

The Solanaceae Genomics Network: Data, Methods, Tools, and the Tomato Genome

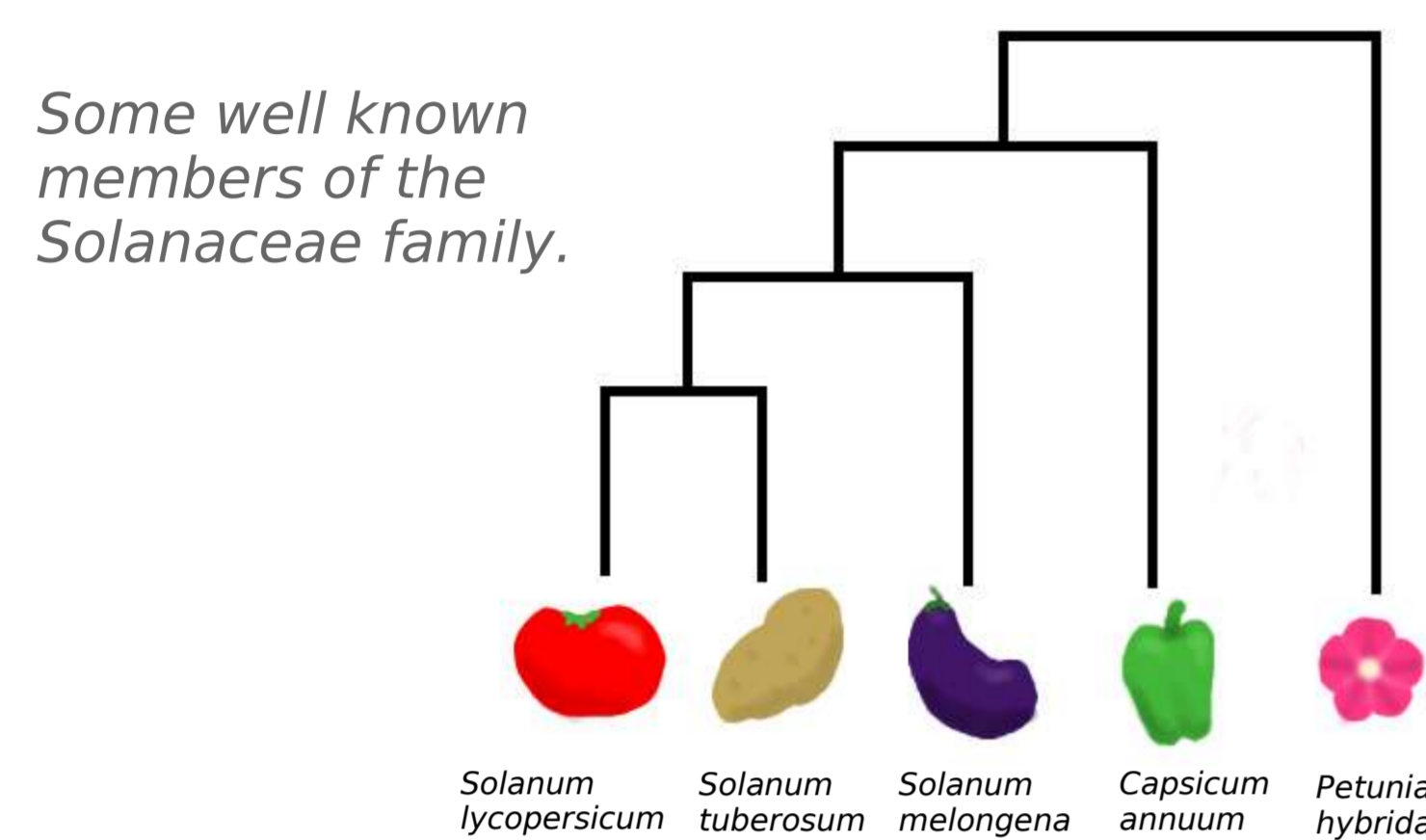
Beth Skwarecki, Nick Taylor, Teri Solow, Chenwei Lin, Eileen Wang, Emil Keyder, Miriam Wallace, Igor Dolgalev, Evan Herbst, Robert Ahrens, Mark Wright, Lukas Mueller, Steven Tanksley

Department of Plant Breeding, Cornell University

What is SGN?

SGN is a rapidly evolving comparative resource for the plants of the Solanaceae family, which includes important crop and model plants such as potato, eggplant, pepper and tomato. SGN houses maps, markers, ESTs, unigenes, and phenotypic data.

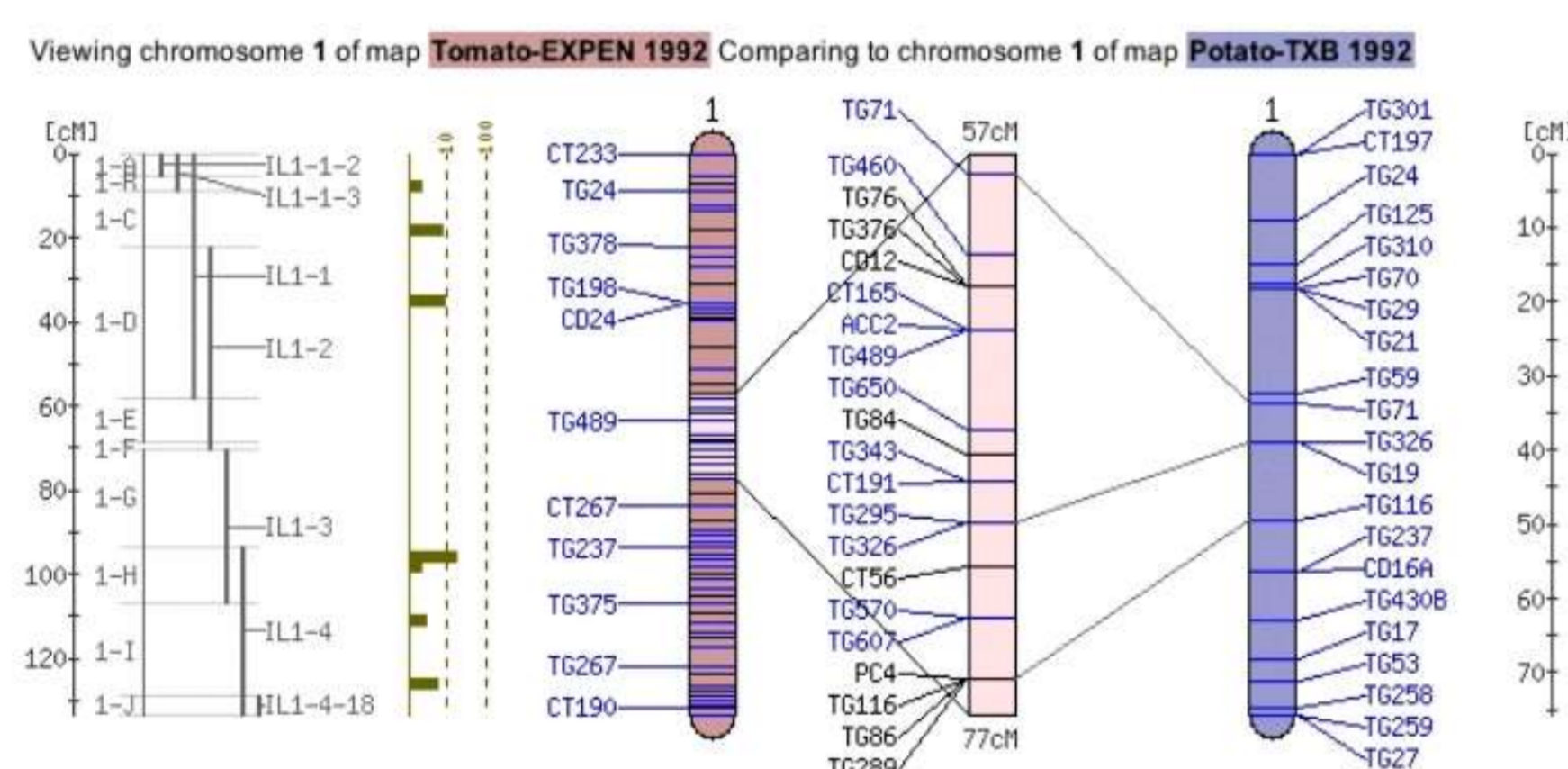
SGN is also part of the bioinformatics platform for the International Solanaceae Project (SOL), which is currently sequencing the euchromatic portion of the tomato genome. As this project progresses, SGN will evolve into a model organism database for tomato.



All data on SGN is available without restrictions. SGN can be accessed on the Web at <http://sgn.cornell.edu>.

Maps and Markers

SGN's collection includes interactive maps for tomato, potato, eggplant and soon pepper. All data for the maps is stored in a database and queried by our interactive map viewer. Through this viewer, a user can easily examine a map, compare it to other maps in the collection, zoom in on a region of interest, access marker information, and view the locations of anchored BACs. Introgression Line (IL) populations can also be seen on this map viewer.

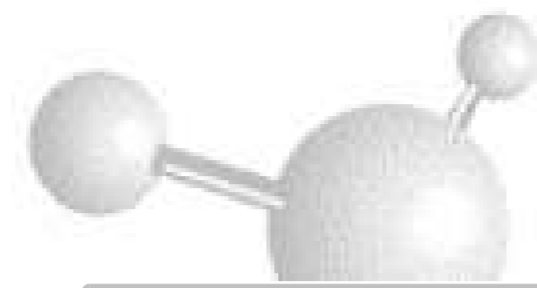


SGN's comparative viewer. Left to right: a centiMorgan ruler, IL lines, anchored BACs, the Tomato-EXPEN 1992 map, an enlarged section of interest, the Potato-TXB 1992 map, and a second ruler.

SGN's growing database currently contains information on over 3,000 markers from 5 interactive maps. Our marker collection includes microsatellite (SSR) markers, RFLP and CAPS based markers, and Conserved Ortholog Set (COS¹ and COSII²) markers for comparative mapping between various Solanaceae and some non-Solanaceae species such as Arabidopsis. The collection also includes markers derived from ESTs and unigenes, from genomic sequences such as BACs, and from genes with known functions in tomato. A powerful and user-friendly marker search is also available, so a user can search markers by type, chromosome, position, map, species and other criteria.

¹ Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14: 1457-1467

² Wu, Feinan, in preparation.



Data Overview

ESTs and Unigenes

	ESTs	Unigenes
Tomato	184,860	30,576
Potato	97,425	24,931
Pepper	20,738	9,554
Petunia	11,479	5,135
Eggplant	3,181	1,841

Maps and Markers

Map	Parents	Markers
Tomato-EXPEN 2000	<i>S. lycopersicum</i> , <i>S. pennellii</i>	1575
Tomato-EXPEN 1992	<i>S. lycopersicum</i> , <i>S. pennellii</i>	553
Tomato-EXHIR 1997	<i>S. lycopersicum</i> , <i>S. hirsutum</i>	135
Tomato-EXPIMP 2001	<i>S. lycopersicum</i> , <i>S. pimpinellifolium</i>	139
Potato-TXB 1992	<i>S. tuberosum</i> , <i>S. berthaultii</i>	178

BACs and the Physical Map

4857 good marker - BAC associations
652 markers are plausibly associated with BACs.
705 plausible contigs comprised of **1880** BACs
2166 BAC singletons

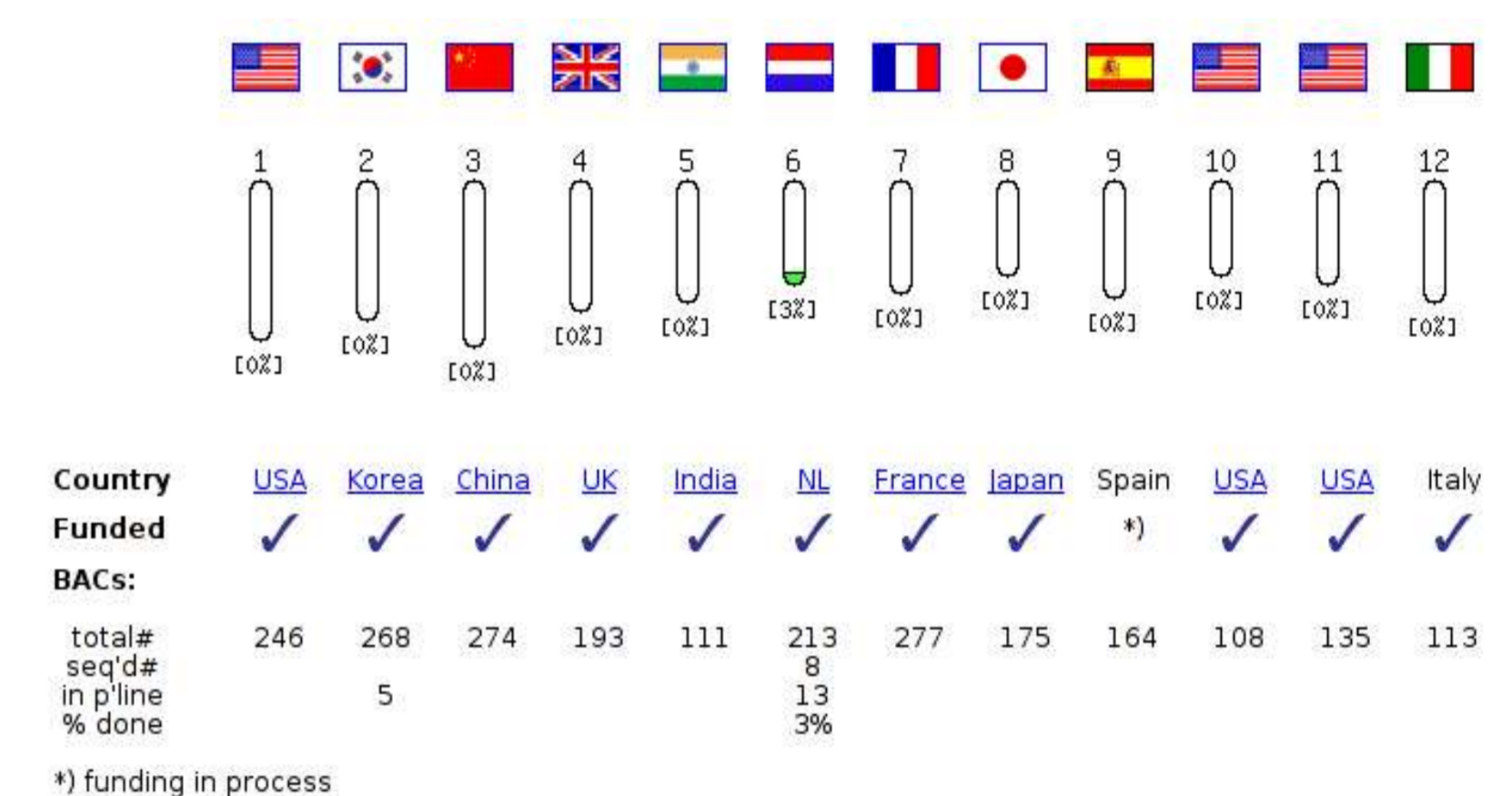
809 BACs are from **425** implausible contigs
3117 implausible marker-BAC associations
7235 ambiguous associations

Datasets are available from our FTP site at <ftp://sgn.cornell.edu>.

The International Solanaceae Genome Project (SOL)

The long-term goal of the SOL project is to create a network of map-based resources and information to address key questions in plant adaptation and diversification. The first cornerstone of this project is the sequencing of the gene-rich euchromatic portion (about 25%) of the 950 Mb tomato genome.

Sequencing will occur on a BAC-by-BAC basis, using an existing physical map and newly sequenced BAC ends to determine a tiling path that covers the desired portion of the genome.

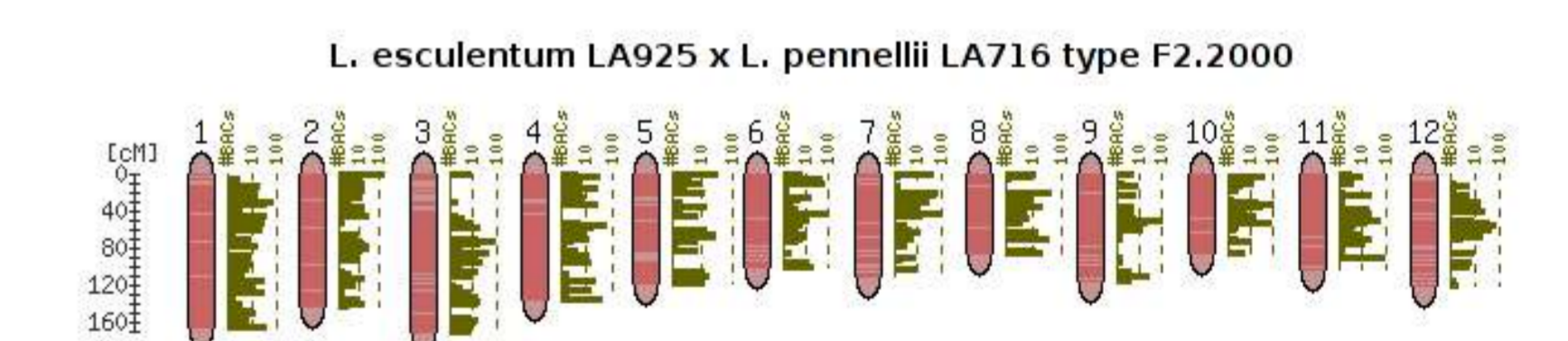


Ten participating countries are sequencing the BACs for their respective chromosomes. This chart can be found at http://sgn.cornell.edu/help/about/tomato_sequencing.html

Putting BACs on the Map

The foundation for the sequencing project is a physical map. Overgo analysis has been used to match BACs to probes based on markers from the Tomato-EXPEN 2000 map.

Analysis of the overgo results has found over 600 markers that unambiguously anchor over 5000 BACs to the genetic map. A Dutch AFLP map will be integrated with these results, bringing the number of anchor points to about 2000. These anchor points will be used as the "seeds" to start the sequencing process. The tiling path can then be extended from bac end sequences and FingerPrint Contigs (FPC).

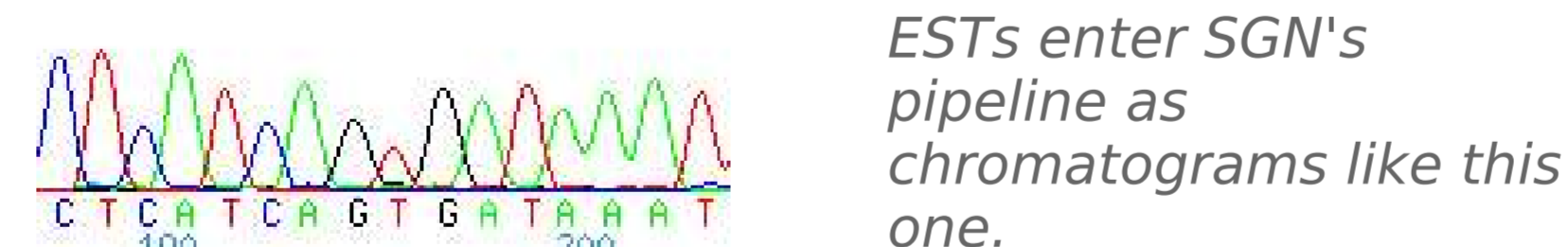


SGN's physical map. Green bars indicate the number of BACs anchored to each point on the Tomato-EXPEN 2000 genetic map.

ESTs and Unigenes

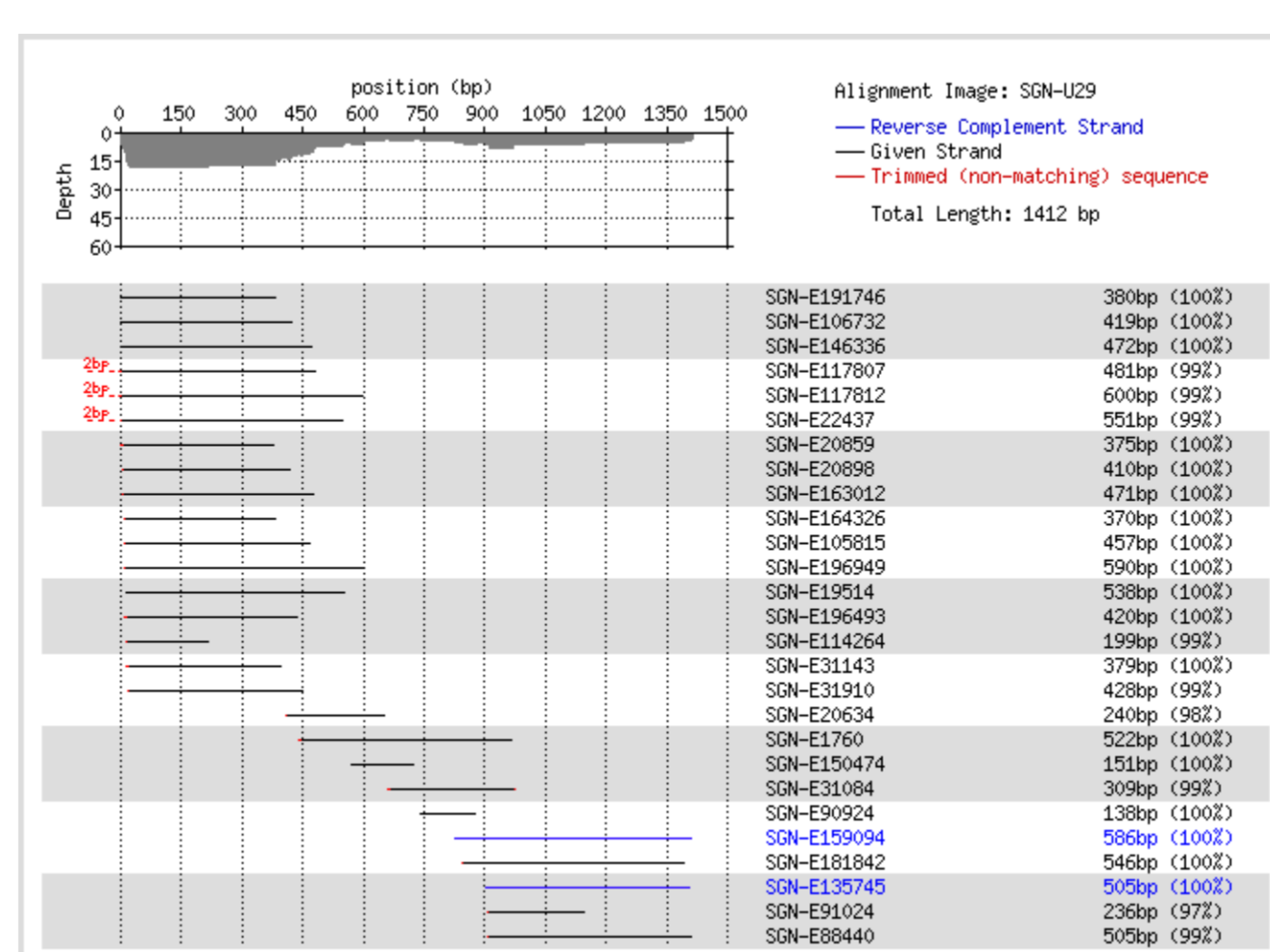
As an important first type of sequence data, Expressed Sequence Tags (ESTs) are collected and assembled using a pipeline that has been developed at SGN.

Sequences are first basecalled from chromatograms with Phred and loaded into the database. Subsequent steps of the custom pipeline trim the sequence according to quality scores and remove vector sequence. Sequences that are contaminated with bacterial or phage DNA, or that appear to be chimeric, are flagged and not used in unigene assembly.



ESTs enter SGN's pipeline as chromatograms like this one.

Unigenes are assembled with Cap-3, after an initial pre-clustering stage that allows the assembly to be run in parallel. Comparisons to tomato mRNAs show that the unigenes are usually of high quality.



An interactive display showing the member sequences of one SGN unigene.

Technical Details

Much of SGN's software is developed in-house and is open source. We write code in Perl and C.

We also use tools like Blast, Phred, and Cap-3. Our systems run Linux, Apache, and MySQL.

Future Plans

SGN is always growing. In the next several months we expect to add new maps and markers, new BAC libraries, BAC sequences and BAC end sequences, and data from FISH experiments. A new annotation pipeline is also being developed, and a registry will keep track of the status of BACs that are being sequenced.

<http://sgn.cornell.edu>