Hunting Down the Papaya Transgenes

Michael Schatz

Center for Bioinformatics and Computational Biology
University of Maryland

January 16, 2008
PAG XVI
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 worldwide
  - 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.

<table>
<thead>
<tr>
<th>Year</th>
<th>Total (× 1,000 lbs)</th>
<th>Puna (× 1,000 lbs)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(virus in Puna) 1992</td>
<td>55,800</td>
<td>53,010</td>
<td>95</td>
</tr>
<tr>
<td>1993</td>
<td>58,200</td>
<td>55,290</td>
<td>95</td>
</tr>
<tr>
<td>1994</td>
<td>56,200</td>
<td>55,525</td>
<td>99</td>
</tr>
<tr>
<td>1995</td>
<td>41,900</td>
<td>39,215</td>
<td>94</td>
</tr>
<tr>
<td>1996</td>
<td>37,800</td>
<td>34,195</td>
<td>90</td>
</tr>
<tr>
<td>1997</td>
<td>35,700</td>
<td>27,810</td>
<td>78</td>
</tr>
<tr>
<td>(transgenic seeds released) 1998</td>
<td>35,600</td>
<td>26,750</td>
<td>75</td>
</tr>
<tr>
<td>1999</td>
<td>39,400</td>
<td>25,610</td>
<td>65</td>
</tr>
<tr>
<td>2000</td>
<td>50,250</td>
<td>33,950</td>
<td>68</td>
</tr>
<tr>
<td>2001</td>
<td>52,000</td>
<td>40,290</td>
<td>77</td>
</tr>
<tr>
<td>2002</td>
<td>42,700</td>
<td>35,880</td>
<td>84</td>
</tr>
</tbody>
</table>

\(^a\) Data were compiled from USDA Statistical Reports of Papaya grown in Hawaii (www.nass.usda.gov/hl).
Papaya Genome Project

- 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

http://cgpbr.hawaii.edu/papaya/
## Sequence Data

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

- 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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total span of scaffolds (Mb)</td>
<td>372</td>
</tr>
<tr>
<td>N50 of scaffolds (Mb)</td>
<td>1.0</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>18,650</td>
</tr>
<tr>
<td>Total length of contigs (Mb)</td>
<td>278</td>
</tr>
<tr>
<td>N50 of contigs (kb)</td>
<td>11</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>48,409</td>
</tr>
<tr>
<td>Total length of anchored scaffolds (Mb)</td>
<td>235</td>
</tr>
<tr>
<td>Total length of anchored and oriented scaffolds (Mb)</td>
<td>161</td>
</tr>
<tr>
<td>Number of anchored scaffolds</td>
<td>291</td>
</tr>
<tr>
<td>Total length of anchored contigs (Mb)</td>
<td>167</td>
</tr>
<tr>
<td>Total length of anchored and oriented contigs (Mb)</td>
<td>117</td>
</tr>
<tr>
<td>Number of anchored contigs</td>
<td>20,636</td>
</tr>
</tbody>
</table>
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 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. $(ycf2)$
   - 1,533 bp
   - 11 total reads, 2 reads across the rearrangement.

3. $(ndhG)$
   - 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

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

* 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.