Hunting Down the Papaya Transgenes

Michael Schatz

Center for Bioinformatics and Computational Biology
University of Maryland

January 16, 2008
PAG XVI

UNIVERSITY OF MARYLAND

Papaya Overview

- *Carica papaya* from the order Brassicales
  - 72 million years apart from the nearest common ancestor with *Arabidopsis*
- Productive food crop grown in tropical and sub-tropical regions world wide
  - One of the most important crops in Hawaii
- Known for its nutritional benefits and medicinal applications
  - Ranked first on nutritional scores for 38 common fruits based on USRDA for a variety of vitamins and minerals
  - Used in a wide range of medical applications including production of papain

Papaya Ringspot Potyvirus

- Plants infected with PRSV lose their photosynthetic capacity and display stunted growth, deformed and inedible fruit, and eventually, plant mortality.
  - When plants are infected at the seedling stage or within two months after planting, the trees will not produce mature fruit.
  - If trees are infected at a later stage, fruit production is reduced and of poor quality because of ringspots on the fruit and a decrease in sugar concentration.

- PRSV in Hawaii
  - Destroyed production on Oahu in the 1950s
  - Force relocation to the Puna district of Hawaii in the 1960s
  - First detected in Puna in 1992
  - By 1995, the industry was in a crisis situation with many fields devastated or abandoned

Images from Dennis Gonsalves
http://www.apsnet.org/online/feature/ringspot/

Pathogen-Derived Resistance

- In 1989, researchers at Cornell attempted a pathogen-derived resistance approach to PRSV
  - Genetically modify the target organism's genome to include genes from the pathogen
  - Interferes with the virus via post-transcriptional gene silencing or RNA-interference
  - Target gene was a coat protein (cp) gene of PRSV HA 5-1, a mild mutant of PRSV.
  - The gene was 'shot' into cultured papaya tissue using a 'gene gun'
- The transgenic line 55-1 was found to be immune to PRSV
  - Evaluated and approved under greenhouse and field conditions
  - Resulted in the development of the homozygous SunUp and heterozygous Rainbow cultivars
  - Commercially released in May 1998, six years after PRSV was discovered in Puna.

SunUp Success

- Papaya production in Puna quickly returned after the introduction of the transgenic lines.
  - Many growers and individuals acknowledge the transgenic lines saved the Hawaiian Papaya Industry.
- The transgenic plants are considered safe for human consumption in the US, in part, because humans have been consuming PRSV infected plants for years
  - Export of transgenic papaya, though, is still restricted because of lingering concern over the exact nature and impact of the papaya transgenes.

Table 1. Fresh papaya production[a] in the state of Hawaii and in the Puna district from 1992-2002.

	Fresh papaya utilization in Hawaii		
Year	Total (× 1,000 lb)	Puna (× 1,000 lb)	%
(virus in Puna) 1992	55,800	53,010	95
1993	58,200	55,290	95
1994	56,200	55,525	99
1995	41,900	39,215	94
1996	37,800	34,195	90
1997	35,700	27,810	78
(transgenic seeds released) 1998	35,600	26,750	75
1999	39,400	25,610	65
2000	50,350	33,950	68
2001	52,600	40,290	77
2002	42,700	35,880	84

[a] Data were compiled from USDA Statistical Reports of Papaya grown in Hawaii (www.nass.usda.gov/hi).

Papaya Genome Project

http://cgpbr.hawaii.edu/papaya/

- In 2004, the University of Hawaii Center for Genomics, Proteomics and Bioinformatics Research Initiative formed an integrative multi-institutional consortium to sequence the papaya tree genome
  - Maui High Performance Computing Center
  - Hawaii Agricultural Research Center
  - US Department of Agriculture
  - Pacific Telehealth & Technology Hui
  - Nankai University, China

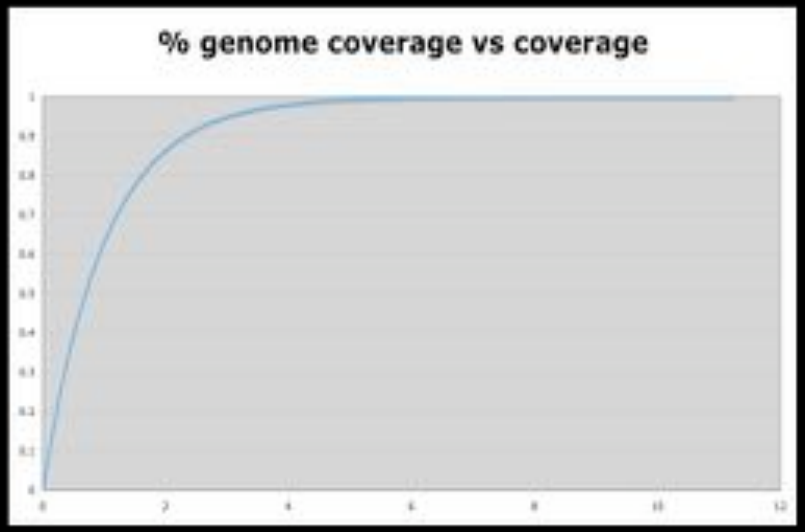
- Goals:
  - Improve the quality and productivity of tropical fruit trees
  - Increase our basic knowledge of higher plant biology
  - Characterize the transgenic insertions

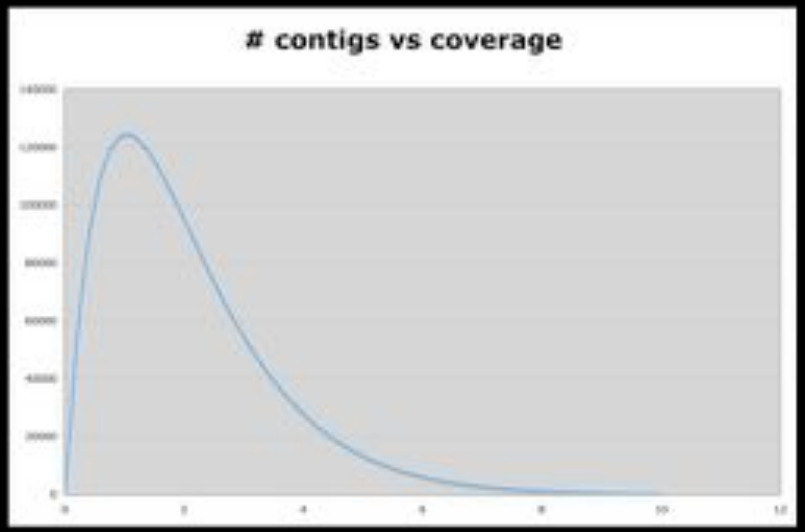
Sequence Data

Insert size (kb)	Genomic library	LUCY trimmed bases (billions)	LUCY trimmed bases after removing organellar sequences (billions)	Number of reads after removing organellar sequences (millions)	Sequence coverage by LUCY trimmed read bases	Fraction of paired reads (%)	Fraction of assembled reads (%)
3	Plasmid	1.01	0.67	0.86	1.80 X	95.2	88.1
6	Plasmid	0.57	0.47	0.62	1.26 X	93.5	87.2
8	Plasmid	0.11	0.04	0.07	0.10 X	90.1	67.4
86	BAC	0.02	0.02	0.03	0.06 X	97.3	80.8
174	BAC	0.02	0.02	0.03	0.06 X	95.7	84.2
total		1.73	1.22	1.61	3.27 X	94.4	86.7

- All reads from a homozygous female SunUp cultivar plant.
- ~11x insert coverage in long range BAC libraries
- Had to exclude 600k reads from the 160kbp chloroplast and 477kbp mitochondrial genomes to improve assembly.

Genome Assembly





- Lander-Waterman statistics predict ~95% of the genome in contigs
  - Relatively high number of contigs
- Assembly Software
  - Arachne:
    - http://broad.mit.edu/wga/
  - Celera Assembler
    - http://wgs-assembler.sourceforge.net
  - LUCY
    - ftp://ftp.tigr.org/pub/software/Lucy/
  - AMOS
    - http://amos.sourceforge.net
  - MUMmer
    - http://mummer.sourceforge.net
- Several iterations of assembly
  - Iteratively improved trimming and scaffolding parameters
  - Celera Assembler is especially sensitive to accurate trimming and uniform coverage

Assembly Statistics

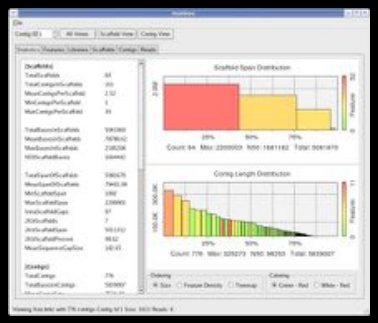
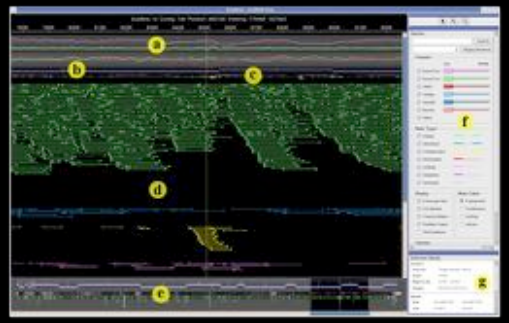
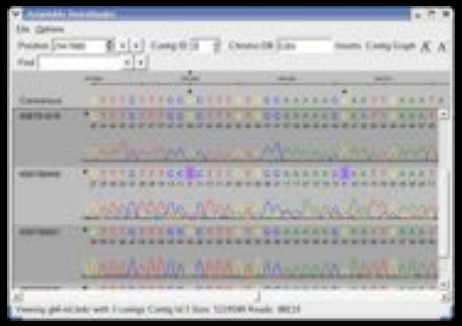
- Final assembly with Arachne
  - Better scaffolds than Celera Assembler
- 278 Mbp in contigs
  - ~75% of the total genome
  - ~90% of the euchromatic regions of the genome
- Contigs and scaffolds anchored and oriented to the 12 papaya linkage groups
  - Utilized 652 of 706 markers in the FPC-based physical map

Total span of scaffolds (Mb)	372
N50 of scaffolds (Mb)	1.0
Number of scaffolds	18,650
Total length of contigs (Mb)	278
N50 of contigs (kb)	11
Number of contigs	48,409
Total length of anchored scaffolds (Mb)	235
Total length of anchored and oriented scaffolds (Mb)	161
Number of anchored scaffolds	291
Total length of anchored contigs (Mb)	167
Total length of anchored and oriented contigs (Mb)	117
Number of anchored contigs	20,636

Assembly Validation

- Compared WGS contigs to finished BACs
  - Error rate for ≥ 3X (74.2% of assembled sequences) is ≤ 0.01%
  - Error rate for 2X (16.3%) is ≤ 0.37%
  - Error rate for 1X (9.5%) is ≤ 0.75%

- Computational methods for finding mis-assembled regions
  - Scan the assembly for suspicious regions with *amosvalidate*
    - Mate Pairs, Depth of Coverage, Repeat K-mers, SNPs, Breakpoints
  - In-depth analysis of flagged regions with *Hawkeye*

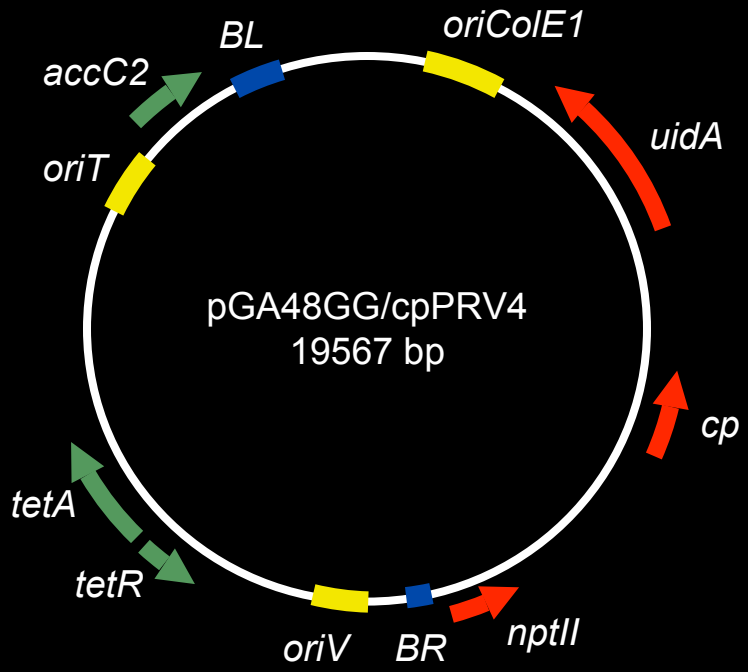




http://amos.sourceforge.net/hawkeye

Transformation Vector

Target transgenes

- *CP:* coat protein gene of PRSV HA 5-1
- *nptII:* neomycin phosphotransferase
- *uidA:* β-glucuronidase (GUS)

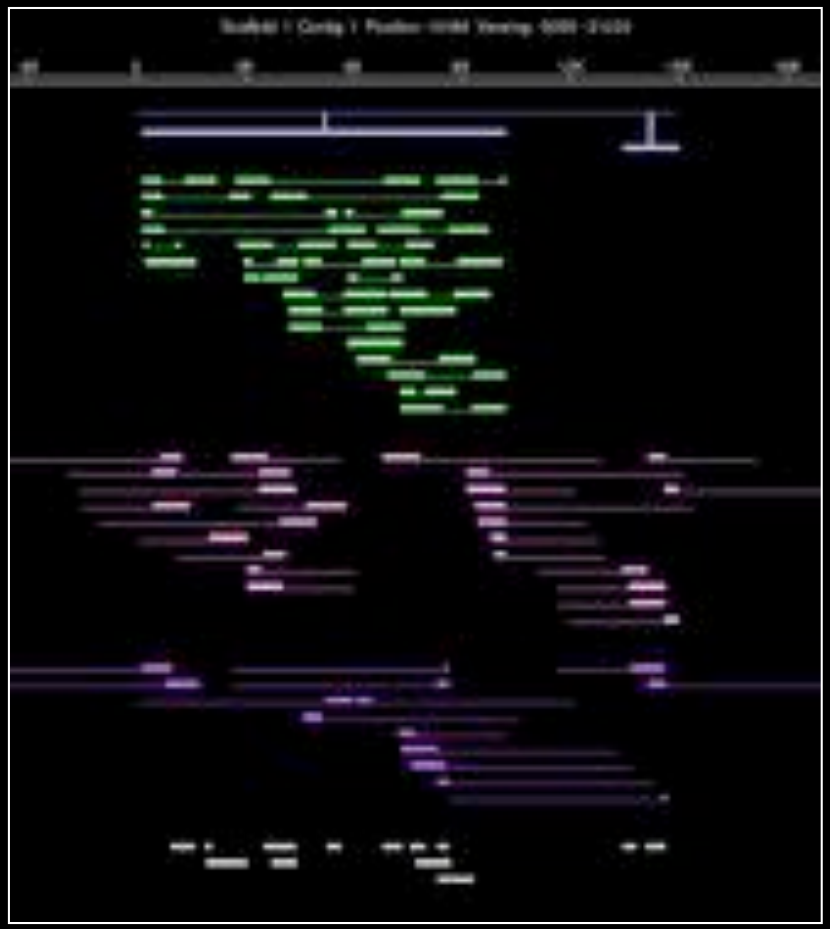


Vector backbone genes

- *tetA*, *tetR:* tetracycline resistance
- *aacC3:* gentamycin resistance
- BL, BR: nonfunctional 5' and 3' halves of β-lactamase
- oriV, oriT, and oriColE1: plasmid replication origins

Transgene Alignments

- Align the sequencing reads to the transformation vector using MUMmer
  - Require at least 20bp exact match, at least 65bp total in alignment (e value < 10^{-31})
  - Avoids potentially mis-assembled contigs and/or singleton reads
- Filter low complexity sequence in the transformation vector to avoid spurious alignments
  - 86bp of T's at 5043-5128
  - 69bp of T's and C' at 5961-6029
- Require alignments extend beyond regions that that are highly similar to sequencing vectors
  - 837-1203, 1324-1502
  - 120-1667, 18498-19255



Transgene Insertions

3 insertions confirmed by Southern blot

1. (*trnS*,*ycf3*) **Functional transgene** (*trnL, trnF*)

9,789 bp

3 reads at 5’ junction, 4 reads at 3’, happy mates spanning

2. **Vector/*tet A*** (*ycf2*)

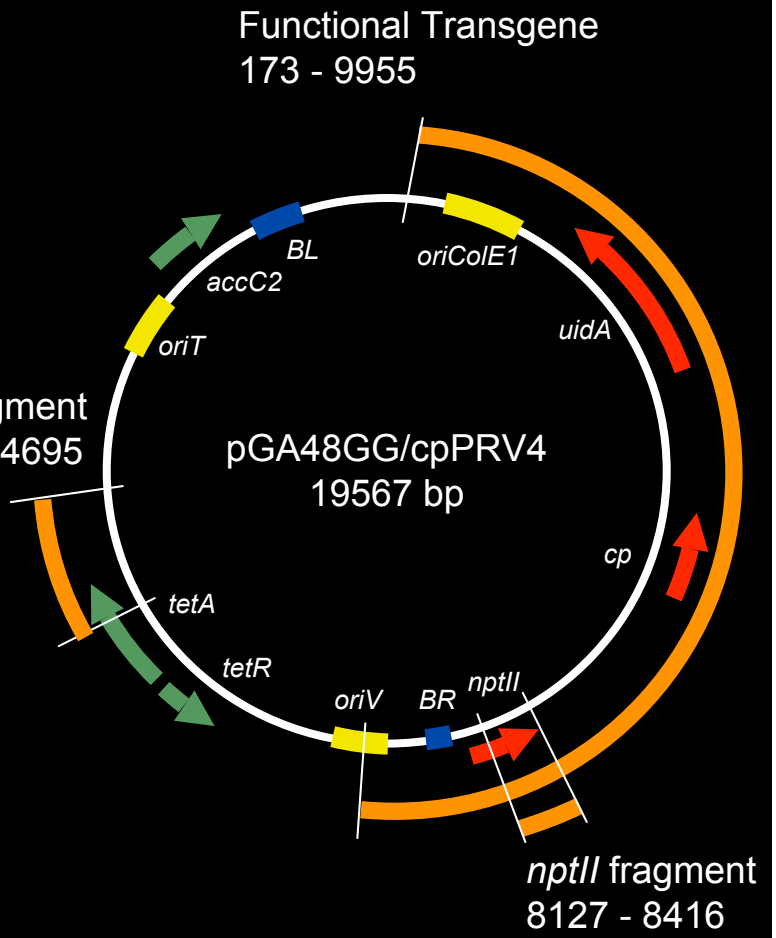
1,533 bp

11 total reads, 2 reads across the rearrangement.

3. ***nptII*** **fragment** (*ndhG*) (*atpB, atpE*)

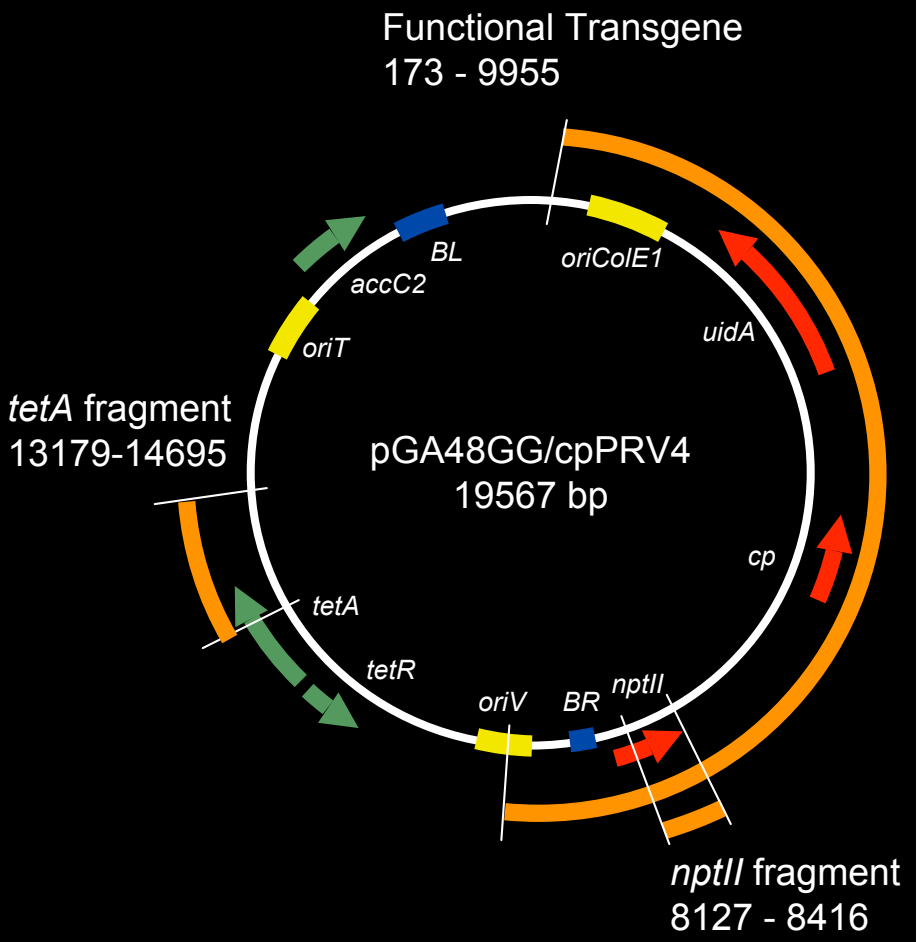
290 bp

Completely spanned by 4 reads



Transgene Insertion Sites

- Analysis also discovered 2 unconfirmed insertions of fragments of the *uidA* (GUS) and coat protein genes.
  - 2 singleton reads span between transgenic sequence and non-transgenic sequence
- Southern blot analysis had weak signal for a potential accC2 insertion
  - A more sensitive alignment found a 21bp exact match in 1 read
- Border sequences related to known transgene insertion features
  - 5/6 are nuclear DNA copies of the AT-rich papaya chloroplast
  - 4/6 junctions are Topo I recognition sites associated with transgene insertion sites.
- No papaya genes were disrupted.



Conclusions

- Draft genome of papaya confirms the presence of 1 functional and 2 non-functional transgenic inserts
  - Most well characterized commercialized transgenic crop
  - SunUp could serve as a transgenic source to breed suitable cultivars throughout the world
- WGS sequencing is an effective method for characterizing transgenic genomes.
- Look for draft Papaya genome in GenBank soon
  - Interesting implications for the ancestral angiosperm

Acknowledgements

- In collaboration with the Papaya Genome Project
  - Maqsudul Alam, University of Hawaii at Manoa
  - Ray Ming, University of Illinois at Urbana-Champaign
  - Dennis Gonsalves, USDA-ARS
  - Steven Salzberg, University of Maryland

- Funding support by
  - University of Hawaii
  - US DoD
  - Maui High Performance Computing Center
  - Hawaii Agriculture Research Center
  - Nankai University
  - USDA
  - University of Illinois
  - NSF Plant Genome Research Program
  - Tianjin Municipal Special Fund for Science and Technology

Genome Characteristics

	Carica papaya	*Arabidopsis thaliana*	*Populus trichocarpa*	*Oryza sativa (japonica)*	*Vitis vinifera*
Size (Mbp)	~325	125	485	389	487
Chromosomes	9	5	19	12	19
G+C content (%)	35.3	35.0	N/A	43.0	36.2
Gene number	23,151*	31,114[+]	45,555	37,544	30,434
Avereage gene length (bp)	2,373	2,232	2,300	2,821	3,399
Average intron length (bp)	501	165	379	412	213
Transposons (%)	51.9	14	42	34.8	41.4

* The number of genes was extrapolated to account for the unassembled regions of the genome.

[+] The gene number of *Arabidopsis* is based on the 27,873 protein coding and RNA genes from the TAIR website (http://www.arabidopsis.org/portals/genAnnotation/genome_snapshot.jsp) and recently published 3,241 novel genes.