Multiplexed *Fragaria* Chloroplast Genome Sequencing

W. Njuguna
Department of Horticulture
Oregon State University
Corvallis, Oregon 97331
USA

A. Liston
Department of Botany and Plant Pathology
Oregon State University
Corvallis, Oregon 97331
USA

R. Cronn
United States Forest Service
Pacific Northwest Research Station
Corvallis, Oregon, 97331
USA

N.V. Bassil
USDA/ARS
National Clonal Germplasm Repository
Corvallis, Oregon 97333
USA

Keywords: high-throughput sequencing, polyploidy, reference genome, sequence alignments, microreads

Abstract

A method to sequence multiple chloroplast genomes using ultra high throughput sequencing technologies was recently described. Complete chloroplast genome sequences can resolve phylogenetic relationships at low taxonomic levels and identify informative point mutations and indels. The objective of this research was to sequence multiple *Fragaria* chloroplast genomes using the Illumina Genome Analyzer. Sixty-three PCR fragments from 22 species were sequenced in four multiplex sequencing runs. Plastome sequences were assembled using a combination of de novo and reference guided assembly with *F. vesca* 'Hawaii 4' providing the reference. De novo assembly resulted in plastome coverage of 43-74%, and reference guided assembly increased the plastome coverage to 76-82%. The alignment of sequenced chloroplast genomes was 96,438 bp and contained 319 parsimony-informative sites which will be useful in identifying chloroplast genome regions of high sequence variation for testing evolutionary relationships.

INTRODUCTION

Strawberry, *Fragaria* L., is in the Rosoideae subfamily in Rosaceae (Potter et al., 2007). The genus *Fragaria* was classified in the Fragariinae subtribe, Potentilleae tribe in the Rosodae superfamily. *Fragaria* species include thirteen diploids ($2n=2x=14$), five tetraploids ($2n=4x=28$), one hexaploid ($2n=6x=42$), four octoploids ($2n=8x=56$) (Staudt, 2008) and one decaploid ($2n=10x=70$) (Hummer et al., 2009). *F. iturupensis*, whose octoploid forms were previously reported (Staudt, 1989; Staudt and Olbricht, 2008), is also represented by decaploid forms (Hummer et al., 2009), which constituted the first reported naturally occurring decaploid strawberry species. Resolution of *Fragaria* phylogeny is not only useful for identification of sources of useful genes (Bringhurst and Voth, 1984; Lawrence et al., 1990), but also for verification of current species designations by inference from predicted species relationships. The need to verify *Fragaria* species designations became evident following flow cytometry (Hummer and Bassil, 2008; Hummer et al., 2009), simple sequence repeat (SSR) data analysis (Njuguna and Bassil, 2008), and species introductions and revisions to the strawberry germplasm collection at the USDA/ARS/NCGR in Corvallis, Oregon by strawberry expert Günter Staudt (1928-2008).

In *Fragaria*, phylogenetic analyses have used chloroplast (Harrison et al., 1997) and nuclear genome sequences (Potter et al., 2000; Rousseau-Gueutin et al., 2009), but relationships remain unclear. The suggested *Fragaria* octoploid genome models AAA'A'BBB'B' (Bringhurst, 1990) and YYY'Y'ZZZZ/YYYYZZZZ (Rousseau-Gueutin et al., 2009), suggest the contribution of two to four diploids to the octoploid genome. The diploid donor species are still not known, but evidence based on grouping of octoploid

and diploid nuclear genes (*ADH*, alcohol dehydrogenase; *GBSSI*-2 or *Waxy*; and *DHAR*, dehydroascorbate reductase) points to *F. vesca* L., *F. mandschurica* Staudt sp. *nova* and *F. iinumae* Makino (Davis and DiMeglio, 2004; Harrison et al., 1997; Potter et al., 2000; Rousseau-Gueutin et al., 2009; Senanayake and Bringhurst, 1967) as possible contributors. Based on low copy nuclear gene sequences, Rousseau-Gueutin et al. (2009) grouped *Fragaria* diploids into three clades. However, the diploid species within clade X were poorly resolved and the placement of *F. bucharica* remains unclear. The use of nuclear genes for phylogenetic analysis is complicated by polyploidy and recombination (Nishikawa et al., 2002), making the chloroplast genome an attractive option for *Fragaria*.

Further exploitation of the chloroplast genome in *Fragaria* for phylogenetic analysis is warranted due its small size, non-recombinant nature, and high sequence conservation, all factors that reduce the complexity of analysis and interpretation of results. For efficient use of limited chloroplast sequence divergence, large-scale sequencing is required, and is now possible with high throughput sequencing platforms. Sequencing of multiple small genomes to high coverage depth using high-throughput sequencing platforms was recently demonstrated (Cronn et al., 2008). In this study, we used multiplex sequencing to uncover sequence divergences in the chloroplast genomes of *Fragaria* species.

MATERIALS AND METHODS

Plant Material and DNA Extraction

Twenty-one of 22 wild *Fragaria* species, and one *Potentilla villosa* accession, a close relative of *Fragaria* in the Rosaceae family, that have been preserved at the germplasm repository in Corvallis (Table 1) were included in the study. DNA was extracted from actively-growing leaves using a modified protocol based on the PUREGENE® kit (Gentra Systems Inc. Minneapolis, MN). The DNA samples were cleaned further with Nucleon PhytoPure™ resin, a component found in the illustra Nucleon Phytopure™ kits for plant and fungal DNA extraction (GE Healthcare UK Limited, Buckinghamshire, UK). DNA quality and quantity was analyzed with a Wallac 1420 VictorV microplate reader (PerkinElmer, Waltham, MA). DNA concentration was adjusted to 3 ng/µl for PCR.

Primer Selection, Design and PCR

A total of 203 chloroplast primers (108 forward, 95 reverse) were screened in various logical combinations in four species to identify primer pairs that amplify >2500 bp fragments and provide maximum coverage of the chloroplast genome. Where possible, primer pairs that amplified single bands and that amplified in most or all of the species were chosen. Of the 203 primers, 141 had been used to amplify the *Cucumis sativus* L. chloroplast genome (Chung et al., 2007), 25 were designed from the complete genome sequence of *Morus indica* 'K2' (Ravi et al., 2006), and 36 were designed from the nearly complete chloroplast genome of *F. vesca* cv. Hawaii4 (http://strawberry.vbi.vt.edu/tiki-index.php). Sixty-three primer pairs were chosen to amplify the entire chloroplast genomes of 21 *Fragaria* species and one *Potentilla* accession (list available upon request from corresponding author). Long-range PCR was carried out using Phusion™ High-Fidelity DNA polymerase (New England Biolabs, Ipswich, MA). Amplifications were performed in 10 µl total reaction volumes containing 5x Phusion GC buffer, 2.5 mM of each dNTP, 10 µM of each primer, 5 U of Phusion DNA polymerase, 0.05 µl of 3% DMSO and 5 ng of DNA template. PCR product quantification was carried out using the Quant-iT™ PicoGreen® dsDNA quantification protocol (Molecular Probes, Inc. Eugene, OR) manufacturer's specifications. Equimolar amounts of PCR products were pooled for each species to generate 1-5 µg of chloroplast DNA which was prepared for sequencing using the Illumina sample preparation kit (Illumina Inc., San Diego, CA) and the modifications of Cronn et al. (2008). Each PCR pool was sheared using nitrogen from a

compressed source at 42 psi for two minutes. The fragments were repaired to remove 3' and fill 5' overhangs before adding 'A' bases to the 3' ends to allow the ligation of single 'T' bases on the 3' end of the adapters. Illumina adapters modified by the addition of a three base pair barcode were ligated to the fragments; a different barcode was used for each species in a multiplex pool. DNA fragments of approximately 300 bp were isolated by cutting the fragments from a 2% low melting agarose gel and PCR was performed to enrich for the targeted fragment sizes. Fragments from five or six chloroplast genomes were then mixed in equimolar ratios at a final concentration of 10 nM per pool in each of four multiplex pools.

Sequence Data Analysis

After the sequencing run, raw image data for each sequencing cycle was processed into base calls and alignment files through the Illumina/Solexa Pipeline (version 0.2.2.6). Binning was carried out using the three base pair nucleotide tags. After sorting microreads into species-specific bins, the barcodes and adapter tags were removed and resulting 32 bp microreads used for subsequent analysis. Contigs of *F. vesca* 'Baron Solemacher' were assembled de novo using Velvet Assembler 0.4 (Zerbino and Birney, 2008), a hash length of 17, minimum coverage 5x, and minimum contig length of 100. The contigs were then aligned to the 135,946 bp chloroplast genome of *F. vesca* 'Hawaii4'. The alignment was manually checked for errors, with insertions and deletions for a more accurate alignment. Non-target sequences from other cellular genomes were discarded and a consensus sequence was generated in Bioedit. The 'Baron Solemacher' sequence was 132,287 bp in length and served as the reference for subsequent analyses. The reference genome was labeled 'FvRefv2' and annotated using DOGMA (http://dogma.ccbb.utexas.edu/).

Contigs for each of the 22 species were assembled as described for *F. vesca* 'Baron Solemacher'. 'N's were added to the ends of the contigs from each of the 22 species to aid in distinguishing gaps in the sequence from indels in the alignment. The N-ended contigs were then aligned to the FvRefv2 reference using CodonCode (www.codoncode.com) and a consensus sequence was generated in Bioedit 7.0 (www.mbio.ncsu.edu/bioedit/bioedit.html/). Gaps (but not indels) between de novo contigs were replaced with FvRefv2 sequence forming a chimeric sequence which was then used for reference guided assembly (RGA) (Shen and Mockler, in prep; http://rga.cgrb.oregonstate.edu). The resulting 'chimeric reference sequence' was different for each species. The settings used for RGA were ≤2 mismatches per microread, Q-values ≥20, read depth ≥5, error rate of 0.06 and SNP acceptance requiring ≥70% of reads to be in agreement.

The online software MAFFT (http://align.bmr.kyushu-u.ac.jp/mafft/software/) was used to align the assemblies from the RGA output. The MAFFT alignment (134,000 bp) was manually checked for misalignments and sequencing errors. The aligned sequences were searched for the presence of primer sequences, and their positions and amplicon sizes in *Fragaria* were noted. Nucleotide positions corresponding to primer positions were eliminated from the alignment because most of the primers used for PCR were not of *Fragaria* origin. Indels were scored for future phylogenetic analysis and summary statistics related to the number of reads in each amplicon were manually calculated.

RESULTS AND DISCUSSION

The 63 primer pairs used in this study were designed to amplify fragments ranging in size from 871-5317 bp with an average size of 3044 bp. Their genome positions and sizes were based on the completely sequenced chloroplast genome of *Nicotiana tabacum* L. (Shinozaki et al., 1986). Primer sequence positions in FvRefv2 allowed the calculation of amplicon sizes in *Fragaria*. Based on these calculations, the *Fragaria* chloroplast genome is estimated to be 131,860 bp, which is smaller than that of *N. tabacum* (155,844 bp). Based on FvRefv2, the mean amplicon size was 3045 bp, median size was 3014 bp, and the size of amplified fragments ranged from 845-5386 bp. Of the 126 primers used to

amplify the *Fragaria* and *Potentilla* chloroplast genomes, 20 sequences (forward and/or reverse) were not found in FvRefv2. The missing primer sequences are attributed to sequencing failures, the smaller size of the chloroplast genome of *Fragaria*, and mutations in primer target regions. Since PCR success was verified by agarose gel electrophoresis before sequencing, amplified products that were not detected in the sequenced chloroplasts could represent non-target amplification of nuclear and/or mitochondrial genome regions.

The median number of reads obtained for each amplicon in the different species ranged from 0-56 while the average was 16. Amplicons with median numbers ≤5 may indicate low quantity in the PCR amplicon pool for the species due to underestimation of PCR product quantity or amplification of non-target DNA. Amplicons with average median values of <5 reads per base pair corresponded to those flanked by missing primer sequences. This further indicates non-target amplification of some regions of the chloroplast genome. Manual improvement of the MAFFT alignment by removing regions of possible sequencing errors, primer sequences, non-target amplicon sequences, and the addition of insertions and deletion reduced the length of the aligned sequence from 134,000 to 96,390 bp. Forty-eight indels were noted in the aligned chloroplast sequences. The final alignment of sequenced genomes was 96,438 bp including indels, and contained 995 variable sites and 319 parsimony-informative sites. The limited variation we observed in the chloroplast genome of *Fragaria* represents less than 1% of the MAFFT alignment, consistent with the low variability noted in previous studies (Harrison et al., 1997; Potter et al., 2000). However, the few variable sites observed in this study will prove useful in resolving relationships among *Fragaria* species as well as in identifying variable regions in the chloroplast genome that may be used as DNA barcodes.

Summaries of sequencing run output and sequencing analysis results are provided (Table 2). The genome coverage obtained with RGA (36-82%) was significantly higher than that obtained with de novo assembly using Velvet Assembler (version 0.4) (7-74%) (p value=0.00001). RGA aligns microreads to their best match on a reference genome and thus higher genome coverage was expected. The genome coverage based on RGA results for three of the species, *F. bucharica* (69%), *F. moschata* (46%), and *F. nubicola* (36%) was significantly lower than for the remaining species (p value=0.04). The low coverage observed in these three species could be due to errors during PCR product pooling or in preparation for sequencing.

CONCLUSIONS

Multiplex sequencing of *Fragaria* chloroplast genomes resulted in genome coverages ranging from 73-82% in 16 species, and low genome coverage in *F. bucharica*, *F. moschata* and *F. nubicola*. To reduce non-specific and null amplifications and errors in amplicon pooling concentrations, the samples with low coverage will be sequenced directly using genomic DNA preps as will four additional samples including *F. iinumae*, *F. pentaphylla*, *F. chiloensis* and *F. moupinensis* that had genome coverage values ≤76% from RGA. This may help to complete the dataset enabling more comprehensive phylogenetic analysis. A preliminary phylogenetic analysis (not shown) supported the three diploid clades of Rousseau-Gueutin et al. (2009) with bootstrap values ranging from 98-100%. The majority of the remaining nodes were supported by 100% bootstrap values, which are higher than those in previous *Fragaria* phylogenetic studies using molecular markers and nuclear and chloroplast genome sequences (Harrison et al., 1997; Potter et al., 2000; Rousseau-Gueutin et al., 2009). The results of this study will be useful in identifying chloroplast genome regions of high sequence variation for testing evolutionary relationships in future studies such as DNA barcoding.

ACKNOWLEDGMENTS

The authors thank Theodore Bunch, Michael Dossett, April Nyberg and Sarah Sundholm for invaluable technical support. We also thank Matthew Parks for assistance with advice in setting up experiments, Dr. Brian Knaus for laboratory and data analysis

assistance and, Mark Dasenko and Chris Sullivan (OSU, Center for Genome Research and Biocomputing) for sequencing and data management support.

Literature Cited

Bringhurst, R.S. 1990. Cytogenetics and evolution in American *Fragaria*. HortScience 25:879-881.

Bringhurst, R.S. and Voth, V. 1984. Breeding octoploid strawberries. Iowa State Journal of Research 58:371-381.

Chung, S.M., Gordon, V.S. and Staub, J.E. 2007. Sequencing cucumber (*Cucumis sativus* L.) chloroplast genomes identifies differences between chilling-tolerant and -susceptible cucumber lines. Genome 50:215-225.

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R. and Mockler, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36:122.

Davis, T.M. and DiMeglio, L.M. 2004. Identification of putative diploid genome donors to the octoploid cultivated strawberry, *Fragaria* ×*ananassa*. Plant and Animal Genome XII. San Diego, CA, January 10-14. (poster #603).

Harrison, R.E., Luby, J.J. and Furnier, G.R. 1997. Chloroplast DNA restriction fragment variation among strawberry (*Fragaria* spp.) taxa. J. Amer. Soc. Hort. Sci. 122:63-68.

Hummer, K. and Bassil, N. 2008. Unexpected polyploidy in wild Asian strawberries HortScience 43:1187.

Hummer, K., Nathewet, P. and Yanagi, T. 2009. Decaploidy in *Fragaria iturupensis* Staudt (Rosaceae). Amer. J. Bot. 96:713-716.

Lawrence, F.J., Galleta, G.J. and Scott, D.H. 1990. Strawberry breeding work of the United States Department of Agriculture. HortScience 25:895-896.

Nishikawa, T., Salomon, B., Komatsuda, T. and von Bothmer, R. 2002. Molecular phylogeny of the genus *Hordeum* using three chloroplast DNA sequences. Genome. 45:1157-1166.

Njuguna, W. and Bassil, N. 2008. A microsatellite fingerprinting set for strawberry, *Fragaria* L. HortScience 43: 915.

Potter, D., Eriksson, T., Evans, R.C., Oh, S., Smedmark, J.E.E., Morgan, D.R., Kerr, M., Robertson, K.R., Arsenault, M., Dickinson, T.A. and Campbell, C.S. 2007. Phylogeny and classification of Rosaceae. Plant Syst. Evol. 266:5-43.

Potter, D., Luby, J.J. and Harrison, R.E. 2000. Phylogenetic relationships among species of *Fragaria* (Rosaceae) inferred from non-coding nuclear and chloroplast DNA sequences. Syst. Bot. 25:337-348.

Ravi, V., Khurana, J., Tyagi, A. and Khurana, P. 2006. The chloroplast genome of mulberry: complete nucleotide sequence, gene organization and comparative analysis. Tree Gen. Genom. 3:49-59.

Rousseau-Gueutin, M., Gaston, A., Aïnouche, A., Aïnouche, M.L., Olbricht, K., Staudt, G., Richard, L. and Denoyes-Rothan, B. 2009. Tracking the evolutionary history of polyploidy in *Fragaria* L. (strawberry): New insights from phylogenetic analyses of low-copy nuclear genes. Mol. Phylogen. Evol. 51:515-530.

Senanayake, Y.D.A. and Bringhurst, R.S. 1967. Origin of *Fragaria* polyploids. I. Cytological analysis. Amer. J. Bot. 51:221-228.

Staudt, G. 1989. The species of *Fragaria*, their taxonomy and geographical distribution. Acta Hort. 265:24-31.

Staudt, G. 2008. Strawberry biogeography, genetics and systematics. In: VI International symposium, 3-7 March 2008, Huelva.

Staudt, G. and Olbricht, K. 2008. Notes on Asiatic *Fragaria* species V: *F. nipponica* and *F. iturupensis*. Botanische Jahrbücher für Systematik 127:317-341.

Zerbino, D. and Birney, E. 2008. Velvet: algorithms for de novo short read assembly using De Bruijn graphs. Genome Res. 18:821-829.

Tables

Table 1. Taxa, PI numbers, ploidy and origin of *Fragaria* accessions used for chloroplast sequencing.

Taxon	PI number	Ploidy	Origin
F. orientalis	PI 551864	4*x*	Russian Federation
F. iinumae	PI 637963	2*x*	Japan
F. nipponica	PI 637975	2*x*	Japan
F. virginiana	PI 612492	8*x*	Canada
F. iturupensis	PI 641091	10*x*	Russian Federation
F. viridis	PI 616857	2*x*	Sweden
F. ×bifera	PI 616613	2*x*	France
F. nilgerrensis	PI 616672	2*x*	China
F. bucharica	PI 551851	2*x*	Pakistan
F. vesca	PI 551507	2*x*	Germany
F. moschata	PI 551528	6*x*	France
F. daltoniana	PI 641195	2*x*	China
F. chiloensis	PI 612318	8*x*	Ecuador
F. tibetica	PI 651567	4*x*	China
F. pentaphylla	PI 651568	2*x*	China
F. gracilis	CFRA 1908	4*x*	China
F. corymbosa	CFRA 1911	4*x*	China
F. ×ananassa ssp. *cuneifolia*	PI 551805	8*x*	United States
F. chinensis	PI 616583	2*x*	China
F. mandschurica	CFRA 1947	2*x*	Mongolia
F. nubicola	PI551853	2*x*	Pakistan

Table 2. Illumina sequencing output and data summary, including the tag (3 bp barcode used for each sample), the multiplex pool for each sample, the sum of microreads, average and median number of reads for each base pair position, de novo and reference guided assembly (RGA) coverage.

Taxon	Tag	Lane/pool	Microreads	Average reads/bp	Median reads/bp	De novo coverage (%)	RGA coverage (%)
F. iinumae	acg	FragA	607,838	51.62	6	43	76
F. corymbosa	agc		1,600,377	205.47	50	74	82
F. tibetica	ccc		1,193,513	161.60	14	65	80
F. viridis	ctg		962,316	113.77	19	68	80
F. daltoniana	gta		621,757	80.72	15	64	81
P. villosa	tac	FragB	1,133,877	112.33	4	54	73
F. nipponica	gat		579,426	47.95	10	58	80
F. bucharica	tca		295,691	7.37	2	10	69
F. moschata	ggg		871,489	72.70	0	10	46
F. iturupensis	tgc		1,921,668	121.96	13	61	78
F. nubicola	tgc	FragC	1,209,936	92.35	0	7	36
F. ×ananassa ssp. *cuneifolia*	aac		864,943	108.11	12	64	82
F. gracilis	atg		305,281	28.00	8	51	79
F. virginiana	cac		727,536	79.40	14	63	79
F. pentaphylla	gct		757,266	21.93	4	27	75
F. chiloensis	ttg		348,758	19.45	3	22	73
F.×bifera	acg	FragD	470,616	43.80	7	48	79
F. mandschurica	agc		682,461	72.96	10	55	81
F. chinensis	ccc		820,618	107.22	17	69	81
F. vesca	ctg		647,651	87.51	8	51	79
F. orientalis	gta		866,894	97.79	14	63	79
F. nilgerrensis	tac		1,042,414	137.55	13	62	80