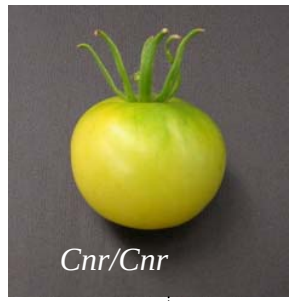
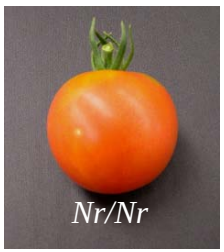
Bioinformatics: Opportunities and Challenges for Data Recovery, Analysis and Sustainability

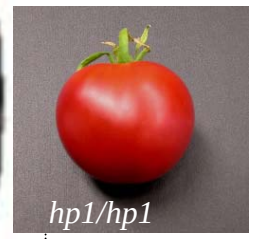
- The changing pace of biology in the genomics era
- The “Systems” explosion
- Role of the informatics specialist
- Challenges of data stability

Jim Giovannoni, Boyce Thompson Institute for Plant Research, USDA-ARS, Cornell, Ithaca NY

Giovannoni *Plant Cell*, 2004







NOR

LeMADS-RIN

LeMADS(?)

CNR

Developmental Regulation

LeACS1A

LeACS4

LeACO1

LeACO3

Autocatalytic synthesis

Ethylene

LeACS2

NR, PG, E8, LeEXP1

Signal Transduction

NR, LeETR4

LeCTR1

LeEINs

Gr

Pathway amplification

LeACS6

PME, LBC

HP1, HP2 (LeDET1), L1,L2

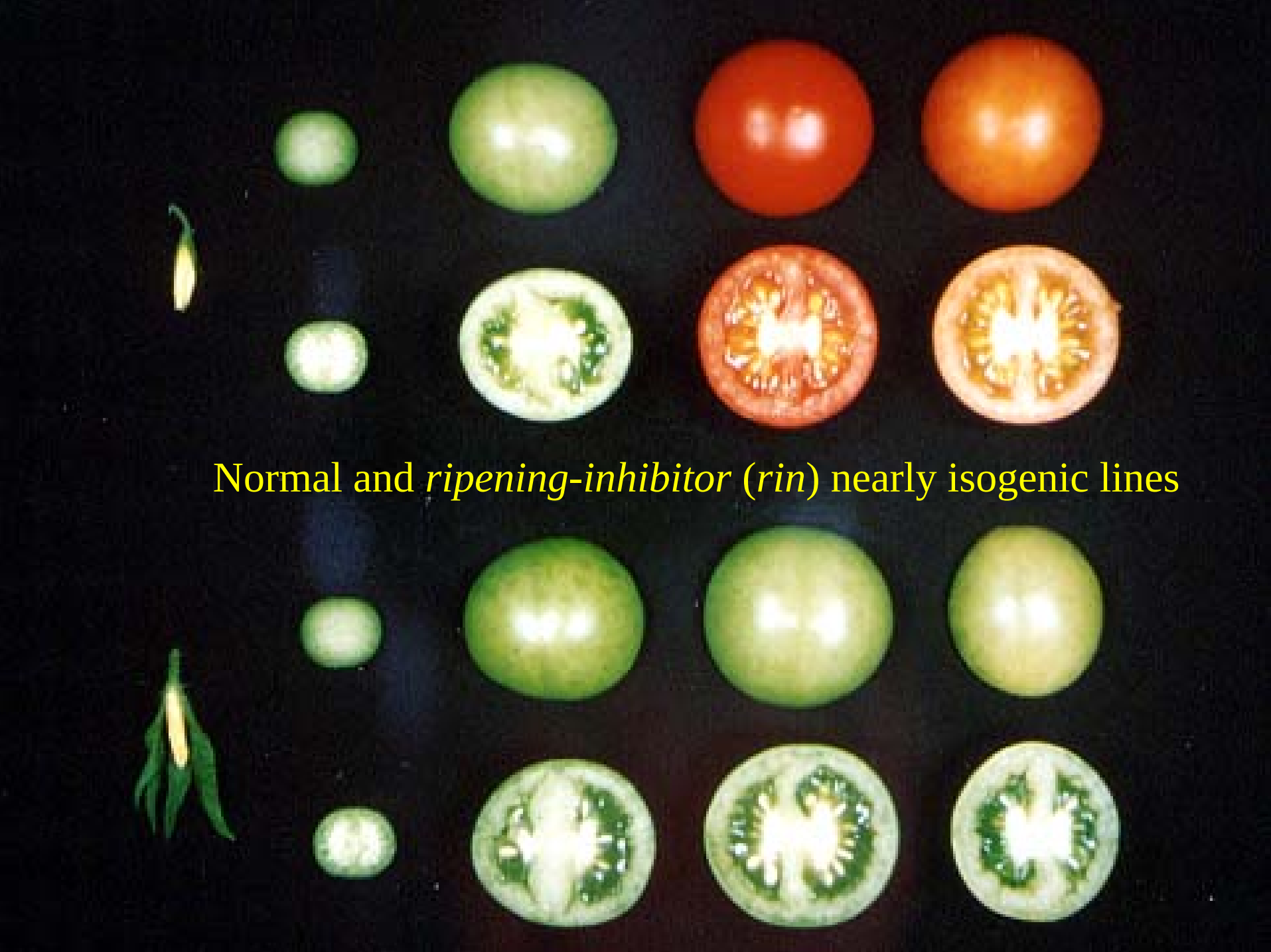
PHYs

Light

PSY, E4, TBG4

(+)

Ripening:
Softening
Flavor/Aroma
Nutrition
Chlorophyll Loss
Carotenoid Accumulation
Pathogen Susceptibility


Normal and *ripening-inhibitor* (*rin*) nearly isogenic lines

The Plant Cell, Vol. 1, 53–63, January 1989, © 1989 American Society of Plant Physiologists

Expression of a Chimeric Polygalacturonase Gene in Transgenic *rin* (Ripening Inhibitor) Tomato Fruit Results in Polyuronide Degradation but not Fruit Softening

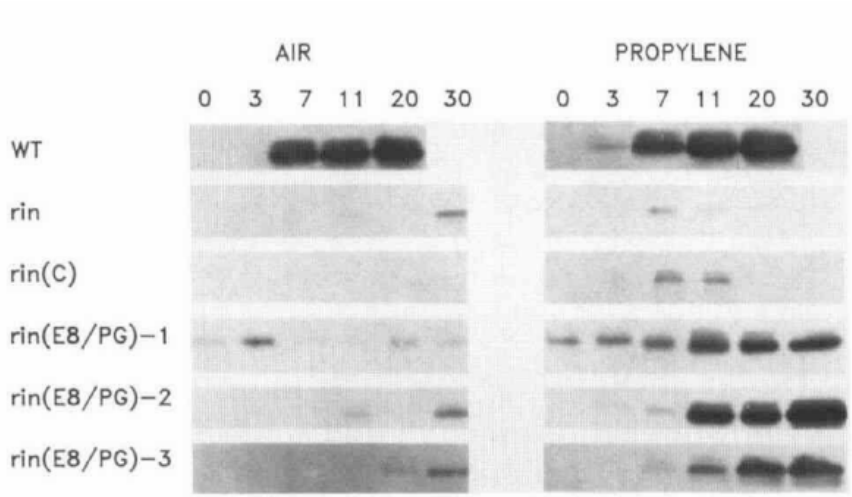
James J. Giovannoni,[a] Dean DellaPenna,[b] Alan B. Bennett,[b] and Robert L. Fischer[a,1]

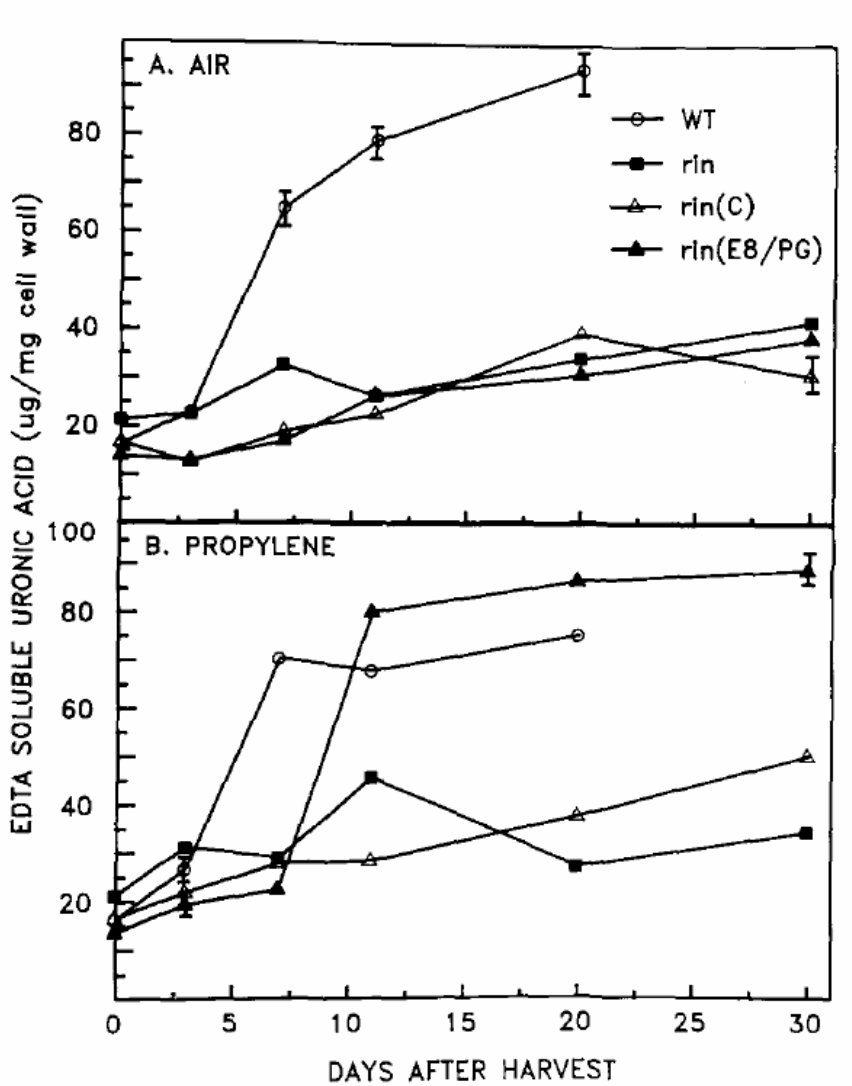
[a]Division of Molecular Plant Biology, University of California, Berkeley, California 94720
[b] Mann Laboratory, Department of Vegetable Crops, University of California, Davis, California 95616

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





The Plant Cell, Vol. 17, 2954–2965, November 2005, www.plantcell.org

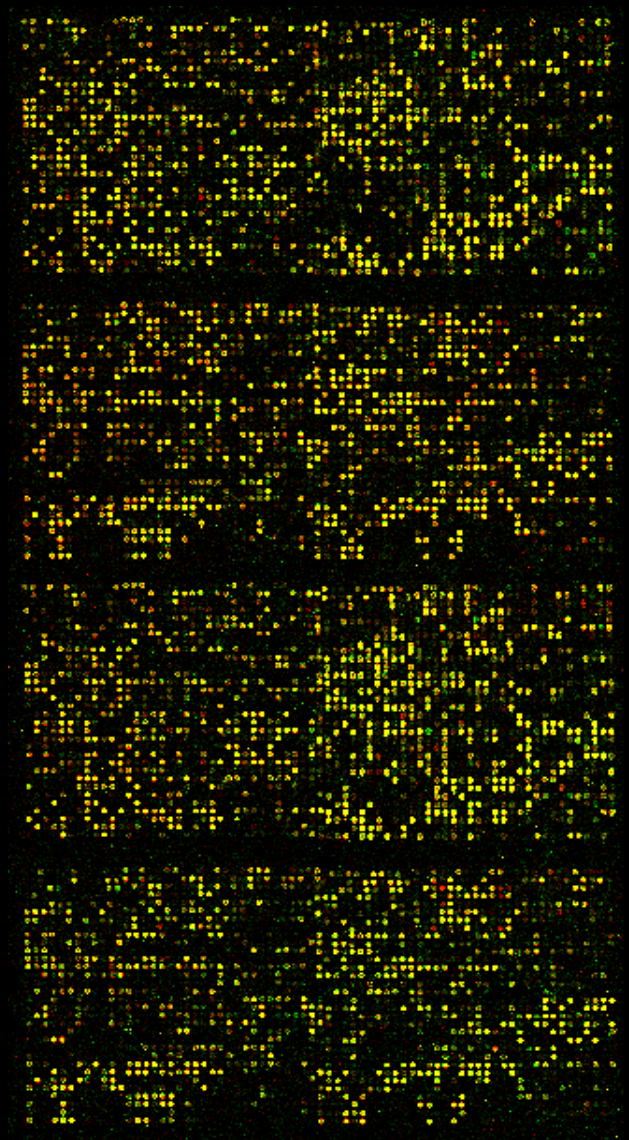
Transcriptome and Selected Metabolite Analyses Reveal Multiple Points of Ethylene Control during Tomato Fruit Development[W]

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 $\leq 1 \times 10^{-10}$) 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:

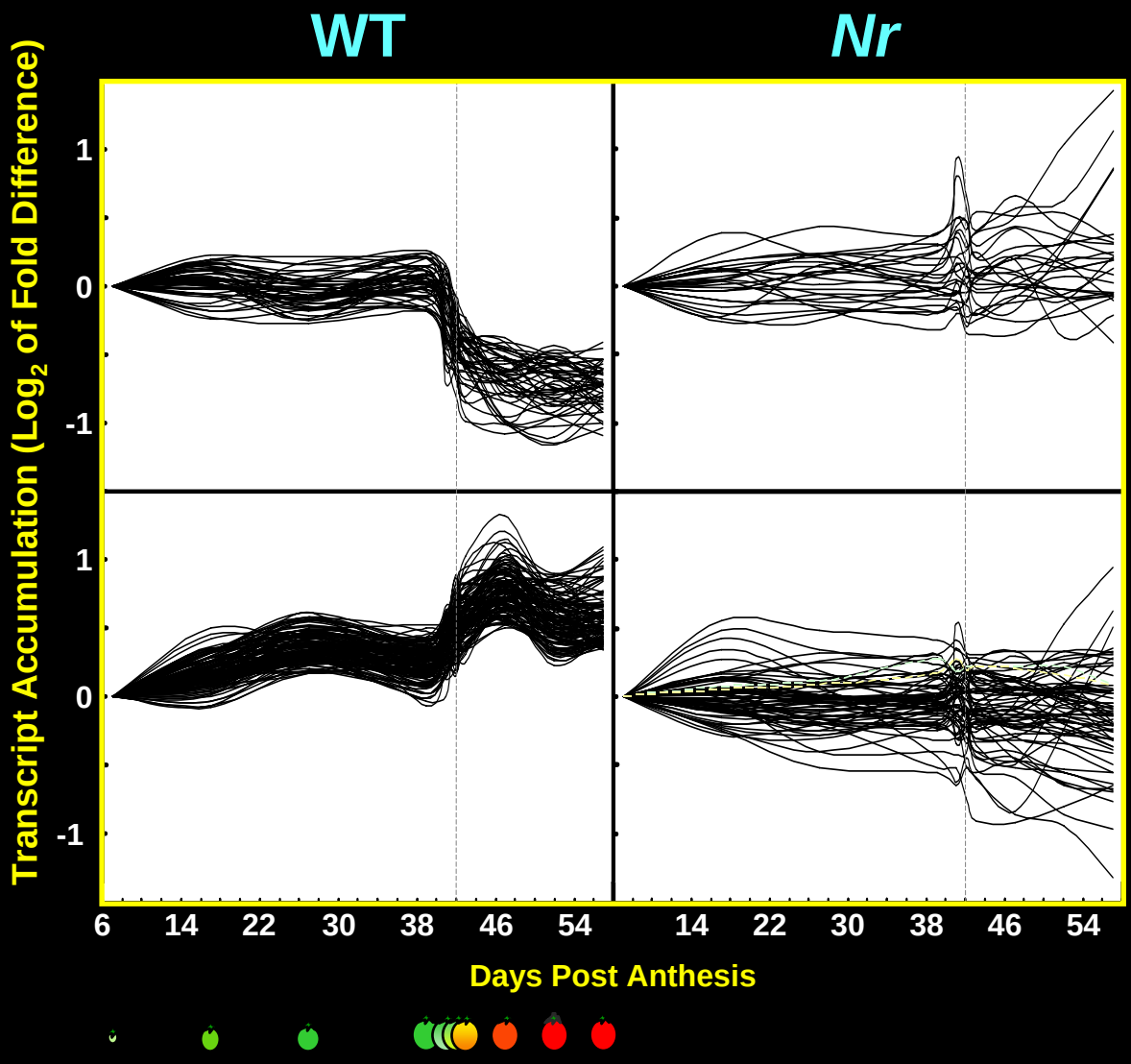
Arrays
www.bti.cornell.edu/CGEP/CGEP.html
Clones
http://ted.bti.cornell.edu/order/index
Sequences
http://www.sgn.cornell.edu
Data
http://ted.bti.cornell.edu/

Experimental Design for Transcriptome Analysis

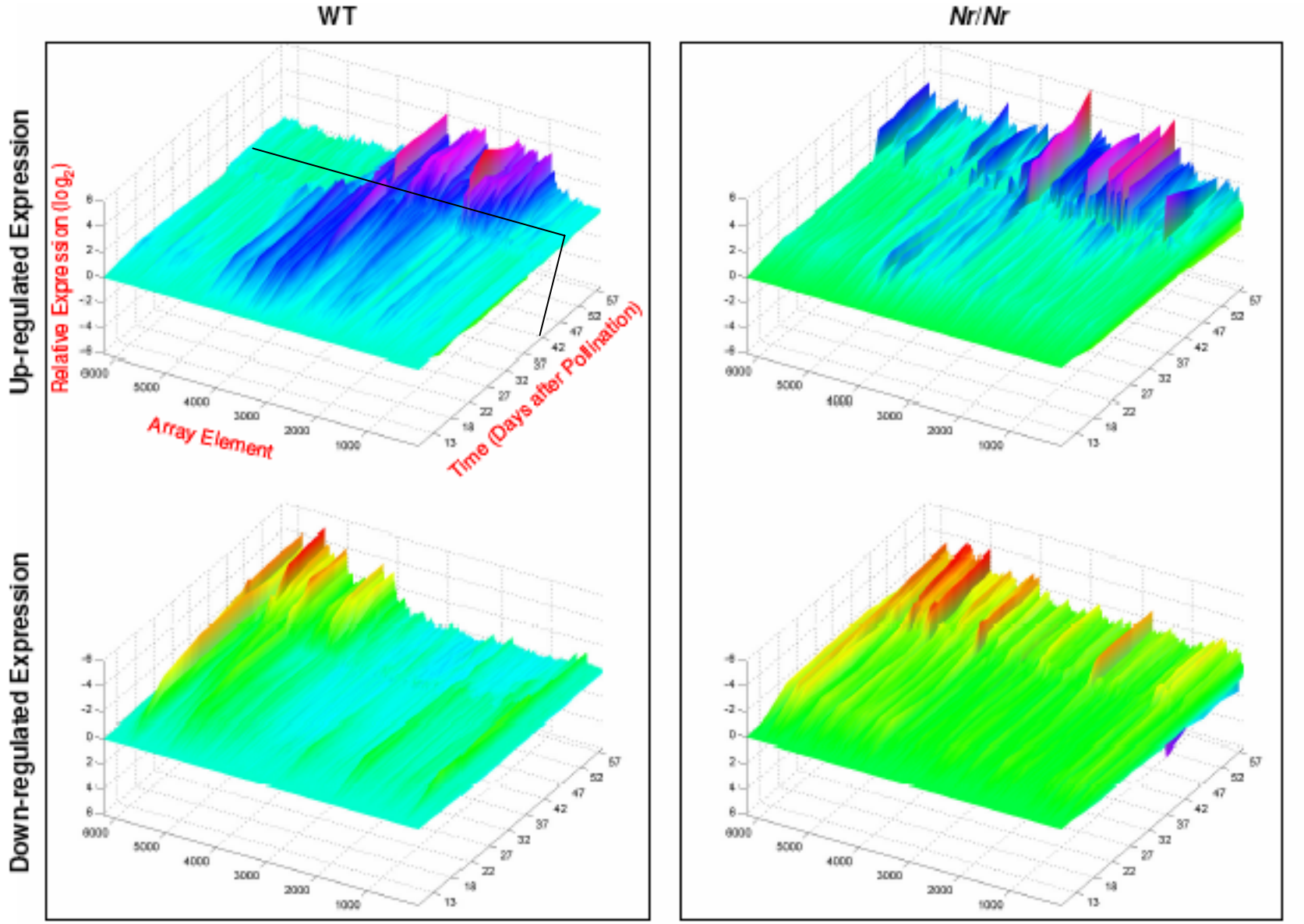
Time Point: 1 2 3 4 5 6 7 8 9 10

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

Expression Ratio: $\frac{17}{7}$ $\frac{27}{17}$


WT
Nr
Transcript Accumulation (Log2 of Fold Difference)
1
0
-1
1
0
-1
6
14
22
30
38
46
54
14
22
30
38
46
54
Days Post Anthesis

1,296,000 gene expression data points



Alba et al., *Plant Cell*, 2005

Transcriptome and Selected Metabolite Analyses Reveal Multiple Points of Ethylene Control during Tomato Fruit Development
Plant Cell Alba et al. 17: 2954

Supplemental Data

Files in this Data Supplement:

- Supplemental Figure 1 - **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 ΔA_{265} 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: A_{265}, absorbance at 265 µm; DAP, days after pollination; FW, fresh weight; *Nr, Never-ripe,* SE, standard error; WT, wild-type.
- Supplemental Figure 2 - **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 ΔA_{265} 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: A_{265}, 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

This Article
- Abstract
- Full Text

Services
- Alert me to new issues of the journal

www.ted.bti.cornell.edu

TOMATO EXPRESSION DATABASE

microarray data warehouse

microarray expression data

digital expression data

TOMET and other tomato links

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 pair-wise 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

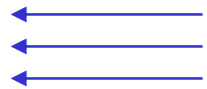
Data analysis (pushing the frontiers)

Database development and management

Digital expression analysis of grape and tomato ripening

Table IV: Genes induced both by tomato ripening and grape ripening

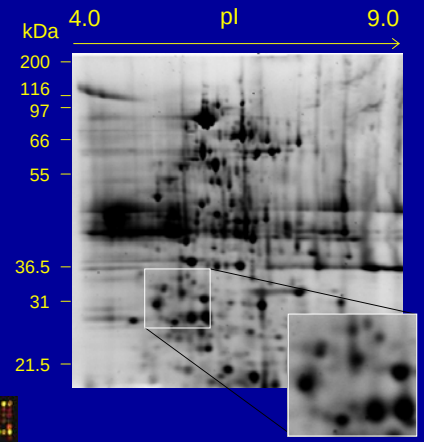
Tomato TC	Grape TC	Annotation
TC125305/TC125359	TC4377	MADS box protein
TC124244/TC124112	TC4730	bZIP transcription factor
TC124196/TC125034	TC9044	zinc finger transcription factor
TC116030	TC4282	xyloglucan endo-1,4-beta-D-glucanase
TC123883/TC124274	TC4394	alcohol dehydrogenase
TC115998	TC4249	Pathogenesis-related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

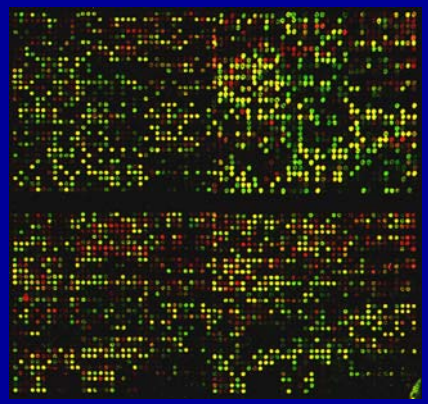


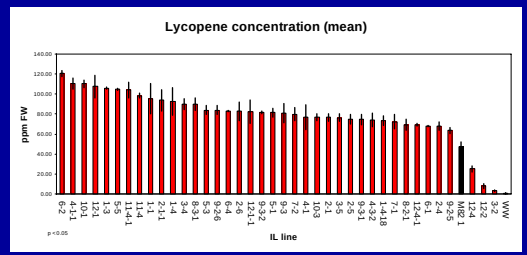
Candidates for RNAi

Challenge: Integration of large scale biological monitoring activities

EST Database ..... Expression Profiling ..... Proteomics ..... Metabalomics







GENES → mRNA → Protein → Phenotype

Modification ◄····· KNOWLEDGE ◄····· Informatics

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

Tomato

Eggplant

Potato

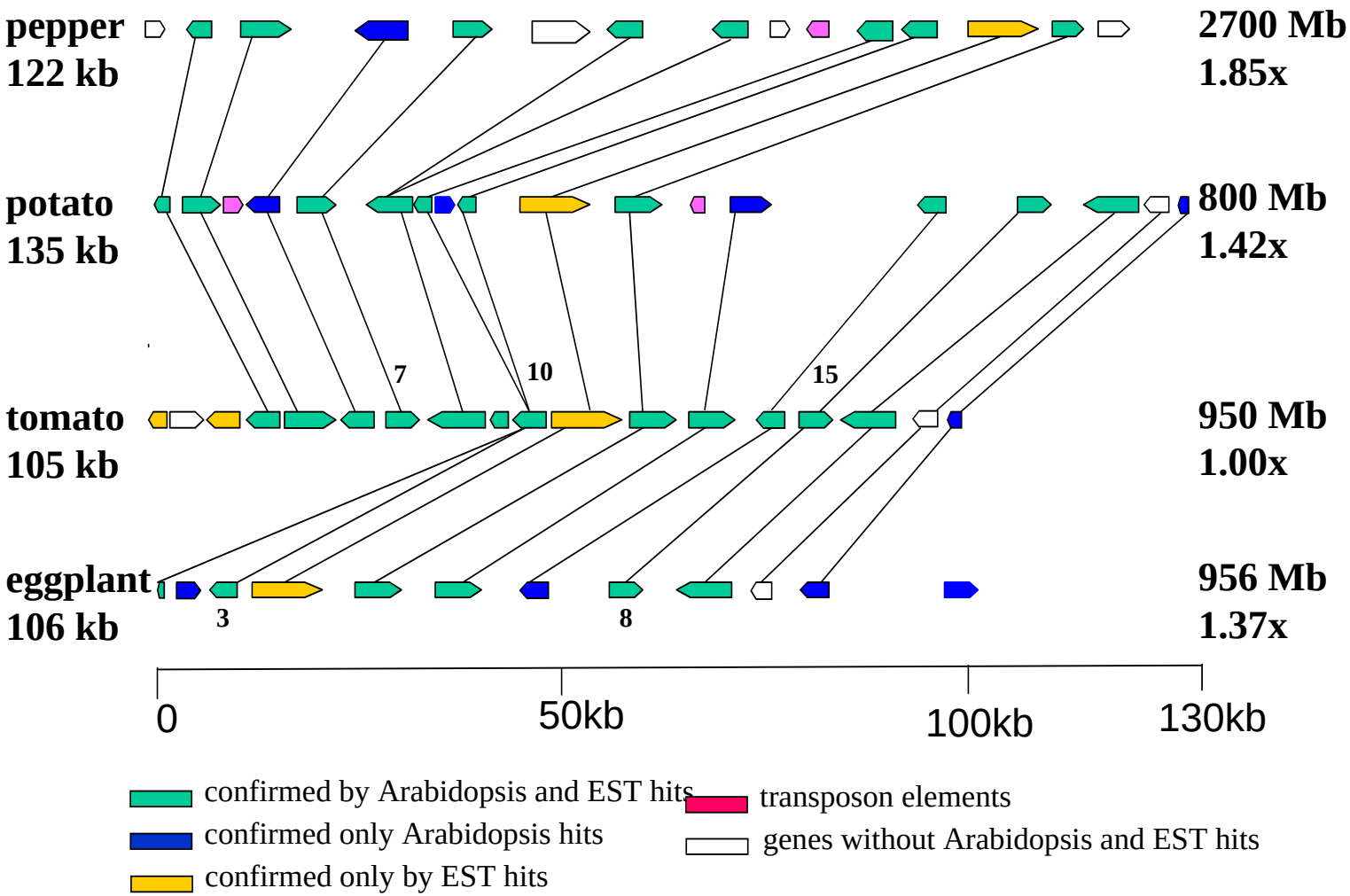
Chili pepper

Pachytene chromosomes

Petunia

Microsynteny amongst solanaceous genomes

Wang et al unpublished results



r/r
Phytoene synthase
(knock-out)

deficient in lycopene
and ß-carotene

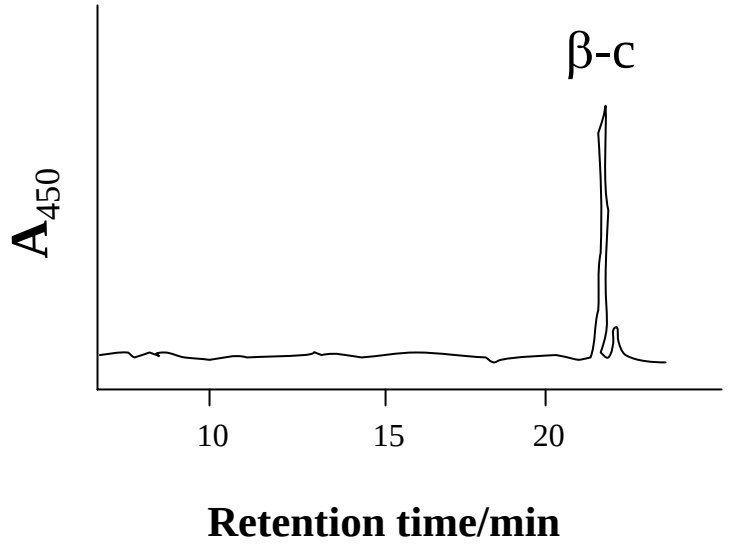
R/R; b/b
normal control

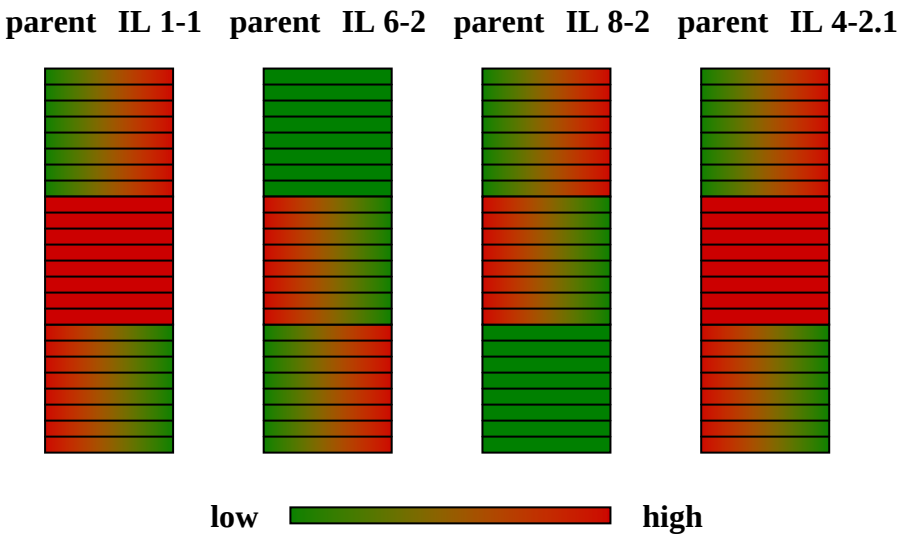
accumulates lycopene
and ß-carotene

B/B
lycopene-ß-cyclase
(over-expresser)

accumulates ß-carotene
at the expense of lycopene

HPLC isolation of lycopene and β-carotene from IL ripe fruit





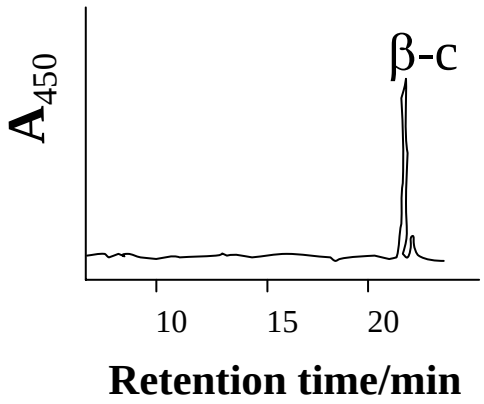
cDNA Array comparisons of IL and parental ripe fruit

cDNAs clustered based on expression patterns

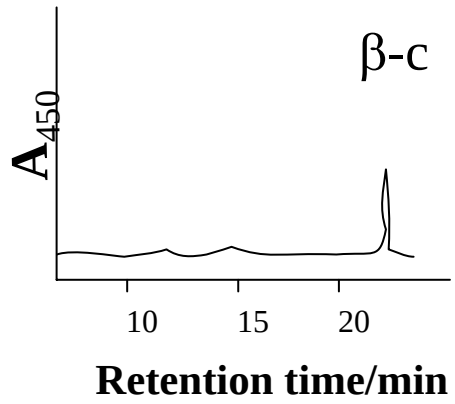
Multiple lines may have similar phenotypic effects

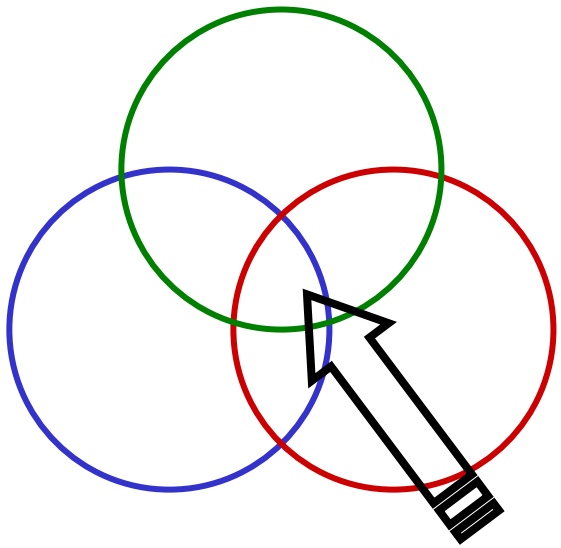
i.e increased B-carotene in comparison to parent

IL lines
A,B,C



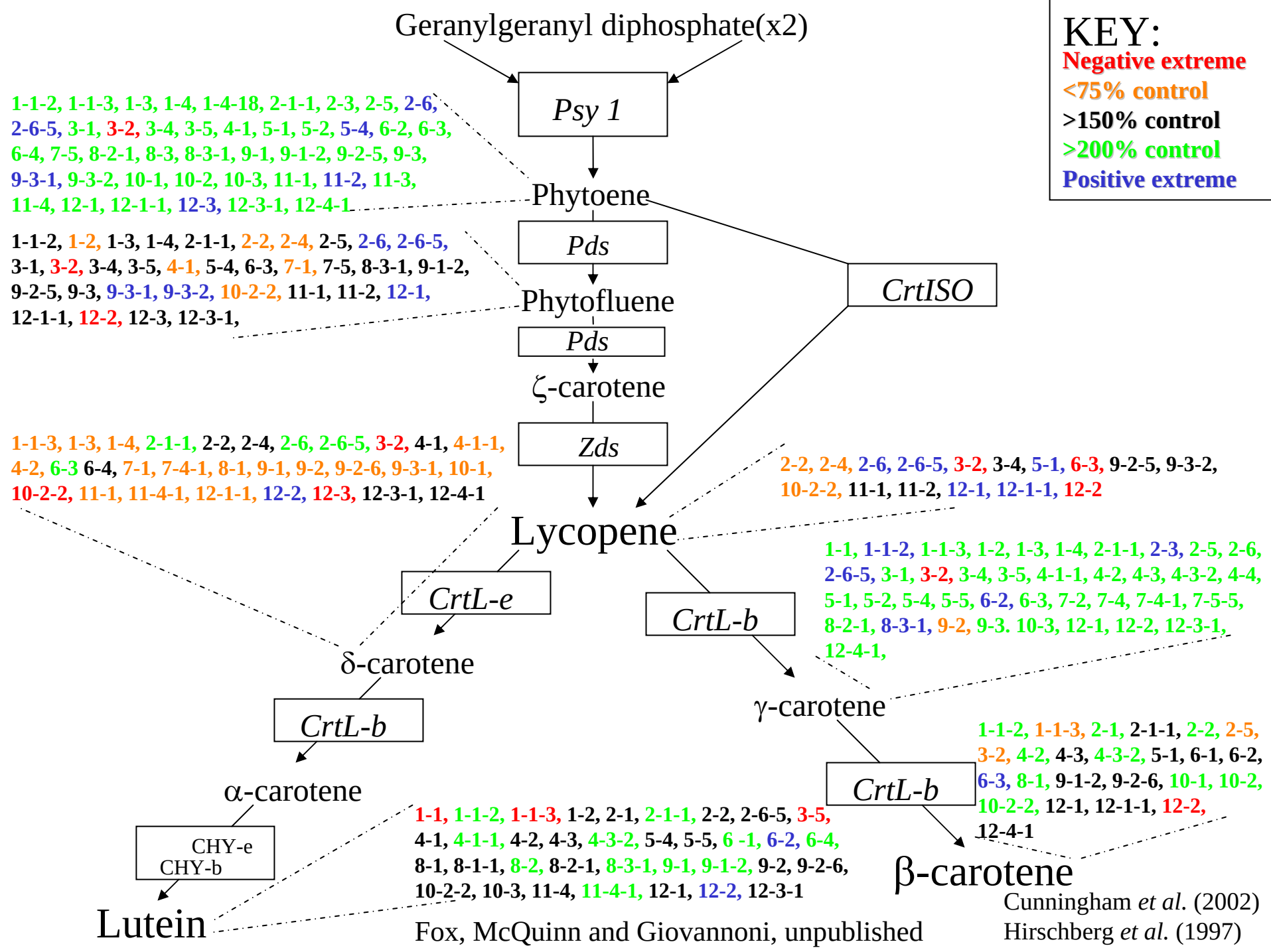
Parental line





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


KEY:
Negative extreme
<75% control
>150% control
>200% control
Positive extreme
Geranylgeranyl diphosphate(x2)
Psy 1
1-1-2, 1-1-3, 1-3, 1-4, 1-4-18, 2-1-1, 2-3, 2-5, 2-6, 2-6-5, 3-1, 3-2, 3-4, 3-5, 4-1, 5-1, 5-2, 5-4, 6-2, 6-3, 6-4, 7-5, 8-2-1, 8-3, 8-3-1, 9-1, 9-1-2, 9-2-5, 9-3, 9-3-1, 9-3-2, 10-1, 10-2, 10-3, 11-1, 11-2, 11-3, 11-4, 12-1, 12-1-1, 12-3, 12-3-1, 12-4-1
Phytoene
Pds
1-1-2, 1-2, 1-3, 1-4, 2-1-1, 2-2, 2-4, 2-5, 2-6, 2-6-5, 3-1, 3-2, 3-4, 3-5, 4-1, 5-4, 6-3, 7-1, 7-5, 8-3-1, 9-1-2, 9-2-5, 9-3, 9-3-1, 9-3-2, 10-2-2, 11-1, 11-2, 12-1, 12-1-1, 12-2, 12-3, 12-3-1,
Phytofluene
CrtISO
Pds
ζ-carotene
Zds
1-1-3, 1-3, 1-4, 2-1-1, 2-2, 2-4, 2-6, 2-6-5, 3-2, 4-1, 4-1-1, 4-2, 6-3 6-4, 7-1, 7-4-1, 8-1, 9-1, 9-2, 9-2-6, 9-3-1, 10-1, 10-2-2, 11-1, 11-4-1, 12-1-1, 12-2, 12-3, 12-3-1, 12-4-1
2-2, 2-4, 2-6, 2-6-5, 3-2, 3-4, 5-1, 6-3, 9-2-5, 9-3-2, 10-2-2, 11-1, 11-2, 12-1, 12-1-1, 12-2
Lycopene
CrtL-e
CrtL-b
1-1, 1-1-2, 1-1-3, 1-2, 1-3, 1-4, 2-1-1, 2-3, 2-5, 2-6, 2-6-5, 3-1, 3-2, 3-4, 3-5, 4-1-1, 4-2, 4-3, 4-3-2, 4-4, 5-1, 5-2, 5-4, 5-5, 6-2, 6-3, 7-2, 7-4, 7-4-1, 7-5-5, 8-2-1, 8-3-1, 9-2, 9-3. 10-3, 12-1, 12-2, 12-3-1, 12-4-1,
δ-carotene
γ-carotene
CrtL-b
CrtL-b
α-carotene
1-1-2, 1-1-3, 2-1, 2-1-1, 2-2, 2-5, 3-2, 4-2, 4-3, 4-3-2, 5-1, 6-1, 6-2, 6-3, 8-1, 9-1-2, 9-2-6, 10-1, 10-2, 10-2-2, 12-1, 12-1-1, 12-2, 12-4-1
CHY-e
CHY-b
1-1, 1-1-2, 1-1-3, 1-2, 2-1, 2-1-1, 2-2, 2-6-5, 3-5, 4-1, 4-1-1, 4-2, 4-3, 4-3-2, 5-4, 5-5, 6 -1, 6-2, 6-4, 8-1, 8-1-1, 8-2, 8-2-1, 8-3-1, 9-1, 9-1-2, 9-2, 9-2-6, 10-2-2, 10-3, 11-4, 11-4-1, 12-1, 12-2, 12-3-1
β-carotene
Lutein
Fox, McQuinn and Giovannoni, unpublished
Cunningham et al. (2002)
Hirschberg et al. (1997)

Tomato Metabolite Database

http://tomet.bti.cornell.edu/

GatorMail Google My Yahoo! Web of Knowledge Harry's Tomato Page News Klee IE Favorites TPJ

Tomato Metabolite Database

Home | Metabolite | IL Lines | Tools | Protocol Reference

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.

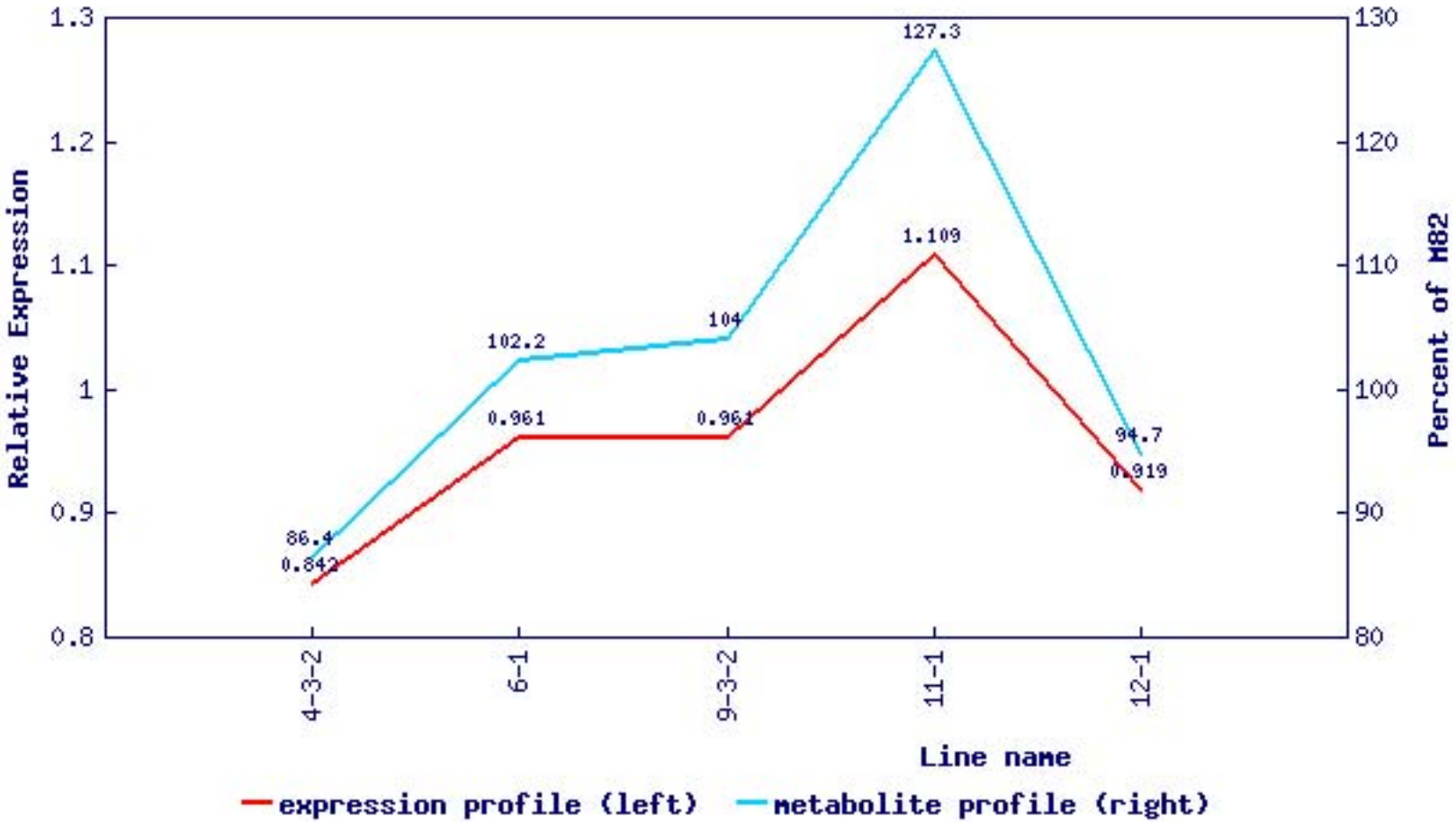
We are also beginning to collect metabolite data from L. hirsutum introgression lines. The metabolite data generated from these lines is now included in the database and ready for view. We plan to expand the function of TOMET to include similar expression profiling data for the L. hirsutum ILs as for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

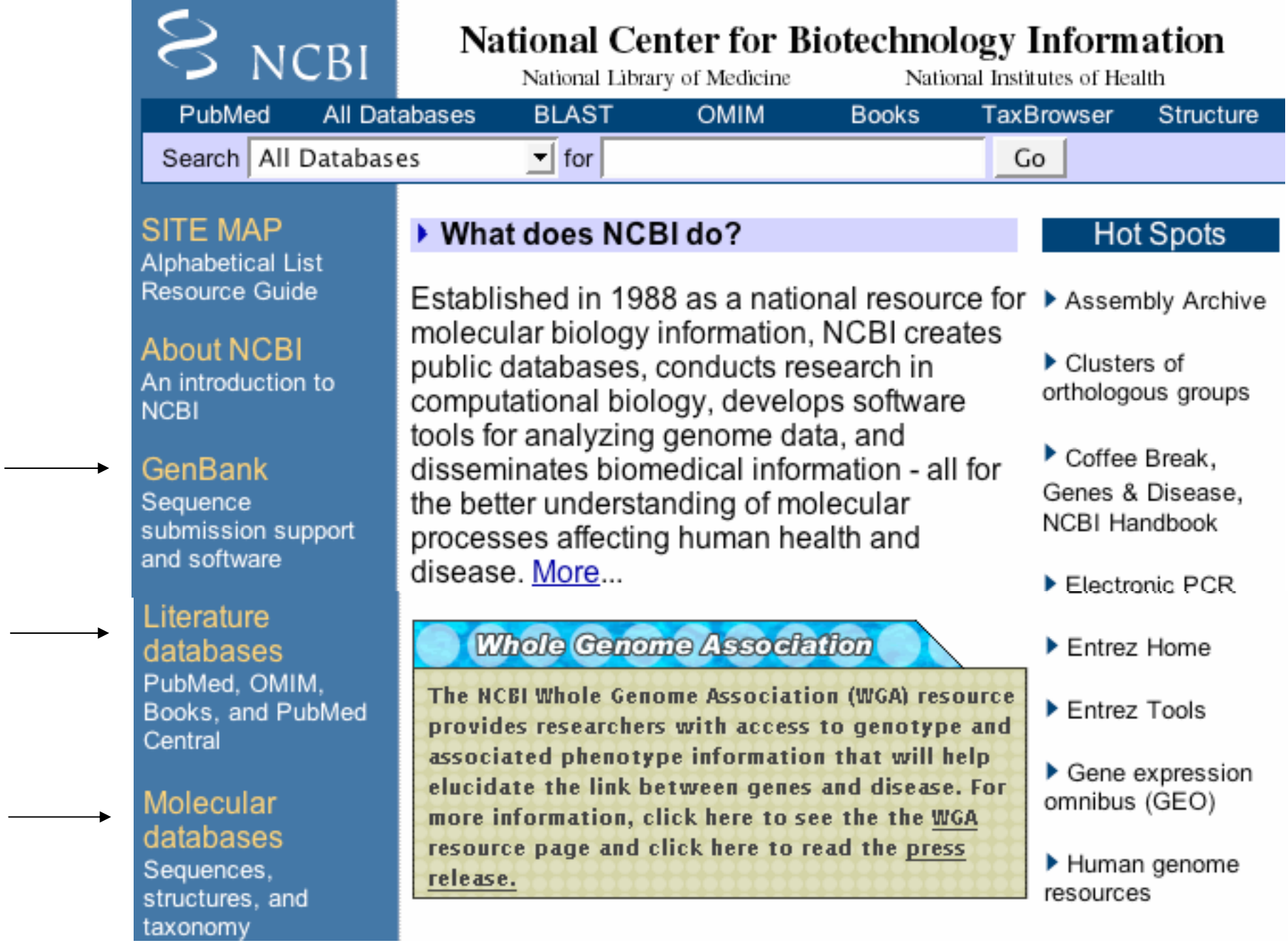
Any questions or comments are welcome and should be directed to the database manager.

Tomato Metabolite Database

Home | Metabolite | IL Lines | Tools | Protocol Reference

Expression profile for 1-1-4.1.19.17 and metabolite profile for reduced ascorbate





http://www.ncbi.nlm.nih.gov/

Home | Login | Forum | Contact | Help

Quick search

sol genomics network

search | maps | sequencing | tools

Getting started

SGN info

- SGN data overview
- More about SGN
- SOL project
- SOL newsletter
- International tomato project
- The ECO-SOL Project

SOL species

Solanaceae

- Tomato
- Pepper
- Potato
- Eggplant
- Petunia
- Solanum nomenclature

Rubiaceae

- Coffee

Tomato genome

- Sequencing progress
- Search BACS
- Overgo plating process Genome browser

Maps and markers

- Available maps
- Search markers
- About COS markers
- About SSR markers
- About COSII markers
- COSII marker spreadsheet

What is SGN?

The SOL Genomics Network contains genomic, genetic and taxonomic information for species in the Euasterid clade, including the families Solanaceae (e.g. tomato, potato, eggplant, pepper, petunia) and Rubiaceae (coffee). Genomic information is presented in a comparative format and tied to the fully-sequenced Arabidopsis genome.

What are Solanaceae?

Why are the Solanaceae being studied?

Who is sequencing the tomato genome?

Calibrachoa parviflora
Courtesy Sandra Knapp

New publications

Characterization of Capsaicin synthase and identification of its gene (*csy1*) for pungency factor capsaicin in pepper (*Capsicum* sp.)

B. C. Narasimha Prasad, Vinod Kumar, H. B. Gururaj, R. Parimalan, P. Giridhar, G. A. Ravishankar

See all publications...

Recent forum topics

Events

Plant GEM 2006

The Plant Genomics European Meeting, Oct 11-14, 2006, Venice, Italy

SOL Genome Workshop 2007

Sept 9-13, 2007
Ramada Hotel Jeju Island, Korea

See all events...

News

EU-SOL website

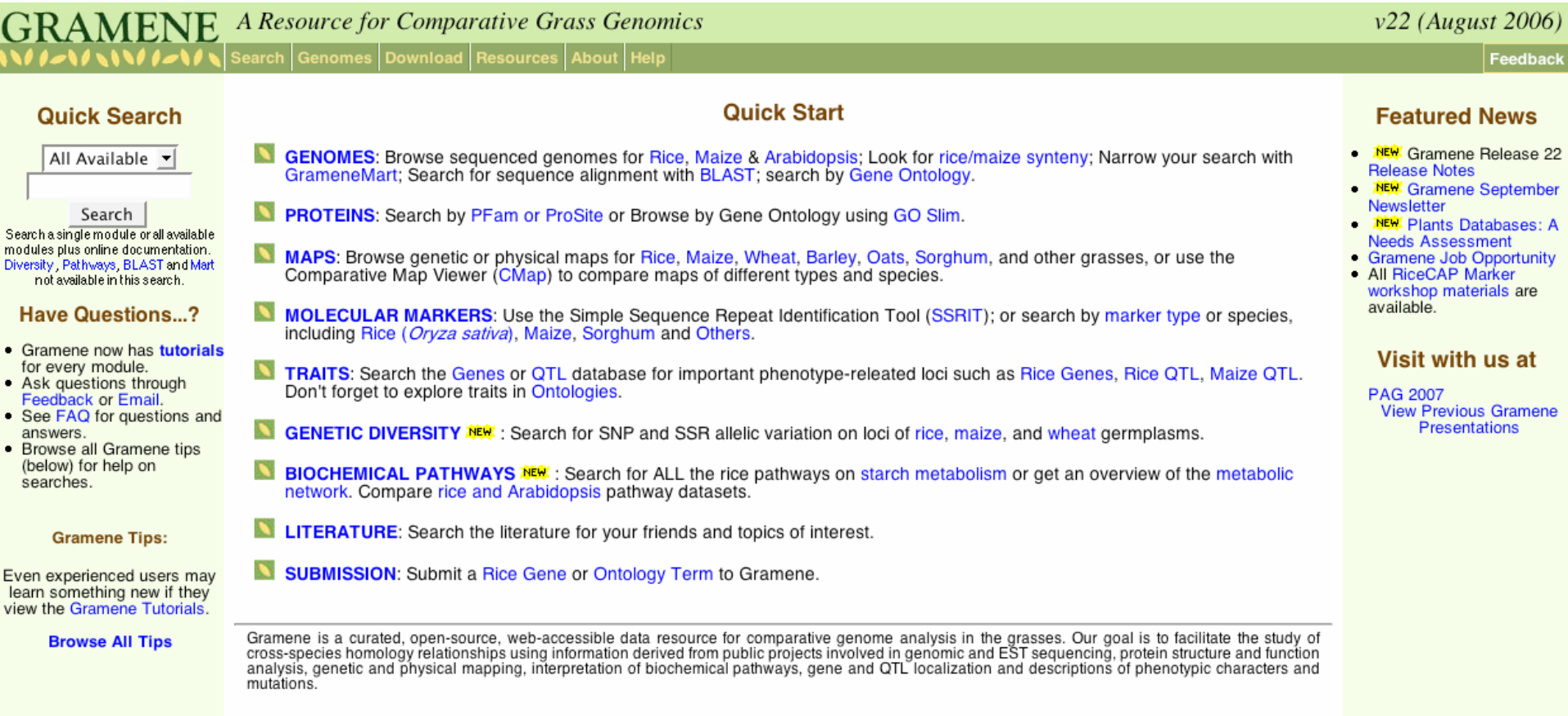
The website for the European SOL project (EU-SOL) is now on-line [Sept 19, 2006].

Sept 2006 SOL Newsletter

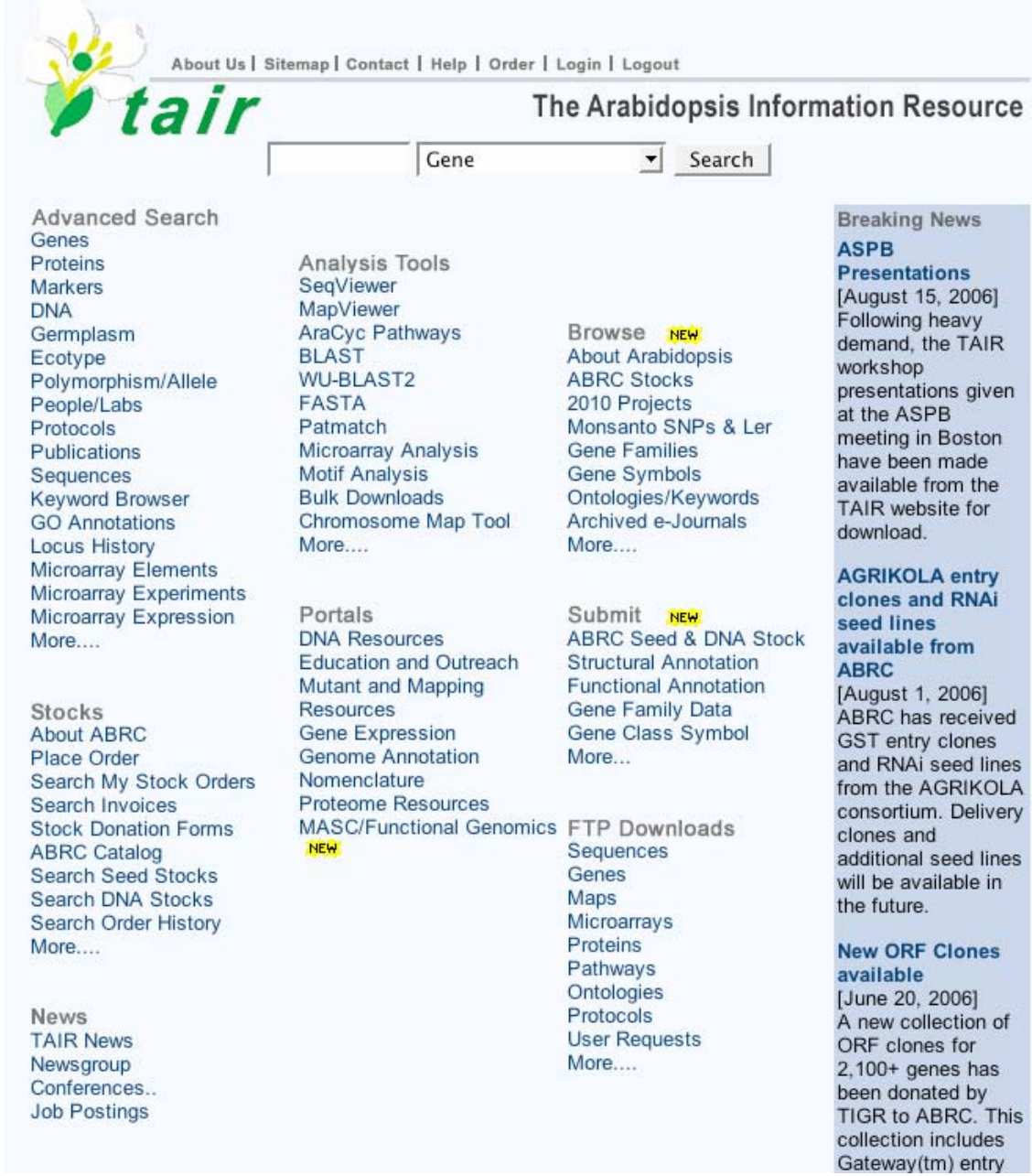
The SOL Newsletter for Sept 2006 is available [Sept 13, 2006].

New BAC extension tool

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



http://www.gramene.org/



http://www.arabidopsis.org/home.html

TIGR

The Institute for Genomic Research

Home | Research Groups | Scientific Programs | Faculty Home

- Database Home
- Comprenhensive Microbial Resources
- Unfinished Microbial Genomes
- Plant Genomics
- Parasites Databases
- Other Eukaryotic Projects
- Gene Indices
- TIGRFAMs
- Fungal Databases
- Fibrolytic Ruminal Bacteria
- TIGR Microbial Observatories
- Genome Properties Database

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.

Challeneges

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