Of tomatoes, genomes and databases

Lukas Mueller
Boyce Thompson Institute for Plant Research

Tomato Disease Workshop 2011
Ithaca, NY
Solanaceae phenotypes
Tomato Genome

- BAC by BAC approach 2004-2009 (~1500 BACs)
- Whole genome shotgun of entire tomato genome, started in 2009
- Technologies used: 454, Solexa (Syngenta), SOLiD
- 454 data assembled using Newbler
- Homopolymer correction using Solexa data
- Integration of BAC sequences
- Ordering and orientation of scaffolds based on genetic maps
- Version 2.4 has most sequence in <100 scaffolds
- Annotated by the ITAG consortium
- Paper submitted!
Note the data access agreement!
International Tomato Genome Sequencing Project

The International Tomato Genome Sequencing Project was begun in 2004 by an international consortium including participants from Korea, China, the United Kingdom, India, the Netherlands, France, Japan, Spain, Italy and the United States. The initial approach was to sequence only the euchromatic sequence using a BAC-by-BAC approach, and in total more than 1,200 BACs have been sequenced. In 2009, a complementary whole-genome shotgun approach was initiated, which in conjunction with other data yielded high quality assemblies. The International Tomato Annotation Group (ITAG) annotates the genome builds generated by this combined sequencing approach.

Please note that all data is released under the data access agreement below.

Data access agreement

PLEASE READ BEFORE ACCESSING THE PRE-PUBLICATION TOMATO GENOME SEQUENCE OR ANNOTATIONS: The International Tomato Genome Sequencing Consortium is pleased to make available a pre-publication draft assembly of the tomato genome for use by public and private research communities as a resource to enable plant biology discovery and improve the human condition through improved agriculture. This assembly was produced by the Dutch/French assembly team and includes both 454 data and Sanger sequence data (BAC-ends, fosmid-ends and Selected BAC Mixture sequences).

We caution you that the current assembly is a "work-in-progress" and as such is subject to modification prior to publication release (anticipated for mid-2011), some of which is likely to be substantial. Therefore we encourage you to carefully and independently validate any conclusions you may draw from this sequence. We will update this resource as improvements in the assembly are made. We welcome any feedback regarding your successes or that may assist us in improving the quality and accuracy of this sequence.

This pre-publication tomato genome data is made available with the understanding that users will respect the rights of those who contributed to this effort to describe the tomato genome in a peer-reviewed publication. This description includes whole genome level analyses on genes, gene families, repetitive sequences etc. We encourage you to review the NIH-NHGRI guidelines on distribution and use of pre-publication genome sequence at http://www.genome.gov/page.cfm?pageID=10506537. Any use of the tomato genome data prior to its publication should credit "The International Tomato Genome Sequencing Consortium". If you are uncertain about how to credit the use of the sequence or its appropriate use please do not hesitate to contact Joyce Van Eck.
Files in ITAG2.3_release

ITAG2.3_assembly.gff3
GFF version 3 file containing the location of SL2.40 contigs and scaffolds on the pseudomolecules.

ITAG2.3_cdna_alignments.gff3
GFF version 3 file containing alignments of existing EST and cDNA sequences to the genome.

ITAG2.3_cdna.fasta
FASTA-format file of the cDNA sequence produced by each gene model.

ITAG2.3_cds.fasta
FASTA-format sequence file of the CDS sequence produced by each gene model.

ITAG2.3_changed_genes.txt
List of genes that have had structural changes since the last release (ITAG2).

ITAG2.3_deleted_genes.txt
List of genes that were in the ITAG2 release, but have been deleted for this release.

ITAG2.3_dropped_features.gff3
GFF version 3 file listing features from the ITAG2 that could not be remapped from the SL2.31 assembly to the SL2.40 assembly.

ITAG2.3_protein_functional.gff3
GFF version 3 file containing functional annotations to protein sequences.

ITAG2.3_gene_list.txt
Text file containing a list of (versioned) gene identifiers present in this release.

ITAG2.3_de_novo_gene_finders.gff3
GFF version 3 file containing predictions from several de novo gene finders. These are intermediate data, used by EuGene to decide the final consensus gene models.

ITAG2.3_genomic.fasta
fasta-format sequence file of genomic pseudomolecule sequences from the SL2.40 assembly.

ITAG2.3_genomic_reagents.gff3
GFF version 3 file containing alignments to other genomic sequences from tomato: genomic clones, other genome builds, etc.

ITAG2.3_infernal.gff3
GFF version 3 file containing Infernal RNA results, see http://infernal.janelia.org/
SGN BLAST
• SGN FTP site:
  • ftp.solgenomics.net/tomato_genome/
Other sequencing in Solanaceae and related Asterids

• Wild tomatoes
  – *Solanum pimpinellifolium* (CSHL), closest wild relative of tomato
  – *Solanum pennellii* (Max Planck)
  – *Solanum galapagense* (BTI), salt tolerant from Galapagos Islands

• Potato, Mimulus

• Pepper (Korea), Nicotiana, Petunia, coffee, are currently being sequenced
Solanum galapagense

Wild relative of tomato

- Phenotypic differences include fruit color, scent, soluble solids, salt tolerance, pathogen response, trichomes, lifecycle (perennial), and leaf morphology.
- Smaller genome than S. lycopersicum

Distribution

- Found only on the Galapagos Islands usually on coastal lava.
- Considered endangered due to introgression from cultivated tomatoes.

http://news.bbc.co.uk
Solanum galapagense

Figure 5  Solanum galapagense S. Darwin & Peralta (based on S. Darwin & Schultz 290, BM).

Darwin et al., 2003
Solanum galapagense

- Sequencing - accession LA436 sequenced using paired-end Illumina HiSeq technology and aligned to the reference *S. lycopersicum* assembly.

- **Results:**
  - Average read depth: 21.83
  - Coverage of ~90% of *S. lycopersicum* genome.
  - ~1.5 million SNPs detected
  - ~400,000 indels
  - 350 genes have no read coverage – missing in *S. galapagense*?
S. galapagense SNP density

(Suzy Strickler)
Comparison of the S. galapagense Pto Locus With Other Sequenced Loci

**Known and Hypothesized Pto Locus Haplotypes**
- differs from VFNT Cherry by a deletion of PtoC
- matches RG-PtoS predicted haplotype

S. galapagense is most similar to the Heinz 1706 sequence
- differs from VFNT Cherry by a deletion of PtoC
- matches RG-PtoS predicted haplotype

Sequence suggests different origin of Pto locus in different susceptible tomato fruit types

(Jesse Munkvold)
Over 100 Solanaceae to be sequenced in the next few years

See http://solgenomics/organism/sol100/view

Initial list created by Sandra Knapp

Add your own species of interest and metadata about the project
Nicotiana benthamiana

- Native to Australia
- Model plant (VIGS)
- Allotetraploid (*Suaveolentes*)
  - *N. sylvestris* (section *Sylvestres*)
  - *N. glauca* (section *Noctiflorae*)
- Large, >3Gb genome
- 19 chromosomes
- Sequenced by a BTI consortium
SGN Phenotype database

- Description of accessions
  - Textual
  - Images
- Traits, including measurements
- Link to genotype data
- Associated to populations
  - Mapping, QTL, mutagenized, etc
- About 22,000 records
  - Zamir mutagenized lines, TGRC, Heirloom lines, SolCAP
**Accession details**

**Accession**: Long John

**Description**: Fruits are of medium size and obovoid in shape. The country of origin is USA. Seeds were obtained from TGRC, UC Davis. This accession belongs to the 'Heirloom' varieties.

**Population**: Cultivars and heirloom lines

**Organism**: Tomato

**Uploaded by**: Esther van der Knaap

**Associated loci**: ovate  sun  [Associate locus]

**Images**

[Images of the accession]

**Phenotype data**

None

**Mapping data**

None

**Known alleles**

<table>
<thead>
<tr>
<th>Locus name</th>
<th>Allele symbol</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SGN accession: 313-105

Accession details

Accession: **313-105**
Description: QTL Tomato Sausage x LA1589 F2
Population: Tomato
Organism: Tomato
Uploaded by: Esther van der Knaap

Associated loci: [Associate locus]

Images

[Add new image]

 Phenotype data

<table>
<thead>
<tr>
<th>Trait</th>
<th>Value</th>
<th>Pop min</th>
<th>Pop max</th>
<th>Pop mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>proximal fruit end blockiness 30%</td>
<td>0.95</td>
<td>0.84</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>distal fruit end blockiness 10%</td>
<td>0.59</td>
<td>0.54</td>
<td>0.58</td>
<td>0.6</td>
</tr>
<tr>
<td>distal fruit end blockiness 20%</td>
<td>0.78</td>
<td>0.76</td>
<td>0.86</td>
<td>0.8</td>
</tr>
<tr>
<td>fruit shape triangle 10%</td>
<td>1.19</td>
<td>0.87</td>
<td>1.33</td>
<td>1.13</td>
</tr>
<tr>
<td>fruit shape triangle 20%</td>
<td>1.1</td>
<td>0.86</td>
<td>1.18</td>
<td>1.06</td>
</tr>
<tr>
<td>fruit shape triangle 30%</td>
<td>1.06</td>
<td>0.86</td>
<td>1.1</td>
<td>1.03</td>
</tr>
<tr>
<td>distal eccentricity index</td>
<td>0.98</td>
<td>0.63</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>fruit length mid-width</td>
<td>187.63</td>
<td>147.5</td>
<td>326</td>
<td>213.91</td>
</tr>
<tr>
<td>fruit mid-height width</td>
<td>199.5</td>
<td>157.38</td>
<td>262</td>
<td>214.6</td>
</tr>
<tr>
<td>distal angle macro 10%</td>
<td>150.03</td>
<td>112.07</td>
<td>158.66</td>
<td>148.16</td>
</tr>
<tr>
<td>proximal fruit end blockiness 5%</td>
<td>0.56</td>
<td>0.37</td>
<td>0.52</td>
<td>0.54</td>
</tr>
<tr>
<td>proximal fruit end height</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Fruit shape QTL populations

- 3 QTL populations generated by Esther van der Knaap (Ohio State University)
- Phenotypic characterization: Fruit shape parameters analyzed using Tomato Analyzer
- Genotyping
- solQTL
Population Details

[Edit] [New QTL Population]
Name: QTL Tomato Howard German x LA1589 F2
Description: QTL Tomato Howard German
Uploaded by: Esther van der Knaap

Phenotype Data and QTLs

<table>
<thead>
<tr>
<th>Trait</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
<th>No. of lines</th>
<th>QTL(s)...</th>
</tr>
</thead>
<tbody>
<tr>
<td>distal angle macro 10%</td>
<td>69.77</td>
<td>172.26</td>
<td>142.52</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>distal angle macro 15%</td>
<td>61.8</td>
<td>145.58</td>
<td>119.24</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>fruit area</td>
<td>224.5</td>
<td>860.93</td>
<td>418.39</td>
<td>113</td>
<td></td>
</tr>
</tbody>
</table>

Download: [Phenotype raw data] [Genotype raw data]

Genetic Map

S. lycopersicum TA492 x S. pimpinellifolium LA1589 (Tomato-EXPIMP 2008)

Literature Annotation

PMID:17283371 Morphological variation in tomato: a comprehensive study of quantitative trait loci controlling fruit shape and development. Show/hide abstract

Isaak Tecle, SGN
solQTL mapping output

Isaak Tecle, SGN
Curated QTLs

This potato consensus map shows the location of QTLs for blight resistance and maturity. [Citations]

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Markers</th>
<th>Marker collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>239</td>
<td>COS</td>
</tr>
<tr>
<td>2</td>
<td>197</td>
<td>COSII</td>
</tr>
<tr>
<td>3</td>
<td>172</td>
<td>KFG</td>
</tr>
<tr>
<td>4</td>
<td>177</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>192</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>152</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>203</td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Protocols</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
http://solcap.msu.edu
Tomato panel (~400 accessions)
Incl. Processing, fresh market, heirloom, wild relatives

Phenotyping for breeder traits:
Tomato Analyzer
Fruit shape
Color
PH
Brix
Vitamin C
Lycopene sugars

Genotyping (Illumina Infinium chip)

Potato panel (~400 accessions)

Phenotyping:
Specific gravity
chip color after cold storage
sucrose/glucose
Skin texture
tuber shape(l/w/h)
eyedepth
skincolor
Flower color
Flesh color
growth habit
total yield etc.

Genotyping (Illumina Infinium chip)
Phenotyping standards

http://solgenomics.net/tools/onto/
Community Gene Curation

- How to annotate genes for 100s of sequenced genomes?
- SGN community curation system for loci
- Researchers can log in and add their own data & annotations
- “locus editors” assigned to each locus
- Locus editors can log in and edit the symbol, name and description
- Trusted users (“submitters”) can add information to every locus
- Figures, sequences, alleles, mutants, publications, ontology terms, comments
- Currently about 150 locus editors for about 400 loci, mainly for tomato
- 75,000 loci in the database
Tomato locus carotenoid isomerase

Locus details

Locus name: carotenoid isomerase
Symbol: crtiso
Gene: carotenoid isomerase
activity
Description: encodes a carotenoid isomerase enzyme required during carotenoid desaturation essential for the biosynthesis of cyclized carotenes and xanthophylls.
Chromosome: 10
Arm: long

Locus synonyms 2: tangerine Solyc10g081650 [Add/Remove]
carotenoid isomerase is a TGRC gene

Locus editors: Tal Isaacson

Created on: 2006-05-10 Last updated on: 2009-12-11 by Tal Isaacson

Tomato-EXPEN 2000
ontologies:

- **biological_process**
  - involved in carotenoid biosynthetic process (GO: 0016117)

- **cellular_component**
  - expressed in plastid (GO: 0009536)
  - located in plastid (GO: 0009536)

- **molecular_function**
  - functions as carotenoid isomerase activity (GO: 0046608)
  - involved in carotenoid isomerase activity (GO: 0046608)

literature:

- PMID:11884678 Cloning of tangerine from tomato reveals a carotenoid isomerase essential for the production of beta-carotene and xanthophylls in plants. Show/hide abstract
- PMID:12399171 Carotenoid isomerase: a tale of light and isomers. Show/hide abstract
- PMID:15557094 Analysis in vitro of the enzyme CRTISO establishes a poly-cis-carotenoid biosynthesis pathway in plants. Show/hide abstract

known alleles:

<table>
<thead>
<tr>
<th>Allele symbol</th>
<th>Allele name</th>
<th>Synonyms</th>
<th>Phenotype</th>
<th>Accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>v</td>
<td>virescent</td>
<td>[add new]</td>
<td>Fruit and flower color typical of t; irregular yellowing near growing point.</td>
<td>10 [Edit]</td>
</tr>
<tr>
<td>1</td>
<td>[add new]</td>
<td>Fruit flesh and stamens orange colored.</td>
<td>0 [Edit]</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>t1^2</td>
<td>Yellowish growing point; light green foliage; resembles t in flower an fruit color</td>
<td>0 [Edit]</td>
<td></td>
</tr>
<tr>
<td><strong>Locus details</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Locus name</strong></td>
<td>carotenoid isomerase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Symbol</strong></td>
<td>crtiso</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gene activity</strong></td>
<td>carotenoid isomerase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>encodes a redox-type, carotenoid isomerase enzyme; required during carotenoid desaturation; essential for the biosynthesis of cyclized carotenones and xanthophylls.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Chromosome</strong></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arm</strong></td>
<td>long</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(* denotes required field.)
SGN Software Development

• SGN software is based on Perl and R and uses many standard modules (BioPerl, Gbrowse, Chado, Pathway Tools, etc)

• Agile software development practices to develop high quality, tested, and re-usable code

• All software is available as open source from http://github.com/solgenomics/
Acknowledgments
Acknowledgments

- Esther van der Knaap
- Dani Zamir
- Roger Chetelat
- SolCAP
  - David Douches
  - David Francis
  - Allen van Deynze
  - Robin Buell
  - Heather Merk
- Tomato Genome
  - Jim Giovannoni
  - Joyce van Eck
  - Stephen Stack
  - Bruce Roe
- Greg Martin
- Jesse Munkvold
- Locus Editors