

A Snapshot of the Emerging Tomato Genome Sequence

Lukas A. Mueller,* René Klein Lankhorst, Steven D. Tanksley, James J. Giovannoni, Ruth White, Julia Vrebalov, Zhangjun Fei, Joyce van Eck, Robert Buels, Adri A. Mills, Naama Menda, Isaak Y. Teclé, Aureliano Bombarely, Stephen Stack, Suzanne M. Royer, Song-Bin Chang, Lindsay A. Shearer, Byung Dong Kim, Sung-Hwan Jo, Cheol-Goo Hur, Doil Choi, Chang-Bao Li, Jiu-hai Zhao, Hongling Jiang, Yu Geng, Yuanyuan Dai, Huajie Fan, Jinfeng Chen, Fei Lu, Jinfeng Shi, Shouhong Sun, Jianjun Chen, Xiaohua Yang, Chen Lu, Mingsheng Chen, Zhukuan Cheng, Chuanyou Li, Hongqing Ling, Yongbiao Xue, **[continued next page.]**

Abstract

The genome of tomato (*Solanum lycopersicum* L.) 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 the larger "International Solanaceae Genome Project (SOL): Systems Approach to Diversity and Adaptation" initiative. The tomato genome sequencing project uses an ordered bacterial artificial chromosome (BAC) approach to generate a high-quality tomato euchromatic genome sequence for use as a reference genome for the Solanaceae and euasterids. Sequence is deposited at GenBank and at the SOL Genomics Network (SGN). Currently, there are around 1000 BACs finished or in progress, representing more than a third of the projected euchromatic portion of the genome. An annotation effort is also underway by the International Tomato Annotation Group. The expected number of genes in the euchromatin is ~40,000, based on an estimate from a preliminary annotation of 11% of finished sequence. Here, we present this first snapshot of the emerging tomato genome and its annotation, a short comparison with potato (*Solanum tuberosum* L.) sequence data, and the tools available for the researchers to exploit this new resource are also presented. In the future, whole-genome shotgun techniques will be combined with the BAC-by-BAC approach to cover the entire tomato genome. The high-quality reference euchromatic tomato sequence is expected to be near completion by 2010.

THE SOLANACEAE, also called nightshades, is a medium-sized flowering plant family of >9000 species, including economically important species such as tomato (*Solanum lycopersicum* L.), potato (*Solanum tuberosum* L.), pepper (*Capsicum annuum* L.), eggplant (*Solanum melongena* L.), tobacco (*Nicotiana tabacum* L.), and petunia (*Petunia ×hybrida* Vilm.) (Knapp et al., 2004). Species of Solanaceae occur on all continents except Antarctica and are very diverse in habit—from trees to tiny annuals—and habitat—from deserts to tropical rainforests. Members of the family also serve as scientific model plants, for the study of fruit development (Gray et al., 1992; Fray and Grierson, 1993; Brummell and Harpster, 2001; Alexander and Grierson, 2002; Adams-Phillips et al., 2004; Giovannoni, 2004; Tanksley, 2004; Seymour et al., 2008), tuber development (Prat et al., 1990; Bachem et al., 1996; Fernie and Willmitzer, 2001), biosynthesis of anthocyanin and carotenoid pigments (Gerats et al., 1985; Giuliano et al., 1993; Mueller

L.A. Mueller, J.J. Giovannoni, R. White, J. Vrebalov, Z. Fei, J. van Eck, R. Buels, A. Mills, N. Menda, I.Y. Teclé, and A. Bombarely, Boyce Thompson Institute, Ithaca, NY 14853; S.D. Tanksley, Dep. Plant Breeding, Cornell Univ., Ithaca, NY 14853; S. Stack, S.M. Royer, S.-B. Chang, and L.A. Shearer, Dep. of Biology, Colorado State Univ., Fort Collins, CO 80523; S.-H. Jo and C.-G. Hur, Plant Genome Research Center, KRIBB, Taejeon 305-600, Korea; B.D. Kim and D. Choi, Seoul National Univ., San 56-1 Shinlim-dong, Gwanak-gu, Seoul 151-742, Korea. Received 21 Aug. 2008. *Corresponding author (lam87@cornell.edu).

[Affiliations continued on next page.]

Published in The Plant Genome 2:78–92. Published 18 Mar. 2009.

doi: 10.3835/plantgenome2008.08.0005

© Crop Science Society of America

677 S. Segoe Rd., Madison, WI 53711 USA

An open-access publication

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

Abbreviations: AGP, Accessioned Golden Path; BAC, bacterial artificial chromosome; COS, conserved ortholog set; EST, expressed sequence tag; FISH, fluorescent in situ hybridization; FPC, fingerprinted contig; HMM, hidden Markov model; HTGS, high-throughput genome sequence; ITAG, International Tomato Annotation Group; LRR, leucine-rich repeats; LZ, leucine zippers; NBS, nucleotide binding sites; NCBI, National Center for Biotechnology Information; PlantGDB, Plant Genome Database; PUT, PlantGDB-assembled unique transcripts; R-genes, resistance genes; SGN, SOL Genomics Network; SSR, simple sequence repeat; TF, transcription factor; Tm, trans-membrane; WGS, whole-genome shotgun.

[Author list continued.]

Ying Wang, Graham B. Seymour, Gerard J. Bishop, Glenn Bryan, Jane Rogers, Sarah Sims, Sarah Butcher, Daniel Buchan, James Abbott, Helen Beasley, Christine Nicholson, Clare Riddle, Sean Humphray, Karen McLaren, Saloni Mathur, Shailendra Vyas, Amolkumar U. Solanke, Rahul Kumar, Vikrant Gupta, Arun K. Sharma, Paramjit Khurana, Jitendra P. Khurana, Akhilesh Tyagi, Sarita, Parul Chowdhury, Smriti Shridhar, Debasis Chattopadhyay, Awadhesh Pandit, Pradeep Singh, Ajay Kumar, Rekha Dixit, Archana Singh, Sumera Praveen, Vivek Dalal, Mahavir Yadav, Irfan Ahmad Ghazi, Kishor Gaikwad, Tilak Raj Sharma, Trilochan Mohapatra, Nagendra Kumar Singh, Dóra Szinay, Hans de Jong, Sander Peters, Marjo van Staveren, Erwin Datema, Mark W.E.J. Fiers, Roeland C.H.J. van Ham, P. Lindhout, Murielle Philippot, Pierre Frasse, Farid Regad, Mohamed Zouine, Mondher Bouzayen, Erika Asamizu, Shusei Sato, Hiroyuki Fukuoka, Satoshi Tabata, Daisuke Shibata, Miguel A. Botella, M. Perez-Alonso, V. Fernandez-Pedrosa, Sonia Osorio, Amparo Mico, Antonio Granell, Zhonghua Zhang, Jun He, Sanwen Huang, Yongchen Du, Dongyu Qu, Longfei Liu, Dongyuan Liu, Jun Wang, Zhibiao Ye, Wencai Yang, Guoping Wang, Alessandro Vezzi, Sara Todesco, Giorgio Valle, Giulia Falcone, Marco Pietrella, Giovanni Giuliano, Silvana Grandillo, Alessandra Traini, Nunzio D'Agostino, Maria Luisa Chiusano, Mara Ercolano, Amalia Barone, Luigi Frusciante, Heiko Schoof, Anika Jöcker, Rémy Bruggmann, Manuel Spannagl, Klaus X.F. Mayer, Roderic Guigó, Francisco Camara, Stephane Rombauts, Jeffrey A. Fawcett, Yves Van de Peer, Sandra Knapp, Dani Zamir, and Willem Stiekema

[Affiliations continued.]

C.-B. Li, J. Zhao, H. Jiang, Y. Geng, Y. Dai, H. Fan, J. Chen, F. Lu, J. Shi, S. Sun, J. Chen, X. Yang, C. Lu, M. Chen, Z. Cheng, C. Li, H. Ling, Y. Xue, and Y. Wang, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China; G. Seymour, Division of Plant Sciences, Univ. of Nottingham, Sutton Bonington, LE12 5RD, UK; G.J. Bishop, S. Butcher, D. Buchan, and J. Abbott, Imperial College London, London, SW7 2AZ, UK; G. Bryan, SCRI Invergowrie, Dundee, DD2 5DA, UK; S. Mathur, S. Vyas, A.U. Solanke, R. Kumar, V. Gupta, A.K. Sharma, P. Khurana, J.P. Khurana, and A. Tyagi, Univ. of Delhi South Campus, New Delhi, 110 02, India; Sarita, P. Chowdhury, S. Shridhar, and D. Chattopadhyay, National Institute for Plant Genome Research, New Delhi, 110 067, India; A. Pandit, P. Singh, A. Kumar, R. Dixit, A. Singh, S. Praveen, V. Dalal, M. Yadav, I.A. Ghazi, K. Gaikwad, T.R. Sharma, T. Mohapatra, and N.K. Singh, NRC on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi, 110 012, India; R. Klein lankhorst, R.C.H.J. van Ham, and W. Stiekema, Centre for BioSystems Genomics, P.O. Box 98, 6700 AB, Wageningen, Netherlands; D. Szinay, H. de Jong, S. Peters, and P. Lindhout, Wageningen Univ., Lab. of Genetics, Arboretumlaan 4, 6703 BD, Wageningen, Netherlands; M. van Staveren, E. Datema, M.W.E.J. Fiers, and R.C.H.J. van Ham, Plant Research International, Droevendaalsesteeg 1, 6708 PB, Wageningen, Netherlands; M. Philippot, P. Frasse, F. Regad, M. Zouine, and M. Bouzayen, UMR990, INRA, chemin de Borde Rouge, 31326 Castanet-Tolosane, France; E. Asamizu, S. Sato, S. Tabata, and D. Shibata, Kazusa, Kisarazu, 292-0818, Chiba, Japan; S. Osorio, A. Mico, and A. Granell, Instituto de Biología Molecular y Celular de Plantas, CSIC/ Universidad Politécnica de Valencia Ciudad Politécnica de la Innovación - Edificio 8E, 46 011 Valencia, Spain; M.A. Botella, Univ. of Málaga, Campus de Teatinos, 29071 Málaga, Spain; G. Falcone, M. Pietrella, and G. Giuliano, ENEA, Casaccia Research Center, Via Anguillarese 301, 00123 Rome, Italy; A. Traini, N. D'Agostino, M.L. Chiusano, M. Ercolano, A. Barone, and L. Frusciante, Dep. of Soil, Plant, Environmental and Animal Production Sciences, Univ. of Naples "Federico II", Via Università 100, 80055 Portici, Italy; D. Zamir, Hebrew Univ., P.O. Box 12, Rehovot 76100, Israel; H. Fukuoka, National Institute of Vegetable and Tea

Science, National Agriculture Research Organization, 360 Kusawa, Ancho, Tsu-shi, Mie 514-2392, Japan (NIVTS); J. Rogers, S. Sims, H. Beasley, C. Nicholson, C. Riddle, and K. McLaren, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK; Z. Zhang, J. He, S. Huang, Y. Du, and D. Qu, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China; L. Liu, D. Liu, and J. Wang, Beijing Genomics Institute, Shenzhen 518083, China; Z. Ye, College of Horticulture and Forestry, Huazhong Agricultural Univ., Wuhan, China; W. Yang, College of Agronomy and Biotechnology, China Agricultural Univ., Beijing 100094, China; G. Wang, Dep. of Horticulture, South China Agricultural Univ., Guangzhou, China; H. Schoof and A. Jöcker, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany; M. Perez-Alonso and V. Fernandez-Pedrosa, Sistemas Genómicos, SL, Avenida Benjamín Franklin, 46980 Paterna, Valencia, Spain; R. Guigó and F. Camara, Centre de Regulació Genòmica, Universitat Pompeu Fabra, Dr. Aiguader, 88, 08003 Barcelona, Spain; S. Humphray, Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Saffron Walden, Essex, CB10 1XL, UK; S. Rombauts, J.A. Fawcett, and Y. van de Peer, VIB/Ghent Univ., Technologiepark 927, 9052 Ghent, Belgium; S. Knapp, Botany Dep., The Natural History Museum, Cromwell Rd., London, SW7 5BD, UK; R. Bruggmann, Rutgers, The State Univ. of New Jersey, Waksman Institute of Microbiology, 190 Frelinghuysen Rd., Piscataway, NJ 08854-8020; M. Spannagl and K.X.F. Mayer, MIPS/Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany; A. Vezzi, S. Todesco, and G. Valle, CRIBI, Univ. of Padua, via U. Bassi, 58/b-35131 Padua, Italy; S. Grandillo, CNR, Institute for Plant Genetics, Portici, Via Università 133, 80055 Portici, Italy; L.A. Mueller, Z. Fei, R. Buels, C.-G. Hur, D. Buchan, S. Mathur, E. Datema, M.W.E.J. Fiers, A. Traini, N. D'Agostino, M.L. Chiusano, H. Schoof, A. Jöcker, R. Bruggmann, M. Spannagl, K.X.F. Mayer, R. Guigó, F. Camara, S. Rombauts, J.A. Fawcett, and Y. van de Peer, International Tomato Genome Annotation Group (ITAG); J.J. Giovannoni and R. White, USDA-ARS, Tower Road, Ithaca, NY 14853, USA.

et al., 2000; Spelt et al., 2002; De Jong et al., 2004; Quattrocchio et al., 2006), and plant defense (Bogdanove and Martin, 2000; van der Vossen et al., 2000; Gebhardt and Valkonen, 2001; Kessler and Baldwin, 2001; Li et al., 2001; Bai et al., 2003; Hui et al., 2003; Pedley and Martin, 2003; Sacco et al., 2007). The Solanaceae have also attracted interest because they produce a number of specialized metabolites that have medicinal properties (Schijlen et al., 2006; Oksman-Caldentey, 2007). The Solanaceae are remarkable in that the gene content of the different species remains similar despite the highly varied phenotypic outcomes (Tanksley et al., 1992; Knapp et al., 2004). This makes the Solanaceae an excellent model for the study of plant adaptation to natural and agricultural environments (Knapp et al., 2004). Most species of the Solanaceae are diploid and share a basic set of 12 chromosomes (Olmstead et al., 1999); recent polyploidizations during the evolutionary history of the family are limited to a few clades such as the potatoes and tobaccos (Clarkson et al., 2005).

A Solanaceae reference genome will be an invaluable resource in addressing two fundamental biological questions: first, how genomes code for extensive phenotypic differences using relatively conserved sets of genes; and second, how phenotypic diversity can be harnessed for the improvement of agricultural products. Sequence data from other species, such as expressed sequence tags (ESTs) (Adams et al., 1991), methylation (Palmer et al., 2003; Whitelaw et al., 2003; Fu et al., 2004), or Cot-filtered sequence (Peterson et al., 2002; Yuan et al., 2003), together with sequencing by novel very high throughput approaches such as 454 sequencing (Margulies et al., 2005) or Solexa sequencing (Shendure et al., 2005) in combination with good comparative maps (Tanksley et al., 1992; Doganlar et al., 2002; Fulton et al., 2002) between many Solanaceae plants (Hoeven et al., 2002; D'Agostino et al., 2007), will enable insights into evolution, domestication, development, response, and signal transduction pathways.

After the sequencing of a number of dicots from the rosid clade (Angiosperm Phylogeny Group, 2003), *Arabidopsis thaliana* L. (AGI, 2000), *Medicago truncatula* Gaertn. (Cannon et al., 2006) using bacterial artificial chromosome (BAC)-by-BAC approaches, and poplar [*Populus trichocarpa* (Torr. & A. Gray)] (Tuskan et al., 2006), grape (*Vitis vinifera* L.) (Jaillon et al., 2007), and others using whole-genome shotgun (WGS) techniques, the sequencing of the first genome in the asterids will shed light on this clade, permitting longer-range evolutionary distance comparisons and provide information about the larger picture of angiosperm evolution.

Ten countries are involved in sequencing the tomato genome and the 12 chromosomes have been allocated among the countries as depicted in Fig. 1. The chloroplast genome was recently completed by a European consortium (Kahlau et al., 2006) and the mitochondrial genome is being sequenced by the Instituto Nacional de Tecnología Agropecuaria in Argentina within the framework of

the EU-SOL project (<http://www.eu-sol.net> [verified 10 Jan. 2009]).

The 950-Mb tomato genome is structured into distal, gene-rich euchromatin and gene-poor pericentromeric heterochromatin. The heterochromatic fraction, consisting mostly of repetitive sequences, will be extremely difficult to sequence. Therefore, the strategy is to initially sequence the euchromatic portions of the genome, which is estimated to make up one-quarter (220 Mb) of the tomato genomic sequence (Peterson et al., 1996) including >90% of the genes (Wang et al., 2006). As a consequence, the effort to sequence the majority of the gene space is less than twice the effort required to sequence the *Arabidopsis* genome at 157 Mb (Bennett et al., 2003).

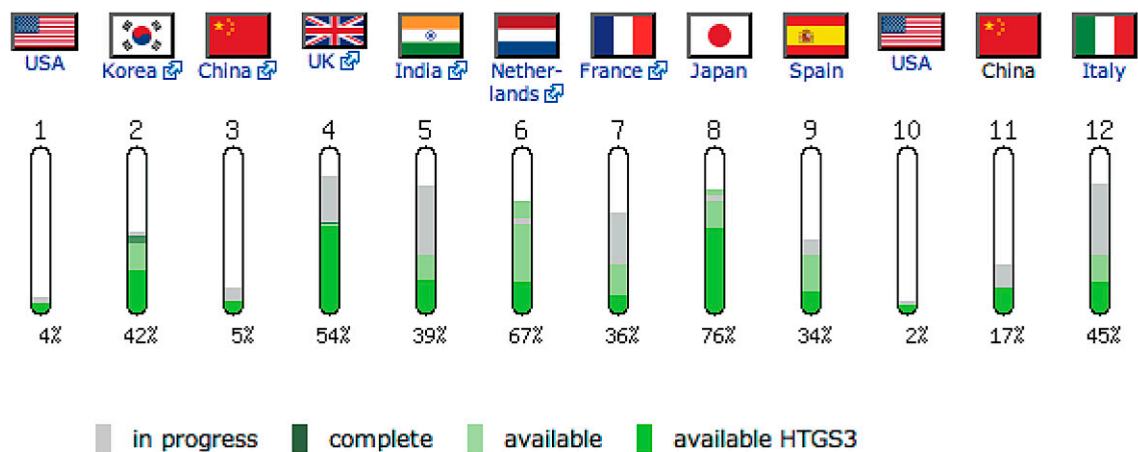
To render the emerging tomato sequence immediately useful to the community, it is being annotated by the International Tomato Annotation Group (ITAG). Annotations are available on the SOL Genomics Network (SGN) website (<http://sgn.cornell.edu/> [verified 10 Jan. 2009]), and a number of Web-based tools have been developed that allow researchers to download and analyze the emerging sequence.

Here, we provide a summary of the status of the project and relevant insights drawn from the annotation of the tomato genome performed to date.

Results and Discussion

To sequence the tomato euchromatin, a BAC-by-BAC approach was chosen in preference to a WGS strategy. This will generate a high-quality “gold standard” sequence, which is essential for use as a reference genome (International Rice Genome Sequencing Project, 2005) and which will serve as the scaffold for the related Solanaceae genomes. In short, the BAC-by-BAC strategy involves the anchoring of BACs or contigs of BACs to a reference genetic map. These anchored BACs are sequenced, and the sequence information is used to extend these BACs and BAC contigs further (“BAC walking”). Gaps between BAC contigs are closed by targeting novel markers or BACs to these gaps, which is then followed by successive rounds of BAC walking.

The high-density F_2 -2000 map (Fulton et al., 2002) is used as a reference genetic map for the sequencing project. This map is based on 80 F_2 individuals from the cross *Solanum lycopersicum* LA925 \times *S. pennellii* Correll LA716 and contains a subset of restriction fragment length polymorphism markers from the Tomato-EXPEN 1992 map (Tanksley et al., 1992). Most of the markers are conserved ortholog set (COS) markers (Fulton et al., 2002; Wu et al., 2006) derived from a comparison of Solanaceae ESTs against the entire *Arabidopsis* genome. Those COS markers selected were single–low copy, having a highly significant match with a putative orthologous locus in *Arabidopsis*. Maps constructed using COS markers can readily be compared and analyzed for chromosome inversions, duplications, and other large-scale genome rearrangements, a characteristic that will



BACs													Total
Chr Total	391	268	274	193	111	213	277	175	164	186	135	113	2,500
In progress	8	2	20	54	48	5	92	5	13	1	18	51	317
Complete	11	159	0	87	20	152	53	133	67	4	6	20	712
Available	19	113	15	105	44	144	100	133	57	4	23	51	808
HTGS 1	5	0	0	0	0	116	63	0	13	0	1	21	219
HTGS 2	0	0	0	0	28	0	36	2	30	0	10	11	117
HTGS 3	14	65	15	103	20	36	21	90	18	4	22	19	427
% Done	3%	47%	4%	56%	35%	55%	28%	69%	35%	2%	13%	34%	

Overall Stats

30% of sequencing is complete

29% of BACs are reported finished

32% of BACs have downloadable sequence

Figure 1. Status of the tomato euchromatin sequence as of September 2008. For each chromosome the responsible country is shown. Progress in the sequencing of each chromosome (Chr) is given, as well as the status and the availability of the bacterial artificial chromosomes (BACs). HTGS, high-throughput genome sequence.

be useful for transferring knowledge from tomato to other species. In addition to COS markers, the map also contains a significant number of simple sequence repeat (SSR) markers, most of which were identified in ESTs (usually in 5' or 3' untranslated regions).

The BACs used in the tomato sequencing project are derived from several libraries, all of which were constructed from the Heinz 1706 tomato line. In addition to a *Hind*III library consisting of 129,024 clones that was available at the outset of the project (Budiman et al., 2000), two additional BAC libraries were generated, an *Eco*RI library of 72,264 clones and an *Mbo*I library of 52,992 clones. Together, these libraries provide more than 25× genome coverage. The BAC libraries have been deep end-sequenced in the United States, with >340,000 high-quality reads equivalent to 20% of the entire genome sequence. The BAC libraries are complemented by a fosmid library. Currently, >180,000 high-quality fosmid end sequences from the Wellcome Trust Sanger Institute and the University of Padua are available, equivalent to 15% of the entire genome sequence. Fosmid libraries are crucial in a genome sequencing project because their narrowly defined insert length can be used as an analytical tool to detect potential misassemblies of BACs, and their generally shorter insert

length is ideal for filling smaller gaps and thereby reducing redundant sequence (Kim et al., 1995). The fosmid library is cut using shearing rather than restriction enzymes to obtain clone coverage in regions low or devoid of the relevant restriction sites.

All BACs from the *Hind*III library and from the *Mbo*I library were fingerprinted and contigs of overlapping BACs were generated using the fingerprinted contigs (FPC) tool (Soderlund et al., 2000). First, an analysis of the BAC fingerprint data yielded 6000 contigs, of which >3500 could be anchored to the genetic map. In an effort to globally reduce the number of contigs, the entire FPC data were reassembled using less stringent assembly criteria (cutoff *E*-value of 1×10^{-12} and tolerance of 7). This resulted in 4360 contigs representing about 658 Mb of sequence. To increase the contig size and to reduce the contig number further, the contigs were manually edited with anchoring information by contig end-search and merging, resulting in 4156 contigs.

Finally, a total of 837 markers were used to anchor the contigs to the tomato genetic map. The anchored contigs represent about 187 Mb of genomic DNA and are mainly composed of euchromatic sequences from the tomato genome.

Validation of the physical map was performed using fluorescent in situ hybridization (FISH) on pachytene complements with entire BAC clones as probes (Chang et al., 2007; Szinay et al., 2008) (see also FISH map on SGN, http://sgn.cornell.edu/cview/map.pl?map_id=13 [verified 10 Jan. 2009]), and by genetic mapping of anchored BACs using panels of tomato introgression line populations (Eshed and Zamir, 1995). The integrated map is available through WebFPC.

Since the current sequencing effort focuses on the tomato euchromatin, determining the chromosomal borders between euchromatin and heterochromatin is essential. Currently, we use FISH to identify BAC inserts from euchromatin–heterochromatin boundaries based on linkage map information and on the specific staining by FISH of the repetitive fraction of the tomato genome (Szinay et al., 2008); see Fig. 2.

In a multinational project, it is important that all participants use the same standards for completing their sequences. The Tomato Genome Project started to develop these standards early on, and they will be maintained and developed when new issues arise. The full quality standards are described in the Tomato Sequencing Guidelines document available online at http://docs.google.com/View.aspx?docid=dggs4r6k_1dd5p56 (verified 5 Feb. 2009).

In summary, the BACs are being sequenced to the following quality standards:

The BAC sequence submitted in high-throughput genome sequence (HTGS) Phase3 consists of a single contig.

All bases of the HTGS Phase 3 consensus sequence must have a Phred quality score of at least 30.

As a result of the shotgun process, the bulk of sequence will be derived from multiple subclones sequenced from both strands. Any regions of unidirectional sequence coverage with a single sequencing chemistry must pass manual inspection for sequence problems but need not be annotated. Regions covered by only a single subclone must be attempted from an alternate subclone or by direct walking on BAC DNA or by BAC polymerase chain reaction. These regions must concur with a restriction digest analysis of the clone. In addition, these regions must be annotated.

At least 99% of the sequence must have less than one error in 10,000 bp as reported by Phrap or other sequence assembly consensus scores. Exceptions must be manually checked and pass inspection for possible problems. Any areas not meeting this standard must be annotated as such.

To date (September 2008), 689 BACs have been sequenced and reported in the SGN BAC registry database (either HTGS Phase 2 or Phase 3) (Fig. 1), representing 74.8 Mb (including overlaps) (available from SGN and GenBank). Of these, 419 are included in the Accessioned Golden Path (AGP) files, which can be viewed in the SGN AGP map representing 44.5 Mb of sequence, representing roughly 20% of the tomato euchromatin. These BACs have been placed into 282 contigs and have been annotated using the ITAG annotation pipeline; see below.

Genome Annotation by ITAG

To render the sequence immediately useful to the community, ITAG is producing a high-quality automated annotation of the tomato genome in a distributed collaborative effort, which involves groups from Europe, Asia, and the United States. The centerpiece of the structural annotation is the EuGene gene prediction platform (Foissac et al., 2008), a powerful predictor capable of integrating a diverse array of inputs, such as evidence-based alignments and ab initio predictions. For the functional annotation, InterPro domains are determined using InterProScan and homology searches are performed. Where possible, other sequence features (i.e., noncoding RNAs) are predicted. An important initial activity of the ITAG group was to generate a training and test set of gene sequences to train gene finders for tomato. Gene finders that are trained or have been trained include EuGene (Foissac et al., 2008), GeneMark (Isono et al., 1994), TwinScan (Korf et al., 2001), and Augustus (Stanke et al., 2008). Results of predicted gene models and their functional annotations are available via the SGN Web site.

In the first batch of annotations partially based on as yet untrained gene finders, the ITAG pipeline has

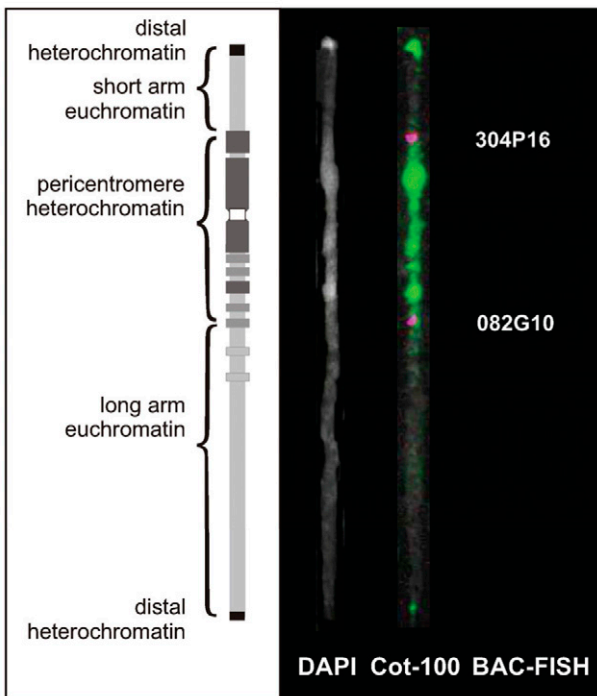


Figure 2. Labeling of the heterochromatic part of tomato chromosome 6 by fluorescent in situ hybridization (FISH) with the Cot-100 genomic DNA fraction (green signal). The differently labeled bacterial artificial chromosome (BAC) clones resident in the heterochromatin–euchromatin borders of the short arm and of the long arm are pseudocolored in red and magenta. DAPI, 4',6'-diamidino-2-phenylindole, dihydrochloride.

identified 7464 protein coding genes longer than 180 nucleotides in 44 Mb of nonredundant sequence. This represents a gene density of approximately one gene per 6 kb, slightly lower than the density of one gene per ~4.5 kb in *Arabidopsis* (AGI, 2000) but is higher than one protein coding gene in 9.9 kb in the rice (*Oryza sativa* L.) genome (International Rice Genome Sequencing Project, 2005). The average coding sequence is 996 bp long and is composed of 3.7 exons. The primary difference between tomato and *Arabidopsis* genes is that tomato genes, including their introns, are longer. The average gene length from this analysis is ~2 kb, with an average intron length of 485 bp and an average exon length of 268 bp, significantly larger than those in *Arabidopsis*. While the lower number of exons per gene almost certainly represents the current lower annotation quality of tomato genes, it is notable that the average intron length is more than twice that in *Arabidopsis*. Assuming a gene density of one gene per 6 kb in the rest of the tomato euchromatin, we can expect that the euchromatin of the tomato genome contains just over 40,000 genes, close to the estimated number of about 35,000 (Hoeven et al., 2002). Obviously, some of these parameters may change with improved tomato genome annotations and the further improvement of trained tomato gene finders. Figure 3 shows the number of tomato genes falling into certain annotation categories, and a comparison to the numbers in the categories found in *Arabidopsis*, rice, and poplar. The numbers in each category are similar between species, indicating that the fraction of the tomato sequence that has so far been sequenced is similar to other plant genomes.

De novo repeat analysis was performed on the available BAC-end sequences, and the resulting repeats were used to analyze both the BAC end sequences as well as the complete BAC sequences. The de novo repeat set masked 57% of BAC-ends and 24% of full BAC sequence, indicating that the BACs selected from the euchromatin contain fewer repeats than the genome as a whole. These results support the recently described distribution of tomato repetitive sequences as determined by FISH (Chang et al., 2008). The fraction of long terminal repeat elements was much higher in BAC-ends (30%) than in the full BAC sequences (12.6%), indicating that there are large differences in the nature of repeats occurring in different genome regions.

The distribution of repeats and gene content on selected chromosomes is shown in Fig. 4, defined by repeat analysis and EST coverage. The information is reported only for those chromosomes for which Tiling Path Format files, which represent the tentative order of the BACs in the chromosome assembly as provided by the sequencing centers, are available at the SGN Web site to date. The following number of BACs were analyzed for each chromosome: chromosome 4, 94; chromosome 5, 35; chromosome 6, 100; chromosome 9, 43; and chromosome 12, 34. This analysis includes a number of BACs that were attributed to heterochromatin but nevertheless have been sequenced. The bars in each panel represent the percentage of nucleotides in a BAC that could be aligned to *Solanum*

lycopersicum ESTs (blue bars) and repeat sequences (red bars). Figure 4 shows that the repeats are much lower in abundance in the euchromatic arms and in some cases form a gradient of increasing density into the heterochromatin, whereas on other arms the transition appears less gradual. Also, in general, the gene-rich BACs have lower repeat content, supporting the general assumption that genes are predominantly present in the relatively repeat-poor euchromatin. The tomato heterochromatin consists of the bulk of the repetitive DNA fraction, which nevertheless also contains some genes as has been described by Yasuhara and Wakimoto (2006).

Transcription factors (TFs) play key roles in regulation of gene expression in various biological processes. The assembled ESTs (Plant Genome Database [PlantGDB]–assembled unique transcripts [PUTs]) of *Solanum lycopersicum* from PlantGDB were searched for putative TFs using hidden Markov model (HMM) profiles, which resulted in the identification of 1463 such PUTs that included 66 of the 71 known TF gene families. Considering that 40,000 genes are predicted in the tomato genome (Hoeven et al., 2002), this indicates that ~3.6% of the total genes in the euchromatic region may be TFs. For *Arabidopsis*, 5.9 to 7% (Riechmann et al., 2000; Riano-Pachon et al., 2007) and rice, 4% (Goff et al., 2002; Riano-Pachon et al., 2007) of the total genes are TFs. Further, 237 PUTs (16%) encoding putative TFs could be mapped on 559 tomato BACs, representing around 56 Mb sequenced tomato genome. On average, one TF gene is present in every 200 kb (assuming average BAC size to be 100 kb); see Table 1. Chromosomes 12 and 11 seem to harbor the highest and lowest density of TF genes, respectively. The major three TF gene families in tomato include AP2-EREBP (APETALA2-ethylene responsive element binding protein), MYB, and bHLH (basic helix-loop-helix) families (not shown).

Sequence analysis of cloned plant disease resistance genes (R-genes) conferring resistance to viral, bacterial, and fungal pathogens has shown that the majority of them possess common sequences and structural motifs. These R-genes can be grouped into three major classes (NBS-LRR type, LZ-NBS-LRR type, or LRR-Tm type) on the basis of their encoded protein motifs such as leucine zippers (LZ), nucleotide binding sites (NBS), leucine-rich repeats (LRR), protein kinases domains, trans-membrane (Tm) domains, and Toll-IL-IR homology regions. We analyzed 48,945 unigene (PUT) sequences of tomato from PlantGDB for the presence of R-gene homologs by a BLASTX analysis against the nonredundant database of the National Center for Biotechnology Information (NCBI) and classified them into the above three categories. The PUT matches to different putative R-genes and LRR motifs only were grouped into the miscellaneous R-gene category. In addition, defense response genes such as glucanases, chitinase, and thaumatin-like proteins were also included in the analysis.

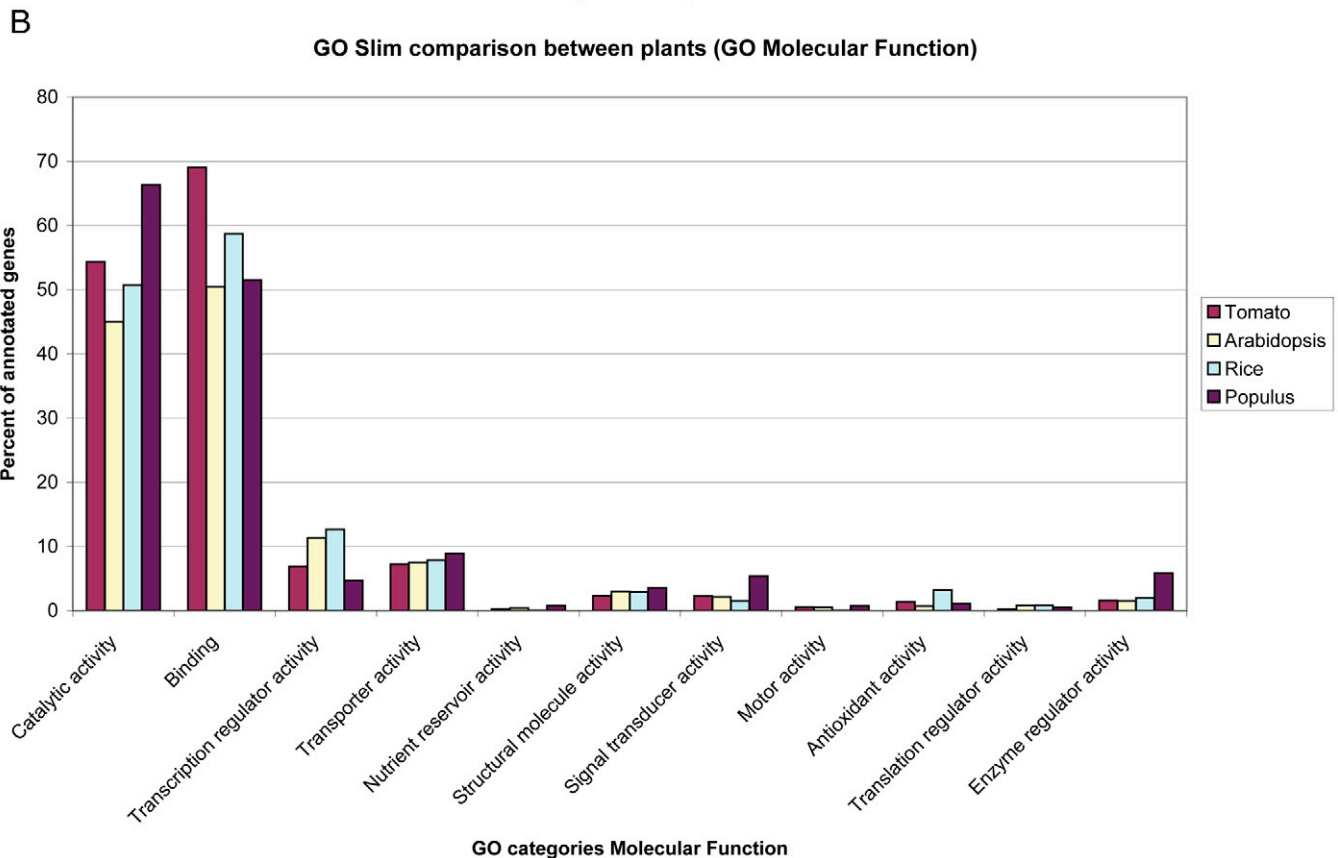
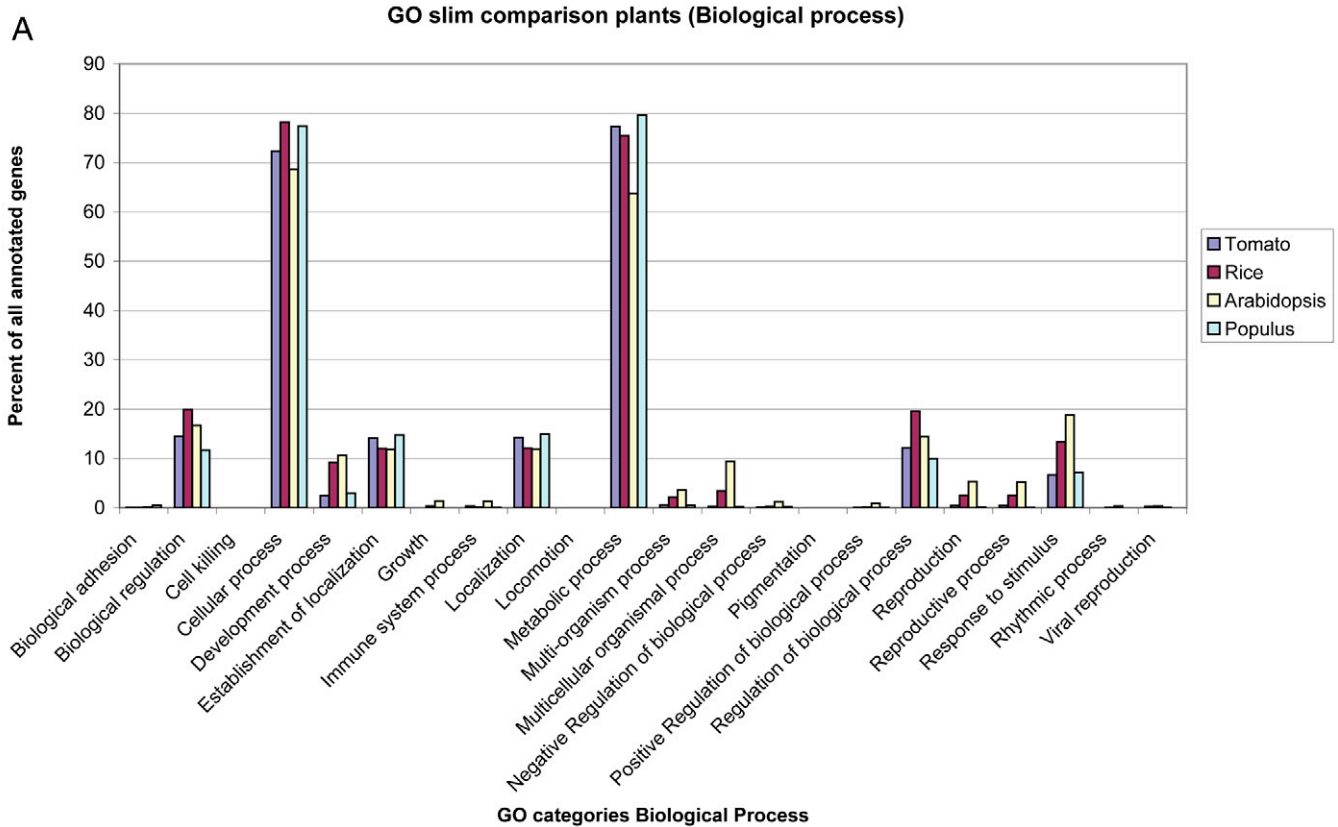


Figure 3. Annotation categories for the annotated tomato genes from the International Tomato Annotation Group annotation pipeline and comparison to categories in *Arabidopsis*, poplar, and rice. (A) Annotation statistics categorized by higher-level gene ontology (GO) biological process terms. (B) Annotation statistics categorized by GO molecular function terms.

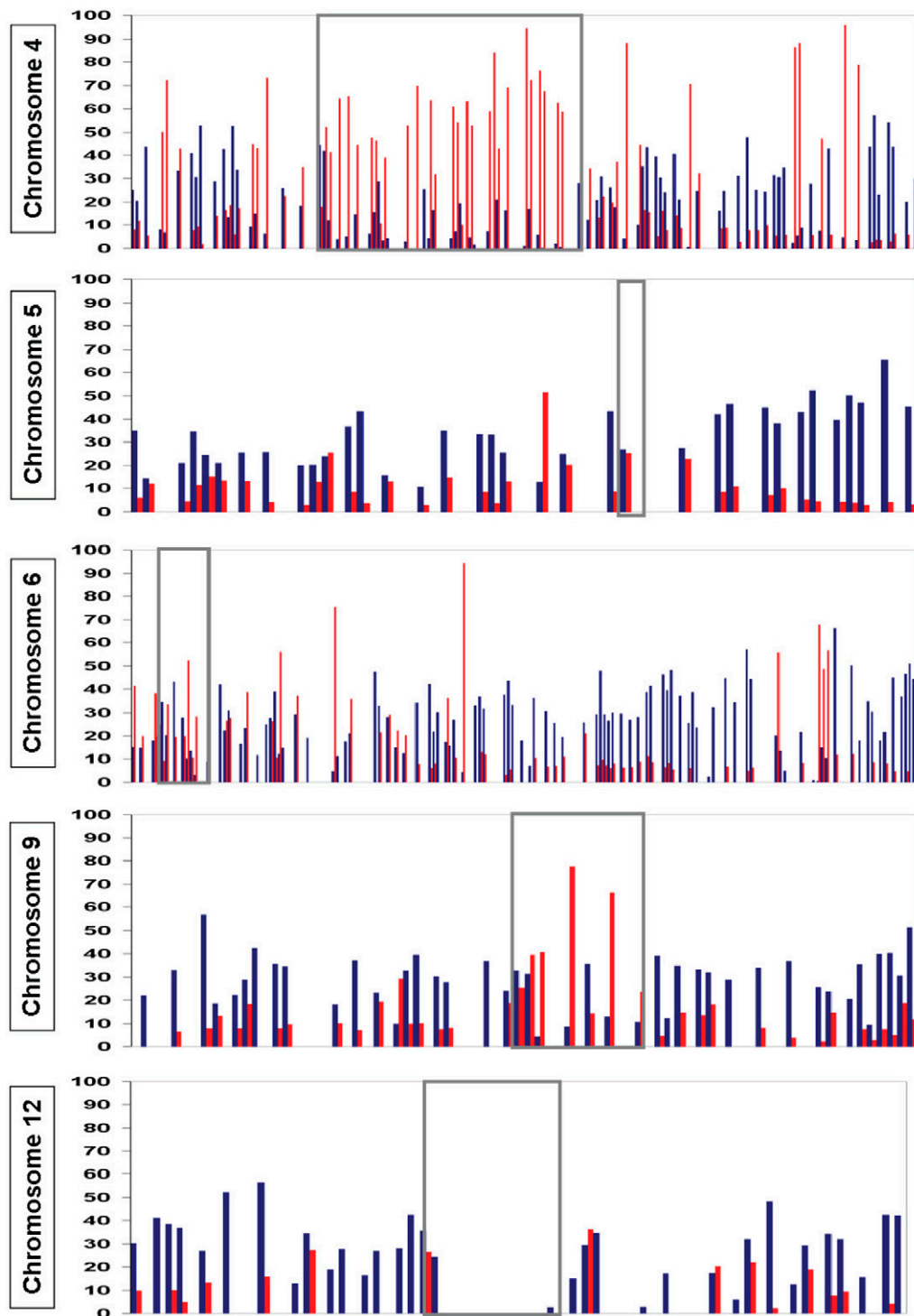


Figure 4. Gene and repeat coverage for selected tomato chromosomes (4, 5, 6, 9, and 12). The bacterial artificial chromosomes (BACs) are arranged in the order they appear along the chromosome. For each BAC, the percentage of expressed sequence tag (blue bars) and repeat (red bars) coverage are shown. The gray rectangle defines the pericentromeric heterochromatic region in each chromosome. The data shown in this figure are available for all the chromosomes under sequencing and are available through the "Genome Overview" at <http://biosrv.cab.unina.it/GBrowse/> (verified 16 Jan. 2009). The data are updated at each new BAC release in GenBank. Updated versions of this figure are provided on unordered BACs and are available at <http://biosrv.cab.unina.it/GBrowse/Graphs/graphall1.html> (verified 16 Jan. 2009).

We found a total of 155 annotations similar to resistance-like genes and 83 annotations showed homology to the defense-response-like genes (Fig. 5).

These R-gene and defense-response gene homologs were mapped in silico onto the sequenced BACs of the different chromosomes to find their physical locations, resulting in the localization of 59 R-gene homologs and

Table 1. Distribution of transcription factors on different tomato chromosomes.

	1	2	3	4	5	6	7	8	9	10	11	12
Chromosome size (Mb)	108	85.6	83.6	82.1	80	53.8	80.3	64.7	81.8	88.5	64.7	76.4
No. of BACs [†] analyzed	9	86	14	88	34	76	87	85	45	4	16	15
No. of transcription factors	4	41	4	31	10	42	34	31	21	3	3	13
No. of transcription factors per BAC	0.44	0.47	0.28	0.35	0.29	0.55	0.40	0.36	0.46	0.75	0.18	0.86

[†]BACs, bacterial artificial chromosomes.

of 21 defense-response gene homologs (see Table 2). Thus, the mapped resistance-like and defense-response genes represent about one-third of all expressed PUTs assembled from the tomato EST database. Since the number of BACs analyzed per chromosome varied considerably, we normalized the frequency of these genes per BAC clone to evaluate their relative distribution on different tomato chromosomes. Based on this analysis, chromosomes 4, 9, and 11 seem to harbor a larger than average number of R-gene homologs per BAC, whereas chromosome 5 has the largest number of defense-response genes per BAC. However, this may change as more sequence data become available, particularly from chromosomes 1, 3, 10, and 11, which were underrepresented when this analysis was undertaken.

Comparison to Potato Sequence

An initial effort was made to compare the gene and repeat content of the tomato and potato genomes, based on the available BAC-end sequences for both species (Datema et al., 2008). The BAC-end sequence comparison is of particular interest as it provides a picture for the complete genome, including both euchromatic and heterochromatic sequence. Comparison using only sequenced tomato BACs will mainly provide a comparison between the euchromatin of tomato and potato. In total, 310,580 BAC-end sequences representing ~19% of the 950-Mb tomato genome were compared to 128,819 potato BAC-end sequences representing ~10% of the 840-Mb potato genome. It is important to note that while most potato varieties used in agriculture are tetraploid, the potato line being sequenced is diploid (van Os et al., 2006).

The tomato genome has a higher overall dispersed repeat content than the potato genome, with the majority of dispersed repeats in both species belonging to the

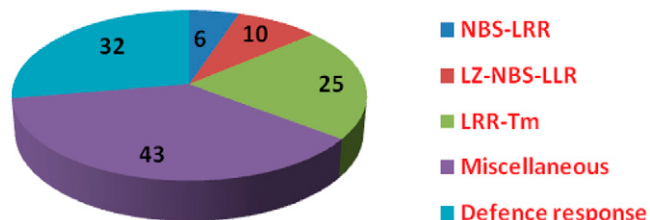


Figure 5. Different categories of disease-resistance-like genes in the tomato unigene set. These genes can be grouped into three major classes (NBS-LRR type, LZ-NBS-LRR type, or LRR-Tm type) on the basis of their encoded protein motifs such as leucine zipers (LZ), nucleotide binding sites (NBS), leucine-rich repeats (LRR), and trans-membrane (Tm) domains.

Gypsy and Copia retrotransposon families. Specifically, the Copia:Gypsy ratio is higher in tomato than in potato, suggesting that the retrotransposon amplification associated with the genome expansion in tomato is predominantly the

result of additional Copia elements. On the other hand, simple sequence repeats (SSRs) motifs are more abundant in potato than in tomato. In both genomes penta-nucleotide repeats are the most common form of SSRs, and AAAAT is the predominant repeat motif. This is in contrast to previously studied plant species, in which di- and penta-nucleotide repeats generally occur least frequently (Asp et al., 2007).

The potato BAC-end sequences have a 1.5- to 1.6-fold higher protein coverage than tomato when aligned to the NCBI nonredundant protein database, and a 1.3- to 1.4-fold higher coverage when compared with the species-specific EST data. Taking into account the difference in genome size and assuming that tomato has ~40,000 genes, potato appears to contain up to 6400 more putative coding regions than tomato. Moreover, the P450 superfamily appears to have expanded dramatically in both species compared with *Arabidopsis thaliana* (Datema et al., 2008), suggesting an expanded network of specialized metabolic pathways in the Solanaceae.

Tomato Genome Tools Available for Researchers

A number of tools have been created for the tomato genome sequencing project that are also useful to the larger research community.

SGN Database, FTP Site, and BLAST Data Sets

All data, sequences, mapping information, and project statistics can be found on <http://sgn.cornell.edu/>.

The SGN database keeps track of the status of each BAC in the sequencing pipeline. The BACs can be searched at SGN (http://sgn.cornell.edu/search/direct_search.pl?search=bacs [verified 13 Jan. 2009]).

The Tomato Genome Browser displays the annotation for each BAC (<http://sgn.cornell.edu/gbrowse/>). All data sets can be downloaded from the SGN File Transfer Protocol (FTP) site (ftp://ftp.sgn.cornell.edu/tomato_genome/ [verified 16 Jan. 2009]), including BAC and contig sequences, BAC-end sequences, annotations in gff3 and GAME XML format, chromatograms and assembly files, and FPC raw data. The BAC-end and full BAC sequences generated in the tomato genome project, as well as tomato transcript sequences generated through other projects, are available in the SGN BLAST tool (<http://sgn.cornell.edu/tools/blast/> [verified 16 Jan. 2009]). The SGN comparative map viewer (<http://sgn.cornell.edu/cview/> [verified 16 Jan. 2009]) (Mueller et al., 2008) displays a number of genetic and physical maps for the tomato genome project.

Tomato and Potato Assembly Assistance System

The Tomato and Potato Assembly Assistance System was developed to automate the assembly and scaffolding of contig sequences for tomato chromosome 6 (Peters et al., 2006).

Morgan2McClintock

A tomato-specific data set was added to the Morgan-2McClintock tool (Lawrence et al., 2006). This tool was implemented at the MaizeGDB database (<http://www.maizegdb.org/>) and initially used the maize Recombination Nodule map (Anderson et al., 2003, 2004) to calculate approximate chromosomal positions for loci given a genetic map for a single chromosome in maize. With the new data set (Chang et al., 2007), the tool can also be used for queries related to tomato.

U Padua PABS (Platform Assisted BAC-by-BAC Sequencing)

The Platform Assisted BAC-by-BAC Sequencing pipeline (Todesco et al., 2008) is an informatics pipeline to optimize BAC-by-BAC sequencing projects.

ISOLA

An Italian SOLanaceae genomics resource, ISOL@ (<http://biosrv.cab.unina.it/isola/> [verified 16 Jan. 2009]), was designed to provide full Web access to details of the genome annotation based on experimental evidence as derived from EST–full-length cDNA sequences (Chiusano et al., 2008).

Summary and Outlook

Recently, the Tomato Genome Sequencing Project has made highly significant progress toward its goal of sequencing 220 MB of euchromatin space of the tomato genome, which has been predicted to contain the majority of tomato genes. In total, more than 950 BACs have been sequenced, representing over one-third of the targeted genome space. Sequences are being deposited at GenBank (<http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=9509> [verified 5 Feb. 2009]) and the SGN database (<http://sgn.cornell.edu/>), and are being annotated using a pipeline established by an international group (ITAG)

of bioinformatics centers. A number of tools have been created that allow both researchers and tomato breeders to work with the emerging sequence. Through the extensive comparative maps that are available, much of the information from the tomato sequence can readily be transferred to other Solanaceae and related asterids such as coffee (*Coffea canephora* L.) (Gentianales, Rubiaceae) or mint (*Mentha*) (Lamiales, Lamiaceae).

A BAC-by-BAC sequencing approach was chosen to sequence the tomato genome because it provides the highest possible sequence quality. However, since the project was started, novel “next generation” sequencing technologies have become available that are now being applied to WGS sequencing for complex genomes. The BAC-by-BAC approach has inherent advantages, and yields insights beyond sequence space as the approach is based on careful evaluation of BAC positions by genetic mapping and by FISH. For example, several inversions could be identified between the cultivated tomato and its wild relative parent used in the reference map (Tang et al., 2008). The main drawback of the BAC-by-BAC approach is that it is relatively more expensive and slower than the WGS approach. Recently, the grape genome was sequenced using a shotgun approach, resulting in >2000 unordered contigs. However, it was estimated that >95% of grape gene sequences were recovered in the sequence (Velasco et al., 2007). Thus, in the future, a hybrid approach for sequencing the tomato genome will be pursued by using WGS as an additional resource for finishing the euchromatic part of the genome and for obtaining sequence for the heterochromatic part of the genome.

A preliminary annotation of about 11% of the total assembled euchromatic space of tomato gives a gene density of one gene per 6 kb, which corresponds to an extrapolated gene count of just over 40,000 genes for the entire euchromatin, consistent with previous estimates. Notably, certain well-known tomato genes have been recovered in the genome sequence, such as R-gene alleles at the *Mi* resistance locus, the fruit shape locus *ovate*, and the *phytoene synthase 1* gene involved in carotenoid biosynthesis.

The tomato genome is repeat-rich, and analyses of BAC-end sequences, which sampled sequence from both the heterochromatin and euchromatin, revealed that about 70% of the sequence was masked and hence largely represent heterochromatin repeats. In full BAC

Table 2. Disease resistance-like and defense response-like unigenes (Plant Genome Database–assembled unique transcripts [PUTs]) mapped on the sequenced bacterial artificial chromosomes (BACs) of the 12 tomato chromosomes.

	1	2	3	4	5	6	7	8	9	10	11	12
Chromosome size (Mbp)	108	85.6	83.6	82.1	80	53.8	80.3	64.7	81.8	88.5	64.7	76.4
No. of BACs sequenced (available at SGN ¹)	19	91	15	105	42	126	100	127	57	4	18	50
Disease-resistance-like genes (PUTs) mapped	1	6	0	12	3	9	4	11	7	0	3	3
No. of resistance-like genes per BAC	0.05	0.07	0.00	0.11	0.07	0.07	0.04	0.09	0.12	0.00	0.17	0.06
No. of defense-response-like unigenes mapped	0	5	0	0	7	1	2	1	1	0	0	4
No. of defense-response-like genes per BAC	0	0.05	0	0	0.17	0.01	0.02	0.01	0.02	0	0	0.08
Total mapped resistance-like and defense-response-like genes: 59												

¹SGN, SOL Genomics Network.

sequences, which were biased toward euchromatin, only 24% of the sequence was repeat masked, confirming earlier results from FISH analyses that the repeat content of hetero- and euchromatic regions are significantly different.

In some chromosomes whose sequencing is advanced, difficulties were encountered in finding new seed BACs in the gap regions. A number of initiatives have been put in place to increase the number of seed BACs, such as additional screening of BAC library filters and markers not used in the overgo process, computational mapping of BAC ends to marker sequences, and mapping of BACs on tomato chromosomes using introgression lines (Eshed and Zamir, 1995). To find novel cleaved amplified polymorphic sequences markers, BACs were selected containing open reading frames or unique sequences at their ends. Nearly 41% of these BACs have been successfully mapped to specific tomato chromosomes in preliminary screening of a set of 120 BACs. The procedure proposed requires minimum cost and efforts to generate new CAPS markers, and identified BACs can be directly used for sequencing. The 200,000-fosmid end sequences currently available have already proven to be extremely valuable for increasing the possibilities of extensions from other sequenced BACs.

Considerable synergies will be derived from the ongoing potato genome sequencing project. Potato, another important food staple in *Solanum*, is being sequenced by another, but similarly structured consortium (<http://www.potatogenome.net/> [verified 16 Jan. 2009]). The first sequences should be available this year. Within *Solanum* tomato and potato are closely related, both are members of the same phylogenetically similar group of species, and only five major pericentromeric inversions have been observed between these two species (Tanksley et al., 1992). Because of their phylogenetic proximity, we expect that it will be possible to close sequence gaps in the tomato genome based on potato data and vice versa. The two projects have a good working relationship and regularly meet at the SOL genome workshops held once a year. All data related to the tomato genome sequencing project can be found on SGN (<http://sgn.cornell.edu/>) and BAC sequences are deposited to GenBank (<http://www.ncbi.nlm.nih.gov/>). We expect that the euchromatin sequence will be close to finished in 2010.

Experimental Procedures

Sequencing

Data Availability and Sequencing Statistics

All data, including BAC and BAC-end sequences, chromatograms, assembly files, FISH localizations, overgo results, and mapping data are available on the SGN Web site (<http://sgn.cornell.edu/>). Sequence data are also available from GenBank (<http://www.ncbi.nlm.nih.gov/>). To track the progress of the project, a BAC registry database is run as a central resource on the SGN website. The sequencing teams have special log-in accounts that allow

them to assign BACs to their projects and then adjust the status of each BAC in their sequencing pipeline. Based on this information, the summary statistics about project progress are calculated and displayed in real time on the International Tomato Sequencing Project overview page at http://sgn.cornell.edu/about/tomato_sequencing.pl [verified 16 Jan. 2009].

Genome Annotation

Repeat Database

A comprehensive repeat database specific for tomato was generated by running RepeatScout (Price et al., 2005) on the BAC-end sequences of each library. The three different repeat collections (one per BAC library) were assembled into one library using the cap3 program. The resulting set was assayed for repeat frequency in the entire BAC-end database, and repeats occurring fewer than 30 times were discarded. This set, referred to as the unirepeat set, was annotated using BLAST against different databases (The Institute for Genomic Research repeat set and GenBank Nonredundant), and was used to assess repeat content in BAC-ends and in full BAC sequences.

ITAG Genome Annotation Pipeline

The ITAG annotation pipeline operates on batches of contigs composed of one or more BACs. These contigs are generated at SGN from the AGP files and the BAC sequences. Analyses such as repeat masking, EST alignment, and gene predictions using different gene finders such as GeneID (Parra et al., 2000), GeneMark (Isono et al., 1994), and Augustus (Stanke et al., 2008) are performed on those BACs. To generate a consensus annotation, these data are combined with homology to protein or genomic sequences from other species (BlastX, TblastX), and fed into the combiner software called EuGene (Foissac et al., 2008). The resulting gene models are then functionally annotated based on homology searches (BlastP), protein domain searches (Interpro) (Mulder et al., 2003), and gene ontology assignment (Ashburner et al., 2000). Non-coding RNAs were identified using the Infernal program (Griffiths-Jones et al., 2003).

Estimation of Transcription Factors in Tomato Genome Using Expressed Sequence Tags

To search putative TFs in the EST data sets of *Solanum lycopersicum*, the assembled ESTs from PlantGDB, version161a, September 2007 release (257,093 ESTs assembled into 48,945 PUTs) was downloaded and translated using ESTScan-3.0.2 (Iseli et al., 1999). These translated PUTs were categorized into TF gene families based on the classification process defined by two plant transcription databases—PlnTFDB (Riano-Pachon et al., 2007) and PlantTFDB (<http://planttfdb.cbi.pku.edu.cn/> [verified 16 Jan. 2009]). A list of domains necessary for classifying a TF into a particular gene family was prepared and the available HMM profiles from PFAM (v22.0 [Finn et al., 2008]) were downloaded. The HMM profiles for the remaining domains were created using the protein

alignments available at PlnTFDB. HMMER searches (<http://hmmer.janelia.org/> [verified 16 Jan. 2009]) were performed on translated PUTs using HMM profiles and hits having E -values of $\leq 10^{-2}$ were selected. Further, these putative TFs were localized on 559 tomato BACs (finished and unfinished BAC sequences downloaded from SGN [bacs205]) by performing BLASTN with selection criteria of $\geq 90\%$ identity and 80% length coverage.

Analysis of Resistance and Defense-Response-Like Genes

We analyzed the 48,945 PUT sequences of tomato downloaded from the PlantGDB (Duvick et al., 2008). All the PUTs were used for BLASTX search with the NCBI nonredundant database (<http://www.ncbi.nlm.nih.gov/>) and top hits of all the genes were extracted in a tabulated form. Each gene showing homology to the above-mentioned three major classes of R-genes, that is, NBS-LRR type, LZ-NBS-LRR type, and LRR-Tm type together with other putative resistance proteins and defense-response genes, making five total categories, were tabulated in Microsoft Excel (Microsoft, Redmond, WA) format. These R-gene and defense-response gene homologs were then mapped in silico on 754 sequenced BACs of respective chromosomes to find their physical locations.

Acknowledgments

Financial sources: Sequencing of chromosome 2 in Korea is supported by Crop Functional Genomic Center, a Frontier 21 Project of the MOEST of Korean government. Chromosome 3 is being sequenced with support of the Chinese Academy of Sciences. Chromosome 4 is being sequenced at the Wellcome Trust Sanger Institute in the United Kingdom with the support of BBSRC/DEFRA and RERAD. The Wellcome Trust Sanger Institute is funded by the Wellcome Trust. Biodiversity work in *Solanum* at the NHM is supported by the NSF PBI program through award DEB-0316614 "PBI Solanum—a worldwide treatment." Chromosome 5 is sequenced by the "Indian Initiative on Tomato Genome Sequencing (IITGS)" funded by Department of Biotechnology, Government of India and supported by Indian Council of Agricultural Research, New Delhi. Chromosome 6 is sequenced with the support of the European Commission (EU-SOL Project PL 016214) and by the Centre for BioSystems Genomics (CBSG), which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research. Chromosome 7 sequencing is funded by the National Institute of Agronomic Research (INRA, France) and the National Research Funding Agency (ANR, France). Chromosome 8 sequencing is supported by the Chiba Prefecture, Japan. Chromosome 9 sequencing is supported by Genoma España. Chromosome 11 is supported by the Chinese Academy of the Sciences. Chromosome 12 is sequenced with the support of the Italian Ministry of Agriculture (Agronanotech Project), the Italian Ministry of Research (FIRB Project), and the EU (EU-SOL project). The U.S. group is supported by the National Science Foundation, USA, grants DBI-0421634 and DBI-0606595. We would like to acknowledge the contribution of the following people at the Wellcome Trust Sanger Institute: Matthew Jones (Shotgun Library Construction), Karen Oliver (Fosmid End Sequencing), Sarah Sims (Shotgun Data Production), Stuart McLaren (Automated Sequence Improvement), and Christine Lloyd (Finishing Quality Control).

References

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, and R.F. Moreno. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252:1651–1656.

- Adams-Phillips, L., C. Barry, and J. Giovannoni. 2004. Signal transduction systems regulating fruit ripening. *Trends Plant Sci.* 9:331–338.
- AGI. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
- Alexander, L., and D. Grierson. 2002. Ethylene biosynthesis and action in tomato: A model for climacteric fruit ripening. *J. Exp. Bot.* 53:2039–2055.
- Anderson, L.K., G.G. Doyle, B. Brigham, J. Carter, K.D. Hooker, A. Lai, M. Rice, and S.M. Stack. 2003. High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. *Genetics* 165:849–865.
- Anderson, L.K., N. Salameh, H.W. Bass, L.C. Harper, W.Z. Cande, G. Weber, and S.M. Stack. 2004. Integrating genetic linkage maps with pachytene chromosome structure in maize. *Genetics* 166:1923–1933.
- Angiosperm Phylogeny Group. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linnean Soc.* 141:399–436.
- Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: Tool for the unification of biology—The Gene Ontology Consortium. *Nat. Genet.* 25:25–29.
- Asp, T., U.K. Frei, T. Didion, K.K. Nielsen, and T. Lubberstedt. 2007. Frequency, type, and distribution of EST-SSRs from three genotypes of *Lolium perenne*, and their conservation across orthologous sequences of *Festuca arundinacea*, *Brachypodium distachyon*, and *Oryza sativa*. *BMC Plant Biol.* 7:36.
- Bachem, C.W., R.S. van der Hoeven, S.M. de Bruijn, D. Vreugdenhil, M. Zabeau, and R.G. Visser. 1996. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development. *Plant J.* 9:745–753.
- Bai, Y., C.C. Huang, R. van der Hulst, F. Meijer-Dekens, G. Bonnema, and P. Lindhout. 2003. QTLs for tomato powdery mildew resistance (*Oidium lycopersici*) in *Lycopersicon parviflorum* G1.1601 co-localize with two qualitative powdery mildew resistance genes. *Mol. Plant Microbe Interact.* 16:169–176.
- Bennett, M.D., I.J. Leitch, H.J. Price, and J.S. Johnston. 2003. Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25% larger than the *Arabidopsis* Genome Initiative estimate of ~125 Mb. *Ann. Bot. (Lond.)* 91:547–557.
- Bogdanove, A.J., and G.B. Martin. 2000. AvrPto-dependent Pto-interacting proteins and AvrPto-interacting proteins in tomato. *Proc. Natl. Acad. Sci. USA* 97:8836–8840.
- Brummell, D.A., and M.H. Harpster. 2001. Cell wall metabolism in fruit softening and quality and its manipulation in transgenic plants. *Plant Mol. Biol.* 47:311–340.
- Budiman, M.A., L. Mao, T.C. Wood, and R.A. Wing. 2000. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res.* 10:129–136.
- Cannon, S.B., L. Sterck, S. Rombauts, S. Sato, F. Cheung, J. Gouzy, X. Wang, J. Mudge, J. Vasdevani, T. Schiex, M. Spannagl, E. Monaghan, C. Nicholson, S.J. Humphray, H. Schoof, K.F. Mayer, J. Rogers, F. Quetier, G.E. Oldroyd, F. Debelle, D.R. Cook, E.F. Retzel, B.A. Roe, C.D. Town, S. Tabata, Y. Van de Peer, and N.D. Young. 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci. USA* 103:14959–14964.
- Chang, S.B., L.K. Anderson, J.D. Sherman, S.M. Royer, and S.M. Stack. 2007. Predicting and testing physical locations of genetically mapped loci on tomato pachytene chromosome 1. *Genetics* 176:2131–2138.
- Chang, S.B., T.J. Yang, E. Datema, J. van Vugt, B. Vosman, A. Kuipers, M. Meznikova, D. Szinay, R. Klein Lankhorst, E. Jacobsen, and H. de Jong. 2008. FISH mapping and molecular organization of the major repetitive sequences of tomato. *Chromosome Res.* 16:919–933.

- Chiusano, M.L., N. D'Agostino, A. Traini, C. Licciardello, E. Raimondo, M. Aversano, L. Frusciante, and L. Monti. 2008. ISOL@: An Italian SOLANACEAE genomics resource. *BMC Bioinformatics* 9:57.
- Clarkson, J.J., K.Y. Lim, A. Kovarik, M.W. Chase, S. Knapp, and A.R. Leitch. 2005. Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytol.* 168:241–252.
- D'Agostino, N., A. Traini, L. Frusciante, and M.L. Chiusano. 2007. Gene models from ESTs (GeneModelEST): An application on the *Solanum lycopersicum* genome. *BMC Bioinformatics* 8(Suppl. 1):S9.
- Datema, E., L.A. Mueller, R. Buels, J.J. Giovannoni, R.G. Visser, W.J. Stiekema, and R.C. van Ham. 2008. Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato. *BMC Plant Biol.* 8:34.
- De Jong, W.S., N.T. Eannetta, D.M. De Jong, and M. Bodis. 2004. Candidate gene analysis of anthocyanin pigmentation loci in the Solanaceae. *Theor. Appl. Genet.* 108:423–432.
- Doganlar, S., A. Frary, M.C. Daunay, R.N. Lester, and S.D. Tanksley. 2002. Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. *Genetics* 161:1713–1726.
- Duvick, J., A. Fu, U. Muppurala, M. Sabharwal, M.D. Wilkerson, C.J. Lawrence, C. Lushbough, and V. Brendel. 2008. PlantGDB: A resource for comparative plant genomics. *Nucleic Acids Res.* 36:D959–D965.
- Eshed, Y., and D. Zamir. 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.
- Fernie, A.R., and L. Willmitzer. 2001. Molecular and biochemical triggers of potato tuber development. *Plant Physiol.* 127:1459–1465.
- Finn, R.D., J. Tate, J. Mistry, P.C. Coghill, S.J. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L. Sonnhammer, and A. Bateman. 2008. The Pfam protein families database. *Nucleic Acids Res.* 36:D281–D288.
- Foissac, S., J.P. Gouzy, S. Rombauts, C. Mathé, J. Amselem, L. Sterck, Y. Van de Peer, P. Rouzé, and T. Schiex. 2008. Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinformatics* 3:87–97.
- Fray, R.G., and D. Grierson. 1993. Molecular genetics of tomato fruit ripening. *Trends Genet.* 9:438–443.
- Fu, Y., A.P. Hsia, L. Guo, and P.S. Schnable. 2004. Types and frequencies of sequencing errors in methyl-filtered and high c0t maize genome survey sequences. *Plant Physiol.* 135:2040–2045.
- Fulton, T.M., R. Van der Hoeven, N.T. Eannetta, and S.D. Tanksley. 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14:1457–1467.
- Gebhardt, C., and J.P. Valkonen. 2001. Organization of genes controlling disease resistance in the potato genome. *Annu. Rev. Phytopathol.* 39:79–102.
- Gerats, A.G., E. Vrijlandt, M. Wallroth, and A.W. Schram. 1985. The influence of the genes An1, An2, and An4 on the activity of the enzyme UDP-glucose:flavonoid 3-O-glucosyltransferase in flowers of *Petunia hybrida*. *Biochem. Genet.* 23:591–598.
- Giovannoni, J.J. 2004. Genetic regulation of fruit development and ripening. *Plant Cell* 16(Suppl.):S170–S180.
- Giuliano, G., G.E. Bartley, and P.A. Scolnik. 1993. Regulation of carotenoid biosynthesis during tomato development. *Plant Cell* 5:379–387.
- Goff, S.A., D. Ricke, T.H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B.M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W.L. Sun, L. Chen, B. Cooper, S. Park, T.C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R.M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalima, A. Oliphant, and S. Briggs. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296:92–100.
- Gray, J., S. Picton, J. Shabbeer, W. Schuch, and D. Grierson. 1992. Molecular biology of fruit ripening and its manipulation with antisense genes. *Plant Mol. Biol.* 19:69–87.
- Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna, and S.R. Eddy. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* 31:439–441.
- Hoeven, R. Van der, C. Ronning, J. Giovannoni, G. Matin, and S. Tanksley. 2002. Deductions about the number, organization, 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.
- Hui, D., J. Iqbal, K. Lehmann, K. Gase, H.P. Saluz, and I.T. Baldwin. 2003. Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Spingidae) and its natural host *Nicotiana attenuata*: V. Microarray analysis and further characterization of large-scale changes in herbivore-induced mRNAs. *Plant Physiol.* 131:1877–1893.
- International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* 436:793–800.
- Iseli, C., C.V. Jongeneel, and P. Bucher. 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. p. 138–148. *In Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. Am. Assoc. Artificial Intelligence, Menlo Park, CA.
- Isono, K., J.D. McIninch, and M. Borodov. 2007. Proceedings of the International Conference on Intelligent Systems for Molecular Biology. 1994. Characteristic features of the nucleotide sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer program GeneMark. *DNA Res.* 1:263–269.
- Jaillon, O., J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisy, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Huguency, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M.E. Pe, G. Valle, M. Morgante, M. Caboche, A.F. Adam-Blondon, J. Weissenbach, F. Quetier, and P. Wincker, and French-Italian Public Consortium for Grapevine Genome Characterization. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
- Kahlau, S., S. Aspinall, J.C. Gray, and R. Bock. 2006. Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. *J. Mol. Evol.* 63:194–207.
- Kessler, A., and I.T. Baldwin. 2001. Defensive function of herbivore-induced plant volatile emissions in nature. *Science* 291:2141–2144.
- Kim, U.J., H. Shizuya, J. Sainz, J. Garnes, S.M. Pulst, P. de Jong, and M.I. Simon. 1995. Construction and utility of a human chromosome 22-specific Fosmid library. *Genet. Anal.* 12:81–84.
- Knapp, S., L. Bohs, M. Nee, and D.M. Spooner. 2004. Solanaceae: A model for linking genomics with biodiversity. *Comp. Funct. Genomics* 5:285–291.
- Korf, I., P. Flicek, D. Duan, and M.R. Brent. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* 17(Suppl. 1):S140–S148.
- Lawrence, C.J., T.E. Seigfried, H.W. Bass, and L.K. Anderson. 2006. Predicting chromosomal locations of genetically mapped loci in maize using the Morgan2McClintock Translator. *Genetics* 172:2007–2009.
- Li, L., C. Li, and G.A. Howe. 2001. Genetic analysis of wound signaling in tomato. Evidence for a dual role of jasmonic acid in defense and female fertility. *Plant Physiol.* 127:1414–1417.
- Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg.

2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
- Mueller, L.A., C.D. Goodman, R.A. Silady, and V. Walbot. 2000. AN9, a petunia glutathione S-transferase required for anthocyanin sequestration, is a flavonoid-binding protein. *Plant Physiol.* 123:1561–1570.
- Mueller, L.A., A.A. Mills, B. Skwarecki, R.M. Buels, N. Menda, and S.D. Tanksley. 2008. The SGN comparative map viewer. *Bioinformatics* 24:422–423.
- Mulder, N.J., R. Apweiler, R.K. Attwood, A. Bairoch, and D. Barrell. 2003. The Interpro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* 31:315–318.
- Oksman-Caldentey, K.M. 2007. Tropane and nicotine alkaloid biosynthesis—novel approaches towards biotechnological production of plant-derived pharmaceuticals. *Curr. Pharm. Biotechnol.* 8:203–210.
- Olmstead, R.G., J.A. Sweere, R.E. Spangler, L. Bohs, and J.D. Palmer. 1999. Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA. p. 111–137. *In* M. Nee, D.E. Symon, R.N. Lester, and J.P. Jessop (ed.) *Solanaceae IV, Advances in biology and utilization*. Royal Botanic Gardens, Kew, UK.
- Palmer, L.E., P.D. Rabinowicz, A.L. O’Shaughnessy, V.S. Balija, L.U. Nascimento, S. Dike, M. de la Bastide, R.A. Martienssen, and W.R. McCombie. 2003. Maize genome sequencing by methylation filtration. *Science* 302:2115–2117.
- Parra, G., E. Blanco, and R. Guigo. 2000. GeneID in *Drosophila*. *Genome Res.* 10:511–515.
- Pedley, K.F., and G.B. Martin. 2003. Molecular basis of Pto-mediated resistance to bacterial speck disease in tomato. *Annu. Rev. Phytopathol.* 41:215–243.
- Peters, S.A., J.C. van Haarst, T.P. Jesse, D. Woltinge, K. Jansen, T. Hesselink, M.J. van Staveren, M.H. Abma-Henkens, and R.M. Klein-Lankhorst. 2006. TOPAAS, a tomato and potato assembly assistance system for selection and finishing of bacterial artificial chromosomes. *Plant Physiol.* 140:805–817.
- Peterson, D.G., S.R. Schulze, E.B. Sciarra, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, and A.H. Paterson. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* 12:795–807.
- Peterson, D.G., S.M. Stack, H.J. Price, and J.S. Johnston. 1996. DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes. *Genome* 39:77–82.
- Prat, S., W.B. Frommer, R. Hofgen, M. Keil, J. Kossmann, M. Koster-Topfer, X.J. Liu, B. Muller, H. Pena-Cortes, and M. Rocha-Sosa. 1990. Gene expression during tuber development in potato plants. *FEBS Lett.* 268:334–338.
- Price, A.L., N.C. Jones, and P.A. Pevzner. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* 21(Suppl. 1):i351–i358.
- Quattrocchio, F., W. Verweij, A. Kroon, C. Spelt, J. Mol, and R. Koes. 2006. PH4 of Petunia is an R2R3 MYB protein that activates vacuolar acidification through interactions with basic-helix-loop-helix transcription factors of the anthocyanin pathway. *Plant Cell* 18:1274–1291.
- Riano-Pachon, D.M., S. Ruzicic, I. Dreyer, and B. Mueller-Roeber. 2007. PlnTFDB: An integrative plant transcription factor database. *BMC Bioinformatics* 8:42.
- Riechmann, J.L., J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O.J. Ratcliffe, R.R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J.Z. Zhang, D. Ghandehari, B.K. Sherman, and G. Yu. 2000. *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* 290:2105–2110.
- Sacco, M.A., S. Mansoor, and P. Moffett. 2007. A RanGAP protein physically interacts with the NB-LRR protein Rx, and is required for Rx-mediated viral resistance. *Plant J.* 52:82–93.
- Schijlen, E., C.H. Ric de Vos, H. Jonker, H. van den Broeck, J. Molthoff, A. van Tunen, S. Martens, and A. Bovy. 2006. Pathway engineering for healthy phytochemicals leading to the production of novel flavonoids in tomato fruit. *Plant Biotechnol. J.* 4:433–444.
- Seymour, G., M. Poole, K. Manning, and G.J. King. 2008. Genetics and epigenetics of fruit development and ripening. *Curr. Opin. Plant Biol.* 11:58–63.
- Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–1732.
- Soderlund, C., S. Humphray, A. Dunham, and L. French. 2000. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* 10:1772–1787.
- Spelt, C., F. Quattrocchio, J. Mol, and R. Koes. 2002. ANTHOCYANINI of petunia controls pigment synthesis, vacuolar pH, and seed coat development by genetically distinct mechanisms. *Plant Cell* 14:2121–2135.
- Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24:637–644.
- Szinay, D., S.B. Chang, L. Khurasteva, S. Peters, E. Schijlen, Y. Bai, W.J. Stiekema, R.C. van Ham, H. de Jong, and R.M. Klein Lankhorst. 2008. High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. *Plant J.* 56:627–637.
- Tang, X., D. Szinay, C. Lang, M.S. Ramanna, E.A. van der Vossen, E. Datema, R. Klein Lankhorst, J. de Boer, S.A. Peters, C. Bachem, W. Stiekema, R.G. Visser, H. de Jong, and Y. Bai. 2008. Cross-species BAC-FISH painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. *Genetics* 180:1319–1328.
- Tanksley, S.D. 2004. The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *Plant Cell* 16(Suppl.):S181–S189.
- Tanksley, S.D., M.W. Ganal, J.P. Prince, M.C. de Vicente, M.W. Bonierbale, P. Broun, T.M. Fulton, J.J. Giovannoni, S. Grandillo, and G.B. Martin. 1992. High density molecular linkage maps of the tomato and potato genomes. *Genetics* 132:1141–1160.
- Todesco, S., D. Campagna, F. Levorin, M. D’Angelo, R. Schiavon, G. Valle, and A. Vezzi. 2008. PABS: An online platform to assist BAC-by-BAC sequencing projects. *Biotechniques* 44:60, 62, 64.
- Tuskan, G.A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R.R. Bhalerao, R.P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G.L. Chen, D. Cooper, P.M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroove, A. DeJardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J.C. Leple, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D.R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryabov, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C.J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer, and D. Rokhsar. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596–1604.
- van der Vossen, E.A., J.N. van der Voort, K. Kanyuka, A. Bendahmane, H. Sandbrink, D.C. Baulcombe, J. Bakker, W.J. Stiekema, and R.M. Klein-Lankhorst. 2000. Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: A virus and a nematode. *Plant J.* 23:567–576.
- van Os, H., S. Andrzejewski, E. Bakker, I. Barrena, G.J. Bryan, B. Caromel, B. Ghareeb, E. Isidore, W. de Jong, P. van Koert, V. Lefebvre, D. Milbourne, E. Ritter, J.N. van der Voort, F. Rousselle-Bourgeois, J. van Vliet, R. Waugh, R.G. Visser, J. Bakker, and H.J. van Eck. 2006. Construction of a 10,000-marker ultradense genetic

- recombination map of potato: Providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* 173:1075–1087.
- Velasco, R., A. Zharkikh, M. Troglio, D.A. Cartwright, A. Cestaro, D. Pruss, M. Pindo, L.M. Fitzgerald, S. Vezzulli, J. Reid, G. Malacarne, D. Iliev, G. Coppola, B. Wardell, D. Micheletti, T. Macalma, M. Facci, J.T. Mitchell, M. Perazzoli, G. Eldredge, P. Gatto, R. Ozyerski, M. Moretto, N. Gutin, M. Stefanini, Y. Chen, C. Segala, C. Davenport, L. Dematte, A. Mraz, J. Battilana, K. Stormo, F. Costa, Q. Tao, A. Si-Ammour, T. Harkins, A. Lackey, C. Perbost, B. Taillon, A. Stella, V. Solovyev, J.A. Fawcett, L. Sterck, K. Vandepoele, S.M. Grandó, S. Toppo, C. Moser, J. Lanchbury, R. Bogden, M. Skolnick, V. Sgaramella, S.K. Bhatnagar, P. Fontana, A. Gutin, Y. Van de Peer, F. Salamini, and R. Viola. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2:e1326.
- Wang, Y., X. Tang, Z. Cheng, L. Mueller, J. Giovannoni, and S.D. Tanksley. 2006. Euchromatin and pericentromeric heterochromatin: Comparative composition in the tomato genome. *Genetics* 172:2529–2540.
- Whitelaw, C.A., W.B. Barbazuk, G. Pertea, A.P. Chan, F. Cheung, Y. Lee, L. Zheng, S. van Heeringen, S. Karamycheva, J.L. Bennetzen, P. SanMiguel, N. Lakey, J. Bedell, Y. Yuan, M.A. Budiman, A. Resnick, S. Van Aken, T. Utterback, S. Riedmuller, M. Williams, T. Feldblyum, K. Schubert, R. Beachy, C.M. Fraser, and J. Quackenbush. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302:2118–2120.
- Wu, F., L.A. Mueller, D. Crouzillat, V. Petiard, and S.D. Tanksley. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: A test case in the Euasterid plant clade. *Genetics* 174:1407–1420.
- Yasuhara, J.C., and B.T. Wakimoto. 2006. Oxyoron no more: The expanding world of heterochromatic genes. *Trends Genet.* 22:330–338.
- Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2003. High-Cot sequence analysis of the maize genome. *Plant J.* 34:249–255.