

The French contribution to the multinational Solanaceae Genomics Project as integrated part of the European effort

C. Delalande, F. Regad, M. Zouine, P. Frasse, A. Latche, J.C. Pech, M. Bouzayen*

UMR 31990 INRA/INP-ENSA Toulouse “Génomique et Biotechnologie des Fruits”, Avenue de l’Agrobiopole, BP 32607, 31326 Castanet-Tolosan cedex; France

*E-mail: bouzayen@ensat.fr Tel: +33-5-62-19-35-71 Fax: +33-5-62-19-35-73

Received December 24, 2006; accepted January 15, 2007 (Edited by K. Hiratsuka)

Abstract The Solanaceae family comprises many species of prime agronomical importance among which tomato and potato are the most important food source crops. In despite of their tremendous morphological diversity, the Solanaceae are closely related at the genetic level and display remarkable similarity in gene content and order. The Solanaceae Genome Project emerged in recent years as an international initiative aiming at generating genomic resources on the Solanaceae species and at coordinating national research efforts across the world. European countries made a substantial contribution to the activity of the Solanaceae International Consortium either through their respective national programs or with the advent of the EU-SOL integrated project. The present paper gives an overview on the specific contribution of the French Solanaceae Programs and on how they fit as an integral part of the European and international initiatives.

Key words: EU-SOL, France, Solanaceae genomics, *Solanum lycopersicum*.

The Solanaceae Genome Project (SOL) is an international initiative aiming at studying biodiversity through a “Systems Approach to Diversity and Adaptation” (Mueller et al. 2005a). The global objectives are to establish a network of information, resources and scientists worldwide to ultimately tackle two major generic questions in plant biology and agriculture: (i) how do a wide variety of morphologically and ecologically distinct organisms can arise from a common set of genes? and, (ii) how does the genetic basis of plant diversity can be exploited to better answer future needs of the human society in an environmentally-friendly and sustainable manner? Because of their remarkable natural diversity, the Solanaceae are well suited to address these questions and the Solanaceae Genome Network (SGN, <http://www.sgn.cornell.edu/>) was set up to drive the coordinated efforts of a large number of laboratories across the world. While the SOL project is well in its track now, it is time to recall that the first seeds were sown during the Euro-American tomato meeting hold in France (Toulouse, June 2001) which launched the idea of an international tomato consortium that could stimulate concerted actions aiming at building generic resources on the tomato model. The official kick off of the SOL consortium took place at the Washington meeting in November 2003 and since then the consortium has been enlarged to a number of new partners among which Japan, China and India decided to participate to the

tomato genome sequencing project. More recently, the activity of the SOL consortium has been enriched by an European initiative in the form of the EU-SOL project, which started in 2006. During the two last years, a number of other SOL national initiatives emerged all over the three continents.

Ongoing European Solanaceae programs mostly focus on tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*) but address a wide range of biological topics from natural biodiversity through plant development to responses to biotic and abiotic stresses. An important part of the European effort has been dedicated to the generation of new resources necessary for the implementation of structural and functional genomic approaches and for the characterisation of genetic diversity. In this context, the French Solanaceae program has been designed as an integral part of the European effort as well as a contribution to the activity of the international consortium. Beside the sequencing of the gene-rich regions of tomato chromosome 7, the main tasks developed by the French partners are dealing with the generation of tools and resources for functional genomics such as: (i) sequencing of the tomato unigene ESTs, (ii) the production of tomato DNA microarrays, and (iii) the construction of TILLING (Colbert et al. 2001) facilities on the tomato. A substantial part of the French effort is also dedicated to the exploration and characterisation of the genetic diversity with the aim to

define QTLs controlling sensory and nutritional qualities of tomato fruit (Causse et al. 2002). The National Institute of Agronomic Research (INRA) played a pioneering role in taking the first initiatives of the Solanaceae genomics programs in France and most of these projects are now funded by INRA, ANR (National Research Funding Agency) and GENOPLANTE, the federative programme for plant genomics research in France. It is also important to mention that the French effort on tomato genomics has benefited from the Trilateral initiative raised through merging plant genomics programs from France, Germany and Spain.

From the sequencing of the unigene set of tomato ESTs to the construction of the new generation of tomato oligo-chips

Unravelling global changes in gene expression associated with developmental processes in the tomato appeared as a prerequisite step towards understanding the complex regulation mechanisms underlying these processes. Therefore, designing the technical tools necessary for high throughput expression studies became a major task of the tomato functional genomics program. The French-US initiative aiming at sequencing the unigene collection of tomato ESTs was the first French contribution to the activity of the SGN consortium. Building on the US tomato EST effort (Fei et al. 2004, see in this issue the contribution by J. Giovannoni), a unigene set of 27000 non-redundant ESTs was selected by *in silico* analysis that comprises novel clones isolated by INRA Bordeaux and Toulouse. This enriched unigene set was then arrayed and transferred to the French partner for sequencing from both 5' and 3' ends. The sequencing effort was mainly supported by INRA with the contribution of Italy (ENEA, Rome). This initiative allowed successful sequencing from both ends of 66% of the clones and among those 53% gave overlapping sequences. Compared to the sequences available at TIGR database, new sequence information are achieved for 49 % of the clones. The newly sequenced unigene set was used for the construction of tomato DNA microarrays printed with 9000 PCR-amplified unique cDNA clones (Alba et al. 2004, Moore et al. 2005, Alba et al. 2005, Fei et al. 2006). Both generic and fruit-oriented tomato microarrays produced at INRA Toulouse were distributed to groups from France and Europe. This centralised platform was also used to perform hybridization and subsequent processing of the microarrays for the need of different groups. More recently, a joint effort between four countries (France, Spain, Italy and USA) led to the construction of an oligo-based new generation of tomato chips. A complete list of 70 mer long oligonucleotides was first derived from the 27000 individual clones composing the unigene collection of tomato ESTs. Thereafter, 12000 clones were selected for

printing in the first version of oligo DNA-chips named EU-TOM1. These state of the art oligo-microarrays are currently being used by the partners of the EU-SOL program but are also available for participants of the SOL international consortium outside Europe. A Laboratory Information Management Systems (LIMS) for microarrays production and traceability including the set up of a dedicated tomato microarrays database providing access to the gene ID of the EU-TOM1 clones has been installed by the French partner (<http://bioinfo.genopole-toulouse.prd.fr/eusol/base/>). The printing of the second set of tomato oligo-based DNA chips (EU-TOM2) has been scheduled within the EU-SOL program for 2007 and will include the remaining tomato unigenes estimated at 15000 clones.

Constructing TILLING resources for high throughput reverse genetic approaches in the tomato

High-throughput functional identification of candidate genes requires reverse genetics technologies that can match the high flow discovery of new genes arising from the new transcriptomic approaches. Until recently, the implementation of this strategy was only possible on model species for which Flanking Sequence Tags (FST) resources associated to collection of tagged insertional mutants are available. However, large collections of insertional mutants are still lacking for the tomato or if they exist they are not publicly accessible. In order to fill this gap, a French initiative led by INRA Bordeaux (http://www.competences.u-bordeaux1.fr/fiche_structure.php?struct=TILLING-Tomate) and Evry (<http://www.evry.inra.fr/public/index.html>) took place in the last two years that aims to build up facilities for TILLING (Targeting Induced Local Lesions IN Genomes) approaches. Newly generated saturated EMS mutant populations were constituted on the tomato species, the corresponding seed stocks constituted and DNA pools were generated from these mutant populations. Efficient protocols for high throughput mutation detection were established and validated. The French initiative has been subsequently integrated within the EU-SOL program which decided to build up centralised TILLING facilities aiming at creating a service platform that provides knockout mutant lines in a particular gene to all the project partners. In addition of allowing the isolation of mutations in a target gene, the TILLING strategy has the advantage of giving access to allelic series for any gene that provides a variety of effects on the gene of interest. Furthermore, when applied to the screening of natural genetic variation, the TILLING approach is called ECOTILLING. Hence, the ECOTILLING strategy opens new prospects for the survey of germoplasm collections towards the identification of alleles of agronomic importance.

Exploring natural diversity and search for fruit quality QTLs

Natural biodiversity within the Solanaceae is still largely unexplored and hence these under-exploited sustainable resources represent a rich reservoir for genetic improvement of commercial Solanaceae crops with novel alleles that improve productivity, quality and adaptation. Current commercial tomato varieties have already largely benefited from natural genetic variation through the introgression of loci bringing unique single gene mutations that affect plant architecture and fruit quality traits. Since over a decade, the European academic groups are contributing to the generation of efficient tools that enable the best use of the available genetic resources. More recently, the integrated EU-SOL project defined its final objective as to extract the under-exploited natural biodiversity present in Solanaceae to improve consumer-driven and environmentally-directed quality of tomato fruits and potato tubers. In that prospect, it is intended to capture multiallelic effects related to organoleptic quality, consumer's health and agrotechnical traits. The French effort, led by INRA-Avignon in this domain, is fully integrated to the European initiative and is dedicated to the development of DNA markers that allow QTL identification, mapping, cloning of the underlying genes and the use of the novel variation in marker-assisted breeding. Part of this effort specially concentrates on the identification of QTLs controlling tomato fruit quality traits and the study of their interaction with different genetic backgrounds (Causse *et al.* 2004, Chaib *et al.* 2006). QTLs involved in several tomato quality traits such as fruit weight, sugar content, fruit texture, aroma composition, have been introgressed from a contrasted intra-specific donor line into elite lines with variable commercial characteristics. The resulting QTL-NILs are used to study the relationships among QTLs and between QTLs and genetic background. Fine mapping of these QTLs and their co-localisation with candidate genes are investigated. The characterisation of the putative candidate genes is also performed by studying their polymorphism in unrelated material. Both French and European programs aim at identifying genes responsible for quality traits and at dissecting the molecular mechanisms underlying these traits by applying state-of-the-art knowledge and innovative technologies. This effort is expected ultimately to open new leads for integrated breeding strategies consisting in the assembly of the validated genes within new elite genotypes with high quality traits.

The sequencing of the tomato genome

The sequencing of the tomato (*Solanum lycopersicum*) genome is the flag project of the SOL initiative not only because the tomato is a reference species for the whole

Solanaceae family but also because it gives concrete expression to the expected collaboration between different members of the international consortium (Mueller *et al.* 2005b). Indeed, the tomato genome sequencing project is an international effort involving 12 different countries including some of the so-called emerging nations such as China, India, Korea and Argentina (the mitochondrial genome). Moreover, in addition to the national contribution of the European countries (France, The United Kingdom, The Netherlands, Spain, Italy), each being in charge of one specific chromosome, the European Union *per se* is now bringing its own contribution through the support given by the EU-SOL program to the tomato genome sequencing project. Currently, the 12 tomato chromosomes are split up between the countries as follows: Korea (chromosome 2), China (chromosome 3), United Kingdom (chromosome 4), India (chromosome 5), the Netherlands (chromosome 6), France (chromosome 7), Japan (chromosome 8), Spain (chromosome 9), Italy (chromosome 12) and the United States (chromosomes 1, 10 and 11). Argentina is in charge of sequencing the mitochondrial genome, while Germany is undertaking the chloroplast genome. The organellar genome sequence information will be important to understand the flow exchange of genetic material between the nuclear and organellar genomes and to assess its impact on evolution within the Solanaceae family. Accessorily, it will also help during the assembly phase of the nuclear genome sequence to distinguish true genomic insertions of organellar sequences from organellar contamination contained in the BAC libraries.

The total size of the tomato genome is estimated to approximately 950 Mb of DNA, more than 75% of which is heterochromatic, rich in repetitive sequences and largely devoid of genes (Wang *et al.* 2006). The majority of genes are found in long contiguous stretches of gene-dense euchromatin located on the distal portions of the chromosome arms. Taking into account this observation, the consortium agreed on a strategy that specifically targets the sequencing effort to the euchromatic portion of the genome in order to cover most of the gene space with a minimum effort. Three BAC libraries, namely *HindIII*, *MboI* and *EcoRI* libraries, were generated by the US tomato genomics program and distributed to the different partners of the international sequencing project by J. Giovannoni's group (Cornell). To set up the tools for the genome sequencing, BAC clones are identified by S. Tanksley's group (Cornell, USA) that cover the minimal tiling path spanning the approximate 220 Mb corresponding to the euchromatin portion. Seed BACs used as starting points for sequencing are anchored to the genetic map by overgo markers, derived from approximately 1500 markers of the tomato high density F2-2000 genetic map (<http://sgn.cornell.edu/>). These seed BACs are thereafter used to walk into the tiling path

using BAC-end sequence data. In this regard, the sequencing project benefits from the deep end-sequencing of the three BAC libraries performed within the US program and from the use of the multi-species introgression lines (ILs) provided by D. Zamir (Hebrew University, Israel).

The French effort devoted to sequencing of the gene-rich portion of tomato chromosome 7 is led by the Laboratory of Genomics and Fruit Biotechnology (INRA, Toulouse) and involves Genome Express (Meylan, France) as a main sequencing partner. The Centre of National Resources for Plant Genomics (CNRGV, Toulouse) provides a great support to the project by ensuring the management of the BAC libraries including duplication, conservation, re-arraying and high throughput screening. The Molecular Cytogenetic Platform (INRA, Rennes) also contributes to the project through the implementation of FISH (Fluorescence In Situ Hybridization) technology to check the assignment of the selected BACs to chromosome 7. Some of the FISH mapping data of the seed-BACs on chromosome 7 are provided by the groups of Steve Stack (Colorado, USA) and Zhukuan Cheng (Bejin, China). Finally, the project also takes advantage from the collaboration with the group led by Dr. D. Jones (The Australian National University, Canberra) who gave access to specific markers on chromosome 7 (Hemming et al. 2004).

In short, the sequencing strategy comprises the following steps: (i) for each marker, one of the many anchored BACs is selected as a “seed”, (ii) verify the identity of the given BAC clone by PCR using primers on the BAC-end sequence and/or on the corresponding marker, (iii) verify the location of the seed BAC by mapping on D. Zamir’s introgressed lines (Eshed and Zamir 1995) or by FISH analysis (De Jong 1998), and (iv) proceed to the sequencing when the BAC satisfied the above mentioned criteria. Upon sequencing and assembly of a given seed BAC, identification of the overlapping BAC to be the next sequencing target is based on the combined use of the BAC-end sequence database and by the three different fingerprint contigs physical maps that were generated by the US (www.genome.arizona.edu/fpc/tomato), China and UK partners. Before entering the sequencing pipeline, the selected BAC is submitted to a final verification by mapping on the introgressed lines.

Finally, it is obvious that in a large project such as the tomato genome sequencing, bioinformatics plays a crucial role. The sequence quality standards and the annotation have to be uniform across the sequences generated through different national projects, and efficient access to the data has to be provided to the partners of SOL community. In order to ensure that the delivered results are comparable in the entire tomato genome, a SOL bioinformatics committee comprised

of representatives of countries participating to the sequencing project, generated a guideline document (<http://sgn.cornell.edu/solanaceae-project/>) describing the standard and the procedures to be followed by all members of the international sequencing project. The French sequencing project benefits from the Bioinformatics Platform of INRA-Toulouse (<http://bioinfo.genopole-toulouse.prd.fr/>) which gives access to both their expertise and their powerful informatics infrastructure for data storage and high throughput sequence analyses required for tomato chromosome 7. In addition, the Bioinformatics Group of INRA Toulouse is actively involved in gene nomenclature and functional gene annotation through the implementation of EuGene (<http://www.inra.fr/internet/Departements/MIA/T/EuGene/>), the gene finder software for eukaryotic organisms designed locally by the Toulouse’s group of Bioinformatics (Foissac et al. 2003). In this regard, it is important to mention that the international tomato genome sequencing consortium choose the EuGene as unified tool for the annotation of the whole tomato genome. A substantial effort is also dedicated to the development of bioinformatic resources within the EU-SOL project where the most prominent roles in this domain will be devoted to the following centres: Munich Information Center of Protein Sequences (MIPS, <http://mips.gsf.de/>), Flanders Institute for Biotechnology (VIB, Ghent, Belgium, <http://www.bits.vib.be/>) and Wageningen University (The Netherlands).

Perspective and up-coming initiatives

The sequencing of the tomato genome opens exciting new perspectives for the understanding of the genetic basis of plant morphological and physiological diversity. It is also expected that the comparative sequence information will give clues to uncover the underlying mechanisms driving plant evolution. However, efficient exploitation of the genome sequence data requires deeper understanding on how a specific subset of the genome information is recruited in a coordinated manner to drive defined developmental processes which in turn determine plant phenotypic diversity. Furthermore, high throughput proteomics and metabolomics programs will likely bring some lacking but crucial information. Metabolomics programs aiming at performing comparative metabolic profiling to reveal important components of tomato fruit (sugars, amino acid, organic acids, isoprenoids and volatiles) are currently being developed (Schauer and Fernie 2006) in the framework of the EU-SOL program mainly by groups from Max Planck Institute (Golm, Germany). Tomato proteomics initiatives are also in progress in France that attempt to identify proteins related to fruit quality through large proteome analysis of introgressed tomato QTL-NILs (INRA, Avignon). Finally, because more than 80% of

organellar proteins are encoded by the nuclear genome, the sequencing of tomato mitochondrial and chloroplastic genomes will only provide partial information on proteins acting in these compartments. In order to fill this gap, the group of INRA Toulouse is now preparing a proteomics program targeting the tomato mitochondria and chromoplast proteomes.

Acknowledgements

The French tomato genomics program is supported by the National Institute of Agronomic Research (INRA), the National Research Funding Agency (ANR), the Federative Program for Plant Genomics Research in France (GENOPLANTE) and the Framework Program 6 of the European Union through the EU-SOL project. We are grateful to Syngenta for providing tomato genetic markers of chromosome 7.

References

- Alba R, Fei Z, Liu Y, Debbie P, Gordon J, Rose J, Martin G, Tanksley S, Bouzayen M, Jahn M, Giovannoni J (2004) ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *Plant J* 39: 697–714
- Alba R, Payton P, Fei Z, McQuinn R, Debbie P, Martin G, Tanksley S, Giovannoni J (2005) Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *The Plant Cell* 17: 2954–2965
- Causse M, Saliba-Colombani V, Lecomte L, Duffe P, Rousselle P, Buret M (2002) QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. *J Exp Bot* 53: 2089–2098
- Causse M, Duffe P, Gomez MC, Buret M, Damidaux R, Zamir D, Gur A, Chevalier C, Lemaire-Chamlet M, Rothan C (2004) A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. *J Exp Bot* 55: 1671–1685
- Chaib J, Lecomte L, Buret M, Causse M (2006) Stability over genetic backgrounds, generations and years of quantitative trait locus (QTLs) for organoleptic quality in tomato. *Theor Appl Genet* 112: 934–944
- Colbert T, Till BJ, Tompa R, Reynolds S, Steine MN, Yeung AT, McCallum CM, Comai L, Henikoff S (2001) High-throughput screening for induced point mutations. *Plant Physiol* 126: 480–484
- De Jong JH (1998) High resolution FISH reveals the molecular and chromosomal organization of repetitive sequences in tomato. *Cytogenet Cell Genet* 81: 104
- Eshed Y, Zamir D (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* 141: 1147–1162
- Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannoni JJ (2004) Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. *Plant J* 40: 47–59
- Fei Z, Tang X, Alba R, Giovannoni J (2006) Tomato Expression Database (TED): a suite of data presentation and analysis tools. *Nucleic Acids Res* 34: 766–770
- Foissac S, Bardou P, Moisan A, Cros MJ, Schiex T (2003) EuGene'Hom: A generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res* 31: 3742–5
- Hemming MN, Basuki S, McGrath DJ, Carroll BJ, Jones DA (2004) Fine mapping of tomato I-3 gene for fusarium wilt resistance and elimination of a co-segregating resistance gene analogue as a candidate for I-3. *Theor Appl Genet* 109: 409–418
- Moore S, Payton P, Wright M, Tanksley S, Giovannoni J (2005) Utilization of tomato microarrays for comparative gene expression analysis in the Solanaceae. *J Exp Bot* 56: 2885–2895
- Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, Herbst EV, Keyder ER, Menda N, Zamir D, Tanksley SD (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol* 138: 1310–1307
- Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, Herbst EV, Keyder ER, Menda N, Zamir D, Tanksley SD (2005a) The DOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol* 138: 1310–1317
- Mueller LA, Tanksley SD, Giovannoni JJ, van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C, Ling H, Xue Y, Seymour G, Bishop G, Bryan G, Sharma R, Khurana J, Tyagi A, Chattopadhyay D, Singh NK, Stiekema W, Lindhout P, Jesse T, Klein Lankhorst R, Bouzayen M, Shibata D, Tabata S, Granell A, Botella MA, Giuliano G, Frusciante L, Causse M, Zamir D (2005b) The tomato sequencing project, the first cornerstone of the International Solanaceae Project (SOL). *Comparative and Functional Genomics* 6: 153–158
- Schauer N and Fernie AR (2006) Plant metabolomics: towards biological function and mechanism. *Trends in Plant Sci* 11: 508–516
- Wang Y, Tang X, Cheng Z, Mueller L, Giovannoni J, Tanksley SD (2006) Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. *Genetics* 172: 2529–2540