The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL)

†Lukas A. Mueller1*, Steven D. Tanksley1, Jim J. Giovannoni2, Joyce van Eck2, Stephen Stack3, Doil Choi4, Byung Dong Kim5, Mingsheng Chen6, Zhukuan Cheng6, Chuanyou Li6, Hongqing Ling6, Yongbiao Xue6, Graham Seymour7, Gerard Bishop8, Glenn Bryan9, Rameshwar Sharma10, Jiten Khurana11, Akhilesh Tyagi11, Debasis Chattopadhyay12, Nagendra K. Singh13, Willem Stiekema14, P. Lindhout15, Taco Jesse16, Rene Klein Lankhorst17, Mondher Bouzayen18, Daisuke Shibata19, Satoshi Tabata19, Antonio Granell20, Miguel A. Botella21, Giovanni Giuliano22, Luigi Frusciant23, Mathilde Causse24, and Dani Zamir25

†A full list of affiliations appears at the end of this paper

*Correspondence to:
Lukas A. Mueller, Department of Plant Breeding, Cornell University, Ithaca, NY, USA.
E-mail: Lam87@cornell.edu

Abstract

The genome of tomato (Solanum lycopersicum) is being sequenced by an international consortium of 10 countries (Korea, China, the United Kingdom, India, The Netherlands, France, Japan, Spain, Italy and the United States) as part of a larger initiative called the ‘International Solanaceae Genome Project (SOL): Systems Approach to Diversity and Adaptation’. The goal of this grassroots initiative, launched in November 2003, is to establish a network of information, resources and scientists to ultimately tackle two of the most significant questions in plant biology and agriculture: (1) How can a common set of genes/proteins give rise to a wide range of morphologically and ecologically distinct organisms that occupy our planet? (2) How can a deeper understanding of the genetic basis of plant diversity be harnessed to better meet the needs of society in an environmentally friendly and sustainable manner? The Solanaceae and closely related species such as coffee, which are included in the scope of the SOL project, are ideally suited to address both of these questions. The first step of the SOL project is to use an ordered BAC approach to generate a high quality sequence for the euchromatic portions of the tomato as a reference for the Solanaceae. Due to the high level of macro and micro-synteny in the Solanaceae the BAC-by-BAC tomato sequence will form the framework for shotgun sequencing of other species. The starting point for sequencing the genome is BACs anchored to the genetic map by overgo hybridization and AFLP technology. The overgos are derived from approximately 1500 markers from the tomato high density F2-2000 genetic map (http://sgn.cornell.edu/). These seed BACs will be used as anchors from which to radiate the tiling path using BAC end sequence data. Annotation will be performed according to SOL project guidelines. All the information generated under the SOL umbrella will be made available in a comprehensive website. The information will be interlinked with the ultimate goal that the comparative biology of the Solanaceae — and beyond — achieves a context that will facilitate a systems biology approach. Copyright © 2005 John Wiley & Sons, Ltd.

Keywords: Tomato; Solanum lycopersicum; Solanaceae; SOL project
**Introduction**

The Solanaceae, also called nightshades, are a large family of more than 3000 species, including the tuber-bearing potato, a number of fruit-bearing vegetables (tomato, eggplant, peppers), ornamental plants (petunias, *Nicotiana*), plants with edible leaves (*Solanum aethiopicum, S. macrocarpon*) and medicinal plants (e.g. *Datura, Capsicum*) [1]. The Solanaceae are the third most economically important plant taxon, and the most valuable in terms of vegetable crops. They are also the most variable of crop families in terms of agricultural utility. The closely related coffee (Rubiaceae) is a highly valuable commodity worldwide. In addition to their role as important food sources, many solanaceous species play a role as model plants, such as tomato and pepper for the study of fruit development [2–9], potato for tuber development [10,11], petunia for the analysis of flavonoids, and tomato and tobacco for plant defence [12–16]. The nightshades have also attracted interest because they produce a number of secondary metabolites, some of which have medicinal properties. The Solanaceae are remarkable in that the gene content of the different species is similar despite the markedly different phenotypic outcomes, making the Solanaceae an excellent model for the study of adaptation to natural and agricultural environments [17]. Many Solanaceae share a basic set of 12 chromosomes and are diploid, indicating an absence of large genome duplications and polyploidizations during the evolutionary history of this family. These intrinsic features of the Solanaceae and closely related species such as coffee make them well suited to address fundamental questions in plant biology.

In November 2003, the International Solanaceae Project (SOL) was initiated at a meeting near Washington, where goals were established for Solanaceae research for the next decade ([http://sgn.cornell.edu/solanaceae-project/](http://sgn.cornell.edu/solanaceae-project/)). The ultimate aim of the SOL project is to address two of the most significant questions in plant biology and agriculture: (a) how can a common set of genes/proteins give rise to a wide range of morphologically and ecologically distinct organisms that occupy our planet?; and (b) how can we deepen our understanding of the genetic basis of plant diversity be harnessed to better meet the needs of society in an environmentally friendly and sustainable manner? To answer these questions in the context of the Solanaceae, it is necessary to link traits to sequence, requiring both extensive phenotyping [18] and sequence information. It would be desirable to obtain as many full solanaceous genome sequences as possible for direct comparison. Due to the high cost of sequencing complete genomes at high quality, this is not feasible at present. The alternative is to fully sequence a high quality reference genome, and to map onto it ‘cheaper’ sequence data, such as ESTs, methyl-filtered sequences [19–21] or low Cot sequences [22,23] from other species. The availability of good comparative maps [24–26] between many solanaceous plants and the large numbers of EST sequences already available [27] is a great benefit in this approach. Thus, sequencing the gene-rich regions of the tomato genome will be the first cornerstone of the SOL project. After sequencing two rosids, *Arabidopsis* [28] and *Medicago* [29], sequencing a solanaceous plant will shed light on a genome of the more distantly related asterid clade, which will permit comparisons between genomes at longer evolutionary distances and thereby help define a larger view of plant evolution. All the information generated under the SOL umbrella will be made available in a comprehensive website, where all information will be interlinked such that, ultimately, the comparative biology of the Solanaceae will become available in a context that will facilitate a systems biology approach to understanding genome evolution, plant development and plant responses to the environment.

**Tomato genome structure and sequencing method**

Approximately three-quarters (730 Mb) of the 950 Mb tomato genome exists as pericentromeric heterochromatin. The remaining one-quarter (220 Mb) of the tomato genome consists of the distal, euchromatic segments of the chromosomes. The DNA found in heterochromatin is rich in repetitive sequences and poor in genes, making it difficult to sequence. The euchromatin is thought to contain mostly single copy sequences and includes more than 90% of the genes, making it relatively easy to sequence. Therefore, the strategy is to sequence only the euchromatic portion of the genome to cover most of the gene space. Sequencing 220 Mb of the tomato genome is therefore a little less
than twice the effort of sequencing the *Arabidopsis* genome at 150 Mb [30].

The ordered BAC approach was chosen for sequencing the tomato genome. All sequencing will be done using standard BAC libraries: the well-characterized *HindIII* library [31], plus two additional libraries that the US part of the sequencing project will provide; all libraries are or will be deep BAC end-sequenced by the US project, for a total of 400,000 BAC end sequences. All sequence will be derived from *Solanum lycopersicum* var Heinz 1706, as this was the basis for the original *HindIII* library. Sequencing will be based on the F2-2000 map ([http://sgn.cornell.edu/](http://sgn.cornell.edu/)), which has been used as the basis for anchoring 1500 markers by overgo (overlapping oligo) hybridization [32]. Currently, more than 650 unambiguous anchor points are available, but further analysis will increase this number to an estimated 800–1000 anchor points. In addition, the F2-2000 map is being combined with an AFLP map, containing more than 1200 markers, generated at Keygene in The Netherlands, providing even more anchor points. For each anchor point, one of the many anchored BACs will be selected as a ‘seed’ and sequenced. The tiling path will then be generated by walking out from this seed in both directions, using the deep BAC end-sequencing data available for the BAC libraries. Information from the fingerprint contig (FPC) map available for the *HindIII* library will also be used. FISH analysis will be used to confirm chromosome mappings and delineate the euchromatin/heterochromatin boundaries [33]; anchored BACs will also be mapped on IL lines to verify location. Other methods available for full genome sequencing have also been assessed. Full genome shotgun sequencing is not a cost-effective way to sequence a fraction of the genome. Methyl and Cot filtering are both methods providing a bias for coding sequence and therefore euchromatic sequence. All these methods do not by themselves provide gene order, which will be critical for a reference genome.

Currently, 10 countries are involved in sequencing the tomato genome. The 12 chromosomes have been split up between the countries as follows (see Figure 1): Korea (chromosome 2), China (3), UK (4), India (5), The Netherlands (6), France (7), Japan (8), Spain (9), Italy (12) and USA (1, 10, 11). The US project will also generate the additional BAC libraries, perform BAC end-sequencing, and build a data repository as a resource for the entire project. Argentina will sequence the mitochondrial genome, and Italy the chloroplast genome. The organellar genome sequences will be important for distinguishing genomic insertions of organellar sequences from organellar contamination contained in the BAC libraries.

**Bioinformatics**

In such a large project, bioinformatics plays a crucial role. In particular, the sequence quality standards and the annotation have to be uniform across the different chromosomes generated through different national projects, and efficient data access has to be provided to the scientific community. A SOL bioinformatics committee comprised of participating sequencing country representatives is working on a bioinformatics guideline document that should be followed by all project members to ensure that results are comparable in the entire genome. The guidelines are available from [http://sgn.cornell.edu/solanaceae-project/](http://sgn.cornell.edu/solanaceae-project/). In addition, the following important conventions have been adopted: (a) the original data has to be stored for all data types, such as all chromatograms and assembly files from the BAC sequencing, all supporting data for gene calls etc; (b) all data has to be traceable to the submitters and data generators; and (c) a unified web resource will be created that makes all data accessible to the users without delay and in an intuitive format (‘one-stop-shop’). Several centres will work together to implement the portal, including MIPS ([http://mips.gsf.de/](http://mips.gsf.de/)), VIB Ghent, Wageningen University and the Research Centre, KRI BB ([http://www.kri bb.re.kr](http://www.kri bb.re.kr)), Kazusa ([http://www.kazusa.or.jp](http://www.kazusa.or.jp)), SGN and centres in other countries. All sequence data will be submitted to GenBank, and to the Solanaceae Genomics Network (SGN; [http://sgn.cornell.edu/](http://sgn.cornell.edu/)), which will serve as a repository and access point for the data.

**Conclusions**

Sequencing the tomato genome offers exciting new perspectives and opportunities for plant biology. The sequence will be compared to other
Figure 1. Overview of the International Tomato Sequencing Project. The 12 tomato chromosomes are sequenced by an international consortium of 10 countries [Korea (chromosome 2), China (3), UK (4), India (5), The Netherlands (6), France (7), Japan (8), Spain (9), Italy (12) and USA (1, 10, 11)], as shown. This overview is available from the SGN website (http://sgn.cornell.edu/help/about/tomato_sequencing.html) and will be continuously updated as sequencing progresses. More information on the project is available on the page.

sequenced genomes such as Arabidopsis and rice to inform us of the evolutionary history of these plants. In conjunction with sequence data for other solanaceous plants, the tomato sequence will be the basis for investigating the phenotypic diversity and comparative biology in the Solanaceae, one of the main aims of the SOL project. This will shed light on mechanisms of gene regulation, evolution, signalling, disease resistance and defence, and fruit development and quality, and finally contribute to the improvement of agriculture through the phenotypic diversity found in nature.

Full List of Author Affiliations

1. Department of Plant Breeding, Cornell University, Ithaca, NY, USA
2. Boyce Thompson Institute, Ithaca, NY, USA
3. Department of Biology, Colorado University, Fort Collins, CO, USA
4. KRIBB, Taejon, Korea
5. Seoul National University, Seoul, Korea
6. Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
7. Warwick HRI, Wellesbourne, UK

Copyright © 2005 John Wiley & Sons, Ltd.
References

1. Knapp, S. 2002. Tobacco to tomatoes: a phylogenetic perspective on fruit development in the Solanaceae. J Exp Bot. 53(377): 2001–2022.

2. Tanksley SD. 2004. The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. Plant J. 34(2): 249–255.

3. Hamilton AJ, Fray RG, Grierson D. 1995. Sense and antisense inactivation of fruit ripening genes in tomato. Curr Top Microbiol Immunol. 197: 77–89.

4. Gray J, Picton S, Shabbeer J, Schuch W, Grierson D. 1992. Molecular biology of fruit ripening and its manipulation with antisense genes. Plant Mol Biol. 19(1): 69–87.

5. Giovannoni JJ. 2004. Genetic regulation of fruit development and ripening. Plant Cell. 16(suppl): S170–180.

6. Brummell DA, Harpster MH. 2001. Cell wall metabolism in fruit softening and quality and its manipulation in transgenic plants. Plant Mol Biol. 42(1–2): 311–340.

7. Rome, Italy

8. Imperial College, London, UK

9. SCRI Invergowrie, Dundee, UK

10. University of Hyderabad, Hyderabad, India

11. University of Delhi South Campus, New Delhi, India

12. National Centre For Plant Genome Research, New Delhi, India

13. NRC on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi, India

14. Centre for BioSystems Genomics, Wageningen, The Netherlands

15. University of Wageningen, Wageningen, The Netherlands

16. Keygene, NV., Wageningen, The Netherlands

17. Plant Research International, Wageningen, The Netherlands

18. Institut National Polytechnique de Toulouse, Toulouse, France

19. Kazusa, Kisarazu, Chiba, Japan

20. Instituto de Biología Molecular y Celular de Plantas Valencia, Spain

21. University of Malaga, Malaga, Spain

22. ENEA, Rome, Italy

23. U Naples, Naples, Italy

24. INRA, Avignon, France

25. Hebrew University, Jerusalem, Israel

Copyright © 2005 John Wiley & Sons, Ltd. Comp Funct Genom 2005; 6: 153–158.
...and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* **14**: 1441–1456.

28. AGL. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**(6814): 796–815.

29. May GD, Dixon RA. 2004. *Medicago truncatula*. *Curr Biol* **14**(5): R180–181.

30. Johnston JS, Pepper AE, Hall AE, *et al*. 2005. Evolution of genome size in brassicaceae. *Ann Bot (Lond)* **95**(1): 229–235.

31. Budiman MA, Mao L, Wood TC, Wing RA. 2000. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res* **10**(1): 129–136.

32. Cai WW, Reneker J, Chow CW, Vaishnav M, Bradley A. 1998. An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization. *Genomics* **54**(3): 387–397.

33. Peterson DG, Lapitan NL, Stack SM. 1999. Localization of single- and low-copy sequences on tomato synaptonemal complex spreads using fluorescence *in situ* hybridization (FISH). *Genetics* **152**(1): 427–439.