Review

Tomato genome sequencing: deciphering the euchromatin region of the chromosome 8

Erika Asamizu*

Department of Plant Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan
*E-mail: asamizu@kazusa.or.jp Tel: +81-438-52-3935 Fax: +81-438-52-3934

Received December 20, 2006; accepted January 9, 2007 (Edited by M. Sekine)

Abstract As a member of the Solanaceae Genomics Network International Tomato Sequencing Project, we launched on sequencing of the chromosome 8. Our task is to sequence the euchromatin region of the chromosome, the estimated size of which is 17 megabases. BAC-by-BAC strategy is adopted for sequencing. We initially received BAC clone candidates anchored to overgo probes developed from 33 markers mapped on chromosome 8. For confirmation of the BAC candidates, we analyzed the sequence of PCR product amplified from the BAC DNA with primers designed on the marker sequence. Twenty-five BAC clones were verified and subjected to shotgun sequencing. As of Nov. 2006, we finished 40 clones to Phase 3 (total non-redundant length 4,429,168 bases) of 25 are seed and 15 are extended clones. In order to find additional seed points, we performed PCR screening of markers against 3D DNA-pool of LE-HBa, SL-MboI and SL-EcoRI BAC libraries, and succeeded in obtaining 9 new seeds. We report the current status of our project and future perspectives toward the final goal, which include development of new microsatellite markers from tomato EST sequences and Selected BAC Mixture shotgun sequencing for the gap filling.

Key words: BAC mixture, genome sequencing, microsatellite marker, tomato chromosome 8.

Genome sequencing of higher plants is completed for the dicot and monocot model, *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000) and *Oryza sativa* (International Rice Genome Sequencing Project 2005, Goff et al. 2002; Yu et al. 2002), and a draft sequence of the poplar genome is now available (Tuskan et al. 2006). Genomics research has been newly emerged as an outcome, and contributed to accelerate gene function analyses. To understand the complex genetic system of diverse higher plant species, it is necessary to analyze plants in different taxa with characteristic features.

The Solanaceae is phylogenetically distinguished from the two model plants (Soltis et al. 1999). It includes more than 3,000 species and is one of most important crop species besides grasses (rice, maize, wheat) and legumes (soybean, pea, alfalfa) and is most valuable in terms of vegetable species (e.g. tomato, pepper, eggplant and potato).

Comparative genomic mapping in grasses, crucifers and legumes showed a significant level of conservation of gene content and order within same taxon. The Solanaceae family is characterized by its highly conserved genome organization; most species possess same number of chromosomes (2n=2x=24) since genome duplication event did not take place in Solanaceae until relatively recently which resulted in several polyploidy species (e.g. tetraploid potato and tetraploid tobacco) (Knapp et al. 2004; Doganlar et al. 2002; Livingstone et al. 1999; Tanksley et al. 1992).

Tomato was selected as a reference model of Solanaceae family since it enables simple diploid genetics with characteristics like short generation time, amenability to transformation, and availability of abundant genetic and genomic resources such as high-density genetic map (Fulton et al. 2002; Tanksley et al. 1992) and over 200,000 EST sequences. The genome size of tomato is estimated to be 950 Mb, which is among the smallest in Solanaceae family. Cytogenetic analyses of the pachytene chromosomes have revealed a largely contiguous euchromatin structure, which comprises approximately 25% of the entire genome (Wang et al. 2006; Zhong et al. 1998).

Currently the Solanaceae Genomics Network (SGN) organizes the International Tomato Sequencing project (Mueller et al. 2005). The chromosome 8 is allotted to Japanese team organized by Kazusa DNA Research Institute and the National Institute of Vegetable and Tea Science. The entire project is aimed at deciphering the

Abbreviations: BAC, bacterial artificial chromosome; cM, centimorgan; EST, expressed sequence tag; FISH, fluorescence *in situ* hybridization; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism.

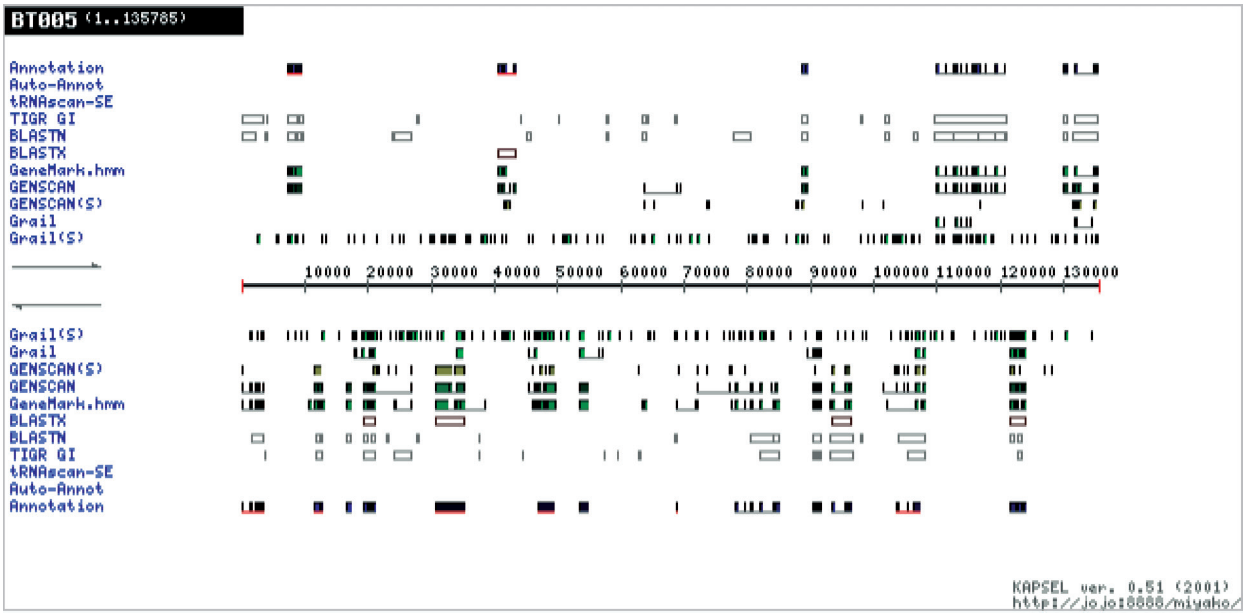

Figure 1. An example of gene prediction on tomato BAC sequence.
Output of the automated genome annotation pipeline. The results of gene prediction programs and database searches are shown.

gene-rich euchromatin estimated primarily to be 220 Mb, and sequencing was initiated from a total of 650 seed BAC clones anchored to overgo probes developed from Conserved Orthologous Set (COS) markers on a genetic map Tomato-EXPEN 2000 developed from F2 population of *S. lycopersicum* LA925×*S. pennellii* LA716 (http://www.sgn.cornell.edu/cview/map.pl?map_id=9). Our basic strategy is BAC-by-BAC sequencing utilizing the BAC end sequences. We report here the current status of the project and discuss the future prospects necessary to target the "end" of the project.

Sequencing status

We launched on the project by verifying the candidate BAC clones selected and sent by the SGN. They developed overgo probes for Conserved Orthologous Set (COS) markers and performed hybridization against BAC clones densely arrayed on filters. We received BAC candidates corresponding to 33 markers and verified the clones by sequence-analyzing the PCR products amplified with primers designed on marker sequences. We succeeded in verifying 25 of them and subjected them to sequencing.

Sequencing of BAC clones is performed by the shotgun strategy. Isolated BAC DNA is fragmented by sonication and fragments ranging from 2.5 to 4.0 kb are subcloned into the pUC118 vector (Takara Bio Inc.). The subcloned fragments are sequenced from both ends, followed by base calling and assembling by the Phred-Phrap program (Ewing et al. 1998; Ewing and Green 1998). Total length of the reads accumulated is equivalent to 5-times the original BAC clone.

Table 1. Preliminary result of tomato BAC annotation.

Number of BAC clones	5
Total length (bp)	637,486
GC content (%)	32.84
Number of predicted genes	92
Length of spliced gene (bp)	138–8250
Number of intron in every one gene	0–23
Gene density (kb)	7.3

To gain first insight into the tomato genome, we ran an automatic annotation pipeline developed in our *A. thaliana* and *Lotus japonicus* genome sequencing projects on finished tomato BAC sequences (Figure 1). The summary of preliminary result is indicated in Table 1. The result suggested that significant number of genes might be encoded in the BAC sequences. A higher gene density of the tomato genome (one gene per every 7.3 kbp) compared to *L. japonicus* (one gene per every 10.7 kbp) with much compact genome (472.1 Mb) (Sato and Tabata 2006) was obtained. We need to close examine the result to remove "false" signals such as of pseudogenes. SGN is now developing an annotation pipeline specifically trained for tomato, which would improve accuracy of the tomato gene prediction and enable discussion on the tomato gene repertoire.

In order to develop additional seed points along the chromosome, we made three-dimensional DNA-pool of the BAC libraries. As indicated in Table 2, this pool enables 4.5-deep screening of the genome. Using the pool, we performed screening of markers failed in the candidate BAC verification and also framework markers on Tomato-EXPEN 2000 genetic map. As a result, we succeeded in obtaining 9 new seed points.

Table 2. Three-dimensional DNA-pool of the tomato BAC libraries.

Library	Number of 384 plate	Insert size average (kb)	Length (Mb)	Genome coverage
LE_HBa	48	117	2156.5	2.27x
SL_MboI	24	135	1244.1	1.30x
SL_EcoRI	24	95–100	875.5	0.92x
Total	96			4.49x

For the first screening, the 384 BAC-pool arrayed in a 96-well plate was used. Genome coverage is based on the estimated genome size of 950 Mb.

Table 3. Status of the BAC-by-BAC extension.

Genetic distance (cM)	Marker	Marker type	Extension
2	TG176	RFLP	Stopped
2	T1123	COS	
3.5	C2_At4g31130	COSII	Stopped
11	T0721	COS	
18	cLEX-11K1B	EST	
20	T0721	COS	
20	T0718	COS	
20	cLET-3-O9	EST	
20	T1358	COS	
20	T1179	COS	Stopped
21.5	cLET-3-M1	EST	
22	CT92	RFLP	
22.7	SSR15	SSR	
23	cLER-5-P17	EST	
23.5	CT228	RFLP	
24	T1435	COS	
30	CT77	RFLP	
34	TG513	RFLP	
36	T1434	COS	Stopped
36.5	TG624	RFLP	Stopped
42	T1530	COS	
45	CT88	RFLP	
49	CT64	RFLP	Stopped
51	T1341	COS	Stopped
63	T1581	COS	Stopped
67	CT148	RFL	Stopped
69	T0487	COS	Stopped
70	T0608	COS	
71	T0337	COS	
72.2	C2_At4g11560	COSII	Stopped
75	T1522	COS	
81	T1433	COS	
84	CT252	RFLP	
87	CT68	RFLP	TGR1

Extension has been exhausted at 11 out of 34 points, and the euchromatin/heterochromatin border has been reached at the distal end of the long arm.

Table 3 indicates the genetic distance of each seed point and the extension status. We seem to have reached to the euchromatin/heterochromatin border in the distal end of the long arm. Extension from a marker at 87 cM has led us to encounter a clone with long stretch of 162-bp repeat which is the previously identified TGR1, tomato subtelomere-specific repeat (Zhong et al. 1998). In some parts, we are currently facing exhaustion of BAC clones to extend. Such un-extendable regions tend to locate near heterochromatic regions e.g. subtelomeric region at 2 cM and pericentromeric region at 20 cM. Several seed points on the long arm are now unable to extend. Concentration of dispersed type of repeat on the long arm of chromosome 8 was shown by a FISH analysis (Chang 2004), which might be one possible reason for the extension difficulty.

Complementary effort

Development of new markers

In order to develop additional seed points which is an urgent need to continue BAC-by-BAC sequencing, we initiated development of microsatellite markers. We searched for simple sequence repeats (SSR) of di-, tri- or tetra-nucleotide repeats of more than 15 bases in the 26,363 tomato EST unigene dataset and 57,422 full-length cDNA sequences in MiBASE (http://www.kazusa.or.jp/jsol/microtom/indexj.html). Using the FINDPATTERNS program (Accelrys GCG; http://www.accelrys.com/products/gcg/) by allowing one-base mismatch, we identified 2627 SSR among which 522 matched with mapped markers, thus leaving the remaining 2105 as new microsatellite marker candidates. Primers were designed using the Primer3 program (Rozen and Skaletsky 2000) by setting the size range of the product to 100–300 bp. We are planning to analyze the polymorphism of SSR within two *S. pennellii* LA716-introgressed lines (IL) of *S. lycopersicum* LA925 (Fulton et al. 2002) and *S. lycopersicum* M82 (Eshed et al. 1992). Polymorphism is going to be detected by electrophoresing the PCR product either on 10% polyacrylamide gel or a capillary sequencer.

An effort for developing additional seed points is also planned by the SGN. They will identify BAC clones whose end sequences contain a gene sequence, develop markers and map them using the *S. pennellii* IL. Their goal is to increase the seed points on each chromosome to up to 100.

New strategy for filling gaps in gene spaces

As it is anticipated that even after new seed points were added, significant part of the chromosomes may be left as gaps unable to be covered by the BAC extension. Gaps in gene spaces may result from dispersed type of repeat which would hamper those regions from being cloned efficiently. It is necessary to design complementary methods in order to fill gaps left in gene spaces in later stage of the sequencing project.

One way is to use other type of libraries. The US team

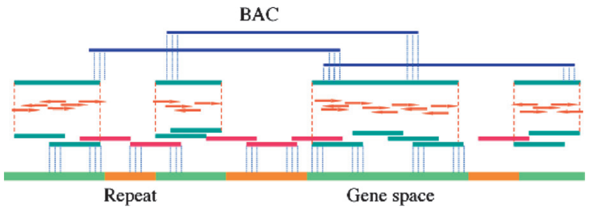

Figure 2. Outline of the Selected BAC Mixture (SBM) shotgun method. In this method, low-copy BAC clones i.e. whose end sequences do not contain repeat sequences are selected by database search, selected clones are mixed and shotgun libraries are constructed using the mixture. Assembled sequences will be anchored to BAC sequences proximal to gaps in gene spaces.

has developed a fosmid library comprised of 200,000 clones with an average insert size of 40 kb. They are also end sequencing the fosmid clones. The US team also plans to make available the cosmid library, average insert size of which is around 20 kb. To remove repeat-containing clones, they have arrayed the clones and performed hybridization with labeled genomic DNA, and selected 25,000 clones that do not show strong hybridization signal, which would give 2-times the euchromatin coverage.

As a complementary method for the gap filling, we have launched on Selected BAC Mixture (SBM) shotgun sequencing. The outline of this procedure is shown in Figure 2. Using the end sequences of 170,408 BAC clones as query, we performed search against the repeat dataset using the BLASTN program (Altschul et al. 1997); 14,229 repeat sequences in TIGR_SolAth_repeat, mips_repeat_collection, and SGN repeat collection (http://www.sgn.cornell.edu/bulk/input.pl?mode=ftp). As a result, 44,226 clones had repeat sequence on both ends, 74,866 clones had repeat on one end, and both end sequences of 54,831 clones did not show similarity to repeat. Among these low-copy BAC clones, 20,000 clones were pooled and shotgun libraries are being constructed. We are planning to accumulate 2.8 million reads from the SBM shotgun libraries, which should give 5-times the euchromatin coverage. This effort will be continued from 2007 to 2008.

Acknowledgments

This work is supported by the Kazusa DNA Research Institute Foundation and grants from National Institute of Vegetable and Tea Science (NIVTS) Priority Research Programme.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402

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

Chang SB (2004) Cytogenetic and molecular studies on tomato chromosomes using diploid tomato and tomato monosomic additions in tetraploid potato. *PhD Dissertation*. Wageningen University, Wageningen, The Netherlands

Doganlar S, Frary A, Daunay MC, Lester RN, Tanksley SD (2002) A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the solanaceae. *Genetics* 161: 1697–1771

Eshed Y, Abu-Abied M, Saranga Y, Zamir D (1992) *Lycopersicon* esculentum lines containing small overlapping introgressions from *L. pennellii*. *Theor Appl Genet* 83: 1027–1034

Ewing B, Hillier L, Wendl MC Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175–185

Ewing B and Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194

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* 4: 1457–1467

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92–100

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800

Knapp S, Bohs L, Nee M, Spooner DM (2004) Solanaceae-a model for linkage genomics with biodiversity. *Comp Funct Genom* 5: 285–291

Livingstone KD, Lackney VK, Blauth JR, van Wijk R, Jahn MK (1999) Genome mapping in capsicum and the evolution of genome structure in the solanaceae. *Genetics* 152: 1183–1202

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: 1310–1317

Rozen S and Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds.), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp 365–386. Source code available at http://fokker.wi.mit.edu/primer3/

Sato S and Tabata S (2006) *Lotus japonicus* as a platform for legume research. *Curr Opin Plant Biol* 9: 128–132

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402: 402–404

Tanksley SD, Ganal MW, Prince JP, de Vincente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB, Messeguer R, Miller JC, Mille L, Patersom AH, Pineda O, Roder MS, Wing RA, Wu W, Young ND (1992) High density molecular linkage maps of the tomato potato genomes. *Genetics* 132: 1141–1160

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604

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

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79–92

Zhong XB, Fransz PF, Wennekes-Eden J, Ramanna MS, van Kammen A, Zabel P, Hans de Jong J (1998) FISH studies reveal the molecular and chromosomal organization of individual telomere domains in tomato. *Plant J* 13: 507–517