Telomere-to-telomere genome assembly of an allotetraploid pernicious weed, Echinochloa phyllopogon

AUTHOR(S):

Sato, Mitsuhiko P; Iwakami, Satoshi; Fukunishi, Kanade; Sugiura, Kai; Yasuda, Kentaro; Isobe, Sachiko; Shirasawa, Kenta

CITATION:

Sato, Mitsuhiko P ...[et al]. Telomere-to-telomere genome assembly of an allotetraploid pernicious weed, Echinochloa phyllopogon. DNA Research 2023, 30(5): dsad023.

ISSUE DATE:

2023-10

URL:

http://hdl.handle.net/2433/286022

DNA Research, 2023, **30**, 1–10
https://doi.org/10.1093/dnares/dsad023
Advance access publication 7 November 2023
Resource Article: Genomes Explored

OXFORD

Resource Article: Genomes Explored

Telomere-to-telomere genome assembly of an allotetraploid pernicious weed, *Echinochloa phyllopogon*

Mitsuhiko P. Sato[1], Satoshi Iwakami[2,*], Kanade Fukunishi[2], Kai Sugiura[2], Kentaro Yasuda[3], Sachiko Isobe[1] and Kenta Shirasawa[1,*]

[1]Department of Frontier Research and Development, Kazusa DNA Research Institute, Chiba 292-0818, Japan,
[2]Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan, and
[3]Agri-Innovation Education and Research Center, Akita Prefectural University, Akita 010-0451, Japan

*To whom correspondence should be addressed. Tel. +81 75 753 6064. Fax. +81 75 753 6062. Email: iwakami.satoshi.2v@kyoto-u.ac.jp (S.I.); Tel. +81 438 52 3935. Fax: +81 438 52 3934. Email: shirasaw@kazusa.or.jp (K.S.)

Abstract

Echinochloa phyllopogon is an allotetraploid pernicious weed species found in rice fields worldwide that often exhibit resistance to multiple herbicides. An accurate genome sequence is essential to comprehensively understand the genetic basis underlying the traits of this species. Here, the telomere-to-telomere genome sequence of *E. phyllopogon* was presented. Eighteen chromosome sequences spanning 1.0 Gb were constructed using the PacBio highly fidelity long technology. Of the 18 chromosomes, 12 sequences were entirely assembled into telomere-to-telomere and gap-free contigs, whereas the remaining six sequences were constructed at the chromosomal level with only eight gaps. The sequences were assigned to the A and B genome with total lengths of 453 and 520 Mb, respectively. Repetitive sequences occupied 42.93% of the A genome and 48.47% of the B genome, although 32,337, and 30,889 high-confidence genes were predicted in the A and B genomes, respectively. This suggested that genome extensions and gene disruptions caused by repeated sequence accumulation often occur in the B genome before polyploidization to establish a tetraploid genome. The highly accurate and comprehensive genome sequence could be a milestone in understanding the molecular mechanisms of the pernicious traits and in developing effective weed control strategies to avoid yield loss in rice production.

Key words: genome assembly, polyploidy, telomere-to-telomere, weed

1. Introduction

Echinochloa phyllopogon ($2n = 4x = 36$) [=*Echinochloa oryzicola*] is a member of the Poaceae family, close to *Setaria italica*, an autogamous plant, and is a noxious weed in flooded rice worldwide. While this species is sometimes referred to as *E. oryzicola*, based on morphological characteristics, cross-compatibility, and chromosome number, *E. oryzicola* and *E. phyllopogon* are considered synonymous.[1] Among the *Echinochloa* genus, it is widely recognized as the species that best adapts to paddy fields.[2] *Echinochloa phyllopogon* was found only in watered environments, although populations have recently been discovered in paddy levees, roadsides, and other places.[3] While these populations display distinct morphological characteristics, they are classified as *E. phyllopogon* based on chromosome number. Herbicides have been used to manage this species owing to their huge impact on rice yields. However, repeated use of herbicides has resulted in the evolution of herbicide resistance in rice fields, posing a serious threat to agriculture.

This species often exhibits resistance to multiple herbicides, which has been attributed to the overexpression of herbicide-detoxifying enzymes such as cytochrome P450 monooxygenases.[4–6] However, the precise molecular mechanisms underlying resistance are not yet fully understood. A highly accurate genome sequence would help elucidate these mechanisms and provide a deeper understanding of how multiple herbicide resistance occurs.

Despite the complex genome structures, including polyploidy, several species of genome sequences in *Echinochloa* were determined.[7,8] Wu et al.[8] revealed the complex and reticulate evolution in the speciation of *Echinochloa* polyploids and reported the chromosome-level genome sequences (945 Mb in length) in *E. phyllopogon*, for which Continuous Long Reads (PacBio), paired-end and mate-pair reads (Illumina), and Hi–C techniques were employed. However, the assembly is shorter than the estimated genome size of 1.0 Gb[7] and included 627 gaps (ca. 50 Mb in length), in which sequences were undetermined. Gapped genome sequences that do not cover the entire genome might miss complex genome structures, genetic bases of agriculturally important traits, and evolutionarily important variations. Owing to recent advanced long-read technology, telomere-to-telomere (T2T) and gap-free genome sequences have been reported in humans,[9] chickens,[10] fungi,[11,12] and plankton.[13,14] Herein, the chromosome-level assembly of the allotetraploid genome of *E. phyllopogon* was reported. The assembly included 12 T2T

Received 24 August 2023; Revised 27 September 2023; Accepted 25 October 2023

and gap-free sequences of the 18 chromosomes in addition to six sequences connected at the chromosome level with only eight gaps. The genome information from this study would contribute to weed controls to avoid yield loss in rice production and to understand weed adaptation and propagation systems.

2. Methods

2.1. Plant materials

single line of *E. phyllopogon*, R511, sampled from California, United States,[15] was used for *de novo* genome assembly. The R511 line crossed the S401 line (CA, USA)[15] to generate an F5 mapping population (n = 118).[4] In addition, seven Japanese lines were used for whole genome re-sequencing analysis: Eoz1804, Eoz1813, Eoz1814, AEC09-14, and AEC01-01 were from paddy fields, while AEC12-91 and AEC09-23 were from vacant land, and paddy levee, respectively.[3,6] All the lines were from the northeastern mainland of Japan.

2.2. *De novo* whole genome assembly

To estimate the genome size of R511, the genome DNA of R511 was extracted using the DNeasy Plant Mini Kit (Qiagen, Tokyo, Japan), and the sequence library was constructed with the MGIEasy PCR-Free DNA Library Prep Kit (MGI Tech, Shenzhen, China) and sequenced on the DNBSEQ-G400 (MGI Tech). The genome size of line R511 was estimated using k-mer distribution analysis of short-read sequences (k = 17) with Jellyfish software (v2.3.0).[16]

For long-read sequencing, high-molecular-weight genomic DNA was extracted from the leaves of R511 using NucleoBond HMW DNA (MACHEREY-NAGEL, Dueren, Germany). Genomic DNA was prepared using the SMRTbell Express Template Prep Kit (PacBio, Menlo Park, CA, USA). Long-read sequence data were obtained using a Sequel IIe system (PacBio). All HiFi reads were assembled using Hifiasm version 0.16.1[17] with default parameters. In parallel, a subset of HiFi reads randomly sampled from the data was assembled using Hifiasm, as described above. The two assemblies from the all datasets and the subsets were aligned with MUMmer4[18] to compare structures and search for contig sequences in one assembly that bridged separated sequences in another assembly.

2.3. Chromosome-level scaffolding via genetic mapping

A genetic map of *E. phyllopogon* was established with a double-digest restriction site-associated DNA sequencing (ddRAD-Seq) technique.[19] Genomic DNA was extracted from the leaves of the F5 mapping population and its parental lines using the DNeasy Plant Mini Kit (Qiagen) and subjected to ddRAD-Seq library construction using the PstI and MspI enzymes.[20] ddRAD-Seq reads were obtained using HiSeq 4000 (Illumina, San Diego, CA, USA), and their low-quality bases (<10 quality value) and adaptor sequences (AGATCGGAAGAGC) were trimmed with PRINSEQ and fastx_clipper in the FASTX-Toolkit, respectively. The cleaned reads were mapped onto primary contigs constructed using all reads as a reference in Bowtie2,[21] and sequence variants were called using BCFtools.[22] High-confidence SNPs were selected using VCFtools[23] (parameters: minDP 5, minQ 999, maf 0.2, max-maf 0.8, and max-missing 0.5) and subjected to linkage analysis using Lep-Map3.[24] The resulting map was merged with the genome assembly using ALLMAPS.[25] The contigs assembling all reads were scaffolded manually based on the downsampled assembly and genetic map, for which 100 Ns were placed between the scaffolded contigs to generate pseudomolecule sequences. Assembly quality was evaluated using BUSCO v5 with the embryophyte_odb10 data[26] and telomere sequences (TTTAGGG) were searched using telomere_finder (https://github.com/MitsuhikoP/telomere_finder). Genetic distance was calculated using the alignment-free genetic distance estimation software, Mash.[27]

2.4. Repetitive sequence analysis and gene prediction

Repetitive sequences were detected with RepeatMasker v4.1.2 (https://www.repeatmasker.org) using repeat sequences obtained from the pseudomolecule sequence using RepeatModeler v2.0.2 (https://www.repeatmasker.org) and from a dataset registered in Repbase.[28]

Potential protein-coding genes were predicted using Braker version 2.1.5[29] with the protein sequences of *Oryza sativa*,[30] *Zea mays*,[31] and *E. phyllopogon* (eo_v2).[8] To assign confidence to predicted genes, homologous genes against the eggNOG 5.0[32] and UniProtKB databases (12 August 2022)[33] were searched using eggNOG-Mapper 2.1.8[34] and DIAMOND 2.0.14,[35] respectively. The genes that hit against eggNOG and UniProtKB were classified as high-confidence (HC) genes; however, those that hit keywords related to transposable elements were classified as TE. The other genes were classified as low-confidence (LC) genes. Gene clustering was performed using OrthoFinder.[36] The mapping annotation of previously reported chromosomal sequences of *E. phyllopogon* (eo_v2)[8] was confirmed using Liftoff.[37] Enrichment analyses for gene ontology of the A and B genome-specific gene were performed using topGO in the R package[38] with *elim* algorithm[39] and multiple corrections were performed with false discovery rate.

2.5. Comparative genome structure analysis

Chromosome-level pseudomolecule sequences were aligned using minimap2 (v2.24)[40] and compared with closely related species using D-GENIES,[41] pafr in the R package (https://github.com/dwinter/pafr). Synteny and collinearity of the predicted genes were detected using MCScanX[42] and visualized using SynVisio.[43]

2.6. Whole-genome resequencing analysis

Genomic DNA was extracted from the leaves of the two United States and seven Japanese lines using a DNeasy Plant Mini Kit (Qiagen). Genomic DNA libraries for short-read sequencing were prepared and sequenced, as described above. In addition, short-read data for 84 *E. phyllopogon* lines were obtained from a public database[8] (accession numbers CRA005291 and CRA005559), which contained two United States, 55 Chinese (including nine northeast China, 39 southeast China, and seven Hainan), 25 Italian, one Korean, and one Malaysian line. The sequence reads for 93 lines were treated as described above, and clean reads were mapped to the pseudomolecule sequence, using Bowtie2[21]; sequence variants, for example, SNPs and InDels, were called using BCFtools.[22] The variants were filtered using VCFtools[23] (parameters: minDP 5, minQ 10, maf 0.05, max-maf 0.95, and max-missing 0.8). The effects of the variants on gene function were annotated using SnpEff v4.3.[44] Principal component analysis (PCA) was performed for all SNPs using PLINK1.9.[45] The maximum

likelihood phylogenetic relationship was inferred with synonymous SNPs using RAxML[46] with the GTRGAMMA model and 1,000 bootstraps. Population structure analysis was performed using the ADMIXTURE ver. 1.3.0[47] with 1,000 bootstraps.

3. Results

3.1. Genome sequence and *de novo* assembly

Based on the *k*-mer frequency analysis using short-read sequences (23.3 Gb), the genome size of *E. phyllopogon*, R511 was estimated as 1.03 Gb (Fig. 1). Subsequently, 4.5 million HiFi sequence reads (68.9 Gb; 66.9 × coverage of the estimated genome size, N50 = 14.9 kb) obtained from two single-molecule real-time cells, were assembled into primary contigs (EPH_r1.0) and haplotigs. The assemblies consisted of 1.00 Gb of primary contigs (EPH_r1.0, including 827 sequences with an N50 length of 45 Mb, Table 1) and 50.9 Mb haplotigs (including 1,647 sequences with an N50 length of 33 kb, Supplementary Table S1). The presence of a single peak in the *k*-mer distribution and the observation of short haplotigs suggests a high level of homogeneity in the genome. In the primary assembly, 38 telomere repeat sequences (TTTAGGG) were detected at the ends of 26 contigs. Of these, 12 and 14 contigs had telomere repeat sequences at both ends and at one end, respectively (Table 2). These results suggested that the 12 contigs were assembled at a gapless telomere-to-telomere (T2T) level.

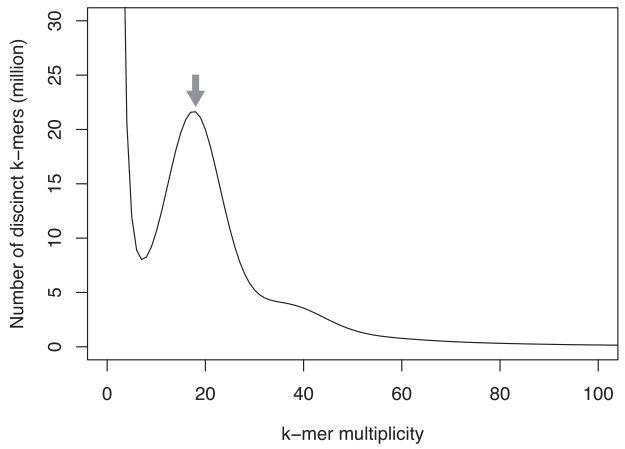

Figure 1. Estimation of the genome size of *E. phyllopogon*, based on *k*-mer analysis (*k* = 17) with given multiplicity values. The grey arrow indicates the peak used by estimation.

Table 1. Statistics of the genome assembly of *E. phyllopogon*

	EPH_r1.1	EPH_r1.0
Total scaffold size (bp)	1,002,757,761	1,002,676,961
Number of scaffolds	19	827
Number of chromosomes	18	18
Scaffold N50 length (bp)	58,069,259	45,224,446
Longest sequence size (bp)	72,873,931	67,065,931
No. of gap	808	–
gap length (bp)	80,800	–
Complete genome BUSCOs	98.9	98.9
Number of HC genes predicted	64,805	–
Complete HC protein BUSCOs	95.1	–

It is expected that higher coverage of HiFi reads will result in increased assembly contiguity. Whereas PacBio recommends coverages of 10 to 15-fold per haplotype for *de novo* assembly using HiFi reads (https://www.pacb.com), given very high coverage, ultra-low frequency of errors may negatively impact the contiguity.[17] A downsampling strategy was tested to determine whether low-depth reads could provide gapless T2T assemblies. Of the 68.9 Gb Hifi reads, we sampled 3 million (45.8 Gb) and established another assembly, EPH_r0.1 (Supplementary Table S1). This assembly contained 580 contigs spanning 1.00 Gb with an N50 of 43 Mb. The contigs of EPH_r0.1 were aligned with those of EPH_r1.0. Six EPH_r0.1 contigs participated in scaffold 10 EPH_r1.0 contigs to generate four scaffold sequences (Supplentary Fig. S1). Of these, three and one scaffold sequences contained telomere sequences at both ends and one end, respectively (Table 2).

3.2. Chromosome-level pseudomolecule sequence construction

A genetic mapping approach was used to construct the remaining two chromosome-level sequences. First, a genetic map was constructed based on the SNPs identified using ddRAD-seq analysis (Supplementary Table S2). An average of 6.0 M ddRAD-seq reads per sample were obtained, of which 95.0% were mapped onto the EPH_r1.0 assembly. After filtering by SNP-calling quality, 14,420 high-confidence SNPs were identified in 26 primary contigs; however, they were not detected in the remaining 801 contigs. A total of 12,643 SNPs were separated into 18 linkage groups and ordered to cover 2,271.6 cm in length (Supplementary Table S3 and Supplementary Fig. S2).

Based on genetic mapping, 18 T2T or chromosome-level contig sequences were fully constructed with the five categories (Table 2): (i) one linkage group supported one contig (in nine sequences); (ii) two linkage groups covered one contig because of the absence of SNPs in the middle of the chromosomes (in three sequences); (iii) one linkage group corresponded to two contigs probably because of a misjoining of the linkage map (in two sequences); (iv) one linkage group corresponded to two contigs to join them into a chromosome-level contig (in two sequences); and (v) no linkage groups were constructed for contigs owing to a lack of SNPs on the entire chromosome (in two sequences). Sequences composed of multiple contigs were connected with 100 Ns. The 801 contigs without SNPs were also connected to 100 Ns to build ch00. The resulting assembly (EPH_r1.1) spanned 1,002.8 Mb in length (Table 1). The complete BUSCO score of EPH_r1.1 showed 98.9%, and the single-copy and duplicated BUSCO scores showed 9.4% and 89.5%, respectively.

To identify the A and B genomes of the tetraploid *E. phyllopogon*, 18 chromosomal sequences were compared with those of a diploid relative, *E. haploclada*.[7] As expected, the nine pairs of the sequences of *E. phyllopogon* corresponded to nine chromosome sequences of *E. haploclada* (Supplementary Fig. S3A). In accordance with genetic distance, sequences close to and distant from *E. haploclada* were named A and B genomes, respectively (Supplementary Fig. S3B). The nomenclature and direction of the scaffold were based on nine chromosomes of foxtail millet (*Setaria italica*)[48] (Fig. 2A), which are phylogenetically close and have the same chromosome number as *E. phyllopogon*.[8] The established pseudomolecule sequences

Table 2. Chromosome scaffolding status by downsampling and genetic map

Chromosome	LG	No. contigs of EPH_r1.0	T2T	Gap	Category
ch1A	8	2	Y	Y	(1)
ch2A	5	1	Y	N	(1)
ch3A	11,12	2	Y	Y	(2)
ch4A	17,18	1	Y	N	(2)
ch5A	3	1	Y	N	(1)
ch6A	1	1	Y	N	(3)
ch7A	14	2	N	Y	(4)
ch8A	7	1	Y	N	(1)
ch9A	2	1	Y	N	(1)
ch1B	9	3	N	Y	(4)
ch2B	15,16	1	Y	N	(2)
ch3B	NA	1	Y	N	(5)
ch4B	10	3	Y	Y	(1)
ch5B	6	1	Y	N	(1)
ch6B	1	1	Y	N	(3)
ch7B	4	1	Y	N	(1)
ch8B	NA	1	Y	N	(5)
ch9B	13	2	Y	Y	(1)

Category: (1) one linkage group supported one contig; (2) two linkage groups covered one contig because of the absence of SNPs in the middle of the chromosomes; (3) one linkage group corresponded to two contigs probably because of a misjoining of the linkage map; (4) one linkage group corresponded to two contigs to join them into a chromosome-level contig; and (5) no linkage groups were constructed for contigs owing to a lack of SNPs on the entire chromosome.

corresponded one-to-one with previously reported chromosome sequences of *E. phyllopogon* (eo_v2)[8] (Fig. 2B).

3.3. Repeat sequence analysis and gene prediction

Repeat sequences occupied 455 Mb of 1.0 Gb in EPH_r1.1 (45.40%) (Table 3). In *E. phyllopogon*, the dominant repetitive sequences were long terminal repeat (LTR) retroelements (16.46%), followed by DNA transposons (4.36%). Notably, repeat content was different between the A genome (194 Mb, 42.93%) and the B genome (252 Mb, 48.47%). The distribution of repeat sequences was similar to that of previously reported genome sequences (eo_v2)[8] except for unclassified sequences that may be unique repetitive sequences in *Echinochloa* (Table 3). This unclassified repeat in the current assembly was ~40 Mb longer than that in eo_v2 and clustered in the middle of the chromosomes (Supplementary Fig. S4) that were not represented in eo_v2.

A total of 132,212 potential protein-coding sequences were identified in the current genome assembly (Supplementary Table S1), based on *ab initio* prediction and amino acid sequence homology among three *Poaceae* species, *O. sativa* (IRGSP-1.0),[49] *Z. mays* (B73_v4),[31] and *E. phyllopogon* (eo_v2).[8] In subsequent gene annotation analysis, 64,805 genes with gene descriptions were assigned as HC genes, 4,809 genes with TE-related terms were assigned as TE-related genes, and the remaining 62,598 genes were classified as LC genes (Supplementary Table S4). BUSCO analysis of all HC genes indicated that the scores for complete BUSCOs were 96.8% and 95.1%, respectively (Supplementary Table S1). Of the HC genes, 97.8% (=63,378/64,805) sequences hit the predicted genes of *E. phyllopogon* (eo_v2).[8]

A total of 32,337 HC, 2,334 TE, and 30,544 LC genes were predicted in the A genome, 30,889 HC, 2,458 TE, and 31,791 LC genes were found in the B genome (Supplementary Table S1). The genes predicted in the A and B genomes of *E. phyllopogon* were clustered among the *Poaceae* species, *O. sativa* and *S. italica* (Fig. 3). In total, 27,120 clusters were identified. Of these, 16,144 clusters (59.5%) were shared among the four sets. A total of 1,832 clusters (343 + 1,225 + 264) were unique to *E. phyllopogon*, of which 343 and 264 clusters were unique to A genome and B genome, respectively.

Of 1,663 A genome-specific clusters (343 + 458 + 682 + 180) and 1,223 B genome-specific clusters (264 + 270 + 531 + 158) (Fig. 3), 7 and 20 GO terms were significantly enriched after in the A and B genomes, respectively (Supplementary Table S5). The GO term 'response to cold' represented by genes encoding CBL-interacting protein kinase (CIPK7) was enriched in the A genome. A total of 34 copies of *CIPK7* genes were found in EPH_r1.1 (32 and 2 copies in the A and B genomes, respectively). Interestingly, 32 copies of the 34 genes in the A genome were clustered on the chromosome 9A. On the other hand, in eo_v2, the copy number of *CIPK7* was only 10 and no clusters were found in the chromosome 9A.

3.4. Comparative analysis of the genome structures of the three species

Based on sequence similarity, the structures of the A and B genomes (EPH_r1.1) and that of *S. italica* were well conserved (Fig. 2). Furthermore, we analysed synteny based on orthologous gene orders (Fig. 4A). In sequence similarity and synteny analyses, potentially large inversions between the *E. phyllopogon* and *S. italica* genomes were observed on chromosomes one (inversion size of 10 Mb), four (6 Mb), five (15 Mb), six (10 Mb), and seven (10 Mb) (Figs. 2A and 4B). To clarify which genome had inversion events, the *E. phyllopogon* and *S. italica* were compared with those of *O. sativa* (Fig. 4). In

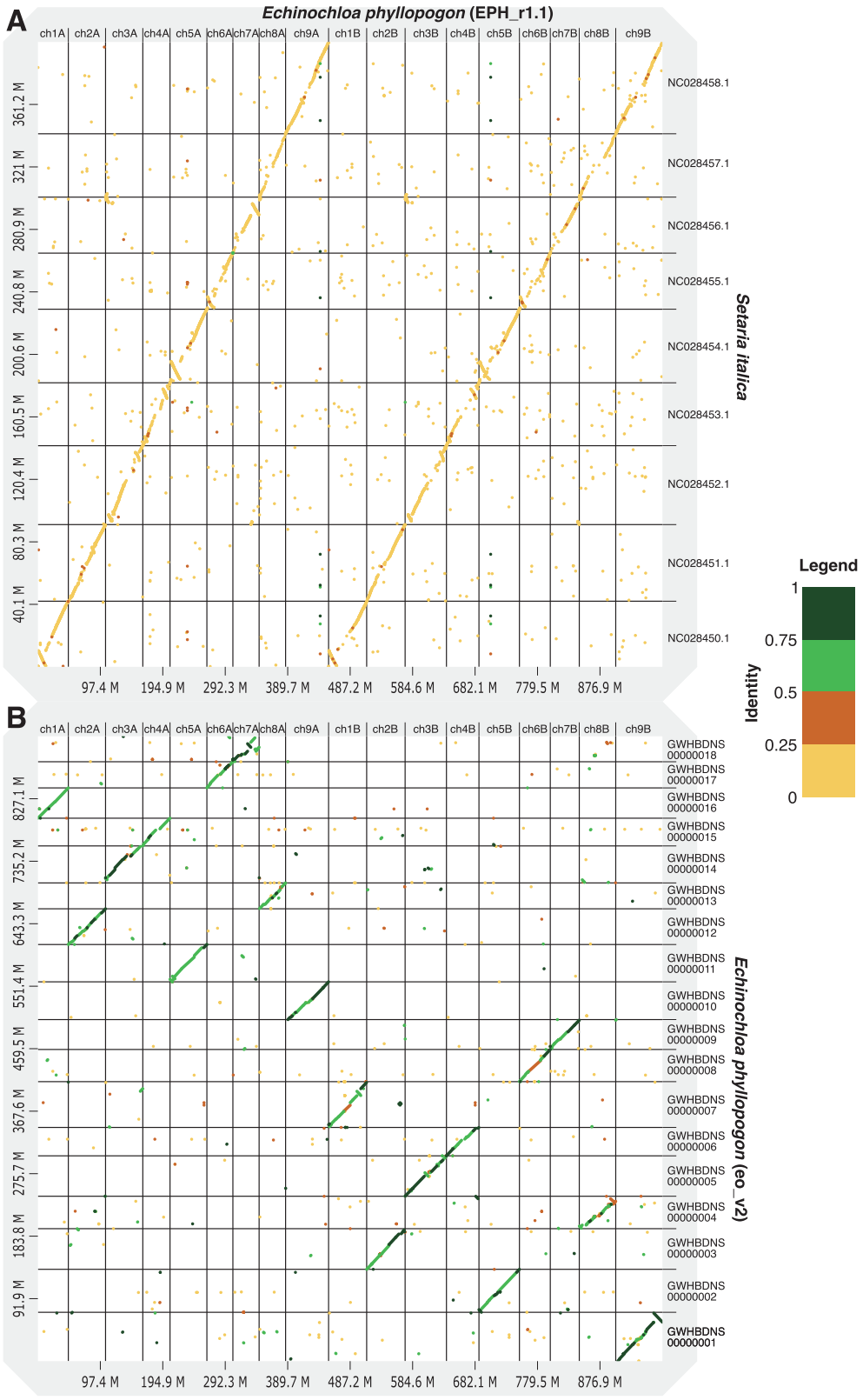

Figure 2. Comparative analysis of the genome sequence and structure among *E. phyllopogon* (EPH_r1.1), (A) *S. italica*, and (B) *E. phyllopogon* (eo_v2). Dots and colours indicate genome structure and sequence similarity, respectively.

total, 503 and 458 synteny regions were found in A and B genomes of *E. phyllopogon*, respectively, whereas 489 synteny regions were observed in *S. italica*. The synteny blocks were well conserved between *E. phyllopogon* and *S. italica*; however, some chromosomal rearrangements for *O. sativa* were found. A potential inversion on chromosome five was found in the A and B genomes of *E. phyllopogon*, whereas inversions on chromosomes one, four, and six were found in *S.*

Table 3. Repetitive sequences in the *E. phyllopogon* and *E. oryzicola* genomes

	E. phyllopogon			*E. oryzicola*		
Repeat type	No. of repetitive elements	Length (bp)	%	No. of repetitive elements	Length (bp)	%
SINEs	11,066	1,681,855	0.17	10,813	1,642,459	0.17
LINEs	35,800	16,389,936	1.63	35,819	16,450,317	1.74
LTR elements	240,701	165,030,695	16.46	243,265	168,554,201	17.83
DNA transposons	162,401	43,757,832	4.36	160,524	43,490,420	4.6
Unclassified	430,222	209,510,768	20.89	436,432	169,542,473	17.93
Small RNA	16,425	8,536,289	0.85	11,916	2,078,326	0.22
Satellites	4,342	659,998	0.07	3,614	505,057	0.05
Simple repeats	172,531	8,308,353	0.83	172,562	8,742,020	0.92
Low complexity	21,409	1,076,893	0.11	20,793	1,077,297	0.11
Total	1,094,897	454,952,619	45	1,095,738	412,082,570	44

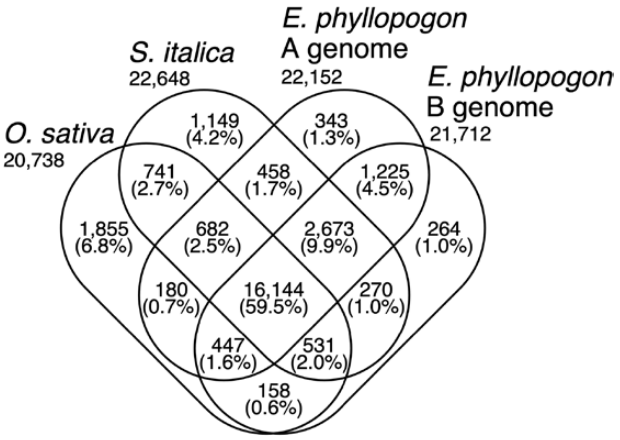

Figure 3. Venn diagram of the numbers of gene clusters in the A and B genomes of *E. phyllopogon* and two Poaceae species.

italica. Another inversion was found on chromosome seven, which was only in the A genome of *E. phyllopogon*, and not in the B genome of *S. italica*.

3.5. Phylogenomic and population structure analysis

Phylogenomic and population structure relationships of the 93 lines were investigated using whole genome sequencing data. A mean of 131 M reads (19.71 Gb) were mapped to EPH_r1.1 as a reference, with an average map rate of 97.11% (Supplementary Table S6). A total of 5,756,135 SNPs were detected across 93 lines, of which 105,949 were synonymous.

The PCA, in which the proportions of variance were 38.58% on PC1 and 9.86% on PC2, showed three clusters (Fig. 5A): (i) seven Hainan lines, (ii) 38 Southern Chinese and one Malaysian line, and (iii) 25 Italian, nine northeast Chinese, one Southern Chinese, five Japanese, four United States, and one Korean line. The two Japanese lines, both of which were from outside of paddy fields, were not included in any cluster. These clusters corresponded to those reported in a previous study[8]: (i) var. *hainanensis* (HN); (ii) lower latitude (LL); and (iii) higher latitude (HL). Therefore, clusters A, B, and C were named HN, LL, and HL, respectively. The results further support that *E. phyllopogon* and *E. oryzicola* are indistinguishable from each other.

The phylogenetic tree had three main lineages—HN, LL, and HL—corresponding to the three PCA clusters (Fig. 5B). The HN group included seven Hainan lines, together with two Japanese lines that were not included in any cluster in the PCA. The HL group consisted of nine northeast Chinese, five Japanese, one Korean, 25 Italian, and four United States lines, whereas the LL group consisted of 39 lines from 38 Southern Chinese and one Malaysian line. The US line R511, for which we determined the genome sequence in this study, belongs to HL, whereas eo_v2 belongs to LL. The population genetic structure examined using ADMIXTURE revealed that the optimal number of clusters was five (Supplementary Fig. S5A). This result agrees with the PCA classifications and the phylogenetic tree. It was suggested that the two Japanese lines close to HN were potential hybrids between the lines from Hainan, Japan, and Southern China (Fig. 5C and Supplementary Fig. S5).

4. Discussion

Here, we present the genome assembly of the tetraploid pernicious weed, *E. phyllopogon* at the chromosome level (Tables 1 and 2). High coverage and high-quality HiFi reads could contribute to complete *de novo* assembly (Table 1). Telomeric repeat analysis revealed that 12 of the 18 pseudomolecule sequences were constructed as T2T and gap-free contigs, whereas the remaining six were constructed as chromosome-level contigs with eight gaps (Table 2). Recently, T2T genome assemblies have been reported in humans,[9] chickens,[10] fungi,[11,12] and plankton.[13,14] In plants, gapless T2T genome assemblies have been reported in some chromosomes of maize,[50] *Arabidopsis thaliana*,[51] and a complete set of chromosomes of watermelon (*Citrullus lanatus*).[52] Here, we reported the chromosomal sequences of *E. phyllopogon* at the T2T level (Tables 1 and 2), even though *E. phyllopogon* has a larger and more complex tetraploid genome than those reported thus far.

Constructing high-quality genome assembly at the chromosome- and T2T-level is still laborious work, requiring additional experiments and data processing following the initial assembly of sequence reads. To achieve high-quality genome assembly, genetic mapping, optical mapping, and/or Hi-C are selectively utilized depending on plant materials and the initial assembly results.[53] Genetic linkage mapping has been widely employed to resolve misassemblies and to assign contigs to maps to build chromosome-level sequences

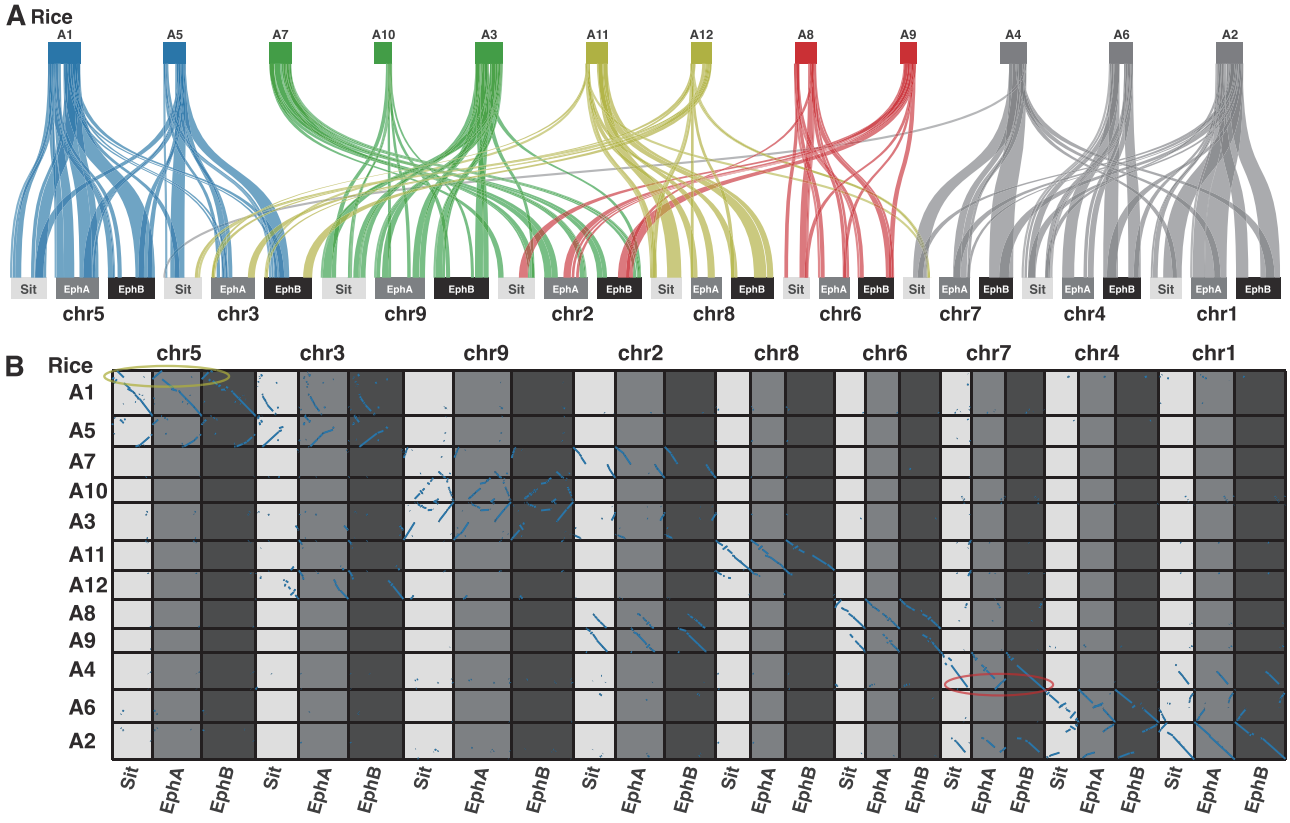

Figure 4. Synteny and collinearity of genes in the A and B genomes of *E. phyllopogon* (EphA and EphB) and *S. italica* (Sit) against rice chromosomes on the top. Colours of rice chromosomes indicate common ancestral genomes of Poaceae species. The yellow circle indicates *E. phyllopogon*-specific inversion and the red circle indicates A genome-specific inversion.

so far.[54–57] Our T2T-level assembly based on HiFi-reads was established using the downsampling strategy as well as genetic mapping. The accuracy of the assembly was verified by the homology to the previous assembly, eo_v2, and synteny conserved in the *Poaceae* (Figs. 2 and 4). Assembly methods together with the long-read sequencing technologies implemented in Revio system (PacBio) and PromethION (Oxford Nanopore Technologies, UK) would contribute to further high-quality chromosome- and T2T-level genome assembly.

The T2T genome enhances our understanding of the evolutionary processes underlying intricate genome structures, including repeat sequences, which have been historically underestimated. The 18 pseudomolecule sequences were grouped into A and B genomes in accordance with their genetic distances from a diploid relative, *E. haploclada* (Supplementary Fig. S3). Comparative analysis of the A and B genomes indicated that the B genome (520 Mb) was longer than the A genome (453 Mb), whereas the number of HC genes in the B genome (30,889) was smaller than that in the A genome (32,337) (Supplementary Table S1). This difference was observed across all chromosomes (Supplementary Fig. S4). Events of genome extension and gene disruption caused by repeat sequence accumulation (Table 3, 42.93% in A genome but 48.47% in B genome) may often occur in the B genome before polyploidization to establish the tetraploid genome of *E. phyllopogon*.

The structure of the genome assembly obtained in this study was consistent with that reported in a previous study (eo_v2) (Fig. 2 and Supplementary Fig. S4). The genome assembly in this study (1.0 Gb) was longer than the eo_v2 assembly (945 Mb), which was close to the estimated size of the genome of *E. phyllopogon* (1.0 Gb) (Fig. 1). The difference in length was derived from the complete assemblies of the repetitive sequences in the middle of the chromosomes, which were absent in the previously reported sequence eo_v2 (Supplementary Fig. S4). Additionally, we identified structural variations between the sequences from this study and those from a previous study (Fig. 2B and Supplementary Fig. S4). As the two lines belong to different clades (Fig. 5B), there may be structural polymorphisms in *E. phyllopogon*. Moreover, in a comparative analysis of genome structures among *Poaceae* species (Fig. 4), chromosome structure variations were uniquely found in *E. phyllopogon* and only A genome.

Phylogenomic analyses revealed at least three groups (HN, HL, and LL) of *E. phyllopogon* (Fig. 5). In HL, northeast China, Japan, the United States, and Italian lines were included, and the Italian lines were located at the base of the lineage (Fig. 5B). This suggests that the ancestor of the HL lineage first invaded Italy and subsequently expanded into northeast China, Japan, and the United States, rather than parapatric differentiation within China. In the HN group, which has been reported as a new variety,[8] the two Japanese lines may be hybrids of Hainan, Japan, and Southern China. The two Japanese lines possessed morphological characteristics, such as plagiotropic (or prostrate) tillers and small seed size in paddy fields,[3] which were distinguishable from those of other Japanese lines belonging to the HL group. Although it is unclear how the potential hybrids were generated and inhabited Japan, weed control to prevent migration and expansion of the novel pernicious weeds is required inside and outside paddy fields.

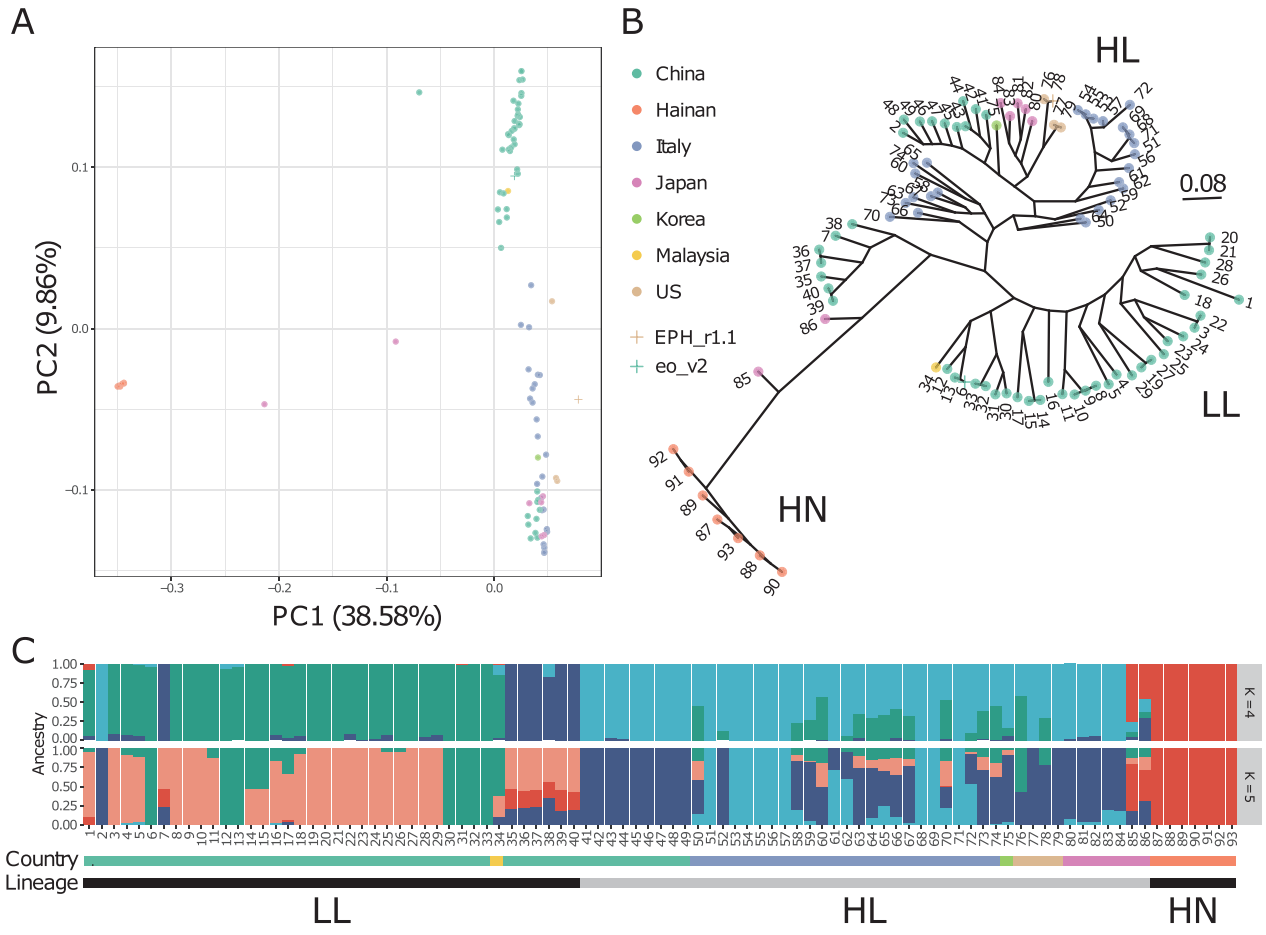

Figure 5. Genetic relationship for *E. phyllopogon*. (A) PCA, (B) phylogenetic tree using the maximum likelihood method, and (C) the genetic structure of $K = 4$ and 5. Dot colours in (A) and (B) were consistent with the horizontal bar colours in (C).

Subgenome-specific enrichment analysis showed that *CIPK7* with the GO term of 'response to cold' was clustered in chromosome 9A of EPH_r1.1, whereas this cluster was not found in eo_v2. CIPK7 interacts with calcineurin B-like (CBL) protein one to regulate cold signalling transduction of the phosphorylation cascade, which leads to freezing tolerance and expression of cold-responsive genes in *Arabidopsis*.[58,59] CIPK7 is also a key regulator of response to the ratio of available carbon and nitrogen nutrients in *Arabidopsis*.[60] CBL and CIPK consist of a large number of gene family members in *Arabidopsis* and rice, and it suggests a rather complex signalling network of potential interactions.[61] Therefore, the A genome-specific CIPK7 cluster in EPH_r1.1 might enable *E. phyllopogon* to adapt to high-latitude regions with severe low temperatures and weak light conditions such as Europe, the United States, and northeastern Asia. We hypothesized that the copy number variations as well as sequence variants of *CIPK7* genes potentially impact the cold adaptation of *E. phyllopogon*.

The highly accurate genome information of *E. phyllopogon* obtained in this study provides insights into the molecular mechanisms underlying multiple resistance to herbicides to avoid serious crop yield loss. Furthermore, *E. phyllopogon* genomics substantially contributes to the understanding of weed adaptation to farms and the pathways of worldwide weed invasion.

Acknowledgements

We thank Y. Kishida, S. Nakayama and A. Watanabe (Kazusa DNA Research Institute) for their technical assistance.

Funding

This work was supported by JSPS KAKENHI grant numbers 19H02955, 22H02347, 22H05172, 22H05181, and 23K18025 and the Kazusa DNA Research Institute Foundation.

Data availability

Raw sequencing reads and assemblies were deposited in the DNA Data Bank of Japan (DDBJ) under the accession number PRJDB14855 and BTCW00000000.1. Genomic information generated in this study is available from Plant GARDEN (https://plantgarden.jp/).

Supplementary Data

Supplementary data are available at *DNARES* online.

References

1. Yabuno, T. 1981, Cytological relationship between *Echinochloa oryzicola* Vasing and the french strain of *E. phyllopogon* stapf subsp oryzicola (Vasing) Koss, *Cytologia*, **46**, 393–6.

2. Yamasue, Y. 2001, Strategy of *Echinochloa oryzicola* Vasing for survival in flooded rice, *Weed Biol. Manag.*, **1**, 28–36.
3. Yasuda, K., Mori, K., and Nakayama, Y. 2020, A tetraploid *Echinochloa* with plagiotropic tillers: its distribution and habitat in the northern part of the main island of Japan, *Weed Biol. Manag.*, **20**, 82–8.
4. Iwakami, S., Endo, M., Saika, H., et al. 2014, Cytochrome P450 CYP81A12 and CYP81A21 are associated with resistance to two acetolactate synthase inhibitors in *Echinochloa phyllopogon*, *Plant Physiol.*, **165**, 618–29.
5. Iwakami, S., Kamidate, Y., Yamaguchi, T., et al. 2019, CYP81A P450s are involved in concomitant cross-resistance to acetolactate synthase and acetyl-CoA carboxylase herbicides in *Echinochloa phyllopogon*, *New Phytol.*, **221**, 2112–22.
6. Suda, H., Kubo, T., Yoshimoto, Y., et al. 2023, Transcriptionally linked simultaneous overexpression of P450 genes for broad-spectrum herbicide resistance, *Plant Physiol.*, **192**, 3017–29.
7. Ye, C.-Y., Wu, D., Mao, L., et al. 2020, The genomes of the allohexaploid echinochloa crus-galli and its progenitors provide insights into polyploidization-driven adaptation, *Mol. Plant*, **13**, 1298–310.
8. Wu, D., Shen, E., Jiang, B., et al. 2022, Genomic insights into the evolution of *Echinochloa* species as weed and orphan crop, *Nat. Commun.*, **13**, 689.
9. Nurk, S., Koren, S., Rhie, A., et al. 2022, The complete sequence of a human genome, *Science*, **376**, 44–53.
10. Huang, Z., Xu, Z., Bai, H., et al. 2023, Evolutionary analysis of a complete chicken genome, *Proc. Natl. Acad. Sci. U.S.A.*, **120**, e2216641120.
11. Bowyer, P., Currin, A., Delneri, D., and Fraczek, M.G. 2022, Telomere-to-telomere genome sequence of the model mould pathogen *Aspergillus fumigatus*, *Nat. Commun.*, **13**, 5394.
12. Kurokochi, H., Tajima, N., Sato, M.P., et al. 2023, Telomere-to-telomere genome assembly of matsutake (*Tricholoma matsutake*), *DNA Res.*, **30**, dsad006.
13. Bliznina, A., Masunaga, A., Mansfield, M.J., et al. 2021, Telomere-to-telomere assembly of the genome of an individual Oikopleura dioica from Okinawa using Nanopore-based sequencing, *BMC Genom.*, **22**, 222.
14. Giguere, D.J., Bahcheli, A.T., Slattery, S.S., et al. 2022, Telomere-to-telomere genome assembly of Phaeodactylum tricornutum, *PeerJ*, **10**, e13607.
15. Tsuji, R., Fischer, A.J., Yoshino, M., Roel, A., Hill, J.E., and Yamasue, Y. 2003, Herbicide-resistant late watergrass (*Echinochloa phyllopogon*): similarity in morphological and amplified fragment length polymorphism traits, *Weed Sci.*, **51**, 740–7.
16. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
17. 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.
18. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. 2018, MUMmer4: a fast and versatile genome alignment system, *PLoS Comput. Biol.*, **14**, e1005944.
19. Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. 2012, Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species, *PLoS One*, **7**, e37135.
20. 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.
21. Langmead, B. and Salzberg, S.L. 2012, Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, **9**, 357–9.
22. Danecek, P., Bonfield, J.K., Liddle, J., et al. 2021, Twelve years of SAMtools and BCFtools, *GigaScience*, **10**, giab008.
23. Danecek, P., Auton, A., Abecasis, G., et al.; 1000 Genomes Project Analysis Group. 2011, The variant call format and VCFtools, *Bioinformatics*, **27**, 2156–8.
24. Rastas, P. 2017, Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data, *Bioinformatics*, **33**, 3726–32.
25. Tang, H., Zhang, X., Miao, C., et al. 2015, ALLMAPS: robust scaffold ordering based on multiple maps, *Genome Biol.*, **16**, 3.
26. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. 2021, BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, *Mol. Biol. Evol.*, **38**, 4647–54.
27. Ondov, B.D., Treangen, T.J., Melsted, P., et al. 2016, Mash: fast genome and metagenome distance estimation using MinHash, *Genome Biol.*, **17**, 132.
28. Bao, W., Kojima, K.K., and Kohany, O. 2015, Repbase update, a database of repetitive elements in eukaryotic genomes, *Mob. DNA*, **6**, 11.
29. 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.
30. Kawahara, Y., de la Bastide, M., Hamilton, J.P., et al. 2013, Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data, *Rice (New York, N.Y.)*, **6**, 4.
31. Jiao, Y., Peluso, P., Shi, J., et al. 2017, Improved maize reference genome with single-molecule technologies, *Nature*, **546**, 524–7.
32. Huerta-Cepas, J., Szklarczyk, D., Heller, D., et al. 2019, eggNOG 50: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, *Nucleic Acids Res.*, **47**, D309–14.
33. UniProt Consortium. 2021, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Res.*, **49**, D480–9.
34. Huerta-Cepas, J., Forslund, K., Coelho, L.P., et al. 2017, Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper, *Mol. Biol. Evol.*, **34**, 2115–22.
35. Buchfink, B., Reuter, K., and Drost, H.-G. 2021, Sensitive protein alignments at tree-of-life scale using DIAMOND, *Nat. Methods*, **18**, 366–8.
36. Emms, D.M. and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol.*, **20**, 238.
37. Shumate, A. and Salzberg, S.L. 2020, Liftoff: accurate mapping of gene annotations, *Bioinformatics*, **37**, 1639–43.
38. Adrian Alexa, J. R. 2017, *topGO*. Bioconductor.
39. Alexa, A., Rahnenführer, J., and Lengauer, T. 2006, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, *Bioinformatics*, **22**, 1600–7.
40. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, **34**, 3094–100.
41. Cabanettes, F. and Klopp, C. 2018, D-GENIES: dot plot large genomes in an interactive, efficient and simple way, *PeerJ*, **6**, e4958.
42. Wang, Y., Tang, H., Debarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.*, **40**, e49.
43. Bandi, V., and Gutwin, C. 2020, Interactive exploration of genomic conservation. In: Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020.
44. 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.
45. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. 2015, Second-generation PLINK: rising to the challenge of larger and richer datasets, *GigaScience*, **4**, 1–16.
46. Stamatakis, A. 2014, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, **30**, 1312–3.
47. Alexander, D.H., Novembre, J., and Lange, K. 2009, Fast model-based estimation of ancestry in unrelated individuals, *Genome Res.*, **19**, 1655–64.

48. Bennetzen, J.L., Schmutz, J., Wang, H., et al. 2012, Reference genome sequence of the model plant Setaria, *Nat. Biotechnol.*, **30**, 555–61.
49. Sakai, H., Lee, S.S., Tanaka, T., et al. 2013, Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics, *Plant Cell Physiol.*, **54**, e6.
50. Liu, J., Seetharam, A.S., Chougule, K., et al. 2020, Gapless assembly of maize chromosomes using long-read technologies, *Genome Biol.*, **21**, 121.
51. Wang, B., Yang, X., Jia, Y., et al. 2022, High-quality *Arabidopsis thaliana* Genome Assembly with Nanopore and HiFi Long Reads, *Genom. Proteom. Bioinform.*, **20**, 4–13.
52. Deng, Y., Liu, S., Zhang, Y., et al. 2022, A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding, *Mol. Plant*, **15**, 1268–84.
53. Rice, E.S. and Green, R.E. 2019, New approaches for genome assembly and scaffolding, *Annu. Rev. Anim. Biosci.*, **7**, 17–40.
54. Fierst, J.L. 2015, Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools, *Front. Genet.*, **6**, 220.
55. Gutiérrez-Valencia, J., Fracassetti, M., Berdan, E.L., et al. 2022, Genomic analyses of the Linum distyly supergene reveal convergent evolution at the molecular level, *Curr. Biol.*, **32**, 4360–4371.e6.
56. Shirasawa, K., Itai, A., and Isobe, S. 2021, Genome sequencing and analysis of two early-flowering cherry (Cerasus × kanzakura) varieties, 'Kawazu-zakura' and 'Atami-zakura', *DNA Res.*, **28**, dsab026.
57. Shirasawa, K., Nishio, S., Terakami, S., Botta, R., Marinoni, D.T., and Isobe, S. 2021, Chromosome-level genome assembly of Japanese chestnut (Castanea crenata Sieb et Zucc) reveals conserved chromosomal segments in woody rosids, *DNA Res.*, **28**, dsab016.
58. Huang, C., Ding, S., Zhang, H., Du, H., and An, L. 2011, CIPK7 is involved in cold response by interacting with CBL1 in *Arabidopsis thaliana*, *Plant Sci.*, **181**, 57–64.
59. Cheong, Y.H., Kim, K.-N., Pandey, G.K., Gupta, R., Grant, J.J., and Luan, S. 2003, CBL1, a calcium sensor that differentially regulates salt, drought, and cold responses in *Arabidopsis*, *Plant Cell*, **15**, 1833–45.
60. Yasuda, S., Aoyama, S., Hasegawa, Y., Sato, T., and Yamaguchi, J. 2017, *Arabidopsis* CBL-interacting protein kinases regulate carbon/nitrogen-nutrient response by phosphorylating ubiquitin ligase ATL31, *Mol. Plant*, **10**, 605–18.
61. Kolukisaoglu, U., Weinl, S., Blazevic, D., Batistic, O., and Kudla, J. 2004, Calcium sensors and their interacting protein kinases: genomics of the arabidopsis and rice CBL-CIPK signaling networks, *Plant Physiol.*, **134**, 43–58.