November 2015. Horticultural Plant Journal, 1 (3): 155–164.

Horticultural Plant Journal

Available online at www.sciencedirect.com
The journal's homepage: http://www.journals.elsevier.com/horticultural-plant-journal

A *de novo* Genome of a Chinese Radish Cultivar

ZHANG Xiaohui[a,1], YUE Zhen[b,1], MEI Shiyong[c,1], QIU Yang[a], YANG Xinhua[b], CHEN Xiaohua[a], CHENG Feng[a], WU Zhangyan[b], SUN Yuyan[a], JING Yi[b], LIU Bo[a], SHEN Di[a], WANG Haiping[a], CUI Na[a], DUAN Yundan[a], WU Jian[a], WANG Jinglei[a], GAN Caixia[c], WANG Jun[b], WANG Xiaowu[a,*], LI Xixiang[a,*]

[a] *Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of Ministry of Agriculture, Beijing 100081, China.*

[b] *The Beijing Genomics Institute, Shenzhen, Guangdong 518083, China.*

[c] *Institute of Economic Crop, Hubei Academy of Agricultural Sciences, Wuhan 430064, China.*

Received 20 September; Received in revised form 20 October 2015; Accepted 1 November 2015

Abstract

Here, we report a high-quality draft genome of a Chinese radish (*Raphanus sativus*) cultivar. This draft contains 387.73 Mb of assembled scaffolds, 83.93% of the scaffolds were anchored onto nine pseudochromosomes and 95.09% of 43 240 protein-coding genes were functionally annotated. 184.75 Mb (47.65%) of repeat sequences was identified in the assembled genome. By comparative analyses of the radish genome against 10 other plant genomes, 2 275 genes in 780 gene families were found unique to *R. sativus*. This genome is a good reference for genomic study and of great value for genetic improvement of radish.

Keywords: *Raphanus sativus*; genome; sequencing

1. Introduction

The radish (*Raphanus sativus*, 2n = 2x = 18) is the most widely cultivated vegetable in China and also an important cash crop worldwide. The radish cultivation history can be dated back to approximately 4 500 years ago in ancient Egypt (Banga, 1976), more than 2 000 years ago in China (Li, 1989) and Europe, and approximately 1 000 years ago in Japan (Crisp, 1995). The long-term artificial selection led to numerous *R. sativus* landraces with different sizes, shapes, colors, and flavors of the edible organs. Beyond these morphological diversity, *R. sativus* also displayed rich multiformity in nuclear genome (Huh and Ohnishi, 2002; Lü et al., 2008; Nakatsuji et al., 2011), mitochondrial (Yamagishi and Terachi, 2003) and chloroplast genomes (Yamane et al., 2005, 2009; Yamagishi et al., 2009). Based on the polymorphism of the mitochondrial and chloroplast genomes, *R. sativus* cultivars were proposed having been domesticated via multi-regional origination (Yamagishi and Terachi 2003; Yamagishi et al., 2009). China is the most important origin center of radish, and has developed hundreds of distinctive landraces.

The draft genomes of a *R. raphanistrum* (wild species) and two assemblies of *R. sativus* 'Aokubi', have been released recently (Kitashiba et al., 2014; Moghe et al., 2014; Mitsui et al., 2015), However, the assembly of *R. raphanistrum* covered only 254.0 Mb (49.3% of the estimated genome) and provided

[1] These authors contributed equally to this work.
* Corresponding author.
E-mail address: lixixiang@caas.cn; wangxiaowu@caas.cn

http://dx.doi.org/10.16420/j.issn.2095-9885.2016-0028

in contigs, which was not assigned to the chromosomes (Moghe et al., 2014). For the two assemblies of 'Aokubi', only 116.0 Mb (21.8% of the estimated genome) (Kitashiba et al., 2014) and 179.8 Mb (31.32% of the estimated genome) (Mitsui et al., 2015) have been assigned to the pseudo-chromosomes, respectively. These three draft genomes are far from meeting the needs as the references in whole-genome screening for the selection signatures of agronomic traits in *R. sativus*.

To construct a fine reference genome for Chinese radish landraces, we generated a high-quality genome of *R. sativus*. This genome assembly will not only benefit to the understanding of how the diverse traits evolved during the domestication but also contribute to the genetic improvement of radish.

2. Materials and methods

2.1. Genome sequencing

Qualified DNA isolated from leaves of tissue-cultured plantlets of *Raphanus sativus* 'Xiangyabai' inbred line (XYB36-2) was used for *de novo* genome sequencing. Paired-end DNA libraries (200, 250, 500 and 800 bp) and mate-paired DNA libraries (2, 5, 10, 20 and 40 kb) were constructed using Illumina sequencing platform standard protocols. Constructed libraries were sequenced on an Illumina HiSeq2000 system following the manufacture's user guide (Li et al., 2010a).

2.2. Genome assembly

SOAPdenovo (version 1.05; http://soap.genomics.org.cn) (Li et al., 2010b) and SSPACE (Boetzer et al., 2011) were used to assemble the genome with the filtered data. For the assembly process, all possible sequences from Illumina reads were assembled using a *de Bruijn* graph methodology, with a k-mer of 127 used as a node and the k−1 bases overlapping between two k-mers defined as an edge. To reduce sequencing errors and limit branches, the ends were trimmed and k-mers with low coverage were removed. The graph was then converted into a contig graph by transformation of the linearly connected k-mers to pre-contig nodes. Dijkstra's algorithm was used to detect bubbles, which were then merged into a single pathway if the branch sequences were similar. Using this method, regions with repeat sequences were merged into consensus sequences.

The assembled contigs were linked to a scaffolding graph based on paired-end reads. Connections between contigs were defined as edges in this graph, and branch lengths were defined as the gap size calculated from the insert size of the paired-end reads. Sub-graph linearization was then applied to convert interleaving contigs into a linear structure. Paired-end reads were applied step by step, with increasing insert sizes of 200, 250, 500 and 800 bp and then 2, 5, 10 and 20 kb used. To fill gaps in the scaffolds, we aligned the paired-end reads and collected those with one end mapped to a contig and the other end falling in a gap, and then performed a local assembly with the retrieved reads. Next, pair-end reads of 40 kb were used to construct super-scaffolds using SSPACE.

2.3. Assessments of accuracy and completeness of the assembly

The quality of the draft genome was comprehensively evaluated by assessing sequencing depth and coverage using fosmid, EST and RNA-Seq sequences. We constructed a Fosmid library harboring 35–45 kb radish (XYB36-2) DNA fragments using MaxPlax lambda packaging extracts (Epicentre). Five fosmid clones were sequenced using an ABI 3730 sequencer. This five fosmid clone sequences were used as reference data to check the coverage rate by map the assembled genome sequence back to them (BLASTn; *E*-value threshold of 1×10^{-5}). The coverage of gene space was further estimated using RNA-seq data (156 501 unigenes from the seven tissues of XYB36-2) and ESTs (26 606 ESTs of *R. sativus* GSK3-1 downloaded from NCBI). The unigenes and ESTs were aligned to the assembled scaffolds using BLAT with an identity cutoff of 90%.

2.4. Anchoring of the assembled scaffolds to pseudochromosomes

To anchor scaffolds onto pseudochromosomes, we constructed a high-density SNP bin-marker-based genetic map using RAD sequencing data from 120 F_2 progenies derived from a cross of *R. sativus* lines KB10Q-22 and XYB36-2. The ordering and orientation of the scaffolds on the pseudo-chromosomes was manually corrected using the linkage map and ortholog blocks between *R. sativus* and *Arabidopsis thaliana*. A few scaffolds exhibiting conflicts, such as those mapping to different linkage groups or containing markers with chimeric genetic distances, were additionally analyzed and resolved on the basis of their paired-end relationships. Gaps between adjacent scaffolds were filled with 5 kb of Ns.

2.5. Gene prediction

De novo gene prediction based on the repeat-masked genome was first performed with Augustus (version 2.4) (Stanke et al., 2004) and GlimmerHMM (version 3.02) (Majoros et al., 2004) under an HMM model. Protein sequences from *Vitis vinifera*, *Carica papaya*, *A. thaliana*, *Schrenkiella parvula* (formerly *Thellungiella parvula*) and *Brassica rapa* were applied in the homolog evidence-based gene annotation using BLASTTN with an *E*-value cutoff of 1×10^{-5}. The *R. sativus* genome sequences were aligned against the homologous protein sequences using GeneWise (Birney and Durbin, 2000) for accurate spliced alignments. *Raphanus* ESTs were aligned