



Resource Article: Genomes Explored

Chromosome-scale assembly of barley cv. ‘Haruna Nijo’ as a resource for barley genetics

Areej Sakkour¹, Martin Mascher ^{2,3}, Axel Himmelbach ²,
Georg Haberer ⁴, Thomas Lux ⁴, Manuel Spannagl ⁴, Nils Stein ^{2,5}, Shoko Kawamoto ⁶, and Kazuhiro Sato ^{1*}

¹Institute of Plant Science and Resources, Okayama University, Kurashiki 710-0046, Japan, ²Department Genebank, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, 06466 Seeland, Germany, ³German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany, ⁴Plant Genome and Systems Biology (PGSB), Helmholtz Center Munich, German Research Center for Environmental Health, 85764 Neuherberg, Germany, ⁵Center for Integrated Breeding Research (CiBreed), Georg-August-University Göttingen, 37075 Göttingen, Germany, and ⁶Department of Informatics, National Institute of Genetics, Mishima 411-8540, Japan

*To whom correspondence should be addressed. Tel: +81-86-434-1244. Fax: +81-86-434-1249. Email: kzsato@okayama-u.ac.jp

Received 26 December 2021; Editorial decision 9 January 2022; Accepted 10 January 2022

Abstract

Cultivated barley (*Hordeum vulgare* ssp. *vulgare*) is used for food, animal feed, and alcoholic beverages and is widely grown in temperate regions. Both barley and its wild progenitor (*H. vulgare* ssp. *spontaneum*) have large 5.1-Gb genomes. High-quality chromosome-scale assemblies for several representative barley genotypes, both wild and domesticated, have been constructed recently to populate the nascent barley pan-genome infrastructure. Here, we release a chromosome-scale assembly of the Japanese elite malting barley cultivar ‘Haruna Nijo’ using a similar methodology as in the barley pan-genome project. The 4.28-Gb assembly had a scaffold N50 size of 18.9 Mb. The assembly showed high collinearity with the barley reference genome ‘Morex’ cultivar, with some inversions. The pseudomolecule assembly was characterized using transcript evidence of gene projection derived from the reference genome and *de novo* gene annotation achieved using published full-length cDNA sequences and RNA-Seq data for ‘Haruna Nijo’. We found good concordance between our whole-genome assembly and the publicly available BAC clone sequence of ‘Haruna Nijo’. Interesting phenotypes have since been identified in Haruna Nijo; its genome sequence assembly will facilitate the identification of the underlying genes.

Key words: *Hordeum vulgare*, full-length cDNA, RNA-Seq, genome sequencing, pseudomolecules

1. Introduction

Cultivated barley is used for many purposes, including animal feed, human food, and malting for brewing. Malting barley has only been cultivated in Japan for ca. 140 years.¹ The founder cultivars were mainly introduced from Europe and crossed with Japanese

landraces, which prior to that had been used for human food. In 1978, the malting barley cultivar ‘Haruna Nijo’ was released from Sapporo Breweries (Tokyo, Japan) and has since been used as a donor of high-quality profiles in Japanese malting barley breeding programs.

At Okayama University (Okayama, Japan), ‘Haruna Nijo’ is used as a key genotype in genetics and genomics studies. ‘Haruna Nijo’ was used for the generation of expressed sequence tags (see also <https://harvest.ucr.edu/> last accessed Jan 19, 2022).² Using these transcript sequences, a high-density genetic map was constructed from a cross between ‘Haruna Nijo’ and the wild barley (*H. vulgare* ssp. *spontaneum*) accession ‘OUH602’,^{3,4} and a set of recombinant chromosome substitution lines was developed.⁵ ‘Haruna Nijo’ was used to generate full-length cDNA (fl-cDNA) sequences,^{6,7} which have been used for the annotation of gene models in the reference genome of the cultivar ‘Morex’.^{8–11} Whole-genome shotgun sequencing was performed for ‘Haruna Nijo’ to enable the estimation of the genic regions of the genome.^{8,12} A BAC library of ‘Haruna Nijo’ was also constructed to isolate the genes responsible for major traits,¹³ such as hull-less caryopsis¹⁴ and seed dormancy.¹⁵ The mitochondrial genome of ‘Haruna Nijo’ was also sequenced and found to be highly similar to that of ‘OUH602’.¹⁶

After the release of a high-quality barley genome assembly generated using BAC-by-BAC sequencing and scaffold alignment,⁹ several whole-genome shotgun assembly techniques were developed for Illumina short reads, such as the DeNovoMAGIC assembly pipeline (NRGene, Ness Ziona, Israel), the TRITEX pipeline,¹⁷ and w2rap-contigger.¹⁸ Using these assembly methodologies, the global landscape of the barley genome (pan-genome)¹⁹ was recently analysed using 20 domesticated and wild accessions¹⁰ based on a selection of 22,000 genomic profiling datasets (GBS) from German gene bank accessions.²⁰

Here, we utilized the TRITEX pipeline to generate a chromosome-scale genome assembly of ‘Haruna Nijo’. We aligned the assembly to the most recently updated assembly, ‘Morex’V3,¹¹ to identify genomic differences among the genotypes. We also aligned the assembly to the published BAC sequences used for gene isolation to estimate the quality of the assembly. A similar sequencing methodology was also recently applied to the wild barley accession ‘OUH602’²¹; however, the assembly of the ‘Haruna Nijo’ genome is desirable for its economic and breeding importance.

The present barley genome annotation, e.g. EnsemblPlants (http://plants.ensembl.org/Hordeum_vulgare/ last accessed Jan 19, 2022), is based on ‘Morex’, which is the North American malting cultivar with a Manchurian landrace pedigree, and differs from malting barleys in other areas of the world. In a recent barley pan-genome analysis,¹⁰ gene projection was performed using informant gene models of ‘Morex’, the German malting cultivar ‘Barke’, and an Ethiopian landrace ‘HOR10350’, which were predicted from transcriptome data and protein homology information using a previously described annotation pipeline.⁹ In addition to this gene projection analysis, we performed *de novo* gene annotation for ‘Haruna Nijo’ using published fl-cDNA sequences and RNA-Seq data. These procedures may provide alternative gene annotation information on the barley genome by characterizing different sources of transcript and protein information from fl-cDNA sequences and RNA-Seq data.

2. Materials and methods

2.1 DNA extraction, library construction, and sequencing

High-molecular-weight DNA was isolated from leaf material of seedlings of ‘Haruna Nijo’²² and size selected for a molecule size of 40 kb or higher. The 440-bp paired-end (PE) libraries were prepared with the Hyper Kapa Library Preparation kit (Kapa Biosystems) with no

polymerase chain reaction amplification. The 8- to 10-kb mate-pair (MP) libraries were constructed with the Nextera Mate Pair library Sample Prep kit (Illumina, San Diego, CA, USA) followed by the TruSeq DNA Sample Prep kit. The 10X libraries were constructed with the Chromium Genome Library Kit & Gel Bead Kit v2 (10X Genomics). Sequencing was performed following Sato et al.²¹ In brief, the 440-bp PE libraries were sequenced for 251 cycles using a NovaSeq 6000 system (Illumina). The 10X and MP libraries were sequenced for 151 cycles from each end of the fragments on the NovaSeq 6000 system. All libraries were prepared and sequenced at the University of Illinois Roy J. Carver Biotechnology Center (Urbana, IL, USA). *In situ* Hi-C libraries were prepared as described by Padmarasu et al.²³ Sequencing data generated from each of the libraries are listed in [Supplementary Table S1](#). The Hi-C data were used to prepare chromosome-scale assemblies using the TRITEX pipeline,¹⁹ which was also used for the contig assembly and scaffolding with the PE, MP, and 10X data ([Supplementary Table S1](#)).

2.2 Transcript sequencing

Published RNA-Seq reads from the seedling root, shoot, spike at flowering, and seeds of ‘Haruna Nijo’¹² were used for the transcript sequencing. An additional RNA sample of a young spike (3 cm in length) from ‘Haruna Nijo’ was also extracted and subjected to an RNA-Seq analysis, as described by Sato et al.¹² These RNA-Seq libraries were sequenced with the MiSeq Reagent Kit V3 (2 × 300 bp cycles) on a MiSeq system (Illumina).

2.3 Gene projection

To derive the projected gene structures for ‘Haruna Nijo’, informant gene models of ‘Morex’, ‘Barke’, and ‘HOR10350’ were employed, which were predicted from transcriptome data and protein homology information¹⁰ using a previously described annotation pipeline.⁹ The projection was based on a stepwise procedure, as previously described.^{10,21} Briefly, BLASTN²⁴ and Exonerate alignments²⁵ of the coding sequences (CDSs) of each of the barley sources of the ‘Haruna Nijo’ genome sequence were computed. The matches were clustered by their genomic loci, and the top-scoring match was selected using a stepwise integration approach. In addition to protein-coding genes, ‘pseudogene’-type mappings were previously projected and included in the CDSs and GFF files but were obviously missing from the protein sequence files.

2.4 *De novo* gene annotation using RNA-Seq and fl-cDNA sequences

A structural gene annotation was performed by combining *de novo* gene calling and homology-based approaches with RNA-Seq, protein, isoseq, and fl-cDNA datasets. Using evidence derived from expression data, RNA-Seq sequences were first mapped against the ‘Haruna Nijo’ genome assembly using STAR²⁶ (version 2.7.8a) and subsequently assembled into transcripts using StringTie²⁷ (version 2.1.5; parameters -m 150-t-f 0.3). Triticeae protein sequences obtained from publicly available datasets (UniProt; <https://www.uniprot.org/> last accessed Jan 19, 2022 accessed 10 December 2021) were aligned against the genome sequence using GenomeThreader²⁸ (version 1.7.1; arguments -startcodon-finalstopcodon -species rice -gcmcoverage 70 -prseedlength 7 -prhdist 4). The fl-cDNAs and isoseq were aligned to the genome assembly using GMAP²⁹ (version 2018-07-04). All RNA-Seq, fl-cDNA, and aligned protein sequences were combined using Cuffcompare³⁰ (version 2.2.1) and

subsequently merged with StringTie (version 2.1.5; parameters –merge -m150) into a pool of candidate transcripts. TransDecoder (version 5.5.0; <http://transdecoder.github.io> last accessed Jan 19, 2022) was used to find potential open reading frames (ORFs) and to predict protein sequences within the candidate transcript set. An *ab initio* annotation was performed using Augustus³¹ (version 3.3.3). GeneMark³² (version 4.35) was additionally used to further improve the structural gene annotation. To avoid potential over-prediction, guiding hints were generated using the above-described RNA-Seq, isoseq, protein, and fl-cDNA datasets and were then trained and optimized using a specific Augustus model for barley, as described by Hoff and Stanke.³¹ Structural gene annotations from different prediction methods were combined using EVIDENCEModeler³³ (version 1.1.1), and the weights were adjusted according to the input source: *ab initio* (Augustus: 5, GeneMark: 2) and homology based (10). Additionally, two rounds of PASA³⁴ (version 2.4.1) were run to identify untranslated regions and isoforms using the above-described fl-cDNA dataset.

BLASTP²⁴ (ncbi-blast-2.3.0+, parameters -max_target_seqs 1 -evalue 1e-05) was used to compare potential protein sequences with a trusted set of reference proteins (UniProt Magnoliophyta, reviewed/Swissprot; downloaded on 3 August 2016; <https://www.uniprot.org> last accessed Jan 19, 2022). This differentiated candidates into complete and valid genes, non-coding transcripts, pseudogenes, and transposable elements. In addition, the PTREP database (Release 19; <http://botserv2.uzh.ch/kellldata/trep-db/index.html> last accessed Jan 19, 2022) was used in the BLASTP analysis; this database of hypothetical proteins contains deduced amino acid sequences in which internal frameshifts have been removed in many cases. This step is particularly useful for the identification of divergent transposable elements with no significant similarity at the DNA level. The best hits were selected for each predicted protein to each of the three databases: UniProt, SwissProt, and PTREP. Only hits with an e-value below 10^{-10} were considered. Furthermore, the functional annotation of all predicted protein sequences was performed using the AHRD pipeline (<https://github.com/groupschoof/AHRD> last accessed Jan 19, 2022).

The proteins were further classified into two confidence classes: high and low. Hits with subject coverage (for protein references) or query coverage (transposon database) above 80% were considered significant. The proteins were classified as high confidence if the sequence was complete and had a subject and query coverage above the threshold in the UniMag database or no BLAST hit in UniMag or PTREP but present in UniPoa. A protein sequence was defined as being low confidence if it was incomplete and had a hit in the UniMag or UniPoa database but not in PTREP. Alternatively, complete protein sequences with no hit in UniMag, UniPoa, or PTREP were also classified as low confidence. In a second refinement step, low-confidence proteins with an AHRD-score of 3* were promoted to high-confidence.

2.5 Repeat and transcript annotation

The final assembly was analysed for repetitive regions using RepeatMasker³⁵ (version 4.0.9) with the TREP repeat library³⁶ (trep-db_complete_Rel-19; downloaded from <http://botserv2.uzh.ch/kellldata/trep-db/downloadFiles.html> last accessed Jan 19, 2022 on 13 September 2020). The repetitive regions were changed to lowercase (-xsmall parameter). The output of RepeatMask was condensed using the perl script ‘one-code-to-find-them-all’³⁷ with the parameters -strict and -unknown.

2.6 Data validation and quality control

Benchmarking Universal Single-Copy Orthologs³⁸ (BUSCO; version 3.0.2) was used with the plant dataset (embryophyta_odb10) to validate the assembly and gene models. For gene prediction, BUSCO uses Augustus^{39,40} (version 3.3). For the gene-finding parameters in Augustus, the species was set to wheat and BUSCO was run in genome mode (-m geno -sp wheat).

2.7 Alignment of published BAC sequences

Published ‘Haruna Nijo’ BAC clone sequences of kernel row type *Vrs1*,⁴¹ brittle rachis *Btr1* and *Btr2*,⁴² and quantitative locus seed dormancy 1 *Qsd1*¹⁵ were downloaded from NCBI. Each clone sequence was aligned to the pseudomolecule sequences of ‘Haruna Nijo’ and ‘Morex’V3 using minimap2.⁴³

2.8 Genome browser

Pseudomolecule assembly, gene models of the CDS, and amino acid models were visualized in Jbrowse genome browser (version 1.16.9). The BLAST (version 2.2.18) and BLAT (version 34) servers were also installed to search for target sequences in the pseudomolecules and gene models.

2.9 Data availability

Raw reads have been deposited in the ENA sequence read archive. Bioproject: PRJEB44504 (ERS_ID: paired-end reads: ERS6294308; mate-pair reads: ERS6294309; 10X reads: ERS6294307; Hi-C reads: ERS6294313; assembly: ERS6294316) (Supplementary Table S2).

The reference assembly is available for download or BLAST search from <http://viewer.shigen.info/harunanijo/index.php>. last accessed Jan 19, 2022

3. Results and discussion

3.1 Genome assembly

We generated the genome assembly from PE and MP short reads and 10X reads. Approximately 868 Gb of raw data was generated, providing an estimated 170× coverage of the genome (Supplementary Table S1). An assembly generated using the TRITEX pipeline¹⁷ resulted in a scaffold N50 value of 18.9 Mb (Table 1). We integrated Hi-C data into the assembly, which uses a genomic distance matrix inferred from native chromatin folding to increase the scaffold-level contiguity to full chromosome size (Supplementary Fig. S1). The final pseudomolecule size was 4.28 Gb, comprising 552 scaffolds and a

Table 1. Statistics of ‘Haruna Nijo’ and two versions of ‘Morex’ assemblies

Parameter	‘Haruna Nijo’	‘Morex’V2	‘Morex’V3
Number of scaffolds in pseudomolecules	552	273	103
Pseudomolecule size (Gb)	4.28	4.34	4.20
Scaffold N50 ^a [Mb]	18.9	43.7	118.9
Scaffold N90 [Mb]	2.6	5.9	21.8
Cumulative size of unanchored scaffold (Mb)	154.3	82.9	29.1

^a‘Scaffold’ refers to top-level entities that constitute the pseudomolecules. In ‘Morex’V3, these are Bionano scaffolds of PacBio HiFi contigs; in the other assemblies, superscaffolds were constructed from PE, MP, and 10X data.

cumulative size of unanchored scaffolds of 154.3 Mb. The pseudo-molecule size of ‘Haruna Nijo’ is comparable with that of the pan-genome assemblies of ‘Morex’V2 obtained using similar sequencing platforms but with a smaller scaffold N50 value. The datasets for ‘Morex’V3 showed improved statistics compared with our assemblies due to the use of accurate long-read sequencing by circular consensus sequencing on the PacBio platform in the generation of this assembly.¹¹ The alignment of the pseudomolecules of ‘Haruna Nijo’ to ‘Morex’V3 individual chromosomes revealed some small inversions (Fig. 1); however, the overall contiguity of entire chromosomes was retained between ‘Haruna Nijo’ and ‘Morex’V3.

For easy access, the reference sequence is available in BLAST-searchable form at <http://viewer.shigen.info/harunanijo/index.php> last accessed Jan 19, 2022.

3.2 Quality of assemblies

We used the spectra-cn function from the Kmer Analysis Toolkit (KAT)⁴⁴ to compare k -mer contents in the scaffolds and pseudomolecules. KAT generates a k -mer frequency distribution from the PE and MP reads and identifies how many times k -mers from each part of the distribution appear in the assemblies being compared.²¹ The spectra-cn plot in Supplementary Fig. S2 generated from the contigs shows sequencing errors (k -mer multiplicity <20) in black, as these are not included in the assembly. Most of the content appears in a single red peak, indicating sequences that appear once in the assembly. The black region under the main peak is small, indicating that most of this content from the reads is present in the assembly. The content that appears to the right of the main peak and is present two or three times in the assembly represents repeats. Pseudomolecules may contain more miss-assemblies than scaffolds; this is not obvious in the spectra-cn plot in Supplementary Fig. S2b.

We evaluated the quality of the ‘Haruna Nijo’ assembly using BUSCO.^{38,45} This program assesses the completeness of a genome by identifying conserved single-copy orthologous genes. The scaffold and pseudomolecule stages had complete single-copy genes at a rate

Table 2. BUSCO statistics of ‘Haruna Nijo’

Factor	Scaffolds	Pseudomolecule
Complete BUSCOs	1,403 (97.5%)	1,396 (96.9%)
Complete BUSCOs: single copy	1,382 (96.0%)	1,378 (95.7%)
Complete BUSCOs: duplicated	21 (1.3%)	18 (1.2%)
Fragmented BUSCOs	14 (1.0%)	14 (1.0%)
Missing BUSCOs	23 (1.5%)	30 (2.1%)
Total BUSCO groups searched	1,440	1,440

of 96.0% and 95.7%, respectively (Table 2). These values are very close to those recently published for the ‘Morex’V2 assembly, which had 97.2% single-copy genes.⁴⁶ The differences are mainly due to the greater number of duplicated genes in the scaffolds (1.3%) than the pseudomolecules (1.2%). Only 1.0% of the fragmented sequences were present in both the scaffolds and pseudomolecules.

3.3 Repeat masking

We analysed each chromosome of the ‘Haruna Nijo’ assembly for repetitive regions using RepeatMasker with the TREP repeat library. This analysis identified 72.8% (3.23 Gb) of the ‘Haruna Nijo’ assembly as transposable elements (Supplementary Table S3), almost all of which were retroelements. The same analysis was performed for ‘Morex’V2 and ‘Morex’V3, producing similar results (Supplementary Table S3). The differences from the published results for the ‘Morex’V2 and ‘Morex’V3 assemblies^{11,17} were due to the different repeat libraries used.

3.4 Gene projection

We assessed the gene content of ‘Haruna Nijo’ using a gene projection approach, as described by Jayakodi et al.¹⁰ for the 20 barley pan-genome assemblies. The total number of loci was 47,367, which is within the range of 42,464 to 47,588 reported for the 20 pan-

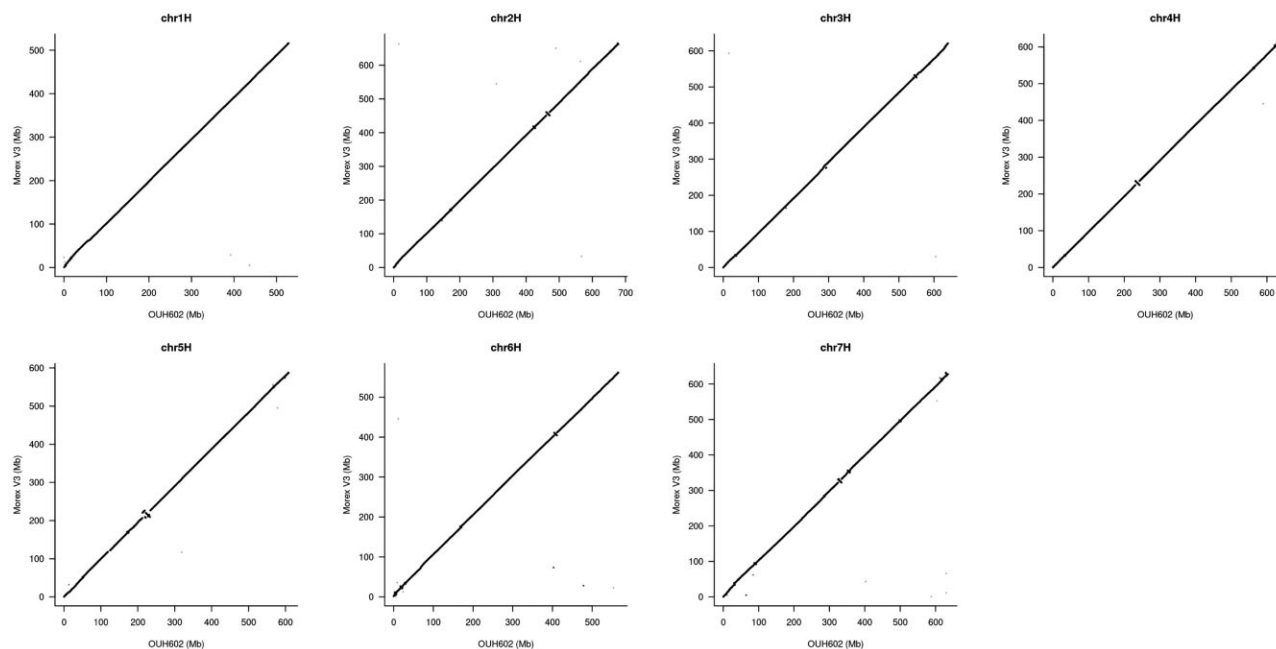
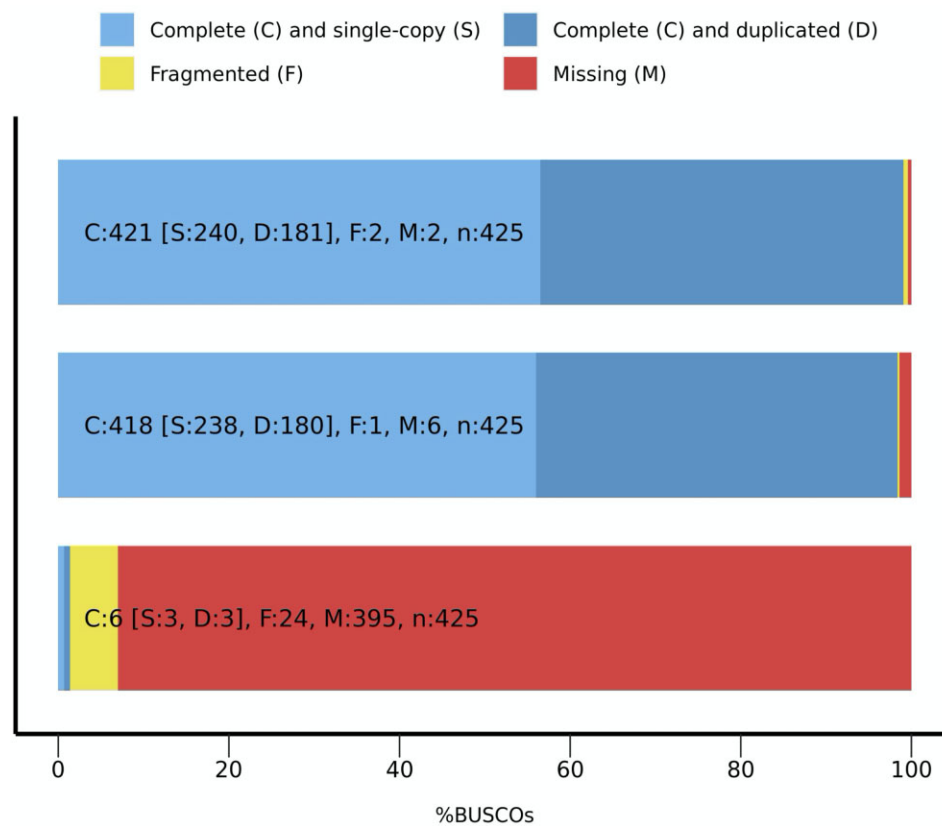


Figure 1. Alignment of pseudomolecules of ‘Haruna Nijo’ to ‘Morex’V3 individual chromosomes.

Table 3. *De novo* gene annotation statistics

Statistics	Complete sequences	High confidence	Low confidence
Number of genes	161,721	49,524	112,197
Number of monoexonic genes	67,724	12,645	55,079
Number of transcripts	181,980	68,751	113,229
Transcripts per gene	1.13	1.39	1.01
cDNA lengths (mRNAs)	1,294	1,696	1,050
CDS lengths (mRNAs)	1,154	1,377	1,018
Exons per transcript (mRNAs)	3.45	5.21	2.38
Exon lengths (mRNAs)	375	326	441
Intron lengths (mRNAs)	675	623	770
CDS exons per transcript (mRNAs)	3.33	4.95	2.35
CDS exon lengths	346	278	434
5' UTR exon number	54,193	48,584	5,609
3' UTR exon number	52,989	44,690	8,299

**Figure 2.** BUSCO assessment results of 'Haruna Nijo' fl-cDNA sequences (upper), high-confidence genes (middle), and low-confidence genes (lower).

genome assemblies. Of the 44,579 protein-coding genes, between 42,800 and 43,211 loci had a BLAST match with an e-value of $<1-30$, and 34,427 and 38,005 were one-to-one reciprocal BLAST orthologs between 'Haruna Nijo' and 'B1K-04-12' or 'Morex'V2, respectively. The overall and orthologous gene content of 'Haruna Nijo' is therefore highly conserved in comparison with other barley lines. Likewise, 15.9% (7,109) of the tandem-repeated genes in 'Haruna Nijo' had similar ranges as were detected for the 20 barley pan-genome assemblies and were located in 2,735 clusters. The gene content statistics above indicate that the 'Haruna Nijo' assembly

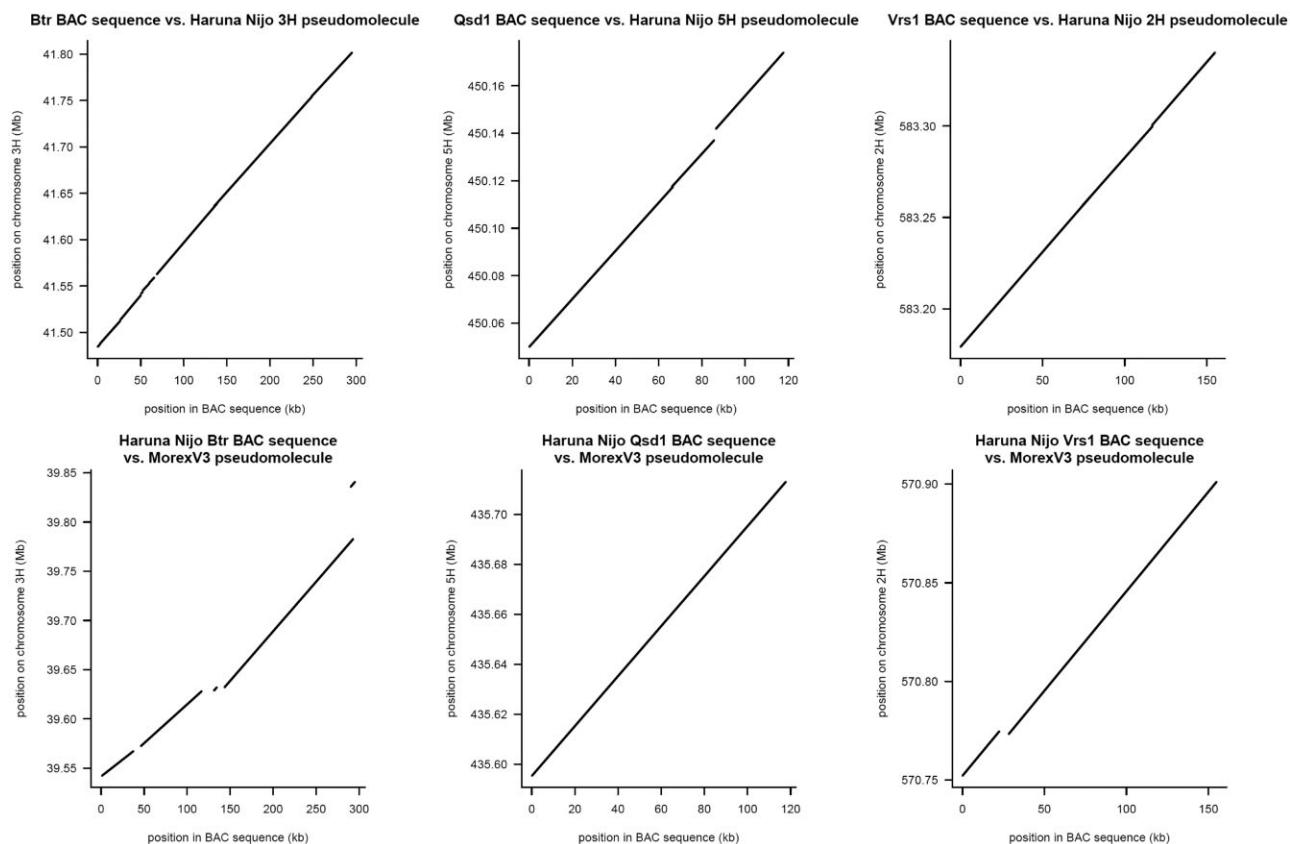
contains a gene set with highly similar characteristics to those reported for the 20 barley pan-genome assemblies.

3.5 *De novo* gene annotation using RNA-Seq and fl-cDNA sequences

A final structural gene annotation yielded 161,721 gene models, including 49,524 high- and 112,197 low-confidence gene models (Table 3). The high number of total gene models is likely due to the *ab initio* prediction step, which was run without the use of

Table 4. BLASTN hits ($<e-100$) among nucleotide sequences of fl-cDNA, gene projection, and *de novo* gene model sequences

Target	Query		
	Full-length cDNA	Gene projection	<i>De novo</i> annotation
Full-length cDNA	22,651	25,977	28,415
Gene projection	19,711	47,367	43,087
<i>De novo</i> annotation	19,636	42,336	49,524
Total hits	19,937	42,753	44,387
Ratio (total hits/number of queries)	0.880	0.903	0.896

**Figure 3.** Alignment of 'Haruna Nijo' BAC sequences of *Btr*, *Qsd1*, and *Vrs1* regions to pseudomolecules of 'Haruna Nijo' and 'MorexV3'.

transposable elements hints; the high number of low-confidence gene models supports this rationale. The BUSCO score of the high-confidence genes was 98.4 (Fig. 2). The average number of transcripts per gene was 1.39 for the high-confidence gene models, which was much higher than 1.01 for the low-confidence gene models.

We next compared our sequences with the fl-cDNA dataset, which consisted of 22,651 sequences generated from 'Haruna Nijo'.^{6,7} These sequences were created from plants grown in 12 different conditions and thus represent a good snapshot of the barley transcriptome. The average insert size of these fl-cDNA sequences was 1,711 bp, which was close to the cDNA length of the high-confidence gene models. Sequence similarities among our data and the fl-cDNA sequences, gene models of gene projection, and *de novo* gene annotations were compared using a BLASTN analysis with a threshold *e*-value of <-100 (Table 3). The 22,651 fl-cDNA query sequences showed high similarity with the sequences from the gene projection (19,771) and *de novo* annotation (19,636) (Table 4).

These numbers are consistent with the number of fl-cDNA sequences with complete ORFs (19,335) reported by Matsumoto et al.⁷; other fl-cDNA sequences had truncated ORF or non-protein-coding sequences. The results also indicated that almost 10% of each gene model did not overlap each other. The amino acid sequences showed a lower level of overlapping than the nucleotide sequences (0.707–0.731; Supplementary Table S4).

3.6 Alignment with BAC clone sequences

We aligned 'Haruna Nijo' BAC clone sequences to pseudomolecules of 'Haruna Nijo' to estimate the contiguity of both sequences (Fig. 3). The BAC clones were analysed using shotgun Sanger sequencing and assembled on an individual clone basis. The BAC clone sequences of *Btr1/Btr2* were composed of several clones and showed apparent discontinuity with the pseudomolecule sequence of 'Haruna Nijo'. The alignment of these BAC sequences with the 'MorexV3' pseudomolecule sequence revealed fragmentation at the

3' region, but the 5' region showed higher contiguity. Another BAC clone sequence, *Qsd1*, which was derived from a single clone, showed more contiguity with the pseudomolecules of 'Haruna Nijo'; however, there was a significant gap between the BAC sequence and the pseudomolecule sequence of 'Morex'V3. The quality of the BAC sequences was comparable with that of 'Morex'V3, but with some structural disorders.

The observed mismatches between BAC clones and pseudomolecules indicate that the pseudomolecules of 'Haruna Nijo' do not have as high of a sequencing quality as those of 'Morex'V3; however, they are useful for examining contiguity in the genome for gene identification.

3.7 Genome browser

The high-performance and user-friendly graphical interface genome browser Jbrowse was used to visualize the pseudomolecule sequence and the gene models. Tracks of *de novo* annotations and gene projections each display the result of the associated annotation (e.g. exon structure, protein names, and transposable elements) to allow a comparison of each gene model. The fl-cDNA sequence track based on the BLAST search result against the pseudomolecule sequence was also provided, showing strict similarity to clones only. In addition to the browser, the user interface of the sequence similarity search programs BLAST and BLAT was also provided. The BLAST search results are directly linked to Jbrowse as a user track, which allows the mapping of query sequences against the reference genome and their comparison with the gene models. The assembled sequence and annotation files can be downloaded from the website (<http://viewer.shigen.info/harunanijo/index.php> last accessed Jan 19, 2022) so that our data can be used in the local user's environment.

3.8 Conclusion

Here, we present an assembly of 'Haruna Nijo' that is of similar quality to the 'Morex'V2 reference.¹⁷ Importantly, it is a European-style Japanese two-row cultivar, expanding barley genomic resources to Japanese and European breeding materials in contrast to the American six-row cultivar 'Morex'. Interesting phenotypes have since been identified in Haruna Nijo; its genome sequence assembly will facilitate the identification of the underlying genes.

Acknowledgements

We thank the National Bioresource Project, Japan, for providing seed samples and BAC clones of 'Haruna Nijo'.

Funding

This work was supported by the JST Mirai Program (grant number 18076896 Japan) to KS, the German Ministry of Education and Research project de. NBI (grant no. 031A536B to MS and GH) and project SHAPE I (grant no. 031B0190A to MM and NS). We thank Anne Fiebig for the data submission.

Accession numbers

Raw reads have been deposited in the ENA sequence read archive. Bioproject: PRJEB44504 [ERS_ID: paired-end reads: ERS6294308; mate-pair reads: ERS6294309; 10X reads: ERS6294307; Hi-C reads: ERS6294313; assembly: ERS6294316].

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at DNARES online.

References

- Seko, H. 1987, History of barley breeding in Japan. In: *Barley Genetics V: Proceedings of the 5th Barley Genetics Symposium*, Okayama, Maruzen Co. Ltd. Okayama Branch, Japan, pp. 915–922.
- Sato, K. 2020, History and future perspectives of barley genomics, *DNA Res.*, **27**, dsaa023.
- Sato, K., Nankaku, N. and Takeda, K. 2009, A high density transcript linkage map of barley derived from a single population, *Heredity (Edinb)*, **103**, 110–7.
- Close, T.J., Bhat, P.R., Lonardi, S., et al. 2009, Development and implementation of high-throughput SNP genotyping in barley, *BMC Genomics*, **10**, 582–94.
- Sato, K. and Takeda, K. 2009, An application of high-throughput SNP genotyping for barley genome mapping and characterization of recombinant chromosome substitution lines, *Theor. Appl. Genet.*, **119**, 613–9.
- Sato, K., Shin-I, T., Seki, M., et al. 2009, Development of 5006 fl-cDNAs in barley: a tool for accessing cereal genomics resources, *DNA Res.*, **16**, 81–9.
- Matsumoto, T., Tanaka, T., Sakai, H., et al. 2011, Comprehensive sequence analysis of 24,783 barley fl-cDNAs derived from 12 clone libraries, *Plant Physiol.*, **156**, 20–8.
- International Barley Genome Sequencing Consortium. 2012, A physical, genetic and functional sequence assembly of the barley genome, *Nature*, **491**, 711–6.
- Mascher, M., Gundlach, H., Himmelbach, A., et al. 2017, A chromosome conformation capture ordered sequence of the barley genome, *Nature*, **544**, 427–33.
- Jayakodi, M., Padmarasu, S., Haberer, G., et al. 2020, The barley pan-genome reveals the hidden legacy of mutation breeding, *Nature*, **588**, 284–9.
- Mascher, M., Wicker, T., Jenkins, J., et al. 2021, Long-read sequence assembly: a technical evaluation in barley, *Plant Cell*, **33**, 1888–906.
- Sato, K., Shin-I, T., Seki, M., et al. 2009, Improvement of barley genome annotations by deciphering the Haruna Nijo genome, *DNA Res.*, **16**, 81–9.
- Saisho, D., Myoraku, E., Kawasaki, S., Sato, K. and Takeda, K. 2007, Construction and characterization of a bacterial artificial chromosome (BAC) library for Japanese malting barley 'Haruna Nijo', *Breed. Sci.*, **57**, 29–38.
- Taketa, S., Amano, S., Tsujino, Y., et al. 2008, Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway, *Proc. Natl. Acad. Sci. U S A.*, **105**, 4062–7.
- Sato, K., Yamane, M., Yamaji, N., et al. 2016, Alanine aminotransferase controls seed dormancy in barley, *Nat. Commun.*, **7**, 11625.
- Hisano, H., Tsujimura, M., Yoshida, H., Terachi, T. and Sato, K. 2016, Mitochondrial genome sequences from wild and cultivated barley (*Hordeum vulgare*), *BMC Genomics*, **17**, 824.
- Monat, C., Padmarasu, S., Lux, T., et al. 2019, TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools, *Genome Biol.*, **20**, 284.
- Clavijo, B., Garcia Accinelli, G., Wright, J., et al. 2017, W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data, *bioRxiv*, doi: 10.1101/110999.
- Jayakodi, M., Schreiber, M., Stein, N. and Mascher, M. 2021, Building pan-genome infrastructures for crop plants and their use in association genetics, *DNA Res.*, **28**, dsaa030.
- Milner, S.G., Jost, M., Taketa, S., et al. 2019, Genebank genomics highlights the diversity of a global barley collection, *Nat. Genet.*, **51**, 319–26.

21. Sato, K., Mascher, M., Himmelbach, A., Haberer, G., Spannagl, M. and Stein, N. 2021, Chromosome-scale assembly of wild barley accession "OUH602", *G3 (Bethesda)*, **11**, jkab244.
22. Dvorak, J., McGuire, P.E. and Cassidy, B. 1988, Apparent sources of the A genomes of wheats inferred from polymorphism in abundance and restriction fragment length of repeated nucleotide sequences, *Genome*, **30**, 680–9.
23. Padmarasu, S., Himmelbach, A., Mascher, M. and Stein, N. 2019, In Situ Hi-C for plants: an improved method to detect long-range chromatin interactions. *Methods Mol. Biol.*, **1933**, 441–72.
24. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
25. Slater, G.S.C. and Birney, E. 2005, Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics*, **6**, 31.
26. Dobin, A., Davis, C.A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, **29**, 15–21.
27. Kovaka, S., Zimin, A.V., Perlea, G.M., et al. 2019, Transcriptome assembly from long-read RNA-seq alignments with StringTie2, *Genome Biol.*, **20**, 278.
28. Gremme, G., Brendel, V., Sparks, M.E. and Kurtz, S. 2005, Engineering a software tool for gene structure prediction in higher organisms, *Inf. Soft. Technol.*, **47**, 965–78.
29. Wu, T.D. and Watanabe, C.K. 2005, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, **21**, 1859–75.
30. Ghosh, S. and Chan, C.K. 2016, Analysis of RNA-seq data using TopHat and Cufflinks, *Methods Mol. Biol.*, **1374**, 339–61.
31. Hoff, K.J. and Stanke, M. 2019, Predicting genes in single genomes with AUGUSTUS, *Curr. Protoc. Bioinformatics*, **65**, e57.
32. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. and Borodovsky, M. 2008, Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training, *Genome Res.*, **18**, 1979–90.
33. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments, *Genome Biol.*, **9**, R7.
34. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.*, **31**, 5654–66.
35. Smit, A., Hubley, R. and Green, P. 2013–2015, RepeatMasker Open-4.0. <http://www.repeatmasker.org/faq.html>, last accessed Jan 19, 2022..
36. Wicker, T., Matthews, D.E. and Keller, B. 2002, TREP: a database for Triticeae repetitive elements, *Trends Plant Sci.*, **7**, 561–2.
37. Bailly-Bechet, M., Haudry, A. and Lerat, E. 2014, "One code to find them all": a perl tool to conveniently parse RepeatMasker output files, *Mob. DNA*, **5**, 13.
38. Waterhouse, R.M., Seppey, M., Simão, F.A., et al. 2018, BUSCO applications from quality assessments to gene prediction and phylogenomics, *Mol. Biol. Evol.*, **35**, 543–8.
39. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004, AUGUSTUS: a web server for gene finding in eukaryotes, *Nucleic Acids Res.*, **32**, W309–W312.
40. König, S., Romoth, L.W., Gerischer, L. and Stanke, M. 2016, Simultaneous gene finding in multiple genomes, *Bioinformatics*, **32**, 3388–95.
41. Komatsuda, T., Pourkheirandish, M., He, C., et al. 2007, Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene, *Proc. Natl. Acad. Sci. U S A.*, **104**, 1424–9.
42. Pourkheirandish, M., Hensel, G., Kilian, B., et al. 2015, Evolution of the seed dispersal system in barley, *Cell*, **162**, 527–39.
43. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics.*, **34**, 3094–100.
44. Mapleson, D., Venturini, L., Kaithakottil, G. and Swarbreck, D. 2017, Efficient and accurate detection of splice junctions from RNAseq with Portcullis. *bioRxiv* (Preprint posted November 10, 2017).
45. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
46. Schreiber, M., Mascher, M. and Wright, J. 2020, A genome assembly of the barley 'transformation reference' cultivar golden promise, *G3 (Bethesda)*, **10**, 1823–7.