Gene Content and Distribution in the Nuclear Genome of *Fragaria vesca*

Ana Clara Pontaroli, Rebekah L. Rogers, Qian Zhang, Melanie E. Shields, Thomas M. Davis, Kevin M. Folta, Phillip SanMiguel, and Jeffrey L. Bennetzen*

Abstract

Thirty fosmids were randomly selected from a library of *Fragaria vesca* subsp. *americana* (cv. Pawtuckaway) DNA. These fosmid clones were individually sheared, and ~4- to 5-kb fragments were subcloned. Subclones on a single 384-well plate were sequenced bidirectionally for each fosmid. Assembly of these data yielded 12 fosmid inserts completely sequenced, 14 inserts as 2 to 3 contiguous sequences (contigs), and 4 inserts with 5 to 9 contigs. In most cases, a single unambiguous contig order and orientation was determined, so no further finishing was required to identify genes and their relative arrangement. One hundred fifty-eight genes were identified in the ~1.0 Mb of nuclear genomic DNA that was assembled. Because these fosmids were randomly chosen, this allowed prediction of the genetic content of the entire ~200 Mb *F. vesca* genome as about 30,500 protein-encoding genes, plus >4700 truncated gene fragments. The genes are mostly arranged in gene-rich regions, to a variable degree intermixed with transposable elements (TEs). The most abundant TEs in *F. vesca* were found to be long terminal repeat (LTR) retrotransposons, and these comprised about 13% of the DNA analyzed. Over 30 new repeat families were discovered, mostly TEs, and the total TE content of *F. vesca* is predicted to be at least 16%.

Published in The Plant Genome 2:93–101. Published 18 Mar. 2009.
doi: 10.3835/plantgenome2008.09.0007
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA
An open-access publication

THE GENUS *Fragaria* is an agriculturally important clade within the family Rosaceae that contains 23 identified species, including the domesticated strawberry (*Fragaria* × *ananassa*), an octoploid derived from accidental hybridization in the early 18th century between *F. chiloensis* and *F. virginiana* (Staudt 1989). The family Rosaceae includes many domesticated fruit and nut crops, such as peach (*Prunus persica*), apple (*Malus domestica*), and almond (*Amygdalus communis*). Nuclear genome sizes in the Rosaceae family tend to be relatively small, and the ~270-Mb peach genome will soon be sequenced and assembled (Sosinski et al. 2008). Given their economic importance and tractable genomes, multiple genome sequencing projects and comparative analyses would be highly justified in this family. One diploid *Fragaria* species, *F. vesca* ($2n = 2x = 14$), contains one of the smallest genomes in any of the flowering plants, about 200 Mb (Akiyama et al. 2001, Folta and Davis 2006), and it is a putative ancestor to the domesticated strawberry. Microsatellite and gene-specific marker genetic maps, expressed sequence tag (EST) data and routine *Agrobacterium*-mediated transformation protocols have been developed in *F. vesca* (reviewed in Folta and Davis 2006).

A.C. Pontaroli, R.L. Rogers, and J.L. Bennetzen, Dep. of Genetics, Univ. of Georgia, Athens, GA 30602; Q. Zhang, M.E. Shields, and T.M. Davis, Dep. of Biological Sciences, Univ. of New Hampshire, Durham, NH 03824; K.M. Folta, Horticultural Sciences Dep., Univ. of Florida, Gainesville, FL 32611; P. SanMiguel, Dep. of Horticulture and Landscape Architecture, Purdue Univ., West Lafayette, IN 47907; A.C. Pontaroli, current address: Estación Experimental Agropecuaria Balcarce, Instituto Nacional de Tecnología Agropecuaria (INTA), CC 276 (7620) Balcarce, Buenos Aires, Argentina; R.L Rogers, current address: Dep. of Organismic and Evolutionary Biology, Harvard Univ., Cambridge, MA 02138. Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. EU024823–EU024872. Received: 6 Sep. 2008. *Corresponding author (maize@uga.edu)

Abbreviations: BAC, bacterial artificial chromosome; EST, expressed sequence tag; LTR, long terminal repeat; TB, transposable element; WGS, whole genome shotgun sequencing.

Hence, *F. vesca* can serve as a model genome for the study of strawberry. Moreover, sequence analysis of the *F. vesca* genome will provide an excellent opportunity to generate insights into plant genome evolution, especially by comparison to other sequenced plant genomes.

Even with current "next generation" technologies (reviewed in Mardis 2008), de novo genome sequence analysis in plants is an expensive, slow, and laborious process. Whole genome shotgun sequencing (WGS) decreases the cost and increases the rate of data generation compared to an ordered clone-by-clone analysis, but it makes assembly and annotation much more challenging. In recent years, virtually all plant genome sequencing programs have opted for the WGS strategy (National Center for Biotechnology Information, 2009) and have accepted the limitation that thousands of gaps (including much of the repetitive DNA) remain in the assembled data. Several gene-enrichment techniques have been developed to increase the efficiency of complete gene discovery and assembly (Rabinowicz et al., 1999; Whitelaw et al., 2003; Yuan et al., 2003; Emberton et al., 2005), and numerous simple approaches allow unambiguous alignment of the gene sequence islands relative to each other and to the genetic and physical maps (Bennetzen et al., 2001; Yuan et al., 2002; Emrich et al., 2004). Still, it would be more satisfying to generate complete assemblies despite their great expense and labor and time requirements, especially because rare cases have been found where gene regulatory components are embedded in the repetitive regions as much as 70 to 100 kb away from the genes (Stam et al., 2002; Clark et al., 2006).

Fortunately, a comprehensive description of the general rules of genome content and organization does not require complete sequence analysis. The GeneTrek technology uses random selection and low-pass sequencing of a small number of clones to describe the overall composition and arrangement of genes, transposable elements (TEs), and other abundant sequences in any genome (Bennetzen 2003, Devos et al. 2005). A GeneTrek analysis of the maize (*Zea mays* L.) genome involving 74 bacterial artificial chromosome (BAC) clones (~0.5% of the nuclear genome) indicated a gene content of about 37,000, with more than 5500 severely truncated pseudogenes (Liu et al. 2007). Using GeneTrek on a great number of genomes, sampled from a phylogenetic perspective, would allow identification of lineages where dramatic changes have occurred in the rates of gene or genome evolution, such as in genome instability, gene loss, or gene duplication, and thereby pinpoint species deserving additional genomic investigation.

We describe here the application of the GeneTrek approach to *F. vesca*. Thirty fosmid clones were analyzed and found to contain 158 predicted genes and 84 identified TEs. The results of this analysis indicate that *F. vesca* contains about 30,500 genes and that they are mostly arranged in gene-rich regions (~1 gene per 5.7 kb) that would be highly amenable to a WGS analysis.

Materials and Methods

Random Fosmid Selection and Fosmid Subclone Library Construction

Following screening by probe hybridization to identify and exclude clones derived from the mitochondrial and chloroplast genomes, 30 fosmids were randomly selected from a library of 33,000 clones containing inserts of mechanically sheared DNA derived from *F. vesca* subsp. *americana* cv. Pawtuckaway (Shields et al., 2005). Fosmid DNA was isolated using the standard alkaline lysis protocol (Sambrook et al., 1989), and DNA was sheared to ~4- to 5-kb fragments using a Hydro-Shear DNA Shearer (Genomic Solutions, Applied Biosystems, Foster City, CA). Dephosphorylated blunt ends were generated with mung bean (*Vigna radiata*) nuclease (30 min at 30°C) followed by incubation with shrimp alkaline phosphatase (60 min at 37°C, plus 15 min enzyme inactivation at 65°C). Adenosine overhangs were added to the 3′ ends with *Taq* DNA polymerase (30 min at 72°C). DNA samples were run in 20-cm-long, 0.8% agarose gels with 0.5X TBE buffer, at 80 V for ~2.5 h. After extraction from the gel with the Qiaex II kit (Qiagen, Valencia, CA), the DNA was ligated into the pCR-4-TOPO vector (Invitrogen, Carlsbad, CA) and subsequently transformed into *Escherichia coli* DH10B cells (ElectroMAX, Invitrogen). Transformants were selected in Luria-Bertani medium + 50 µg mL^{-1} kanamycin. Plasmid DNA for sequencing was isolated from *E. coli* using standard techniques (Sambrook et al., 1989).

Fosmid Sequence Analysis and Assembly

Fosmid clones were sequenced to an average of 14× redundancy (one 384-well plate per fosmid, bidirectionally), using an ABI 3700 capillary sequencer with T3 and T7 primers (5′-ATTAACCCTCACTAAAGGGA-3′ and 5′-TAATACGACTCACTATAGGG-3′, respectively) and ABI PRISM Big Dye Terminator chemistry (Applied BioSystems). Base calling and quality assessment were performed with PHRED (Ewing et al., 1998), and reads were assembled with PHRAP. Assemblies were visually inspected, and contigs were ordered, using CONSED (Gordon et al., 1998).

Fosmid Sequence Annotation

Ab initio gene prediction was performed on each fosmid sequence using FGENESH (Softberry, Inc., Mount Kisco, NY) with the *Arabidopsis* training set. All the obtained predicted genes, as well as the genomic DNA sequence, were used as queries in BlastX searches against the National Center for Biotechnology Information nonredundant protein database for *Viridiplantae* and the *Arabidopsis* protein database (http://arabidopsis.org/servlets/Search?action=new_search&type=protein). In parallel, each of the fosmid sequences was queried in BlastN and tBlastX searches against itself, the entire *F. vesca* sequence dataset used in this study plus an additional set of 20 gene-targeted *F. vesca* fosmids

(Davis, unpublished data; Genbank accession numbers EU024823-EU024872), an in-house plant repeat database and the TIGR Plant Repeat Database (http://www.jcvi.org/). Ab initio repeat identification was performed using LTR_STRUC (McCarthy and McDonald, 2003) and a *Helitron*-finding program (L. Yang and J. Bennetzen, unpublished). All of the outputs generated (from gene- and repeat-finding programs, nucleotide and protein database queries) were layered onto the fosmid genomic sequence using the Apollo Genome Annotation Curation Tool (version 1.9.1; Lewis et al., 2002). FGENESH-predicted genes with homology to known repeat sequences, or to manually annotated TEs in other *F. vesca* fosmids, were removed from subsequent analysis of gene content. All BlastX hits were manually inspected to assess the likelihood of a predicted gene to be "real," based on homology to known or hypothetical *Arabidopsis* proteins (Expect value $\leq 10^{-10}$), the evolutionary range of species with significant hits, and the length of sequence homology compared with the total length of the target gene. Severely truncated genes—that is, those missing more than 30% of their full protein-encoding length compared to the consensus length of the gene from *A. thaliana* and other plant species—were classified as gene fragments and not considered in total gene number estimations. Transposable elements were identified either by homology to known TEs or by structural features, following the process described by Liu et al. (2007).

Modeling Gene Content from the Sampled Data

Total gene number in the *F. vesca* genome was first estimated by simple extrapolation of the value observed in the sample of sequences used in this study (30 fosmids, comprising 1.035 Mb, or 0.54%, of the ~200 Mb *F. vesca* genome). To account for the sampling error derived from the usage of a finite and relatively low number of randomly selected sequences, 1000 iterations of gene number prediction were performed using sampling with replacement for a total of 30 fosmid inserts. Both the full range of gene number predictions and the range within a 95% confidence level were computed.

Results

Thirty fosmids were randomly selected and the inserts shotgun sequenced, using a single 384-well plate of subclones. This approach generated ~480 kb of data, or about 14× redundancy for the inserts (which averaged ~34.5 kb) after vector and low-quality reads were removed. Even without any further finishing reactions to fill gaps or replace sequences with low-quality PHRED scores, 12 fosmids were found to be fully finished at Bermuda quality standards (Human Genome Project Information, 1997). Most fosmids, however, had some sequence gaps, as shown in Table 1. Despite these gaps, all of the fosmids could be fully annotated for gene and TE content, thereby saving the more than twofold additional cost and time that would have been required to fully finish these sequences.

All fosmid inserts were first analyzed for gene content. Gene candidates were discovered initially using FGENESH, and those with homology to annotated TEs were removed from further analysis. The remaining gene candidates were only considered "confirmed" if they exhibited $\leq 10^{-10}$ (e-value) homology at the predicted peptide level to a hypothetical or known gene in *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000). Otherwise, they were classified as "hypothetical genes." This highly conservative approach will miss any genes that may be specific to the *Fragaria* lineage (if such genes exist) or any genes that may have been missed in the >19 Mb of the *A. thaliana* genome that was not initially sequenced (Liu and Bennetzen, 2008). This conservative gene confirmation strategy was used because the gene-finding programs alone can more than double gene number predictions in plants because of the high content of novel open reading frames found in low-copy-number TEs (Bennetzen et al., 2004). Most of these TEs are represented in EST libraries, so an expression criterion is of little use for gene confirmation, but the high rate of TE evolution relative to genes makes cross-species conservation a very good tool for gene confirmation (Frazer et al., 2003, Bennetzen et al., 2004).

Out of the 1.035 Mb of assembled insert data, 182 protein-encoding genes (137 confirmed plus 45 hypothetical) were predicted. Eight of the confirmed genes were interrupted by the boundaries of the fosmids, a disadvantage of using inserts of this size compared to using the 100-kb-plus inserts routinely found in BAC clones. Even a single nucleotide change, leading, for instance, to an altered promoter or a missense mutation within an exon, can turn a gene into a pseudogene, meaning that we cannot predict what percentages of these genes are functional. However, 24 of the predicted genes that are fully internal to the fosmids appear to be severely truncated (missing >30% of their full protein-encoding length compared to the consensus length of the gene from *A. thaliana* and other plant species). Because genes could only be assayed for biological truncation if they were not truncated by the boundary of the clone and if they had a homolog elsewhere that could be characterized for its entire length, the 24 biologically truncated genes must be compared to 137 confirmed genes minus 8 genes (i.e., 129 confirmed genes) that were truncated on the sequenced clones. Thus, the frequency of severely truncated genes was measured as 24/(24 + 129) or 16%.

Given that the *F. vesca* genome is predicted to be ~200 Mb in size, a simple extrapolation from these data suggests 21,840 (confirmed) or 30,530 (confirmed plus hypothetical) genes, plus 4780 pseudogenes, in the entire nuclear genome. However, any finite number of sequenced clones will create some sampling error. Hence, 1000 repetitions of gene number prediction were performed using sampling with replacement for a total of up to 30 fosmid inserts. This generated an average gene number prediction of 30,500 as expected, with a range of ~21,900 to ~39,300. A total of 95% of the predictions were

Table 1. Insert size, contig arrangement, and gene content of the *Fragaria vesca* fosmid sequences investigated

Fosmid clone	Genbank Accession No.	Insert size	Total no. of contigs	Contig arrangement		Number of genes			No. of gene fragments
				Ordered	Unordered	Confirmed	Hypothetical	Confirmed + hypothetical	
		bp							
01L02	EU024824	40,302	1	1		5	4	9	2
05N03	EU024825	34,710	8	7	1[†]	4	0	4	0
11D02	EU024828	37,961	1	1		5	0	5	0
12K04	EU024829	33,108	6		6[†]	0	0	0	0
13I03	EU024830	37,707	1	1		7	0	7	1
15B13	EU024833	23,612	5	5		0	0	0	0
17G19	EU024834	17,219	1	1		2	0	2	1
17O22	EU024835	34,190	2	2		6	0	6	2
18A19	EU024836	40,908	1	1		3	3	6	0
22H18	EU024839	37,851	1	1		4	1	5	0
22L05	EU024840	35,212	2	2		3	3	6	0
27F10	EU024841	37,210	2	2		3	2	5	1
29G10	EU024842	31,881	3	3		3	1	4	3
30I24	EU024843	37,599	3	3		9	2	11	0
32A10	EU024844	33,677	2	2		4	5	9	1
34D20	EU024846	30,034	2	2		4	2	6	1
38H02	EU024848	31,669	1	1		3	1	4	0
38H05	EU024849	32,050	2	2		1	2	3	0
40B22	EU024850	36,230	1	1		8	1	9	0
40M11	EU024851	31,718	1	1		5	0	5	2
43P07	EU024853	43,741	2	2		2	2	4	1
44J07	EU024854	29,836	3	3		2	2	4	4
47H15	EU024855	35,017	3	3		1	2	3	0
49B16	EU024857	40,969	9	6	3[‡]	0	0	0	0
49C17	EU024858	37,266	3	3		4	2	6	1
52E09	EU024862	37,346	1	1		5	4	9	0
63F17	EU024866	27,689	2	2		3	2	5	2
72E18	EU024867	36,293	1	1		8	1	9	1
75H22	EU024869	31,662	1	1		7	1	8	1
84N10	EU024872	40,283	2	2		2	2	4	0

[†] Unordered contigs composed of ribosomal RNA genes.

[‡] Unordered contigs composed of long terminal repeat retrotransposon fragments.

in the range of ~24,600 to ~35,300 (Fig. 1), making this the predicted gene number for *F. vesca* under the gene annotation criteria used.

Transposable elements make up large parts of all plant genomes, even as much as 40% of the tiny (~100 Mb) *Selaginella moellendorffii* genome (Zhu, DeBarry, Yang and Bennetzen, unpublished data), but it is more routinely ~20% in small genomes like *A. thaliana* (~140 Mb) (Liu and Bennetzen, 2008) to >70% in medium-size angiosperm genomes like that of maize (~2400 Mb) (Liu et al. 2007). Because TEs exist in great variety and change sequence more rapidly than genes, it is not possible to describe all of the TEs in a new species merely by homology to TEs known from some other species. Hence, the most comprehensive discovery processes for TEs involve both homology-based and structure-based search criteria.

Each of the fosmid sequences was queried against itself, the entire *F. vesca* sequence dataset used in this study, an additional set of 20 gene-targeted *F. vesca* fosmids, an in-house plant repeat database, and the TIGR Plant Repeat Database. Sequences with hits of Expect value ≤ 10^{-10} were further inspected for TE structural features, and translated sequences were queried against protein databases to detect TE conserved coding domains (see above, "Materials and Methods"). This process uncovered a total of 84 intact or fragmented TEs, which were classified into 33 different TE families. The most abundant recognizable TEs were long terminal repeat (LTR) retrotransposons (13 *copia*-like, 14 *gypsy*-like, and 3 unclassified elements), followed by MITEs (13 elements), CACTA-like transposons (4 elements), non-LTR retrotransposons (4 elements) and 1 *Mu*-like element. Three clones—18A19, 49B16, and 84N10—exhibited blocks of nested LTR retrotransposons, which ranged from 12 to 35 kb in length. In terms of percentage genome composition, the LTR retrotransposons made up 13% of the annotated DNA. Unclassified repeats were as abundant as LTR retrotransposons (32 elements) but constituted only 4.1% of the DNA. In total, this study indicates that >16% of the *F. vesca* genome is composed of identified TEs. The TE and other repeat descriptions are summarized in Table 2.

Three classes of TEs, namely, LTR retrotransposons (Jin and Bennetzen, 1994), Pack-MULEs (Jiang et al., 2004), and *Helitrons* (Morgante et al., 2005), have been observed to acquire fragments of normal nuclear genes. In the annotated *F. vesca* DNA, no *Helitrons* or Pack-MULEs were detected, and none of the LTR retrotransposons discovered showed regions of homology to known nuclear genes. Given this result, it is not surprising that most of the gene fragments detected in this study were found in the vicinity of their corresponding full-length genes. Adjacent repeated genes are expected products of unequal homologous recombination, and the fragmentation (i.e., functional inactivation) of at least one copy would be expected as the most frequent final outcome of this duplication process (Lynch and Connery, 2000; Devos et al., 2002).

Most of the fosmids sequenced were very rich in genes, with few or no identified TEs. On these gene-rich fosmids, predicted gene density averaged 1 gene per 5.7 kb. One such clone is depicted in Fig. 2. Five fosmids contained no or few genes: three of these fosmids were partially (clone 05N03) or totally (clones 12K04 and 15B13) composed of ribosomal RNA genes, another fosmid (clone 49B16) consisted of a cluster of nested transposons, and the remaining fosmid (clone 11D02) had a mitochondrial DNA insertion that made up about half of its sequence length (Table 1). This last clone is likely an outcome of a mitochondrial DNA insertion in the nuclear genome because it has stop codons and various indels in ORFs required for mitochondrial function (data not shown). Organellar DNA insertions into nuclear DNA are common in plant genomes (Noutsos et al., 2005).

Discussion

Despite recent efforts to investigate the sequences of several plant genomes, relatively little is understood concerning their immense range of sizes and compositions. Although most or perhaps all of the major mechanisms of genome rearrangement (polyploidy, TE amplification, ectopic recombination, chromosome breakage, and illegitimate recombination) are now known, it is not known why there are different levels of activity for each of these phenomena in even closely related species (Vitte and Bennetzen, 2006). To begin to understand these processes, many more plant genomes need to be screened for unusual genomic behavior (such as exceptionally high or low levels of TE activity) to identify those species where the rules governing genomic rearrangement activity may be best investigated. The GeneTrek procedure (Bennetzen, 2003), used here on the *F. vesca* genome, allows an efficient search for such species. A GeneTrek analysis also allows prediction of the results, resources required, and best approach to full genome sequence analysis of target model species like *F. vesca* (Devos et al., 2005; Liu et al., 2007).

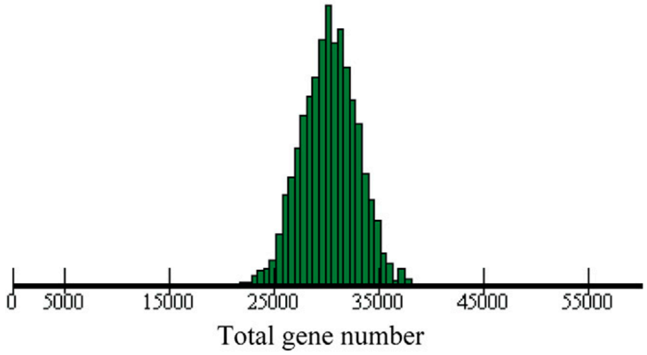

Figure 1. *Fragaria vesca* gene number modeled by sampling with replacement: mean distribution for 1000 iterations of a sample size of 30 fosmid sequences. Mean = 30,530 genes; range = ~21,900 to ~39,300 genes (95% confidence level).

To determine how many genes need to be sequenced in a full genome analysis, and to determine whether gene or TE evolution is unusual in a genome, it is essential to derive reasonably accurate estimations of total gene number and TE composition in a target genome. However, predicting gene and TE content is a process fraught with potential errors. Even fully sequenced genomes can yield gene number assessments that are off by more than twofold if the erroneous contributions of low-copy-number TEs are not taken into account (Bennetzen et al., 2004). Some publications have gone so far as to conclude that an apparent coding sequence cannot be a TE if it does not have a high copy number or does have an identified EST. However, all or virtually all TEs are expressed at some level in some tissue or treatment, and low-copy-number TEs are abundant in any genome where they are sought by structural criteria (SanMiguel and Bennetzen, 1998; McCarthy and McDonald, 2003; Ma and Bennetzen, 2004). Finally, the acquisition of gene fragments by TEs can easily give rise to false gene identification (Fu and Dooner, 2002; Morgante et al., 2005). These issues are now generally recognized, so that future genome annotations should give much more accurate assessments of both gene and TE composition.

Given the exceptionally dynamic nature of angiosperm genomes (Wendel et al., 2002; Wessler, 2006; Bennetzen, 2007), surveys of hundreds of species within a given family would be a useful way to discover those individual taxa with hyper-exceptional rates of genome change. Such hyper-evolving species would serve as appropriate organisms for efficient investigation of the mechanisms by which genome structure and function might shift into a rapid evolution mode, and how this change becomes manifest. Even with current technologies, however, de novo full genome sequencing of thousands of species would cost billions of dollars. Sample sequence analyses, such as the GeneTrek approach (Bennetzen, 2003; Devos et al., 2005) or related strategies (Hawkins et al., 2006; Piegu et al., 2006), can characterize a genome's most abundant components (i.e., genes, satellites, and TEs) in a general way for less than 1% of the cost of a full genome analysis. These approaches require some creativity in data analyses (Devos et al., 2005; Hawkins et al., 2006; Liu et al., 2007; DeBarry et al., 2008) but need only small amounts of sequence information so long as it is generated from randomly selected clones. Smaller insert clones have the advantage that they sample more genomic locations per megabase analyzed, but the disadvantage of a higher ratio of edge effects. A particularly problematic edge effect is that features may not be identified because of the limited extension of data at the edge. For instance, single reads from a small plasmid insert were found to only allow de novo gene identification in *A. thaliana* shotgun sequence data about 66% of the time that a read was inside a known gene (Liu and Bennetzen, 2008).

Table 2. Transposable element (TE) composition of the *Fragaria vesca* fosmid sequences investigated.

TE type	Family	Number of elements		Sequence covered	
		Intact	Fragmented		
				bp	%
LTR retrotransposons[†]		17	13	135,585	13.1
Copia-like		9	4	55,752	5.4
	13I03	1	–	4,146	
	27F10_1	2	–	12,416	
	38H02_1	1	–	3,635	
	47H15	1	–	7,167	
	52E09_1	3	4	21,638	
	84N10_1	1	–	6750	
Gypsy-like		5	9	68,271	6.6
	17G19_1	1	–	3,364	
	18A19_1	1	2	28,266	
	18A19_2	1	6	22,157	
	22H18	1	1	9,437	
	38H05_1	1	–	5,047	
unclassified		3	–	11,562	1.1
	22L05_1	1	–	3,361	
	29G10_1	1	–	5,733	
	43P07_1	1	–	2,468	
Non-LTR retrotransposons		3	1	11,663	1.1
	44J07_1	–	1	482	
	47H15_1	1	–	5,669	
	52E09_1	1	–	1,946	
	72E18_1	1	–	3,566	
CACTA-like transposons		2	2	13,403	1.3
	29G10	1	2	11,668	
	34D20	1	–	1,735	
MITEs		10	3	3,127	0.3
	01L02	4	3	1,622	
	11D02	2	–	258	
	17O22	1	–	83	
	72E18	3	–	1,164	
Mu-like transposons		1	–	4,038	0.4
	40B22	1	–	4,038	
Unclassified repeats		12	20	43,466	4.2
	22H18_1	2	–	197	
	22H18_2	2	–	1,954	
	38H05_1	1	3	655	
	44J07_1	1	4	1,519	
	47H15_1	2	–	659	
	84N10_1	2	6	22,841	
	43P07_2	1	5	14,975	
	63F17_1	1	2	666	
Totals		45	39	211,282	20.4

[†] LTR, long terminal repeat.

The analysis of 30 fosmid clones led to the prediction of ~24,600 to ~35,300 genes in the *F. vesca* genome. This range is an outcome of the number of fosmids analyzed and would both narrow and become more accurate if additional fosmids were studied. Our modeling studies indicate that even a doubling of the number of fosmids would only narrow the range of predictions by 17% (data not shown). Hence, it seems that the current level of precision is an appropriate tradeoff between cost and accuracy.

The gene number prediction for *F. vesca* is in line with the predictions of ~27,000 for *A. thaliana* (Wortman et al., 2003), ~32,000 for rice (*Oryza sativa* L.; Rice Annotation Project, 2007), ~45,000 for black cottonwood (*Populus trichocarpa* L.; Tuskan et al., 2006), ~30,400 for grapevine (*Vitis vinifera* L.; Jaillon et al., 2007) and ~23,000 for papaya (*Carica papaya* L.; Ming et al., 2008). In these other species, however, a comprehensive effort has not been made to predict the number of pseudogenes. In maize, severely truncated pseudogenes comprise about 15% of the sequences that would be routinely authenticated as genes by all other criteria except a discrete search for truncation (Liu et al., 2007). The similar percentage of these pseudogenes predicted in *F. vesca* (16%) is intriguing, as the lower number of TEs in this small genome should generate fewer opportunities for TE-derived gene fragmentation, as with the gene fragments within *Helitrons* and Pack-MULEs. Still, the 16% number should be viewed as a low estimate because many of the other sequences annotated as genes are likely to be inactivated by smaller indels, by promoter mutations, or by single nucleotide changes in coding regions. On the other hand, our conservative requirement that true *F. vesca* genes have a homolog in *Arabidopsis* will underreport those few genes that either were created uniquely in the *F. vesca* lineage or were lost from the *A. thaliana* lineage. Taking all of these results in combination, across all of these species, strongly suggests that a gene number of 25,000 genes per haploid genome is fully sufficient for the functioning of most or all flowering plant genomes.

The TE composition of *F. vesca* is predicted to be at least 16% of the genome, which is in line with the estimated TE content of the *Arabidopsis* genome (Liu and Bennetzen, 2008). The sampling procedure used in this study should have yielded an unbiased picture of the TE landscape in the *F. vesca* genome. Hence, the

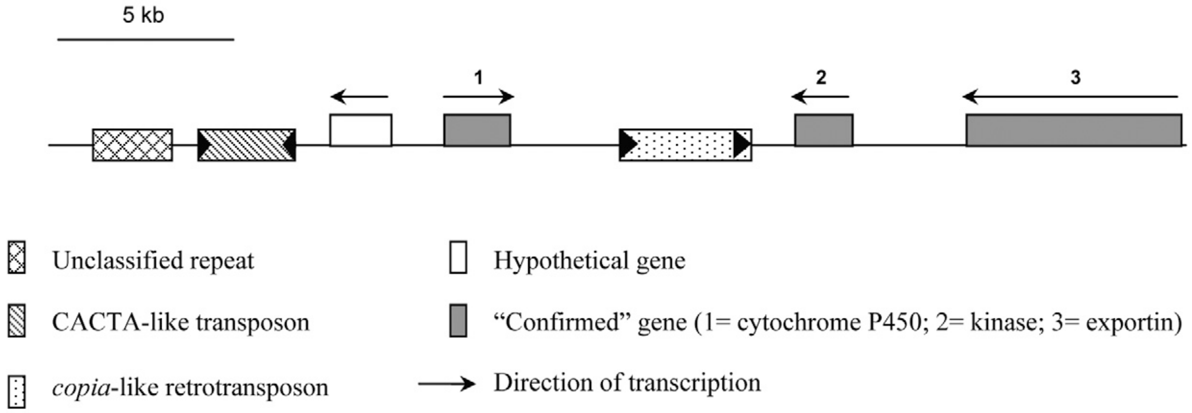
5 kb
1
2
3
Unclassified repeat
CACTA-like transposon
copia-like retrotransposon
Hypothetical gene
"Confirmed" gene (1= cytochrome P450; 2= kinase; 3= exportin)
Direction of transcription

Figure 2. Gene and transposon arrangement on an average *Fragaria vesca* insert in the fosmids analyzed. The fosmid clone shown is 38H02. Boxes with arrows above indicate predicted genes; the gray-filled boxes are confirmed genes (predicted to encode: 1 = a cytochrome P450 enzyme, 2 = a kinase, and 3 = an exportin). Cross-hatched fill indicates an unclassified repeat, the box with opposing arrows indicates a CACTA-like transposon, and the box with the directly repeated arrows indicates a *copia*-like long terminal repeat retrotransposon.

data generated here can serve as a basis for ascertaining whether this small genome is subject to the same mechanisms responsible for genome contraction and expansion in other species, like *Arabidopsis* (Devos et al., 2002) and rice (Ma et al., 2004; Vitte and Panaud, 2003).

Because extant *F. vesca* is a likely descendant of one of the diploid ancestors of the octoploid commercial strawberry, *Fragaria ×ananassa*, it is probable that the sequence of the *F. vesca* genome will be quite informative vis-à-vis the *Fragaria ×ananassa* genome. In general, one expects genomes from the most closely related species to be most alike, but the rapid genomic change that can be initiated after polyploidy (reviewed in Chen, 2007) suggests that *F. vesca* studies may provide an imprecise guide to the *Fragaria ×ananassa* genome. The most important of these changes appears to be "fractionation," whereby some of the homeologous genomes lose copies of many of the genes duplicated by the polyploidy event (Ilic et al., 2003; Lai et al., 2004; Freeling, 2008). In such cases, however, at least one copy of every original gene is retained (on one homeolog or another), such that a diploid ancestor is actually a better indicator of overall gene content in the polyploid than would be the sequence of any single homeologous region in the polyploid itself. A recent study of genetic linkage associations between diploid and octoploid strawberry indicates that the octoploid maintains seven linkage groups that are homeologous to those in *F. vesca*. There is strong agreement between the arrangement of genes in the diploid and the polyploid, suggesting that major arrangements have not occurred in these members of the genus (Rousseau-Gueutin et al., 2008). This serves as another argument for the value of *F. vesca* as a model genome, one that would be of tremendous value even if the commercial strawberry genome were fully sequenced.

The data generated in this study may also be useful for analyzing the local colinearity of genes across the *F. vesca* genome, compared to other sequenced plant genomes. This analysis will be particularly powerful if extended across multiple species of known phylogenetic relatedness, such as the relationships described between *Fragaria* and *Prunus* species (Vilanova et al., 2008). Such studies allow examination of the natures and rates of adjacent gene rearrangement, a process that has been extensively investigated in the grasses (Chen et al., 1998; Tikhonov et al., 1999; Song et al., 2002; Ilic et al., 2003; Wicker et al., 2003; Bowers et al., 2005) but much less so in other flowering plants (Rossberg et al., 2001; Tang et al., 2008). Moreover, a comparative study of this nature has the potential to identify the precise sequences of events, the rates, and the lineage specificities (Ilic et al., 2003) that have shuffled the gene content of plant genomes. Once such properties have been investigated across a wide range of species, it may become possible to generate a unified theory of the evolution of plant genome structure and function, thus allowing prediction of how genomes might be changed to alter important biological processes.

Acknowledgments

This project was supported in part by grant no. 2005-35300-15467 from the National Research Initiative Plant Genome Program of USDA's Cooperative State Research, Education and Extension Service.

References

Akiyama, Y., Y. Yamamoto, N. Ohmido, M. Oshima, and K. Fukui. 2001. Estimation of the nuclear DNA content of strawberries (*Fragaria* spp.) compared with *Arabidopsis thaliana* by using dual-stem flow cytometry. Cytologia (Tokyo) 66:431–436.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

Bennetzen, J.L. 2003. Higher throughput comparative genomics of wheat and related cereals. Proc. Tenth Int. Wheat Genetics Symp. 1:215–220.

Bennetzen, J.L. 2007. Patterns in grass genome evolution. Curr. Opin. Plant Biol. 10:176–181.

Bennetzen, J.L., V.L. Chandler, and P. Schnable. 2001. National Science Foundation-sponsored workshop report: Maize genome sequencing project. Plant Physiol. 127:1572–1578.

Bennetzen, J.L., C. Coleman, J. Ma, R. Liu, and W. Ramakrishna. 2004. Consistent over-estimation of gene number in complex plant genomes. Curr. Opin. Plant Biol. 7:732–736.

Bowers, J.E., M.A. Arias, R. Asher, J.A. Avise, R.T. Ball, G.A. Brewer, et al. 2005. Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proc. Natl. Acad. Sci. USA 102:13206–13211.

Clark, R.M., T.N. Wagler, P. Quijada, and J. Doebley. 2006. A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. Nat. Genet. 38:594–597.

Chen, J. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu. Rev. Plant Biol. 58:377–406.

Chen, M., P. SanMiguel, and J.L. Bennetzen. 1998. Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. Genetics 148:435–443.

DeBarry, J.D., R. Liu, and J.L. Bennetzen. 2008. Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF). Bioinformatics 9:235.

Devos, K.M., J.K.M. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 12:1075–1079.

Devos, K.M., J. Ma, A.C. Pontaroli, L.H. Pratt, and J.L. Bennetzen. 2005. Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. Proc. Natl. Acad. Sci. USA 102:19243–19248.

Emberton, J., J. Ma, Y. Yuan, and J.L. Bennetzen. 2005. Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. Genome Res. 15:1441–1446.

Emrich, S.J., S. Aluru, Y. Fu, T.J. Wen, M. Narayanan, L. Guo, D.A. Ashlock, and P.S. Schnable. 2004. A strategy for assembling the maize (*Zea mays* L.) genome. Bioinformatics 20:140–147.

Ewing, B., L. Hillier, M. Wendl, and P. Green. 1998. Basecalling of automated sequencer traces using phred: I. Accuracy assessment. Genome Res. 8:175–185.

Folta, K.M., and T.M. Davis. 2006. Strawberry genes and genomics. Crit. Rev. Plant Sci. 25:399–415.

Frazer, K.A., L. Elnitski, D.M. Church, I. Dubchak, and R.C. Hardison. 2003. Cross-species sequence comparisons: A review of methods and available resources. Genome Res. 13:1–12.

Freeling, M. 2008. The evolutionary position of subfunctionalization, downgraded. Genome Dyn. 4:25–40.

Fu, H., and H.K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. USA 99:9573–9578.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: A graphical tool for sequence finishing. Genome Res. 8:195–202.

Hawkins, J.S., H. Kim, J.D. Nason, R.A. Wing, and J.F. Wendel. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. 16:1252–1261.

Human Genome Project Information. 1997. Policies on release of human genomic sequence data: Bermuda-quality sequence. Available at http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml (verified 10 Feb. 2009). Oak Ridge National Laboratory, Oak Ridge, TN.

Ilic, K., P.J. SanMiguel, and J.L. Bennetzen. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes. Proc. Natl. Acad. Sci. USA 100:12265–12270.

Jaillon, O., J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467.

Jiang, N., Z. Bao, X. Zhang, S.R. Eddy, and S.R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569–573.

Jin, Y.-K., and J.L. Bennetzen. 1994. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize. Plant Cell 6:1177–1186.

Lai, J., J. Ma, Z. Swigonova, W. Ramakrishna, E. Linton, V. Llaca, B. Tanyolac, Y.-J. Park, O.-Y. Jeong, J.L. Bennetzen, and J. Messing. 2004. Gene loss and movement in the maize genome. Genome Res. 14:1924–1931.

Lewis. S.E., S.M.J. Searle, N. Harris, M. Gibson, V. Iyer, J. Ricter, C. Wiel, L. Bayraktaroglu, E. Birney, M.A. Crosby, J.S. Kaminker, B. Matthews, S.E. Prochnik, C.D. Smith, J.L. Tupy, G.M. Rubin, S. Misra, C.J. Mungall, and M.E. Clamp. 2002. Apollo: a sequence annotation editor. *Genome Biology* 3(12):research0082.

Liu, R., and J.L. Bennetzen. 2008. ENCHILADA REDUX: How complete is your genome sequence? New Phytol. 179:249–250.

Liu, R., C. Vitte, J. Ma, A.A. Mahama, T. Dhliwayo, M. Lee, and J.L. Bennetzen. 2007. A GeneTrek analysis of the maize genome. Proc. Natl. Acad. Sci. USA 104:11844–11849.

Lynch, M., and J.S. Connery. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Ma, J., and J.L. Bennetzen. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA 101:12404–12410.

Ma, J., K.M. Devos, and J.L. Bennetzen. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14:860–869.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133–141.

McCarthy, E.M., and J.F. McDonald. 2003. LTR_STRUC: A novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367.

Ming, R., S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J.H. Saw, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996.

Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo, and A. Rafalski. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. 37:997–1002.

National Center for Biotechnology Information. 2009. Entrez genome project: Properties of eukaryotic genome sequencing projects. Available at http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi?taxgroup=11:|12:Land%20Plants&p3=12:Land%20Plants (verified 10 Feb. 2009). NCBI, Washington, DC.

Noutsos, C., E. Richly, and D. Leister. 2005. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Res. 15:616–628.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Sanyal, H.R. Kim, K. Collura, D.S. Brar, S.A. Jackson, R.A. Wing, and O. Panaud. 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 16:1262–1269.

Rabinowicz, P.D., K. Schutz, N. Dedhia, C. Yordan, L.D. Parnell, L. Stein, W.R. McCombie, and R.A. Martienssen. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat. Genet. 23:305–308.

Rice Annotation Project. 2007. Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. Genome Res. 17:175–183.

Rossberg, M., K. Theres, A. Acarkan, R. Herrero, T. Schmitt, K. Schumacher, G. Schmitz, and R. Schmidt. 2001. Comparative sequence analysis reveals extensive microcolinearity in the *lateral suppressor* regions of the tomato, arabidopsis, and capsella genomes. Plant Cell 13:979–988.

Rousseau-Gueutin, M., E. Lerceteau-Köhler, L. Barrot, D.J. Sargent, A. Monfort, D. Simpson, P. Arús, G. Guérin, and B. Denoyes-Rothan. 2008. Comparative genetic mapping between octoploid and diploid *Fragaria* species reveals a high level of colinearity between their genomes and the essentially disomic behavior of the cultivated octoploid strawberry. Genetics 179:2045–2060.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989 Molecular cloning: A laboratory manual. Cold Spring Harbor Lab. Press, Cold Spring Harbor, NY.

SanMiguel, P., and J.L. Bennetzen. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. (Lond.) 82:37–44.

Shields, M., Q, Zhang, and T. M. Davis. 2005. Large-insert genome library resources for strawberry (*Fragaria*). Poster 050. *In* Proc. of the Plant & Animal Genomes XIII Conf., San Diego, CA, 15–10 Jan. 2005.

Song, R., V. Llaca, and J. Messing. 2002. Mosaic organization of orthologous sequences in grass genomes. Genome Res. 12:1549–1555.

Sosinski, B., A. Abbott, D. Main, and R. McCombie. 2008. Sequencing the peach genome. p. 13. *In* Proc. of the 4th Rosaceae Genomics Conf., Pucon, Chile. 16–19 March 2008.

Stam, M., C. Belele, W. Ramakrishna, J. Dorweiler, J.L. Bennetzen, and V.L. Chandler. 2002. The regulatory regions required for *B′* paramutation and expression are located far upstream of the maize *b1* transcribed sequences. Genetics 162:917–930.

Staudt, G. 1989. The species of *Fragaria*, their taxonomy and geographical distribution. Acta Hortic. 265:23–34.

Tang, H., J.E. Bowers, X. Wang, R. Ming, M. Alam, and A.H. Paterson. 2008. Synteny and collinearity in plant genomes. Science 320:486–488.

Tikhonov, A.P., P.J. SanMiguel, Y. Nakajima, N.D. Gorenstein, J.L. Bennetzen, and Z. Avramova. 1999. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc. Natl. Acad. Sci. USA 96:7409–7414.

Tuskan, G.A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604.

Vilanova, S., D.J. Sargent, P. Arús, and A. Monfort. 2008. Synteny conservation between two distantly-related Rosaceae genomes: *Prunus* (the stone fruits) and *Fragaria* (the strawberry). Plant Biol. 8:67.

Vitte, C., and J.L. Bennetzen. 2006. Analysis of retrotransposon diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad. Sci. USA 103:17638–17643.

Vitte, C., and O. Panaud. 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice (*Oryza sativa* L.). Mol. Biol. Evol. 20(4):528–540.

Wendel, J.F., R.C. Cronn, J.S. Jonhston, and H.J. Price. 2002. Feast and famine in plant genomes. Genetica 115:37–47.

Wessler, S.R. 2006. Transposable elements and the evolution of eukaryotic genomes. Proc. Natl. Acad. Sci. USA 103:17600–17601.

Whitelaw, C.A., W.B. Barbazuk, G. Pertea, A.P. Chan, F. Cheung, Y. Lee, et al. 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science 302:2118–2120.

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z. Liu, J. Dubcovsky, and B. Keller. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^{m} genomes of wheat. Plant Cell 15:1186–1197.

Wortman, J.R., B.J. Haas, L.I. Hannick, R.K. Smith, Jr., R. Maiti, C.M. Ronning, et al. 2003. Annotation of the Arabidopsis genome. Plant Physiol. 132:461–468.

Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2002. Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. Genome Res. 12:1345–1349.

Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2003. High Cot sequence analysis of the maize genome. Plant J. 34:249–255.