Advances of Whole Genome Sequencing in Strawberry with NGS Technologies

Sachiko Isobe*, Kenta Shirasawa and Hideki Hirakawa

Kazusa DNA Research Institute, Kisarazu 292-0818, Japan

Next generation sequencing (NGS) is one of the most impactful technologies to appear in the 21st century, and has already brought important changes to agriculture, especially in the field of breeding. Construction of a reference genome is key to the advancement of genomic studies, and therefore, de novo whole genome assembly has been performed in various plants, including strawberry. Strawberry (Fragaria × ananassa) is an allo-octoploid species (2n = 8x = 56), which has four discriminable subgenomes. Because of its complex genome structure, de novo whole genome assembly in strawberry has been considered a difficult challenge. However, recent advances of NGS technologies have allowed the construction of chromosome-scale de novo whole genome assembly. In this manuscript, we review the recent advances in de novo whole genome sequencing in strawberry and other Fragaria species. The genome structure and domestication history in strawberry is one of the largest questions in genetic and genomic studies in strawberry. Therefore, the domestication history in strawberry is also be reviewed based on comparisons of genes and genome sequences across Fragaria species.

Key Words: Fragaria × ananassa, genus Fragaria, Next generation sequencing.

Introduction

Next generation sequencing (NGS) is one of the most impactful technologies to appear in the 21st century. At present, NGS technology has contributed more to medicine than to agriculture, but important changes to agriculture are clearly commencing, especially in the field of breeding. The term “next generation” is an umbrella term referring to various sequencing technologies related to the Sanger sequencing method (Yang et al., 2014). Briefly, there are two generations of these technologies, known as the 2nd and 3rd generations. The 2nd generation NGS technologies includes platforms such as Illumina (https://www.illumina.com/), Ion Torrent (https://www.thermofisher.com/) and MGI (https://en.mgitech.cn/), and generate large number of sequence reads with short length. Currently, several sequencers appeared in the beginning of 2nd generation were went out of production, such as Roche 454 (https://www.roche.com/) and Life Technologies SOLiD (https://www.thermofisher.com/). The 3rd generation NGS technologies, which include PacBio (https://www.pacbio.com/) and Oxford Nanopore Technologies (ONT, https://nanoporetech.com/), generate long sequence reads with a medium number of reads (Cao et al., 2017). Because the two platforms sequencing a single molecule DNA, the 3rd generation NGS is also called as single molecule sequencing. The quality of de novo whole genome assembly has been improved along with the advances of NGS technologies from the 2nd to 3rd generation, though the 2nd generation technologies are still the main platforms for re-sequencing and gene expression analysis. In addition, chromosome-scale assembly has become less difficult through the assistance of other NGS technologies—namely, optical mapping with the Bionano platform (https://bionanogenomics.com/) and libraries constructed with long DNA fragments such as 10X Genomics (https://www.10xgenomics.com/) and Hi-C (Lieberman-Aiden et al., 2009).

Construction of a reference genome is key to the advancement of genomic studies, and therefore, since the appearance of NGS technologies, de novo whole genome assembly has been performed in various plant species. According to Chen et al. (2019), a total of 181 crop horticultural species had been sequenced by December 31, 2018. Nevertheless, de novo whole genome assembly in polyploid species is still challenging, because homoeologous chromosomes often cause chimeric assembly. The first polyploid genome assembly was amphidiploid paleopolyploid in soybean (Schmutz et al., 2010), after which the assembly of polyploid genomes expanded to other species, including strawberry. Kyriakidou et al. (2018) listed 47 sequenced...
polyploid genomes reported by May 2018. In polyploid species, genome sequencing in ancestor diploid species was often performed prior to the targeted polyploid genome sequencing to construct a reference, such as for banana (D’Hont et al., 2012), cotton (Wang et al., 2012), wheat (Ling et al., 2013), and peanut (Bertioli et al., 2016).

Cultivated strawberry (Fragaria × ananassa) is an octoploid species (2n = 8X = 56), originated in the 18th century by artificial crosses between two octoploid species, F. chiloensis and F. virginiana (Darrow, 1966). The genome structure of strawberry was initially considered to be allo-autopolyploidy with AABBBBCC (Federova, 1946) or AA AAA’A’BBBB models (Senanayake and Bringhurst, 1967) based on cytological studies. An allopolyploidy model (AAA’A’BBB’B’) was then proposed based on isozyme segregation (Bringhurst, 1990) and later supported by the segregation patterns of DNA markers (Isobe et al., 2013; Kunihisa, 2011). Genus Fragaria contains 25 species comprising 13 diploids, 5 tetraploids, 1 hexaploid, 5 octoploids, and 1 decaploid (Hummer et al., 2009; Staudt, 2009). Although discussions regarding the ancestors in strawberry are ongoing, F. vesca and F. inunae are currently considered the most likely diploid progenitors (Rousseau-Gueutin et al., 2009). Therefore, the F. vesca genome was first sequenced as a reference in genetic and genomic analysis (Shulaev et al., 2011). In addition, recent advances in NGS technologies have led to the successful construction of an octoploid strawberry genome. In this manuscript, we review the recent advances in de novo whole genome sequencing in strawberry and other Fragaria species.

**De novo whole genome assembly in F. vesca**

*De novo* whole genome sequencing in *F. vesca* was reported by Shulaev et al. (2011) as a first reference genome in genus *Fragaria*. The genomes of a fourth-generation inbred line, ‘Hawaii 4’, were sequenced by Roche 454, Illumina Solexa and Life Technologies SOLiD with 39x coverage, resulting in the generation of 3,263 scaffolds (fvesca v1.0 scaffolds) (Table 1). The total and N50 lengths of the scaffolds were 201.9 Mb and 1.36 Mb, respectively. The assembled scaffolds were aligned on a diploid *Fragaria* reference map, and 204 scaffolds were successfully anchored with a total length of 198.1 Mb as pseudomolecules (fvesca v1.0 pseudo). Meanwhile, gene prediction was performed on the 3,263 scaffolds, and a total of 34,809 gene models were generated with an average length of 1,160 bp. A chloroplast genome was also constructed and 78 protein, 30 tRNA and 4 rRNA genes on the genome were predicted. The genome sequences and predicted genes are available in the Genome Database for Rosaceae (GDR, https://www.rosaceae.org/species/fragaria/all) (Jung et al., 2019).

The predicted genes were then re-annotated with transcript sequences derived from 25 different tissues with two biological replicates in the *F. vesca* cultivar ‘YW5AF7’ (Darwish et al., 2015). As a result, a total of 32,831 gene models were predicted with a total coding length of 38.8 Mb. The set of revised sequences was named version (v) 1.1. Genome assembly was also revised based on *F. vesca* linkage maps (Tennesen et al., 2013, 2014) constructed based on targeted sequence capture, which enabled mapping of 200 bp sequences including polymorphic sites. The v1.1 scaffold sequences were anchored via captured sequences located on the linkage maps, and conflict v1.1 scaffolds were broken and re-ordered according to the linkage maps. The revised version of *F. vesca* genome was designated v2.0.a1 (Tennesen et al., 2014).

The latest version of *F. vesca* genome at present (December 2019) is v4.0.a1 (Edger et al., 2018). Scaffolds were re-constructed with genomes of ‘Hawaii 4’ sequenced by PacBio. Chromosome-scale assembly was performed by hybrid assembly with Bionano optical mapping. The number and total length of the assembled scaffolds are 31 and 220.8 Mb, respectively, with an N50 length of 36.1 Mb. Of the 31 scaffolds, 9 cover 99.8% of the assembly, and these 9 scaffolds are anchored to 7 pseudomolecules by Bionano optical maps. A total of 24.96 Mb sequences were newly generated on the v4.0.a1 genome. Improved gene prediction and annotation analysis led to identification of an additional 1,496 genes. The pseudomolecule sequences are available on the GDR database.

**De novo whole genome assembly in cultivated strawberry (F. × ananassa)**

*De novo* whole genome assembly was first reported by Hirakawa et al. (2014). Genome sequences of the Japanese strawberry variety ‘Reikou’ were obtained by Illumina GAIIx, HiSeq, and Roche 454. This was a pioneering work in polyploid genome sequencing, and algorithms for the discrimination of multiple heterozygous genomes were not well established when the work started. Therefore, two types of assembly, namely FAN_r1.1 and FANhybrid_r1.2, were constructed (Table 1). FAN_r1.1 is an assembly sequence generated from Illumina reads that represents the entire genome sequence in ‘Reikou’. FAN_r1.1 consists of 625,966 sequences, comprising a total of 697.8 Mb with an N50 length of 2,201 bp. The genome size of ‘Reikou’ was estimated as 692 Mb based on kmer frequency analysis with Illumina reads in the Hirakawa et al. (2014) manuscript. However, our later kmer frequency analysis with additional Illumina reads suggested that the genome size of ‘Reikou’ was 811 Mb (not published). The total length of FAN_r1.1 is almost the same as the estimated genome size based on kmer frequency analysis (811 Mb), suggesting that the assembled genome covers the whole genome of ‘Reikou’. FANhybrid_r1.2 was generated as a virtual reference
Table 1. Statistics of assembled *Fragaria* genomes.

<table>
<thead>
<tr>
<th>Species</th>
<th>F. vesca</th>
<th>F. × ananassa</th>
<th>F. iinumae</th>
<th>F. orientalis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence name</strong></td>
<td>Fvesca_v1.0_scaffolds</td>
<td>FAN_r1.1</td>
<td>FII_r1.1</td>
<td>Hhirakawa et al. (2014)</td>
</tr>
<tr>
<td><strong>Assembled line</strong></td>
<td>Hawaii 4</td>
<td>Reikou</td>
<td>*</td>
<td>—</td>
</tr>
<tr>
<td><strong>Estimated genome size (Mb)</strong></td>
<td>~240</td>
<td>~811</td>
<td>221</td>
<td>323.163</td>
</tr>
<tr>
<td><strong>Assembly level</strong></td>
<td>Scaffolds</td>
<td>Scaffolds</td>
<td>Scaffolds</td>
<td>Scaffolds</td>
</tr>
<tr>
<td><strong>Number of sequences</strong></td>
<td>3,263</td>
<td>811</td>
<td>217,822</td>
<td>3,309</td>
</tr>
<tr>
<td><strong>Total length (Mb)</strong></td>
<td>201.9</td>
<td>173.2</td>
<td>119.6</td>
<td>306.4</td>
</tr>
<tr>
<td><strong>N50 length (bp)</strong></td>
<td>173.2</td>
<td>1,733.50</td>
<td>206.4</td>
<td>1,275</td>
</tr>
<tr>
<td><strong>Sequencing platforms</strong></td>
<td>Roche 454, Illumina Solexa, Life Technologies SOLiD</td>
<td>Fvesca_v1.0_scaffolds with a linkage map</td>
<td>Illumina HiSeq, Hi-C</td>
<td>Illumina HiSeq, 10X Genomics, PacBio, Hi-C</td>
</tr>
<tr>
<td><strong>Reference</strong></td>
<td>Shulaev et al. (2011)</td>
<td>—</td>
<td>—</td>
<td>Edger et al. (2019a)</td>
</tr>
</tbody>
</table>

* Germplasm provided by Prof. Yanagi Tomohiro of Kagawa University, Japan.

The genome (haploid genome) that integrates the genome sequences of homoeologous chromosomes by eliminating heterozygous bases. Both Illumina and Roche 454 sequences were used for FANhybrid_r1.2. The number of sequences in FANhybrid_r1.2 is 211,588, with a total and an N50 length of 173.2 Mb and 5,137 bp, respectively. Although the genome sequences created in this study were fragmented, the first assembled strawberry genome sequences contributed to subsequent molecular genetics and gene expression analyses (Landi et al., 2017; Wada et al., 2017). The two assembled genome sequences are available on Strawberry GARDEN (http://strawberry-garden.kazusa.or.jp/, Fig. 1) and GDR. Gene sequences and marker information are also included in the Strawberry GARDEN. The available functions in Plant GARDEN are keyword search, Blast search and extraction of sequences. Genome and gene sequences are displayed on JBrowse.
of all reads against the genomes of strawberry (FAN_r1.1) and other organisms, including humans, bacteria, and nematodes, in order to estimate the ratio of contamination in each library. As a result, 144 of the 288 libraries were confirmed to include more than 50% of the strawberry reads in the libraries. According to the BLAST search of all reads against the *F. vesca* genome (v2.0a1), the 144 libraries were classified into the 7 pseudomolecules on the *F. vesca* genome. Although there are still a few issues to be solved, such as sequence bias caused by whole genome amplification by PCR, the approach is expected to contribute to elucidation of the structure of the complex strawberry genome.

Chromosome-scale strawberry genome assembly was then reported by Edgar et al. (2019b) based on scaffolds constructed by DenovoMAGIC 3 (NRGene, https://www.nrgene.com/solutions/denovomagic/). DenovoMAGIC is a genome sequence assembly method only available as a commercial service provided by NRGene. The assembly is done based on Illumina reads with paired end (PE), mate pair (MP), and 10X Genomics libraries, and just recently PacBio reads have also been included. DenovoMAGIC has an excellent function that discriminates multiple heterozygous genomes in polyploid species, and hence it has been used for polyploid genome assembly in crops such as wheat (Avni et al., 2017). The genome of a strawberry variety, ‘Camarosa’, was sequenced by Illumina, 10X Genomics and PacBio platforms with a total of 65x coverage for DenovoMAGIC 3 assembly, and Hi-C clustering was performed to generate chromosome-scale scaffolds. The assembled genome sequences (Camarosa_Genome_v1.0.a1) consisted of 28 pseudomolecules with a total and N50 lengths of 805.5 Mb and 27.8 Mb, respectively. The genome size of ‘Camarosa’ was estimated as 813 Mb by flow cytometry analysis, and the assembled genome sequences covered 99% of the genome. A total of 108,087 protein coding genes in 341.4 Mb were predicted on the Camarosa_Genome_v1.0.a1 genome. The sequences are available on the GDR database.

Chromosome-scale scaffolds were also constructed for the ‘Reikou’ genome (Table 1, in preparation). Assembly was performed by DenovoMAGIC 3 with Illumina PE, MP, and 10X Genomics reads. The total length of the assembled sequences was 1,406 Mb, and consisted of 32,715 sequences and an N50 length of 3.9 Mb. Because the scaffolds were generated for all haplotype genomes in ‘Reikou’, so-called “phased genome” in the DenovoMAGIC analysis, the length was almost double the estimated genome size. The assembled genome was designated as FAN_r2.3. The scaffolds were aligned on a high density linkage map constructed by an S$_i$ mapping population of ‘Reikou’ (Nagano et al., 2017) with an IStraw 90K Axiom® Array (Bassil et al., 2015) and SSR markers (Isobe et al., 2013). The ‘Reikou’ linkage map consisted of 31 linkage groups, and a total of 62 (31 × 2 haplotypes) pseudomolecules were developed with a total length of 1,733.5 Mb. The FAN_r2.3 sequence is available on the download site in Strawberry GARDEN.

**De novo whole genome assembly in other Fragaria spp.**

The genomes of three diploids, *F. iinumae, F. nipponica, F. nubicola*, and one tetraploid, *F. orientalis*, were sequenced by Illumina HiSeq (Hirakawa et al., 2014). *De novo* assemblies were performed by SOAPdenovo v1.05 and GapCloser 1.10, and generated scaffolds with a total length of 200–206 Mb in diploid species, and 214 Mb in tetraploid *F. orientalis* (Table 1). The assembled genome sequences are available in the Strawberry GARDEN and GDR databases.

Chromosome-scale genome assembly in *F. iinumae* was recently reported by Edgar et al. (2019a). PacBio reads were obtained with 172x coverage and assembled by FALCON v0.3.0. The generated scaffolds were polished with Illumina reads by Pilon and then anchored to a *F. iinumae* map. As a result, eight chromosome-scale pseudomolecules were generated. The total length of the assembled sequences was 265.6 Mb, and consisted
of eight scaffolds and an N50 length of 33.98 Mb. The assembled genome sequences are available in the GDR databases.

Clarification of domestication history in strawberry with a genomic approach

The genome structure and domestication history are among the major unresolved issues in genetic and genomic studies in strawberry. Tennessen et al. (2014) discussed the evolutionary origins of strawberry based on dense linkage maps constructed with targeted capture sequences by POLiMAPS. By comparing strawberry and diploid Fragaria sequences, they concluded that one of the four subgenomes in strawberry originated from F. vesca, another was from F. iinumae, and the remaining two subgenomes were from an unknown ancestor close to F. iinumae. Their findings agreed with the results of a study by Hirakawa et al. (2014), which compared the strawberry genome (FANhybrid_r1.2) with the genomes of five Fragaria spp., F. vesca, F. iinumae, F. nipponica, F. nubicola, and F. orientalis. Based on the SNP-haplotype comparison, Sargent et al. (2016) also proposed a similar model, A-A, b-b, X-X, in which A, b, and X represented F. vesca-like, F. iinumae-like, and an unknown donor close to the F. iinumae genome, respectively.

Meanwhile, Yang and Davis (2017) performed phylogenetic analysis in Fragaria spp. using the Fluidigm Access Array system and Roche 454 platform by comparing 24 single or low copy nuclear genes in 16 Fragaria species. They hypothesized that genomes in octoploid species were contributed by at least four diploid species (F. vesca, F. iinumae, F. bucharica, F. viridis) and one unknown allele donor.

After construction of a chromosome-scale reference genome, Edger et al. (2019b) performed de novo transcriptome assembly in every described diploid Fragaria species. Phylogenetic analysis based on 19,302 nuclear genes with strawberry and other Fragaria species suggested that F. iinumae and F. nipponica, which are endemic species in Japan, were progenitors. Edger and colleagues hypothesized that the two species (F. iinumae and F. nipponica) donated genomes for generating five tetraploid species in China. The third candidate diploid ancestor is F. viridis, which distributes in Europe and Asia, and partially overlaps its distribution with a hexaploid F. moschata in East Eurasia. It was assumed that F. moschata originated from a hybridization between F. viridis and a tetraploid derived from F. iinumae and F. nipponica. The fourth candidate ancestor is F. vesca subsp. bracteata, which is an endemic species in North America. With these four candidate ancestor diploid species in hand, it was hypothesized that octoploid species were generated in North America by integration of a hexaploid generated in East Eurasia and an F. vesca subsp. bracteata. Edger et al. (2019b) also explored the potential existence of a “dominant” subgenome, which would have greater gene contents and higher gene expression, based on the abundance and distribution of transposable elements (TE), and concluded that a dominant subgenome was contributed by the F. vesca progenitor.

Conclusion

Only a decade ago, the construction of a chromosome-scale reference genome in strawberry was only a dream. Now that dream is a reality, and we have new basic information that can be used to answer a host of questions in this crop. Edger et al. (2019a) put forward a hypothesis in regard to the origin of strawberry, but further studies will be needed to verify their version of the domestication history. For example, Liston et al. (2019) immediately argued against the hypothesis based on the results of a chromosome-scale phylogenomic analysis and phylogenetic analysis based on an F. moschata linkage map. They averred that Edger et al. (2019b) drew the wrong conclusion from their gene-scale phylogenetic analysis, and proposed that while F. vesca and F. iinumae were possible ancestors in strawberry, F. nipponica, F. viridis, and F. moschata were not. However, shortly after this refutation, Edger et al. (2019a) countered with the construction of a chromosome-scale reference genome in F. iinumae for chromosome-scale comparison in multiple Fragaria species. NGS technology will continue to fuel such conversation, and to generate further chromosome-scale assembled genomes and accurate gene prediction. It is expected that the development of high quality assembled genomes will in turn impact evolutionary studies as well as genomics, molecular genetics and breeding in strawberry.

Literature Cited


Bertioli, D. J., S. B. Cannon, L. Froenicke, G. Huang, A. D.


