Genomic Sequence Annotation and Deep Allele Sampling in *Fragaria*

Davis TM[1], Shields ME[1], Zhang Q[1], and Linyuan Li[2].
Departments of Biological Sciences[1] & Mathematics[2], University of New Hampshire

Tombolato DCM and Folta KM
Horticultural Sciences Department, University of Florida

Bennetzen JL and Pontaroli AC
Department of Genetics, University of Georgia

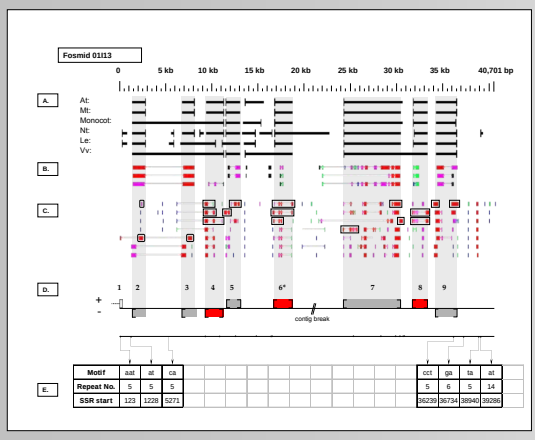
Genomic Sequence Annotation

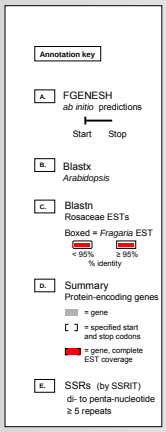
Background

***F. vesca* model species.**
Considerable attention has been devoted to development of *F. vesca* as a diploid model species for *Fragaria*. Its favorable features include: small (~200 Mb) genome size, self-compatibility, short generation time, small plant size, ease of seed and vegetative propagation, availability of useful mutants and inbred lines, well-developed linkage map, and ease of genetic transformation.

***F. vesca* 'Pawtuckaway' fosmid sequences.**
We have constructed an *F. vesca* genomic library in a fosmid vector using physically sheared DNA, and have sequenced 20 gene-targeted and 30 randomly selected clones, generating a total of 1.75 Mb genomic sequence.
The 50 sequences are deposited under GenBank accession numbers EU024823-EU024872.
This poster section describes the current annotation status of these sequences.

Sequence Annotation Sample





Putative stop	Putative identity	# aa	Blastx At/Virid.	Best E value	Locus (At)	Blastn EST 5' end	%ID	Blastn EST Middle(/ALL)	%ID	Blastn EST 3' end	%ID
excluded	Unknown protein	680	NP_194926	na	At4g31980	DY674412	99%				
		680	NP_194926	E=6.1e-73	At4g31980						
1392	unknown protein	448	NP_181184	E=3.3e-45	At2g36430						
		680	NP_194926	E=5.1e-88	At4g31980						
6783	unknown protein	448	NP_181184	E=1.2e-63	At2g36430						
										DV439723	100%
9445	unknown protein	209	NP_568247	E=8.3e-30	At5g11280			(EX684627)	99%	EX688210	100%
	OCP3 (OVEREXPRESSOR OF CATIONIC										
13050	PEROXIDASE 3)	354	NP_196688	E=4.0e031	At5g11270	CX661043	99%			CX661123	99%
	HY5 (ELONGATED HYPOCOTYL 5); DNA										
18732	binding / transcription factor	168	NP_568246	E=2.3e-39	At5g11260	CO816924	97%	(DY667374)	99%	EX686844	98%
i73-21072											
	CESA1 (CELLULOSE SYNTHASE 1);									EX657066	99%
30383	transferase, transferring glycosyl groups	1081	NP_194967	E=7.0e-160	At4g32410	DY674781	99%			DY672977	100%
								(EX684687)	99%		
33342	nucleic acid binding	130	NP_565781	E=2.1e-18	At2g34160			(EX684678)	99%		
	SHS1 (SODIUM HYPERSENSITIVE 1);										
34310	binding / transporter	392	NP_194966	E=1.5e-76	At4g32400	EX672645	99%			DY668003	99%

Annotation Summary

Gene density in gene-rich regions:
~ 1 protein-encoding gene per 5.7 kb

SSR density:
~ 1 SSR per 4.5 kb.

Transposable element content in 30 random clones (~ 1 Mb)

LTR retrotransposons:	13	*copia*-like
	14	*gypsy*-like
	3	unclassified
non-LTR retrotransposons:	4	
MITEs:	13	
CACTA-like transposons:	4	
Mu-like element:	1	

In terms of percent 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 comprised of identified TEs, which is in line with the estimated TE content of the *Arabidopsis* genome (Liu and Bennetzen 2008).

FGENESH predictions vs homology-based inferences.

When Blastx and Blastn homology-based inferences of start and stop codon positions in 20 gene-targeted clones were compared with FGENESH ab initio predictions, homology inferences were matched most closely by the *Medicago* and *Arabidopsis* FGENESH predictions, and least closely by the *Vitis*-based predictions.

Start	Stop	
99	92	Total possible matches
80	72	Matches: Mt
80	68	Matches: At
89	81	Matches: Mt+At
4	3	Additional matches, other than Mt+At
93	84	Matches: All
10	11	Missed by Mt+At
6	8	Missed by all

EST coverage

In the 20 gene-targeted fosmids (~0.75 Mb), only ~16% of all putative genes had complete *Fragaria* EST coverage (with ≥ 98% nucleotide identity), while over 50% had no *Fragaria* EST support at all.

These findings document a substantial need for more *Fragaria* EST sequencing, particularly from flower and fruit libraries and other sources and conditions not represented by existing libraries.

Acknowledgements:
This research is supported by USDA-CSREES NRI Plant Genome Grants 2005-35300-15467, 2008-35300-04411, and New Hampshire Agricultural Experiment Station Project NH00433.

Deep Allele Sampling

Background

The cultivated strawberry genome.
The cultivated strawberry, *Fragaria ×ananassa* is an octoploid (2N = 8X = 56), hybrid species, arising < 300 years ago via hybridization between octoploids *F. chiloensis* and *F. virginiana*.

Proposed octoploid genome models:

Three-genome models

I	**Fedorova (1946):**	AAAABBCC
II	**Senanayake & Bringhurst (1967):**	AAA'A'BBBB

Four-genome model

III	**Bringhurst (1990):**	AAA'A'BBB'B'

Critique: Each model is based on a different, very narrow, germplasm sampling, and no model has been rigorously validated.

Diploid ancestry.
Octoploid genome models imply subgenome contributions from either three (Models I and II) or four (Model III) different diploids. An initial phylogenetic analysis (Davis & DiMeglio, 2004) drew attention to four diploids (2N = 2X = 14) as possible genome donors to the octoploid species:

Candidate ancestral diploids:

primary candidates - *F. vesca*, *F. iinumae;*

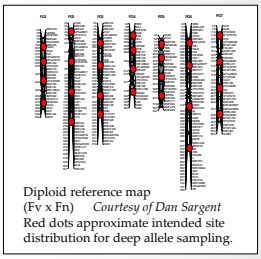
secondary candidates - *F. mandshurica* and *F. bucharica* (accessions formerly known as *F. nubicola*).

Caveat: Results from initial allele sampling at several gene pair loci suggest that an as yet unknown diploid species may also have been a genome contributor to the octoploids.

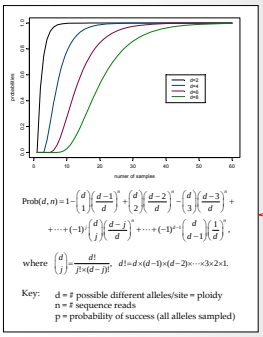
Hypothesis:

Genome composition varies within and between octoploid species.

This hypothesis will be tested by research conducted under a new USDA-NRI Plant Genome Grant: *A multi-dimensional approach to comparative genomics in Fragaria (Rosaceae).* TM Davis, Project Director.



Diploid reference map
(Fv x Fn) *Courtesy of Dan Sargent*
Red dots approximate intended site distribution for deep allele sampling.



Approach:

This project integrates molecular and cytogenetic methodologies.

I. Deep allele sampling

- 28 gene-based primer pairs, generating products in 370-400 base range.
- Loci well-distributed on diploid reference map (left).
- 96 germplasm accessions sampled.
- Pyrosequencing 8 pools of barcoded products (Roche – Titanium platform).

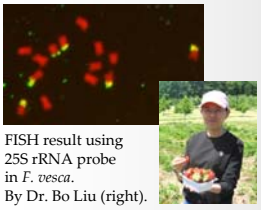
Accessions 1-12 28 primer pairs Barcodes: MID-1 – MID-12	Accessions 13-24 28 primer pairs Barcodes: MID-1 – MID-12	Accessions 25-36 28 primer pairs Barcodes: MID-1 – MID-12	Accessions 37-48 28 primer pairs Barcodes: MID-1 – MID-12
Accessions 49-60 28 primer pairs Barcodes: MID-1 – MID-12	Accessions 61-72 28 primer pairs Barcodes: MID-1 – MID-12	Accessions 73-84 28 primer pairs Barcodes: MID-1 – MID-12	Accessions 85-96 28 primer pairs Barcodes: MID-1 – MID-12

Adaptor Multiplex Identifier Sequences (5'-3')

MID-1 ACGAGTGCGT	MID-7 CGTGTCTCTA
MID-2 ACGCTCGACA	MID-8 CTCGCGTGTC
MID-3 AGACGCACTC	MID-9 TAGTATCAGC
MID-4 AGCACTGTAG	MID-10 TCTCTATGCG
MID-5 ATCAGACACG	MID-11 TGATACGTCT
MID-6 ATATCGCGAG	MID-12 TACTGAGCTA

Sampling strategy

Expect 80-120K reads per pool, 400+ bp/read.
Given 28 primer pairs and 12 accessions/pool:
28 x 12 = 336 combinations/pool.
Assuming 80K reads / 336 combinations,
= 238X coverage per primer pair/accession.
Is that enough??
As calculated, 95% and 99% confidence of complete allele sampling (per gene) in an octoploid requires, respectively, 38 and 51 reads to detect 8 different alleles, but only 16 and 21 reads to detect 4 different allele "types" (from 4 genome types). "Excess" coverage helps compensate for any pooling imbalances.



FISH result using 25S rRNA probe in *F. vesca*.
By Dr. Bo Liu (right).

II. Molecular cytogenetics

Fluorescent in situ hybridization (FISH) and Genomic in situ hybridization (GISH) analysis will be performed on a diverse sampling of octoploids, using probes from candidate diploids.

Concordance between genome composition patterns inferred from allele sampling and molecular cytogenetic analysis will be examined.

Expected Results

Define allelic diversity in 28 strawberry genes (112 loci in octoploids).

Identify SNPs of potential use to breeders at up to 112 loci in a diverse sampling of important cultivars.,

Resolve the question of octoploid genome composition(s).

Establish molecular cytological methods for determination of octoploid genome composition.

Define the octoploids' diploid ancestry, to extent currently possible.