Bioinformatics: Opportunities and Challenges for Data Recovery, Analysis and Sustainability

• The changing pace of biology in the genomics era
• The “Systems” explosion
• Role of the informatics specialist
• Challenges of data stability

Jim Giovannoni, Boyce Thompson Institute for Plant Research, USDA-ARS, Cornell, Ithaca NY
Ripening:
Softening
Flavor/Aroma
Nutrition
Chlorophyll Loss
Carotenoid Accumulation
Pathogen Susceptibility

Giovannoni *Plant Cell*, 2004
Normal and *ripening-inhibitor* (*rin*) nearly isogenic lines
Expression of a Chimeric Polygalacturonase Gene in Transgenic rin (Ripening Inhibitor) Tomato Fruit Results in Polyuronide Degradation but not Fruit Softening

James J. Giovannoni, a Dean DellaPenna, b Alan B. Bennett, b and Robert L. Fischer a, 1

a Division of Molecular Plant Biology, University of California, Berkeley, California 94720
b Mann Laboratory, Department of Vegetable Crops, University of California, Davis, California 95616

Tomato fruit ripening is accompanied by extensive degradation of pectic cell wall components. This is thought to be due to the action of a single enzyme, polygalacturonase, whose activity is controlled, at least in part, at the level of gene expression. At the onset of tomato fruit ripening, polygalacturonase enzyme activity, mRNA levels, and relative rate of gene transcription all increase dramatically. To elucidate the role of polygalacturonase during tomato fruit ripening, we utilized a pleiotropic genetic mutation, rin, that blocks many aspects of ripening, including the activation of polygalacturonase gene transcription. The polygalacturonase structural gene was ligated to a promoter that is inducible in mature rin fruit and inserted into the fruit genome, and plants were regenerated. This allowed expression of the polygalacturonase gene in transgenic rin fruit at a time corresponding to ripening in wild-type fruit. Expression of this gene resulted in the accumulation of active polygalacturonase enzyme and the degradation of cell wall polyuronides in transgenic rin fruit. However, no significant effect on fruit softening, ethylene evolution, or color development was detected. These results indicate that polygalacturonase is the primary determinant of cell wall polyuronide degradation, but suggest that this degradation is not sufficient for the induction of softening, elevated rates of ethylene biosynthesis, or lycopene accumulation in rin fruit.
8 figures

502 data points
- all presented
- many as primary data
Transcriptome and Selected Metabolite Analyses Reveal Multiple Points of Ethylene Control during Tomato Fruit Development

Rob Alba, a Paxton Payton, a,1 Zhanjun Fei, a,2 Ryan McQuinn, b Paul Debbie, a Gregory B. Martin, a,c Steven D. Tanksley, d and James J. Giovannoni a,b,3

a Boyce Thompson Institute for Plant Research, Cornell University Campus, Ithaca, New York, 14853
b U.S. Department of Agriculture, Agricultural Research Service, Plant, Soil, and Nutrition Laboratory, Ithaca, New York, 14853
c Department of Plant Pathology, Cornell University, Ithaca, New York, 14853
d Department of Plant Breeding, Cornell University, Ithaca, New York, 14853

Transcriptome profiling via cDNA microarray analysis identified 869 genes that are differentially expressed in developing tomato (Solanum lycopersicum) pericarp. Parallel phenotypic and targeted metabolite comparisons were employed to inform the expression analysis. Transcript accumulation in tomato fruit was observed to be extensively coordinated and often completely dependent on ethylene. Mutation of an ethylene receptor (Never-ripe [Nr]), which reduces ethylene sensitivity and inhibits ripening, alters the expression of 37% of these 869 genes. Nr also influences fruit morphology, seed number, ascorbate accumulation, carotenoid biosynthesis, ethylene evolution, and the expression of many genes during fruit maturation, indicating that ethylene governs multiple aspects of development both prior to and during fruit ripening in tomato. Of the 869 genes identified, 628 share homology (E-value ≤1 × 10^{-10}) with known gene products or known protein domains. Of these 628 loci, 72 share homology with previously described signal transduction or transcription factors, suggesting complex regulatory control. These results demonstrate multiple points of ethylene regulatory control during tomato fruit development and provide new insights into the molecular basis of ethylene-mediated ripening.
TOM1 cDNA Array

Total elements (spots): 13,440

Non-redundant tomato sequences: 8,700

Re-sequencing: >80% have been sequence verified from both 3’ and 5’ ends (France).

Access:

Arrays
www.bti.cornell.edu/CGEP/CGEP.html

Clones
http://ted.bti.cornell.edu/order/index

Sequences
http://www.sgn.cornell.edu

Data
http://ted.bti.cornell.edu/
Experimental Design for Transcriptome Analysis

Time Point: 1 2 3 4 5 6 7 8 9 10

Stage Name: 7DAP 17 DAP 27 DAP MG B-1 B B+1 B+5 B+10 B+15

Expression Ratio: \( \frac{17}{7} \) \( \frac{27}{17} \)
1,296,000 gene expression data points

Alba et al., *Plant Cell*, 2005
Supplemental Data

Files in this Data Supplement:

- **Supplemental Figure 1** - Ethylene regulates accumulation of reduced ascorbate in pericarp prior to tomato ripening. Ascorbate (ASC) was extracted as described in Methods and measurements were based on ΔA265 after enzymatic remove via ASC oxidase; mean values (N = 10) and SE bars are shown. Gray bars represent WT fruit and white bars represent Nr fruit. Abbreviations: A265, absorbance at 265 μm; DAP, days after pollination; FW, fresh weight; Nr, Never-ripe, SE, standard error; WT, wild-type.
- **Supplemental Figure 2** - Ethylene regulates accumulation of reduced ascorbate in pericarp prior to tomato ripening. Ascorbate (ASC) was extracted as described in Methods and measurements were based on ΔA265 after enzymatic remove via ASC oxidase; mean values (N = 10) and SE bars are shown. Gray bars represent WT fruit and white bars represent Nr fruit. Abbreviations: A265, absorbance at 265 μm; DAP, days after pollination; FW, fresh weight; Nr, Never-ripe, SE, standard error; WT, wild-type.

- Supplemental Table 1
- Supplemental Table 2
- Supplemental Table 3
Tomato Expression Database is part of the Tomato Genome Project organized through collaboration with Dr. Steve Tanksley, Dr. Jim Giovannoni, Dr. Greg Martin, and The Institute of Genome Research (TIGR) and funded by the National Science Foundation (NSF). It includes the tomato microarray data warehouse, tomato microarray expression data and tomato digital expression data. They were created and are maintained in the Giovannoni fruit ripening and nutritional genomics lab.

Tomato microarray data warehouse is for public storage/retrieval of raw microarray data resulting from use of the publicly available tomato microarray also resulting from this same project. Users of this microarray are encouraged to submit their resulting raw data to TMD to facilitate future assessment and analyses. TMD employs the MIAME protocol for microarray experiment characterization to facilitate maximal utility for the research community. Contact the site administrator to receive a login ID and password for data submission.

Tomato microarray expression data contains basic information and microarray data about the ESTs on our current cDNA microarrays. Currently TED contains ~12,000 ESTs representing sequences unique to specific tissues (e.g. fruit, flower), as well as specific genetic, biochemical and physiological functions (e.g. protein kinases, transcriptional factors). All these ESTs were re-sequenced from both ends. The re-sequenced data and sequence annotation information were included in TED and are also available through the Solanaceae Genomics Network (SGN). Currently TED contains normalized and processed microarray data for ten time points with nine pairwise comparisons during wild type and Nr mutant fruit development. TED will continue to be updated with additional expression results and additional ESTs as the size of our microarrays increases.

Tomato digital expression data presents expression analysis resulting from sequence prevalence characterization of over 150,000 tomato ESTs derived from 27 different non-normalized EST libraries. More information about tomato ESTs and tomato genomics in general can be found at the TIGR Tomato Gene Index and SGN websites.

Any questions or comments should be directed to the database manager.
Informatics Specialist:

- capabilities in computer science, statistical analysis and biology

- typically a biologist with training in computer science or a computer scientist with exposure to biology

Roles:

Experimental design
Data analysis (pushing the frontiers)
Database development and management
Digital expression analysis of grape and tomato ripening

<table>
<thead>
<tr>
<th>Tomato TC</th>
<th>Grape TC</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC125305/TC125359</td>
<td>TC4377</td>
<td>MADS box protein</td>
</tr>
<tr>
<td>TC124244/TC124112</td>
<td>TC4730</td>
<td>bZIP transcription factor</td>
</tr>
<tr>
<td>TC124196/TC125034</td>
<td>TC9044</td>
<td>zinc finger transcription factor</td>
</tr>
<tr>
<td>TC116030</td>
<td>TC4282</td>
<td>xyloglucan endo-1,4-beta-D-glucanase</td>
</tr>
<tr>
<td>TC123883/TC124274</td>
<td>TC4394</td>
<td>alcohol dehydrogenase</td>
</tr>
<tr>
<td>TC115998</td>
<td>TC4249</td>
<td>Pathogenesis-related protein</td>
</tr>
<tr>
<td>TC125239</td>
<td>TC4910</td>
<td>calcineurin B like protein</td>
</tr>
<tr>
<td>TC123982</td>
<td>TC9086</td>
<td>Calmodulin</td>
</tr>
<tr>
<td>TC116962/TC116318/TC116319</td>
<td>TC4046/TC4559/TC4193/TC4064/TC4181/TC4034</td>
<td>heat shock protein</td>
</tr>
<tr>
<td>TC124903/TC126297/TC126413</td>
<td>TC4348</td>
<td>heat shock protein</td>
</tr>
<tr>
<td>TC124001</td>
<td>TC9134/TC4581</td>
<td>heat shock protein</td>
</tr>
<tr>
<td>TC115895</td>
<td>TC4236/TC4822</td>
<td>ubiquitin</td>
</tr>
<tr>
<td>TC123771</td>
<td>TC4209</td>
<td>elongation factor 1-alpha</td>
</tr>
<tr>
<td>TC124929</td>
<td>TC9284</td>
<td>heavy metal ion transport protein</td>
</tr>
<tr>
<td>TC124731</td>
<td>TC9962</td>
<td>endoplasmatic reticulum retrieval protein</td>
</tr>
</tbody>
</table>
Challenge: Integration of large scale biological monitoring activities

EST Database  Expression  Proteomics  Metabolomics
   Profiling

GENES  →  mRNA  →  Protein  →  Phenotype

Modification  KNOWLEDGE  Informatics

Lycopene concentration (mean)

ppm FW

p < 0.05
WHY SOLANACEAE? Members of the family share a common genome that led to many diverse outcomes that benefit society.

- Tomato
- Eggplant
- Potato
- Petunia
- Chili pepper

Pachytene chromosomes
Microsynteny amongst solanaceous genomes

*Wang et al. unpublished results*

- **pepper**
  - 122 kb
  - 2700 Mb
  - 1.85x

- **potato**
  - 135 kb
  - 800 Mb
  - 1.42x

- **tomato**
  - 105 kb
  - 950 Mb
  - 1.00x

- **eggplant**
  - 106 kb
  - 956 Mb
  - 1.37x

- **confirmed by Arabidopsis and EST hits**
- **confirmed only Arabidopsis hits**
- **confirmed only by EST hits**
- **transposon elements**
- **genes without Arabidopsis and EST hits**
Phytoene synthase (knock-out) is deficient in lycopene and β-carotene.

Normal control (R/R; b/b) accumulates lycopene and β-carotene.

Lycopene-β-cyclase (over-expresser) (B/B) accumulates β-carotene at the expense of lycopene.
HPLC isolation of lycopene and β-carotene from IL ripe fruit

Retention time/min

A_450

β-c

parent IL 1-1 parent IL 6-2 parent IL 8-2 parent IL 4-2.1

low high

cDNA Array comparisons of IL and parental ripe fruit

cDNAs clustered based on expression patterns
Multiple lines may have similar phenotypic effects
i.e. increased B-carotene in comparison to parent

IL lines A, B, C

Parental line

Retention time/min

These lines may also have similar expression patterns with respect to a subset of cDNAs

Candidate cDNAs for carotenoid regulation will come from these intersections
Tomato Metabolite Database (TOMET) is part of the Tomato Nutrient Project organized through collaboration with Dr. Harry Klee and Dr. Jim Giovannoni and funded by the National Science Foundation (NSF).

We are using a set of Lycopersicon pennellii-derived introgression lines (ILs) that together cover the entire genome in the background of L. esculentum Var. M82. Our collection contains two parent lines (M82 and L. pennellii) and 76 ILs (the 50 original lines and 26 new ILs), each containing a single introgression from L. pennellii (LA 716) in the genetic background of the processing tomato variety M82. Our objective is to correlate changes in metabolite accumulation among introgression lines with changes in gene expression using tomato microarrays.

We are also beginning to collect metabolite data from L. hirsutum introgression lines. The metabolite data generated from these lines is now included in the database and ready for view. We plan to expand the function of TOMET to include similar expression profiling data for the L. hirsutum ILs as for L. pennellii. We are also planning to develop tools for cross and combined analysis of both sets of germplasm.

Any questions or comments are welcome and should be directed to the database manager.
Expression profile for 1-1-4.1.19.17 and metabolite profile for reduced ascorbate
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

The NCBI Whole Genome Association (WGA) resource provides researchers with access to genotype and associated phenotype information that will help elucidate the link between genes and disease. For more information, click here to see the WGA resource page and click here to read the press release.
Quick Start

- **GENOMES:** Browse sequenced genomes for Rice, Maize & Arabidopsis; Look for rice/maize synteny; Narrow your search with GrameneMart; Search for sequence alignment with BLAST; search by Gene Ontology.

- **PROTEINS:** Search by PFam or ProSite or Browse by Gene Ontology using GO Slim.

- **MAPS:** Browse genetic or physical maps for Rice, Maize, Wheat, Barley, Oats, Sorghum, and other grasses, or use the Comparative Map Viewer (CMap) to compare maps of different types and species.

- **MOLECULAR MARKERS:** Use the Simple Sequence Repeat Identification Tool (SSRIT); or search by marker type or species, including Rice (*Oryza sativa*), Maize, Sorghum and Others.

- **TRAITS:** Search the Genes or QTL database for important phenotype-related loci such as Rice Genes, Rice QTL, Maize QTL. Don’t forget to explore traits in Ontologies.

- **GENETIC DIVERSITY:** Search for SNP and SSR allelic variation on loci of rice, maize, and wheat germplasms.

- **BIOCHEMICAL PATHWAYS:** Search for ALL the rice pathways on starch metabolism or get an overview of the metabolic network. Compare rice and Arabidopsis pathway datasets.

- **LITERATURE:** Search the literature for your friends and topics of interest.

- **SUBMISSION:** Submit a Rice Gene or Ontology Term to Gramene.

Visit with us at
PAG 2007
View Previous Gramene Presentations

http://www.gramene.org/
http://www.arabidopsis.org/home.html
Plant Genomics

TIGR Plant Transcript Assemblies represent clustered, assemblies of all transcripts for ~140 plant species and can be accessed here.

The TIGR-NCSU Phytophthora infestans Mitochondrial Genome Haplotyping Database, sponsored by USDA, can be accessed here.

The Comprehensive Phytopathogen Genome Resource provides a centralized resource for accessing genomic data for plant pathogens including viral, bacterial, fungal, oomycete, and nematodes and can be accessed here.

The TIGR Wheat Genome Database
The TIGR Wheat Genome Database provides access to wheat genomic and EST sequences along with other bioinformatic analyses such as alignments to the rice genome.

The TIGR Arabidopsis thaliana Database provides access to genomic sequence data and annotation generated at TIGR and assemblies of Arabidopsis ESTs from world-wide sequencing projects.

The TIGR Rice Database provides links to the USDA-CSREES/NSF/DOE-funded rice genome project at TIGR and includes sequence data, annotation, and links to the Oryza sativa Gene Index.

Potato Functional Genomics Project provides links to the NSF-funded potato genome project at TIGR and includes sequence data, annotation, and links to the Solanum tuberosum Gene Index.

The TIGR Maize Database provides links to the NSF-funded Consortium for Maize Genomics project and includes sequence, assembly and annotation data and links to the Maize Gene Index.

TIGR Plant Repeat Databases is a collection of repetitive sequences for 12 plant genera and four plant families.

The TIGR Loblolly Pine Functional Genomics Project in collaboration with the Institute of Paper Science and Technology and funded by the National Science Foundation, can be accessed here.

The TIGR Medicago truncatula Database provides access to annotations generated at TIGR and Medicago ESTs and BAC sequences from world-wide sequencing projects, here.

www.tigr.org
Systems approaches offer enormous opportunities to understand biology
- medicine
- food security and sustainability
- environmental protection
- renewable energy resources
- basis of the future “bio-based economy”

Informatics, informatics specialists and data managers and curators hold the key to the kingdom for the realization of these promises.

Challenges
- long-term data maintenance and availability
- standard methods of data presentation
- standard tools for basic data analysis
- uniform standards of data quality