Complete (%)	84.94
No. of rRNA	1,416
No. of 18S rRNA	40
No. of 28S rRNA	14
No. of 5.8S rRNA	107
No. of 5S rRNA	1,255
No. of tRNA	7,389

Table 1. Statistics of the assembly and annotation of DHW3006 onion genome.

genome assembly, the annotation of large genomes still remains difficult and the gene models of allium plants tend to be shorter than those of other Asparagales order plants.

In this study, we report a chromosome-level assembly of the onion genome for the double-haploid (DH) line DHW30006, a short-day yellow onion (Table 1, Fig. 1). The assembly was generated using 1,006 Gb of PacBio Sequel II sequences, resulting in a total of 12.77 Gb with 60,552 contigs and an N50 contig length of 307 Kb. From the contigs, 65,730 gene models were predicted. The annotated contigs were ordered into eight pseudochromosomes of 12.07 Gb based on Hi-C data. A total of 9.8 Gb of repetitive sequences (76.9%) were determined, along with 1,258 rRNAs and 12,738 tRNAs, totaling 1.24 Mb, were annotated. The genic regions of the onion genome represented approximately 580 Mb (4.54%), with an average gene length of 8,827 bp, an average exon length of 206 bp, an average intron length of 1,720 bp, and 5.48 exons per gene. These values were found to be similar to those of other Asparagales plant genomes (Table 2). High-quality gene models of this onion genome will aid in genomic analyses of other alliaceous plants and facilitate breeding and research of *Allium* plants.

Methods

Plant materials and sample preparation. For genome sequencing, we selected the doubled haploid (DH) *A. cepa* line DHW30006. It was derived from the female gametophyte of 'Wonye 30006', which was a short-day yellow onion[9]. The onion plants were grown in a greenhouse for 4 months in pots. Leaves from one young onion plant were harvested for DNA extraction (Fig. 2a). For IsoSeq and mRNAseq for gene prediction, DHW30006

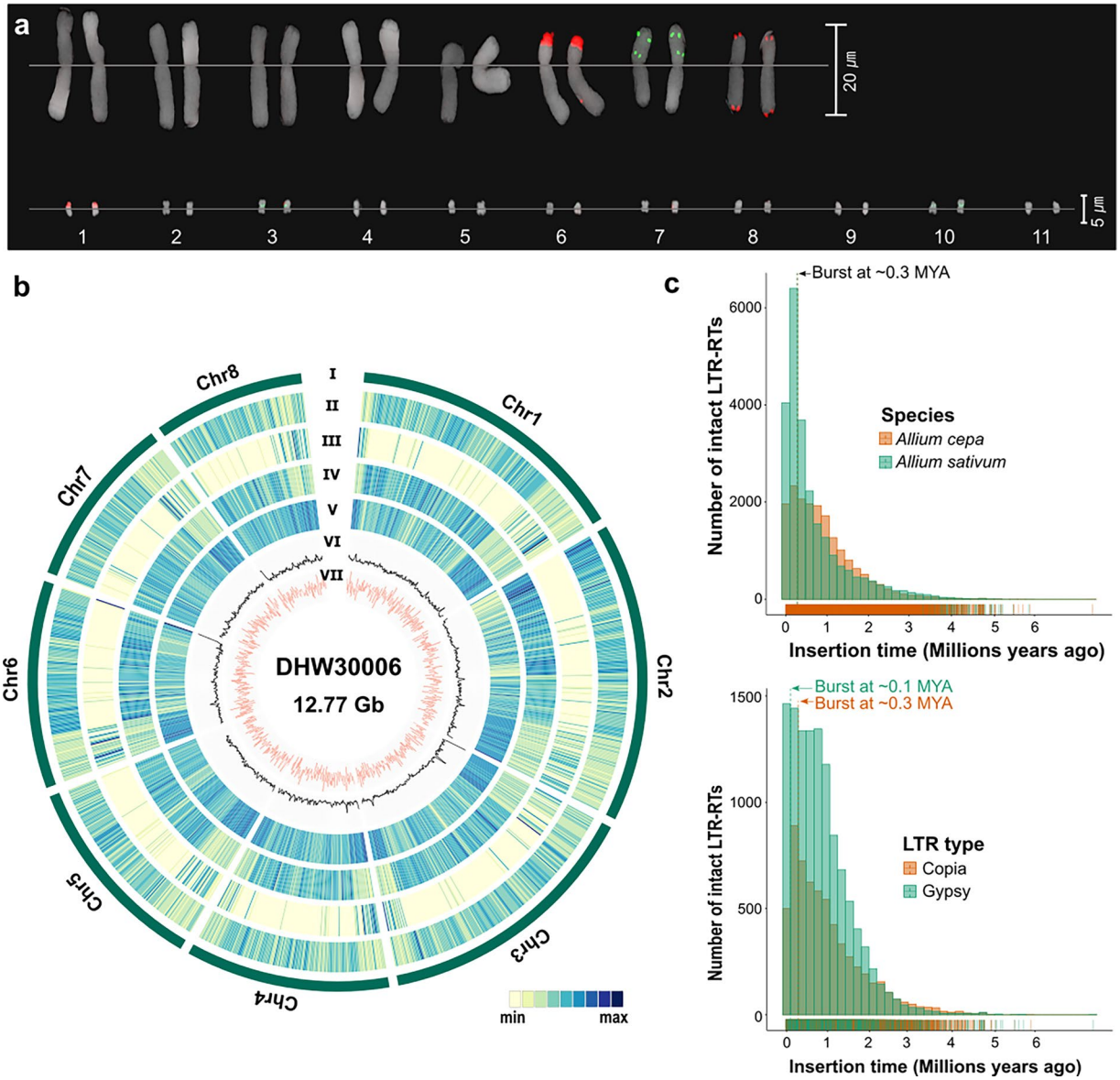

Fig. 1 Onion genome. (**a**) FISH images of onion and wild watermelon at the same magnification. The onion (*A. cepa* L.) genome consists of approximately 16 Gb across 8 chromosomes (upper), and the wild watermelon *Citrullus amarus* genome consists of 357 Mb across 11 chromosomes (lower). The probes were 5S rDNA (red) and 45S rDNA (green). Bars: 20 μm/5 μm. (**b**) Chromosome-level genome overview. The tracks indicate chromosomes (I), gene density (II), transposable elements (III), Copia retrotransposon density (IV), Gypsy retrotransposon density (V), GC content (VI) and GC skews (VII). (**c**) Copy number and age distribution of *Allium cepa* and *Allium sativum* intact LTR retrotransposon families (upper). Copy number and age distribution of Gypsy and Copia retrotransposons identified in *A. cepa* DHW30006 (lower). Dashed lines indicate the peaks of repeats corresponding to the burst age.

onions were sown in September, overwintered, and grown in the field at the Allium Vegetable Research Center (National Institute of Horticultural and Herbal Science, NIHHS), Muan, South Korea (34°58′02.7″N, 126°27′06.8″E). Different tissues were harvested individually: a root, a bulb, and a leaf from a young plant, as well as two developmental stages of roots, bulbs, and flowers from March to May (Fig. 2b). Harvested samples for DNA and RNA extractions were frozen immediately in liquid nitrogen and stored −70 °C deep freezer until extraction. For isolation of high molecular weight genomic DNA, it was extracted from eight grams of young leaves from one young onion plant as previously described AquaPhenol (MPbio, USA) extraction protocol[10]. Total RNAs were extracted using Trizol reagent (ambion, USA) according to the manufactural instruction. To eliminate residual DNA fragments, it was treated with a DNaseI (Qiagen, Germany) and purified with one more Trizol and Chloroform treatment.

	Genome (Mb)	No. genes	Avg. gene length (bp)	Genic region (Mb)	Exons /gene	Avg. exon length (bp)	Avg. intron length (bp)	Reference
Allium sp.								
A. cepa DHW30006	12,773.31	65,730	8,827.07	580.20	5.48	205.56	1,719.74	This study
A. cepa DHCU066619	14,937.43	541,098	2,509.72	1,358.00	3.05	177.69	952.70	Finkers *et al.*[6]
A. cepa DHCU	15,941.26	61,619	3,552.06	218.87	3.67	315.48	897.75	Hao *et al.*[8]
A. sativum	16,462.71	57,183	5,228.54	298.98	3.49	341.30	1,622.32	Hao *et al.*[8]
A. sativum	16,559.45	57,561	5,202.81	299.48	3.64	218.96	1,668.52	Sun *et al.*[5]
A. fistulosum	11,674.85	50,234	6,110.15	306.94	4.07	251.07	1,657.42	Hao *et al.*[8]
A. fistulosum	11,273.88	62,259	5,000.13	311.30	3.93	208.47	1,424.73	Liao *et al.*[7]
Asparagus officinalis	1,187.54	26,460	7,482.49	197.99	5.81	270.55	1,407.10	GenBank GCA_001876935.1
Phalaenopsis equestris	1,064.20	20,081	15,102.43	303.27	5.74	273.46	3,128.06	GenBank GCA_001263595.1
Apostasia shenzhenica	348.73	21,743	6,147.71	133.67	4.50	244.13	1,441.26	GenBank GCA_002786265.1
Dendrobium catenatum	1,104.26	22,566	11,897.63	268.48	5.37	308.9	2,609.26	GenBank GCA_001605985.2

Table 2. Comparison of gene elements among Asparagales plants.

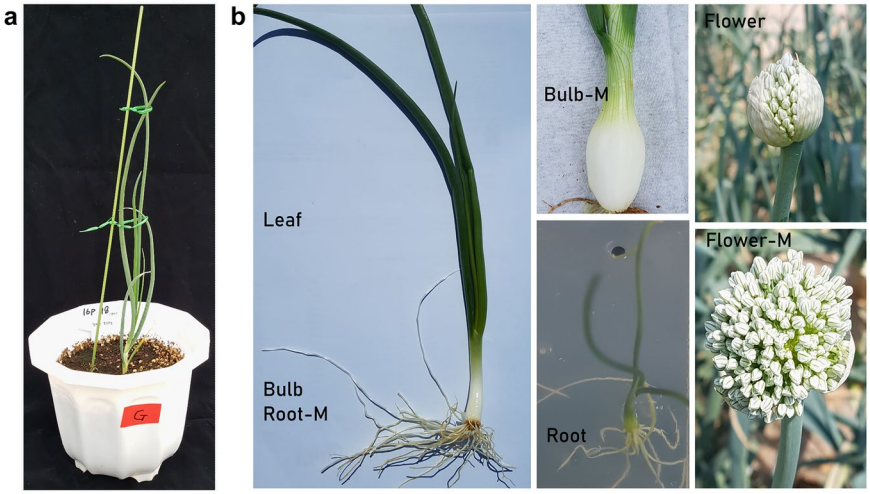

Fig. 2 DHW30006 onion plants used for DNA and RNA sequencing. (**a**) For genome sequencing, DHW30006 onion plants were grown in a greenhouse for four months and young onion leaves were harvested. (**b**) For IsoSeq and mRNAseq, DHW30006 onions were sown in September, overwintered, and grown in the farm of the Allium Vegetable Research Center. And then it was harvested separately by tissues; a root, a bulb and a leaf from a young plant, two developmental stages of roots, bulbs and flowers.

PacBio, Illumina, Hi-C and RNA sequencing. To generate high-quality genome sequences, both long-read sequencing using the PacBio Sequel II system and short-read sequencing using the Illumina NovaSeq 6000 platform were performed. High-molecular-weight genomic DNA was sheared into ~20 kb fragments using a G-tube (Covaris, Woburn, MA, USA), following the manufacturer's recommended protocol for long-read genome sequencing. The SMRTbell™ library was constructed using the SMRTbell™ Template Prep Kit (Pacific Biosciences, Menlo Park, CA, USA). After annealing the sequencing primer to the SMRTbell template, DNA polymerase was bound to the complex using the Sequel™ Binding Kit (Pacific Biosciences). The prepared library was then loaded onto the PacBio Sequel II system for sequencing. A total of one hundred Sequel™ 1 M SMRT cells were used, and each SMRT cell was subjected to a 600-minute movie runtime to capture sequencing data. For short-read sequencing, genomic DNA was randomly fragmented into ~350 bp and ~550 bp fragments using the Covaris S2 system (Covaris Inc., Woburn, MA, USA). The fragmented DNA was end-repaired, and Illumina sequencing adapters were ligated to the processed DNA fragments. The final short-read library was sequenced on an Illumina NovaSeq 6000 system (Illumina Inc., San Diego, CA, USA) using a paired-end (2 × 150 bp) sequencing strategy to generate high-coverage short-read data. PacBio long-read data were produced a total of 1,006,477,019,394 bp (63-fold coverage) with N50 read length 26,289 bp, and Illumina short-read data were produced a total of 855,328,332,552 bp (53.5-fold coverage). For scaffolding, seven CHiCAGO libraries and five Dovetail Hi-C libraries were prepared in a similar manner as described previously[11,12], sequenced on an Illumina Novaseq 6000 by 2 × 100 sequencing, and produced a total of 483,185,565,306 bp and 177,737,984,042 bp, respectively.

Iso-Seq sequencing and analysis were done by the PacBio Sequel sequencing platform and IsoSeq3 pipeline, and produced a total of 3,407,596,089 bp with mean read length 1,679 bp. For RNA sequencing, the libraries

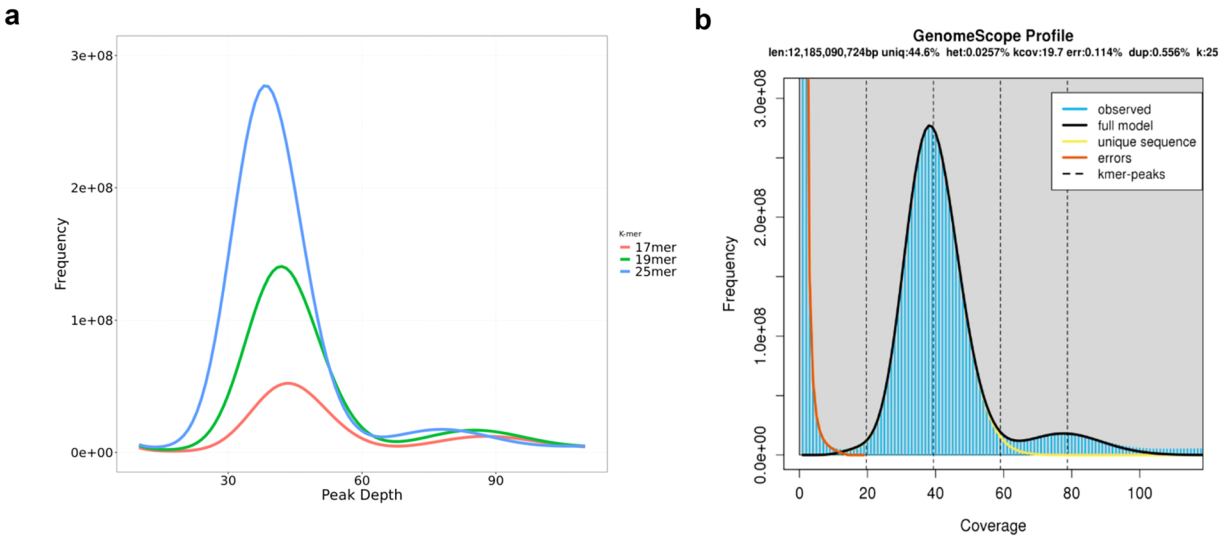

Fig. 3 Onion genome size estimation. (**a**) K-mer histogram plot using Jellyfish by 17 (red), 19 (green) and 25 (blue) mer (**b**) GenomeScope results with total genome length (len), percent of the genome that is unique (uniq), overall rate of hetrozygosity (het), mean k-mer coverage for heterozygous bases (kcov), average rate of duplication (dup) and k-mer size (k).

were prepared according to the manufacturer's instructions (Illumina Truseq stranded mRNA library prep kit). RNA sequencing was performed using an Illumina NovaSeq 6000 system following provided protocols for 2 × 100 PE sequencing and produced a total of 46,525,891,188 bp. RNAseq reads were preprocessed using cutadapt v.2.8[13] and mapped using the aligner STAR v.2.7.1a[14].

Genome size estimation and k-mer analysis. To estimate genome size, we used whole genome sequencing data, k-mer counting by Jellyfish version 2.1.3[15] with the k-mer size set to 17, 19 and 25 were used, and the genome size can be estimated using the following formula: Genome size = total number of nucleotides/peak depth of k-mer frequency distribution. Additionally, we were analyzed using the GenomeScope[16] website to obtain estimates for genome sizes (k = 25, 12.19 Gb), heterozygosity (0.0257%) and duplication levels (0.556%) (Fig. 3).

Genome assembly and scaffolding. *de novo* assembly was conducted using FALCON-Unzip assembler[17] with filtered subreads sequences (Fig. 4). The length cut-off option was specified based on the subreads N50 value 26,289 bp. We got primary assembly by unzip for the phased diploid assembly and polished using Arrow consensus algorithm. We performed error correction with 350-bp and 550-bp Illumina short reads using Pilon[18] with haplotig-merged primary contigs by default parameters to improve the base quality of genome assembly. After error correction, the primary assembly was extracted.

For scaffolding, the input *de novo* assembly, shotgun reads, CHiCAGO library reads, and Dovetail Hi-C library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies. An iterative analysis was conducted. First, Shotgun and CHiCAGO library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of CHiCAGO read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative mis-joins, to score prospective joins, and make joins above a threshold. After aligning and scaffolding CHiCAGO data, Dovetail Hi-C library sequences were aligned and scaffolded following the same method. After scaffolding, shotgun sequences were used to close gaps between contigs. To remove the ambiguity by genomic duplications that interfere with scaffold continuity, we performed purge_dups[19] with whole genome sequencing data by default parameters. The analysis of the scaffolding was performed with HiRise iteratively. The chromosome sequences for the onion were conducted with corrected contigs and scaffolds using RagTag[20]. We finally assembled 12.7 Gb of the onion genome, 94.53% of which is anchored to 8 chromosomes (Table 1a).

Genome annotation. Before annotating the structural and functional features, repeat sequences were modeled with RepeatModeler v2.0.3 and masked using RepeatMasker v4.1.1 (Fig. 4). Further, the repeats classified into their subclasses upon the Repbase v20.08 database (Table 3). A total of 9.8 Gb (76.9%) of sequences were annotated and LTR elements were predominantly abundant, accounting for 5.7 Gb (44.9%), especially those of the Gypsy family (4.5 Gb). Genome was scanned for non-coding RNAs through tRNAscan-SE 2.0[21], RNAmmer v1.2[22] and Infernal v1.1.2[23]. The evaluation of rRNAs and tRNAs revealed 1,416 rRNAs and 7,389 tRNAs, totaling 851 Kb.

The structural annotations for the repeat-masked genome were conducted with five genomic gene models from the Asparagales taxonomical order, proteins from eleven traits of *Allium cepa*, plant transcription factor[24] and R gene[25] datasets from respective databases (Table S1, https://doi.org/10.6084/m9.figshare.28079846.v1)[26].

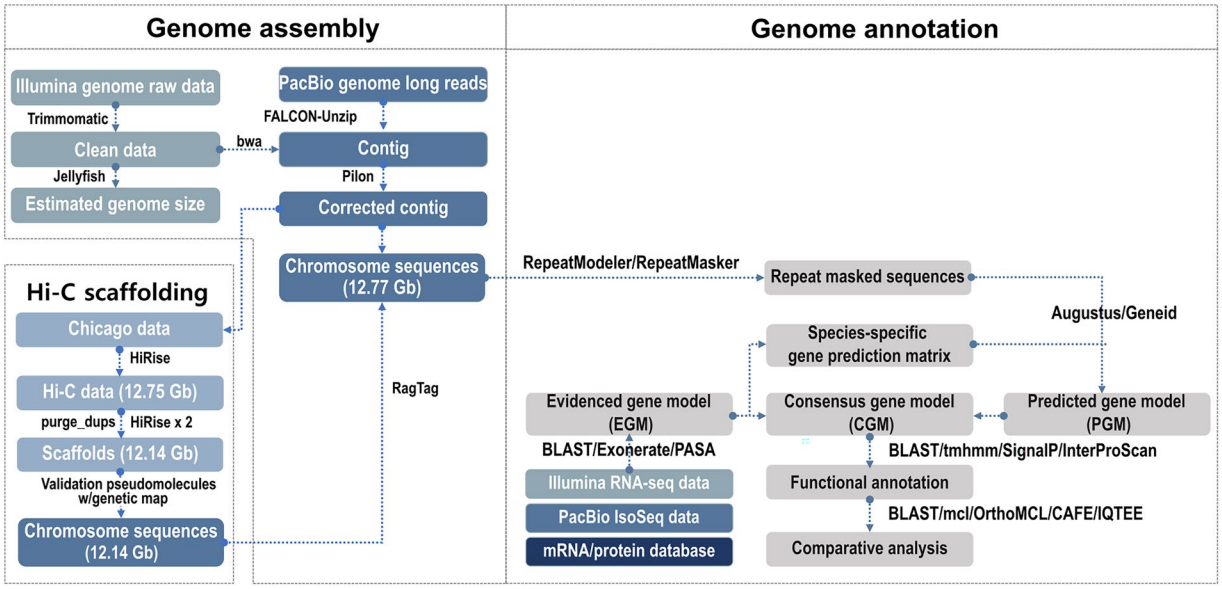

Fig. 4 Flowchart of genome assembly and annotation of *A. cepa* DHW30006.

Types	Copy number	Repeat length (bp)	Genome fraction (%)
SINEs	87,580	38,705,385	0.30
LINEs	285,243	179,212,044	1.40
LTR elements	6,821,112	5,737,338,577	44.94
DNA transposons	906,131	464,030,981	3.63
Unclassified	7,928,553	4,248,356,031	33.27
Satellites	38,658	40,705,065	0.32
Simple repeats	2,078,945	310,205,248	2.43
Low complexity	278,547	15,316,432	0.12
Total		9,819,925,515	76.91

Table 3. Statistics of repetitive sequences in DHW30006 onion genome.

The transcriptome generated in this study, consisting of *de novo* assembled 46.5 Gb of short read RNA-Seq sequences from Illumina and 3.4 Gb of long read full-length transcripts from PacBio IsoSeq method, were also included. The *in-house* pipeline used three modules, as explained in Shin *et al.*[27], including an evidence-based gene modeler (EGM), an *ab-initio* gene modeler called predicted gene model (PGM), and a consensus gene modeler (CGM). For EGM, PASA v2.4.1[28], NCBI-BLAST + v2.2.28+, and exonerate v2.4.0 methods were used to map the listed transcripts and proteins to the masked genome. For PGM, the mapped protein information was used to train *ab-initio* gene prediction methods using AUGUSTUS v3.1.0 and Geneid v1.4[29]. Finally, the consensus model was developed from both models generated from EGM and PGM. A final set of 65,730 genes were obtained; 8,827 bp of average gene length with 5.48 exons per gene. The functional annotation of genes and proteins was performed using various tools, including GO, KEGG pathway, conserved domains, and assigned transcription factors through Trinotate v3.0.1[30], InterProScan v.5.3[31], and PlanTFDB v5.0[24] (Table 4). Additionally, R genes/proteins were manually curated from The Plant Resistance Genes database[25] and RGAugury v1.0[32]. All analyses were performed using default parameters unless stated otherwise. In summary, the genome annotation identified a predominantly repetitive genome, characterized by abundant LTRs, and yielded 65,730 high-confidence genes with detailed functional annotations, contributing valuable resources for further research. We found 65,730 genes, with an average gene length of 8.8 Kb and an average number of exons per gene of 5.48 (Table 1b, Fig. 1b).

Synteny analysis of onion and garlic. We performed the synteny analysis between this onion and the garlic genome[5] with protein sequences and related gene annotation information using GENESPACE[33], and high collinearity was observed (Fig. 5).

Long terminal repeat retrotransposons insertion time estimation. Intact Long terminal repeat retrotransposons (LTR-RT) repeats were retrieved from the garlic[5] and onion (DHW30006) genome assemblies and the time estimation was calculated with LTR-retriver v. 2.9.0[34] following the formula $T = K/2\mu$, where K is the divergence rate and μ is the neutral mutation rate[35]. Frequency graphs of the insertion time were rendered using

Database	Annotation number	Annotation ratio (%)
NCBI nr	58,483	88.97
SwissProt	43,789	66.62
KEGG	32,739	49.81
eggnog	553	0.84
GO	41,804	63.60
GO BP	33,714	51.29
GO CC	33,389	50.80
GO MF	36,564	55.63
No hit	6,856	10.43
Overall	65,730	100.00

Table 4. Functional annotation of protein-coding genes.

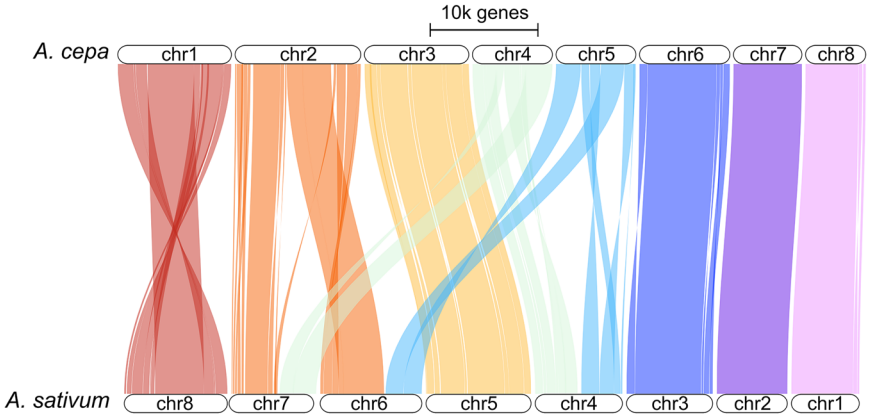

Fig. 5 Synteny map of *A. cepa* and *A. sativum*. The vertical lines indicate the regions of orthologous genes, and the order of *A. cepa* chromosomes is colour-coded. The chromosome (rounded rectangle) is scaled by gene rank orders.

R software. Based on LTR insertion time analysis, the accumulation of LTRs over the last 2 million years might have been responsible for the genome size increase (Fig. 1c).

Fluorescence *in situ* hybridization. The bulbs DHW30006 onion were obtained from the Allium Vegetable Research Center and the seeds of wild watermelon *Citrullus amarus* PI 299379 were obtained from SPRING SEED CO., LTD. FISH was performed with fresh root tips using the procedures described by Lim *et al.*[36]. Pre-labeled oligoprobes (PLOPs) for 5S rDNA and 45S rDNA sequences were labeled following the procedures by Waminal *et al.*[37]. Homologous chromosomes were paired based on their rDNA signals, chromosomal size, centromeric position, and arranged in a decreasing order based on chromosomal lengths (Fig. 1a).

Data Records

The PacBio, Illumina, Hi-C sequencing and RNA sequencing data have been deposited in the NCBI Sequence Read Archive[38], and the assembled genome has been deposited in GenBank[39] at the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) under BioProject number PRJNA912256. And the genome and annotation datasets is also available at the National Agricultural Biotechnology Information Center (NABIC, https://nabic.rda.go.kr/genome/nolog/introductionPage.do?projectNo=141) under accession number NG-1522[40]. The genome and annotations datasets have been deposited in FigShare (https://doi.org/10.6084/m9.figshare.28079846.v1)[26].

Technical Validation

The assembly quality was assessed using BUSCO analysis with three databases: viridiplantae_odb10 (425 single-copy orthologs) and embryophyte_odb10 (1,614 single-copy orthologs). The results showed complete BUSCO scores of 95.6% and 91.4%, respectively (Table 5). For this, BUSCO v5.6.1 (https://busco.ezlab.org/)[41] was used and default parameters were applied. To validate the chromosome sequence based on nucleotide level, we compared the chromosome sequences to the markers used to construct the linkage maps[42] and it showed 84.4% collinearity between the DHW30006 genome and the linkage map (Table S2, https://doi.org/10.6084/m9.figshare.28079846.v1)[26]. To evaluate the k-mer completeness and quality value (QV), we performed Merqury v1.3[43] with illumina whole genome sequencing data. The completeness of the genome assembly was assessed using k-mer analysis, yielding a k-mer completeness score of 88.96% and a quality value (QV) of 35.03. To further evaluate the completeness and assembly quality of the DHW30006 genome, which contains a high

BUSCOs	viridiplantae_odb10		Embryophyta odb10	
Complete	406	95.6%	1,475	91.4%
Complete and single-copy	384	90.4%	1,381	85.6%
Complete and duplicated	22	5.2%	94	5.8%
Fragmented	6	1.4%	39	2.4%
Missing	13	3.0%	100	6.2%
Total	425	100.00%	1,614	100.00%

Table 5. Assessment of genome assembly and gene prediction.

proportion of repetitive sequences, the LTR Assembly Index (LAI) was calculated based on LTR-RTs using LAI[44]. The analysis resulted in an LAI score of 22.91.

Code availability

1) FALCON-unzip
   length_cut_off = 26000
2) LTR_retreiver
   gt suffixerator -tis -suf -lcp -des -ssp -sds -dna
   gt ltrharvest -minlenltr 100 -maxlenltr 7000 -mintsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 1 -seed 20

For the genome annotation, we employed the 'gene_prediction_pipeline.py' script available on the GitHub repository (https://github.com/MyungheeJung/Senna_tora.git) of Kang *et al.*[45].

Received: 2 August 2024; Accepted: 12 February 2025;
Published online: 26 February 2025

References

1. Ekşi, G., Özkan, A. M. G. & Koyuncu, M. Garlic and onions: An eastern tale. *Journal of ethnopharmacology* **253**, 112675 (2020).
2. Ricroch, A., Yockteng, R., Brown, S. & Nadot, S. Evolution of genome size across some cultivated Allium species. *Genome* **48**, 511–520 (2005).
3. Flavell, R., Bennett, M., Smith, J. & Smith, D. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochemical genetics* **12**, 257–269 (1974).
4. Jakše, J. *et al.* Pilot sequencing of onion genomic DNA reveals fragments of transposable elements, low gene densities, and significant gene enrichment after methyl filtration. *Molecular Genetics and Genomics* **280**, 287–292 (2008).
5. Sun, X. *et al.* A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and allicin biosynthesis. *Molecular Plant* **13**, 1328–1339 (2020).
6. Finkers, R. *et al.* Insights from the first genome assembly of Onion (*Allium cepa*). *G3* **11**, jkab243 (2021).
7. Liao, N. *et al.* Chromosome-level genome assembly of bunching onion illuminates genome evolution and flavor formation in Allium crops. *Nature Communications* **13**, 6690 (2022).
8. Hao, F. *et al.* Chromosome-level genomes of three key Allium crops and their trait evolution. *Nature Genetics* **55**, 1976–1986 (2023).
9. Kim, C. *et al.* Md-late male sterile line 'Wonye 30006' for F1 seed production of onion (*Allium cepa* L. *Korean Journal of Breeding Science* **46**, 428–432 (2014).
10. Soundararajan, P. *et al.* Influence of genotype on high glucosinolate synthesis lines of Brassica rapa. *International Journal of Molecular Sciences* **22**, 7301 (2021).
11. O'Connell, B. L. *Developing and Applying Chromatin Proximity Ligation Methods*. (University of California, Santa Cruz, 2017).
12. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289–293 (2009).
13. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011).
14. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
15. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
16. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
17. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods* **13**, 1050–1054 (2016).
18. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).
19. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
20. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome biology* **23**, 1–19 (2022).
21. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**, 9077–9096 (2021).
22. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* **35**, 3100–3108 (2007).
23. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
24. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic acids research*, gkw982 (2016).
25. Osuna-Cruz, C. M. *et al.* PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic acids research* **46**, D1197–D1201 (2018).
26. Lee, S. J. & Cho, H. Onion (*Allium cepa* L.) DHW30006 Genome Information. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.28079846.v1 (2024).
27. Shin, G.-H. *et al.* First draft genome for red sea bream of family Sparidae. *Frontiers in Genetics* **9**, 643 (2018).
28. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654–5666 (2003).

29. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Current protocols in bioinformatics* **18**, 4.3. 1–4.3. 28 (2007).
30. Bryant, D. M. *et al*. A tissue-mapped axolotl *de novo* transcriptome enables identification of limb regeneration factors. *Cell reports* **18**, 762–776 (2017).
31. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
32. Li, P. *et al*. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC genomics* **17**, 1–10 (2016).
33. Lovell, J. T. *et al*. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11**, e78526 (2022).
34. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology* **176**, 1410–1422 (2018).
35. Bowen, N. J. & McDonald, J. F. Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome research* **11**, 1527–1540 (2001).
36. Lim, K.-B. *et al*. Characterization of rDNAs and tandem repeats in the heterochromatin of Brassica rapa. *Mol Cells* **19**, 436–444 (2005).
37. Waminal, N. E. *et al*. Rapid and efficient FISH using pre-labeled oligomer probes. *Scientific reports* **8**, 8224 (2018).
38. Cho, H. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP418719 (2023).
39. Cho, H. *Genbank* https://identifiers.org/insdc.gca:GCA_038502295.1 (2024).
40. Cho, H. NABIC https://nabic.rda.go.kr/genome/nolog/introductionPage.do?projectNo=141 (2024).
41. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
42. Cho, Y., Kim, B., Lee, J. & Kim, S. Construction of a high-resolution linkage map and chromosomal localization of the loci determining major qualitative traits in onion (*Allium cepa* L.). *Euphytica* **217**, 1–12 (2021).
43. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 1–27 (2020).
44. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic acids research* **46**(21), e126–e126 (2018).
45. Kang, S.-H. *et al*. Genome-enabled discovery of anthraquinone biosynthesis in Senna tora. *Nature communications* **11**, 5875 (2020).

Acknowledgements

This research was supported by a grant from the Cooperative Research Program for Agriculture Science and Technology Development (Project title: National Agricultural Genome Program, Project no. PJ013637), Rural Development Administration, Republic of Korea. We would like to thank Dr. Michael J. Havey (University of Wisconsin) for his interest and advice in this study. We also thank Sunggil Kim (Chonnam National University) for providing information of onion genetic map. We thank Phase Genomics (Seattle, USA) for their assistance in generation of the Hi-C data and scaffold assembly.

Author contributions

H.C. and B.O.A. planned and coordinated the project. H.C., B.O.A. and S.H.S. acquired funding. C.W.K., J.W.H., D.S.K., H.E.L. and E.S.L. prepared and selected the D.H. onion. H.C., B.O.A., S.H.S., J.S.K., T.H.L., S.M.L., S.H.K., S.Y.W. and S.C.L. contributed to sequencing strategy decision. J.Y.P., J.W.H., J.S.K. and H.C. grew the plant and prepared samples for DNA and RNA. S.J.L. and H.Y.J. performed genome assembly and structural analysis, and Y.A.B.Z. analyzed LTRs. M.J. and Y.S. performed genome annotation and comparative analysis. Y.J.H. performed karyotype analysis. H.C. and J.Y.P., H.J. and J.H.K. contributed to the data analysis. H.C., Y.A.B.Z., M.J. and S.J.L. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.C. or B.O.A.

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

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