Appendix 1. Technical Document for an International Consortium to Sequence the Tomato Genome

Summary

The tomato genome will be sequenced as the cornerstone of a International Solanaceae Genome Initiative – a project that aims to develop the family Solanaceae as a model for systems biology for understanding plant adaptation and diversification (see International Solanaceae Genome Initiative white paper http://sgn.cornell.edu/solanaceae-project/). The tomato genome is comprised of approximately 950 Mb of DNA – more than 75% of which is heterochromatin and largely devoid of genes. The majority of genes are found in long contiguous stretches of gene-dense euchromatin located on the distal portions of each chromosome arm. We propose to identify and sequence a minimal tiling path of BAC clones through this approximately 220 Mb euchromatin. The starting point for sequencing the genome will be approximately 1500 "seed" BAC clones individually anchored to the tomato high density genetic map based on a single, common *L. esculentum x L. pennellii* F2 population (referred to as the F2.2000; http://www.sgn.cornell.edu/).

Sequencing will proceed on a BAC by BAC basis. Each sequenced anchor BAC will serve as a seed from which to radiate out in the minimum tiling path in either direction. Identification of the correct next BACs in the euchromatin minimum tiling path for sequencing will be based on the use of a BAC end sequence database that will be created as part of this project, as well as a fingerprint contig physical map that is currently being jointly constructed by U.S. and Dutch scientists and will be completed by July 2004. Likewise, identification of the 1500 anchor BACs will have been completed by July 2004 through joint efforts of U.S. and Dutch scientists and will be made freely available to all scientists. A subset of the sequenced BACs will be localized on pachytene chromosomes via FISH (fluorescence in situ hybridization) to help guide the extension of the tiling path through the euchromatic arms of each chromosome and to determine when the heterochromatin and telomeric regions have been reached on each arm. A single bioinformatics portal will be created for this project which will be mirrored at several locations around the world and provide a mechanism by which researchers in different locations can develop and contribute bioinformatics tools and information to the project. Likewise, a common set of standards for BAC sequencing and finishing have been adopted and described herein as well as a common set of standards for gene nomenclature, and structural and functional gene annotation. International participation and coordination of this project is described herein and will be overseen by an International Solanaceae Genome Initiative steering committee and an annual Solanaceae Genomics Meeting to be hosted each year by a different participating country.

Send comments to Steven Tanksley sdt4@cornell.edu.

Rationale for Sequencing the Tomato Genome

Sequencing the tomato genome is the cornerstone of a larger project entitled the "International Solanaceae Genome Project (SOL): A Systems Approach to Understanding Diversity and Adaptation in Plants". (http://sgn.cornell.edu/solanaceae-project/. The purpose of the current document is to describe the technical and strategic details of how the tomato genome will be sequenced by an international consortium of countries and scientists. The strategy outlined in this document was the result of discussions and contributions by scientific representatives from all of the countries/research groups who have committed to contributing to sequencing the tomato genome (Table 1). Feed back and suggestions on this document are welcome and should be e.mailed to S. Tanksley

(sdt4@cornell.edu) J. Giovannoni (jjg33@cornell.edu) or any of the other groups participating in sequencing (see Table 1).

Table 1. List of tomato chromosomes to be sequenced by various countries/research organizations

Chrom	Country	P.I.(s)	Institution	e.mail addresses	Agency for Grant Applications	Target Deadline	Estimated Size Euchromatin (Mb)*	Projected # BACs**
1	USA	S.Tanksley, J.Giovannoni, J.Van Eck, L.Mueller, S.Stack	Cornell U. USDA/ARS Colo St U.	sdt4@cornell.edu, jjg33@cornell.edu, jv27@cornell.edu, lam87@cornell.edu, sstack@lamar.colostate.edu	National Science Foundation	Jan-04	24	246
2	Korea	D.Choi, B-D.Kim	KRIBB, Seoul Natl. U	doil@kribb.re.kr, kimbd@snu.ac.kr	BioGreen21 Project / RDA /Fronteer 21 Project / CFCG Ministry of Science and Technology (MOST)	Feb-04 July-04	26	268
3	China	C.Li, Y.Xue, Z.Cheng, M.Chen, H.Ling	Chinese Acad Sci	lichu@msu.edu, ybxue@genetics.ac.cn, zkcheng@genetics.ac.cn, mschen@genetics.ac.cn, hqling@genetics.ac.cn	Chinese Academy of Science Natural Science Foundation	Mar-04	26	274
4	UK	G.Bishop, G. Seymour	Imperial College, Warwick HRI	gdb@aber.ac.uk, graham.seymour@hri.ac.uk	BBSRC /DEFRA	Jan-04	19	193
5	USA	(see above)	(see above)	(see above)	National Science Foundation	Jan-04	11	111
6	The Netherlands	W.Stiekema P.Lindhout T.Jesse	Centre for BioSystems Genomics, Wageningen U, Keygene	willem.stiekema@CBSG.NL, Pim.Lindhout@wur.nl, Taco.Jesse@keygene.com	funded (in progress)		20	213
7	France	M.Bouzayen	BMPMF	bouzayen@flora.ensat.fr	National Agency for Genome Sequencing	Mar-04	27	277
8	USA	(see above)	(see above)	(see above)	National Science Foundation	Jan-04	17	175
9	Spain	M.Botella	U. Malaga	mabotella@uma.es	Genome Espana	Mar-03	16	164
10	USA	(see above)	(see above)	(see above)	National Science Foundation	Jan-04	10	108
11	USA	(see above)	(see above)	(see above)	National Science Foundation	Jan-04	13	135
12	Italy	G.Giuliano	ENEA	giulianog@casaccia.enea.it	Italian Ministry of Agriculture	May-04	11	113
Total							219	2276

* based on analysis of pachytene chromosomes by deJong and Stack (personal comm)
** assumes average BAC lengh 120 kb and 20% overlap between BACs in minimum tiling path

Objectives

The objective of the tomato sequencing project are to:

1) produce a contiguous sequence of the gene rich, euchromatic arms of each of the 12 tomato chromosomes
2) process and annotate this sequence in a manner consistent and compatible with similar data from Arabidposis, rice and other plant species.
3) create an international bioinformatics portal for comparative Solanaceae genomics which can store, process, and make available to the public the sequence data and derived information from this project and associated genomics activities in other solanaceous plants (for detailed information on these last two goals, see Standards Document)

Why is the tomato genome a good reference for the Solanaceae and for related plant taxa?

The Solanaceae family is unique in that there have been no large-scale duplication events (e.g. polyploidy) early in the radiation of this family. The polyploidy events (e.g. tetraploid potatoes and tetraploid tobacco) are all recent and diploid forms of both of these species are still widely distributed. As a result, macrosyntenty and microsynteny conservation amongst the genomes of tomato, potato, pepper and eggplant is very high (fig 1,2). This allows one to predict regions between genomes that are identical by descent and to study the evolution of sequence and function of orthologous genes – a key to understanding

diversification and adaptation. Most solanaceous species also have the same basic chromosome number (x=12) suggesting that large scale genomic changes have involved chromosomal inversions and/or

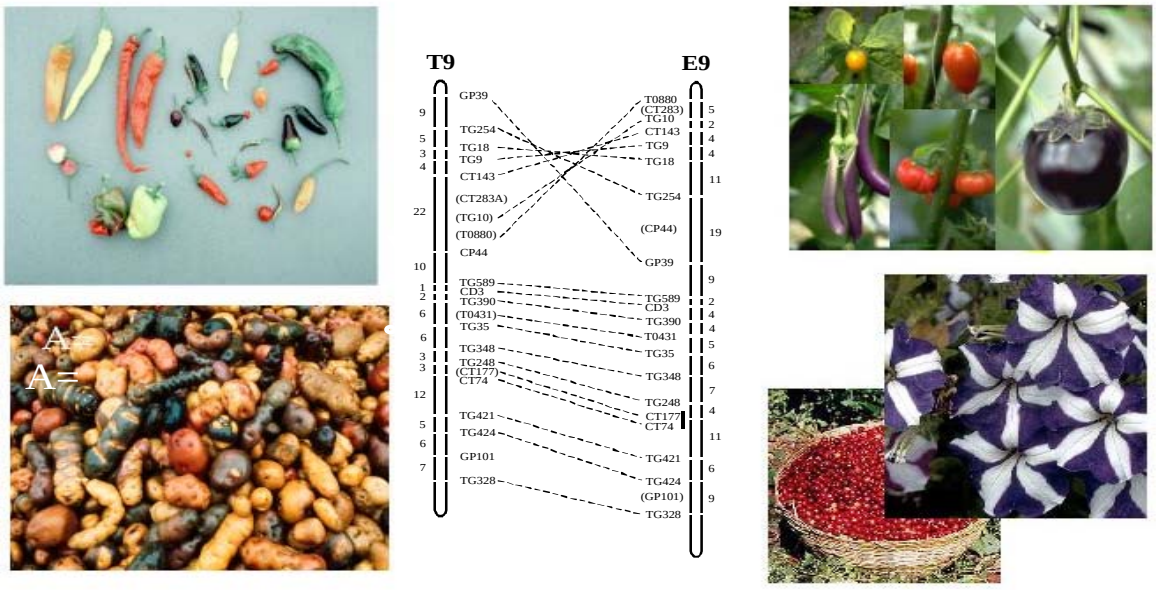

Figure 1. Despite their highly divergent adaptations and plant forms, the genomes of Solanaceous plants are well conserved interms of macrosynteny. Middle shows chromosome 9 or tomato and eggplant which differ bv a single paracentric inversion. Microsynteny is also largely conserved (fig 6) making for ideal studies of orthology and evolution of gene functions across the family.

interchanges -- a prediction that has been largely born out by comparative mapping studies (Tanksley et al. 1992, Livingstone et al. 1999, Doganlar et al. 2002).

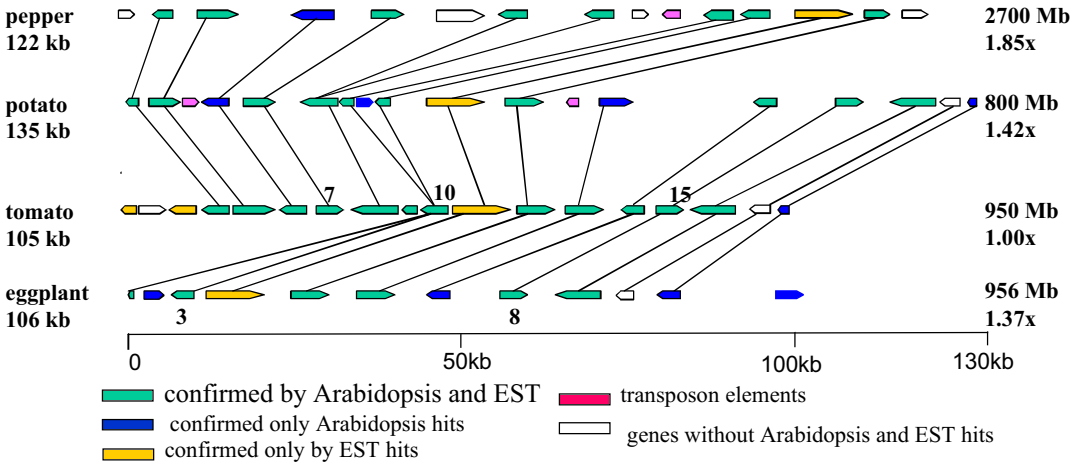

Figure 2. Microsynteny across the solanaceae. BACS from the same region of chromosome 2 from tomato, potato, eggplant and pepper were sequenced . Based on these preliminary results, both the gene content and gene order is conserved amongst these species in this region of the genome. Genes indicated in white boxes have no evidence other than gene finding programs and may be artifacts of the gene finding algorithms. (Wang, Giovannoni, Wing and Tanksley, unpublished results).

The tomato genome is organized in a manner that makes it cost effective to sequence.

The tomato genome contains 950 Mb of DNA which is organized into 12 chromosomes (n=x=12) (Arumuganathan and Earle 1991). Unlike the chromosomes of maize or rice, in which heterochromatin and euchromatin are interspersed, the heterochromatin in tomato is concentrated around the centromeres (fig 3,4). This pericentric heterochromatin is largely devoid of genes, but
constitutes approximately 75% of the DNA (Khush et al. 1964, Rick 1971, Peterson et al. 1996, Van derHoeven et al. 2002). In contrast, the distal portions of each tomato chromosome are comprised of largely contiguous stretches of gene-rich euchromatin which correspond to less than 25% of the DNA (fig 3) (Peterson et al. 1996).

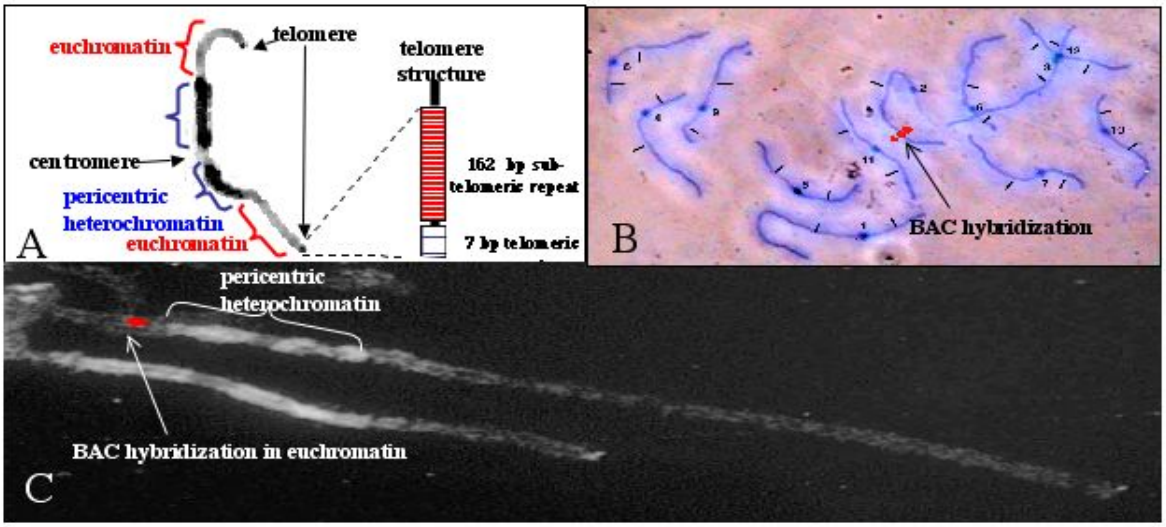

Figure 3 . Tomato pachytene chromosomes. **A**) Acetocarmine stain of chromosome 8 showing differentiation of gene rich euchromatin (light staining), pericentric heterochromatin (dark staining), centromere and telomeres (from Rick 1971). Macrostructure and sequence composition of telomeres are shown at right (see Ganal et al 1991 for details) **B**) Fluorescence *in situ* hybridization (FISH) of BAC clone from the euchromatic portion of chromosome 2. This clone (BAC19) has been sequenced and has a high gene density characteristic of euchromatin (VanderHoeven et al 2002). Sample has been prepared to reveal synaptonemal complexes (SC, dark blue). Light blue halos around each SC is due to DAPI staining of DNA. Kinetochores appear as dark staining spheres approximately 1 um in diameter. Chromosome identities as well as transition points between euchromatin and heterochromatin (indicated by bars) can be determined based on chromosome length, arm ratios, relative position along each arm and SC density (Sherman and Stack 1995, Stack personal comm). **C**) Fluorescence *in situ* hybridization (FISH) of BAC clone from the euchromatic portion of chromosome 6 which borders the centromeric heterochromatin (de Jong and Chang, personal comm). Standards have been established between the Stack and deJong labs for integrating results from FISH of results on pachytene chromosomes.

Rather than sequencing the entire tomato genome (950 Mb), we propose to sequence the approximately 220 Mb of euchromatin that contains the majority of genes (Van derHoeven et al. 2002, Fig 4, Table 1). The advantage of this approach is that we will be able to recover both GENE CONTENT AND GENE ORDER is an essential feature of the proposed euchromatin-sequencing project. Gene order would largely be lost in a whole genome shotgun sequencing approach or a methyl-filtration sequencing approach (Rabinowicz et al. 1999). Having an ordered sequence is essential to: 1) predict both gene content and gene order in other solanaceous species connected through comparative genetic maps (see previous section) 2) facilitate positional gene cloning in solanaceous species and 3) allow the organization of the tomato genome to be compared with the genomes of other sequenced organisms.

Detailed Experimental Plan:

a. Establishment of high density genetic map and anchoring to "seed BACs"

A high density genetic map, based on an L. *esculentum* x *L. pennellii* F2 population (referred to as the F2.2000 population) has been completed in the U.S. via funding from the National Science Foundation and contains 1500 sequenced markers (mostly EST markers and microsatellites; http://www.sgn.cornell.edu). By March 2004, another approximately 1000 AFLP markers will be added to the map (using DNA from the same population) by KeygeneNV in The Netherlands. This map will serve as the basis of the BAC by BAC sequencing project.

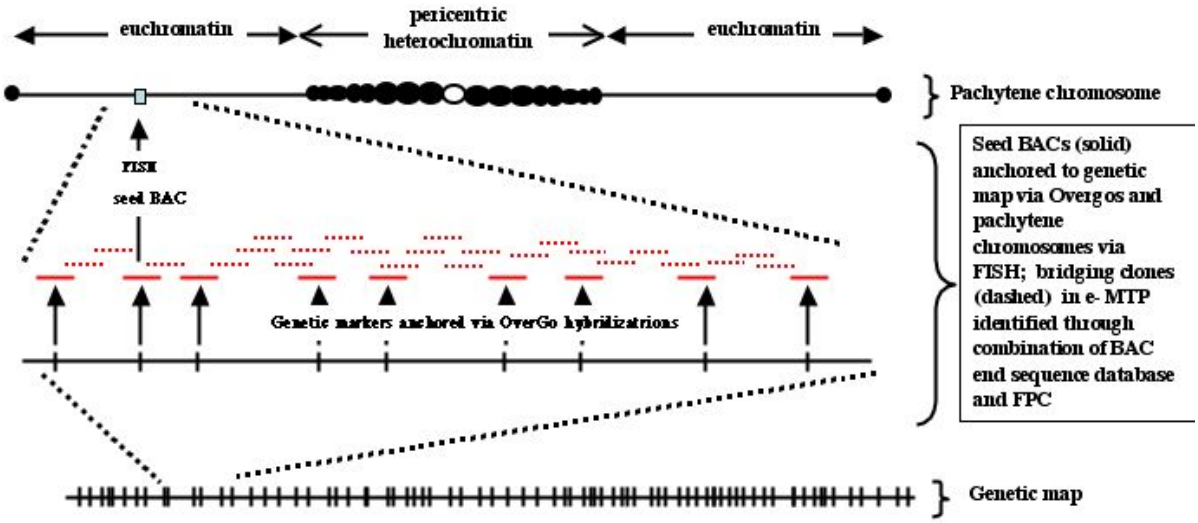

Figure 4. Strategy sequencing of tomato euchromatin. 1500 BACs will be individually anchored to the F2.2000 high density genetic map via overgo hybridization and AFLP matching. This aspect of the project will be completed by July 2003. These "seed BACs" will serve as starting points for sequencing the euchromatic arms of each tomato chromosomes. Floresence in situ hybridization (FISH) will be used to verify localization to euchromatin. Sequencing will extend in both directions from seed BACs through euchromatin minimum tiling path (e-MTP) identified via a BAC end sequence tag connector (STC) database (to be generated as part of this project) and a finger print contig map (FPC) which will be completed by July 2003.

At Cornell University in the U .S., Overgo probes have been constructed for the 1500 sequenced markers. Currently, 500 have been screened on the HindIII BAC library and approximately two-thirds have been definitively assigned to single BACs. Based on the rate of progress, all Overgo probes will have been screened by July 2004 and it is estimated that 1000 unambiguous BAC anchor points (referred to as "seed BACS") will have been identified. Simultaneously, Keygene will screen the AFLP markers from the same map to the same BAC library (fig 5). This is expected to identify another 500 unambiguous anchored seed BACs, making a total of 1500 seed BAC anchor points from which to begin sequencing. All of this work is to be completed by July 2004 and the result to be make publicly available, including through SGN (http://www.sgn.cornell.edu).

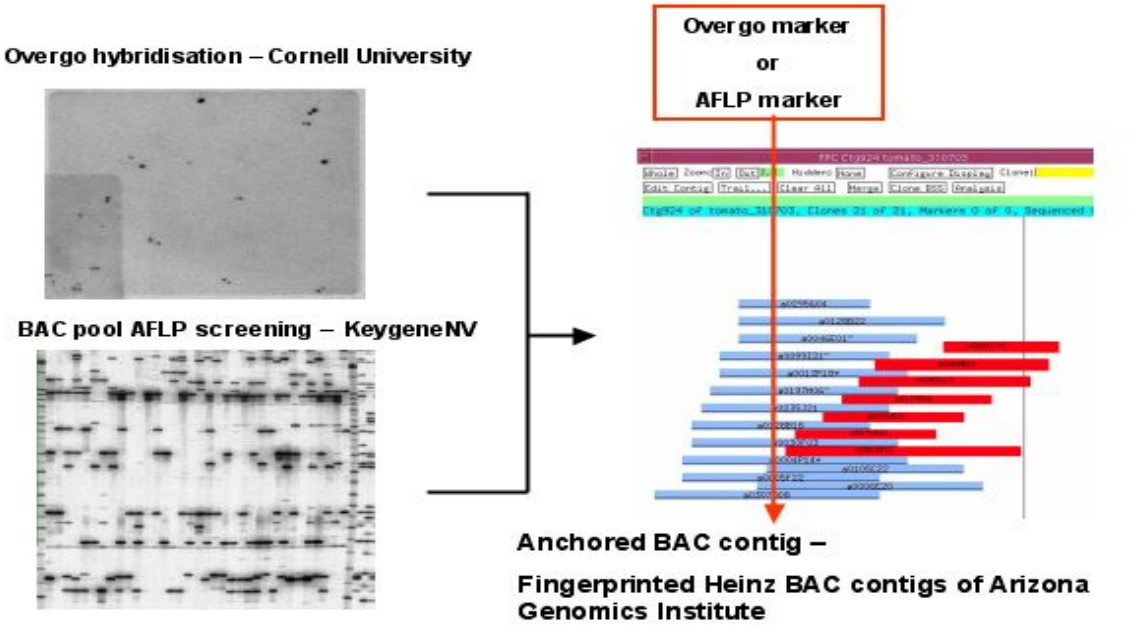

Figure 5. High density genetic map (based on F2.2000) population will be populated by approximately 2500 markers (1500 ESTs from Cornell and 1000 AFLPs from KeygeneNV). Majority of those markers will be have been individually anchored to BACs via Overgo hybridization (top) or AFLP technology (bottom) , providing a minimum of 1500 BACs unambiguously anchored to the genetic map. In turn, many of those BACs will be members of the Fingerprint Contig Map being constructed at U. Arizona. These anchored "seed BACs" will serve as the starting point for sequencing the euchromatic arms of each tomato chromosome. FISH will be applied to a subset of these BACs to determine their positions in the chromosomes relative to telomeres and pericentric heterochromatin.

b. Establishment of sequence tag connector (STC) database from the *Hind*III BAC library and a new *Msp*I BAC library

We propose to end sequence 50,000 clones from the *Hind*III BAC library (6 x coverage) as well as a similar number from the *MspI* I BAC library (see below). Together these sequences will result in an STC database comprised of 200,000 end mate pairs. Each BAC end sequence will be subjected to automated annotation to determine the proportion of ends that are likely to correspond to genic regions (see next section for details). This STC database (*Hind*III and *Mps*I) will be used both to confirm and to extend the euchromatin minimal tiling path (e-MTP) and to select new clones for sequencing from the e-MTP as described below. The availability of the STC database should significantly reduce the degree of overlap of between clones in the e-MTP and thus reduce the amount of redundant sequencing. Use of combined STC and FPC data to select a minimum tiling path has been shown to reduce the overlap between adjacent clones from more than 30% to less than 10%, thus reducing overall sequencing costs. Moreover, having corroborative STC and FPC data can reduce false positives in extending a tiling path for sequencing. We will also determine for each end sequenced clone from both the *Hind*III and MspI libraries, the percentage that contain putative genes at: a) neither end b) one c) both ends. For the *Hind* III library, the frequency of putative genes at the end of each clone will provide an independent estimate of the total gene number for tomato, currently estimated at 35,000 genes (Van der Hoeven et al. 2002). By comparing the statistics between the *Hind*III and *Msp*I libraries we will also be able to determine to what extent MspI clones are enriched for hypomethylated genes.

Generating a *MspI* BAC library enriched for demethylated regions of the genome. A BAC library made with specific restriction enzymes may be biased for or against certain regions of a genome, and in extreme cases, specific regions of the genome may be missing entirely. Libraries made with different restriction

enzymes will have different biases. The success of the proposed project is contingent on having as complete a minimal tiling path across the euchromatin (e-MTP). Hence, we have opted to construct a second tomato BAC library based on partial digests with *MspI* -- a 4-base recognition enzyme sensitive to cytosine methylation. Like other C-methylation sensitive enzymes, *Msp*I cuts in undermethylated, single copy or genic regions (Burr et al. 1988, Miller and Tanksley 1990, Rabinowicz et al. 1999). Thus BACs derived from *Msp*I digests are more likely to contain genes at either end and may be preferentially enriched for gene-rich euchromatic regions. A 15 x BAC library will be constructed from *L. esculentum* Heinz 1706 (same variety used for the *Hind*III library) based on *MspI* partial restriction digests. As described above, 50,000 BACs from the *Msp*I library will also be end-sequenced to contribute to the BAC end sequence database needed for extending walks out from the seed BACs through a minimum tiling path.

c. Shotgun sequencing of euchromatin by radiating out from each "seed" BAC through the euchromatic minimal tiling path (e-MTP)

As described above, approximately 1500 seed BACs will be subjected to shotgun sequencing. The division of sequencing activities between countries will be on a chromosome basis (Table 1). We estimate that the average map distance between these "seed" BACs will be approximately 1 cM. Since most of the genetic anchoring the BACs correspond to genes (or ESTs), the BACs are likely to be biased towards the euchromatic portion of the genome. Since the euchromatic portion of the genome is estimated at approximately 220 Mb (van der Hoeven et al 2002, Table 1), the average physical distance between seed BACs may be as little as 200 kb. A subset of the seed BACs will be subjected to fluorescence in situ hybridization (FISH) on spreads of pachytene synaptonemal complexes (SCs) and pachytene chromosomes to determine where these clones reside on each chromosome with respect to each other (an independent determination of BAC order), telomeres, centromeres, heterochromatin and euchromatin (fig 3,4). FISH analysis will be split between the laboratories of Dr. Steven Stack (Colorado State University, U.S.A.) and Dr. Hans de Jong (Wageningen, The Netherlands, using standardized techniques such that the results from the two labs can be integrated.

The Role of the Fingerprint Contig Physical (FPC) Map in Aiding Sequencing of the Tomato Euchromatin

The 1500 seed BACs will serve as starting points for map-based sequencing of the euchromatin minimal tiling path (e-MTP) (fig 4,5,6). The sequence of each "seed" BAC will be BLASTed against the BAC end STC database to identify BACs with minimal overlap in either direction. To aid in this process, a partial fingerprint contig physical map (FPC) has been constructed at the University of Arizona through funding from the National Science Foundation (http://www.genome.arizona.edu/fpc/tomato/) and at Keygene NV in The Netherlands. Integrated into the FPC will be the same 1500 individually anchored BACs that will serve as seed points for sequencing. The FPC will be exploited to aid the tomato sequencing project in two ways. As described above, the primary manner in which the e-MTP will be extended in either direction from the seed BACs will be through matches with the STC database. However, where possible, the FPC serve as confirmatory data in extending the e-MTP. Second, as described earlier, the high-density genetic map is comprised largely of coding genes (ESTs). Hence the map should be biased towards gene-rich euchromatic regions (Vander Hoeven et al. 2002). It is thus predicted that a large proportion of the FPCs (especially from the *Hind*III BAC library) will not be anchored to the genetic map because they are part of the 75% of the genome comprising the pericentric heterochromatin that contains few, if any genes. To test this hypothesis and to shed light on the portions of the tomato genome not contained in the e-MTP, we will select 10 BACs from each of 10 non-anchored contigs from the FPCs for sequencing, annotation, genetic mapping and FISH on pachytene chromosomes. Genetic mapping will utilize sequence information to design CAPS assays usable on the *L. esculentum x L. pennellii* F2.2000 mapping population, which is the basis of the current high density tomato genetic map (www.sgn.cornell.edu).

5. Related research/sequencing that will tie in with the tomato genome sequence

As pointed out earlier, sequencing the tomato genome is a key component to a larger International project "The International Solanaceae Genome Project (SOL)". In this regard, the tomato genome will serve as the center point by which to compare and tie together genome information across plants in the family solanaceae (and related taxa, e.g. coffee, lettuce, sunflower). It will also offer an opportunity to begin tying together genome informatics across dicots starting with tomato and Arabidopsis (especially through integration with TAIR). Much of the discussion of how this will be approached conceptually and bioinformatically, is in the Standards Document (see next section). Furthermore, a more detailed description of the larger set of research projects in the solanaceae and how they relate to each other and the tomato genome sequence, as part of the "The International Solanaceae Genome Project", can be found in the "The International Solanaceae Genome Project" white paper (http://sgn.cornell.edu/solanaceae-project

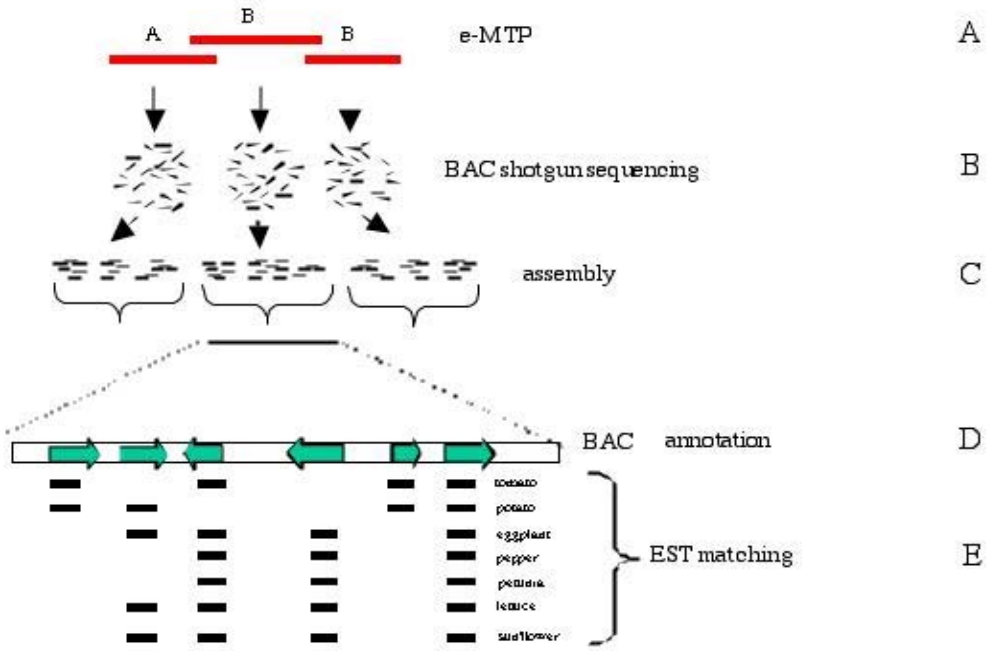

Figure 6. BAC sequencing, assembly and annotation. Each BAC in the e-MTP (A) will be shotgun sequenced, assembled and finished (B,C). End sequences from each BAC will be blasted against the BAC-end STC database to identify the next step in the e MTP. The FPC will be used as confirmatory data. Each BAC will be subjected to gene finding/annotation (D), including matching of orthologous ESTs from related species (E).

6. Bioinformatics: Creation of the International Solanaceae Genome Network (ISGN) to Accommodate an International Tomato Genome Sequencing Project in a Format That Ties Together all Solanaceous Species

Sequencing the tomato genome, and the larger SOL project (http://sgn.cornell.edu/solanaceae-project/), will involve scientists, research organizations and bioinformatics centers. For such a project to succeed and provide maximum utility to the scientific world, it important that the bioinformatics responsibilities be shared by the corresponding national sequencing projects using common protocols and done in a manner that the database and interface can be mirrored and improved on a worldwide basis. For that reason, much emphasis

and discussion has been devoted to this. The next section "Standards Documen" describes in detail in details the plans and starndards for sequencing the tomato genome and the SOL project in general.

Management and Coordination

Formation of Steering Committee for the International Solanaceae Genome Project (SOL). At the November 3 meeting, it was decided that the committee will be comprised of representative scientists from each of the countries actively participating in the tomato sequencing project. The co-chairs of that committee are Drs. Dani Zamir and Marc Zabeau. The membership and responsibilities of this committee will be decided by March 30, 2004. Some of the expected activities are to coordinate activities across the international community to avoid overlap of efforts, to assure sequencing is done in a consistent manner with a minimum set up standards, to assure that information from the sequencing project is freely available to the world community. They will also oversee the bioinformatics activities.

Formation of an International Bioinformatics Steering Committee to oversee bioinformatics for the SOL and especially sequencing of the tomato genome. Dr. Lukas Mueller (curator for SGN – Solanaceae Genome Network database/website) has agreed to organize and serve as the chair of this committee. This committee will be charged with drafting and overseeing both the bioinformatics components of the tomato genome sequencing project as well as the overarching SOL project. This committee will work hand-in-hand with the SOL Steering Committee. This committee will be established by March 30, 2004.

Establishment of an annual International Solanacaeae Genomics Conference. It was agreed that an annual conference will be held devoted to the topic of genomics research in the Solanaceae. The first meeting will be held in J September 2004 and hosted by Holland. The meeting times and locations for the 2005 and 2006 meetings will be decided within the next 6 months. It was suggested that the 2006 meetings might be held at the University of Wisconsin as part of the Solanaceae systematics meetings. Both the SOL Steering Committee and the Bioinformatics Steering Committee will hold discussion sections at these gatherings.

Literature Cited:

Arumuganathan K and Earle E. 1991. Estimation of nuclear DNA content of plants by flow cytometry. Plant Mol Biol Rep. 9:208-218

Burr, B.A., Burr, F.A., Thompson, K.H., Albertson, M.C. & Stuber, C.W. Gene mapping with recombinant inbreds in maize. Genetics 118, 519-526 (1988).

Donganlar et al. 2002. A Comparative Genetic Linkage Map of Eggplant (*Solanum melongena*) and its Implications for Genome Evolution in the Solanaceae Genetics 161: 1697-1711

Khush GS, Rick CM, Robinson RW (1964) Genetic activity in a heterochromatic chromosome segment of the tomato. Science 145: 1432-1434.

Livingstone et al. 1999, Genome mapping in *Capsicum* and the evolution of genome structure in the Solanaceae. Genetics **152:** 1183-1202.

Miller JC, Tanksley SD (1990) Effect of different restriction enzymes, probe source, and probe length on detecting restriction fragment length polymorphism in tomato. Theor Appl Genet 80:385-389

Peterson DG, Price HJ, Johnston JS, Stack SM (1996) DNA content of heterochromatin and euchromatin in tomato (Lycopersicon esculentum) pachytene chromosomes. Genome 39: 77-82

Rabinowicz et al. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet 23:305-308

Rick CM (1971) Some cytogenetic features of the genome in diploid species. Stadler Symposiuim 1:153-174.

Tanksley et al. 1992. High density molecular linkage maps of the tomato and potato genomes. Genetics 132:1141-1160

Van der Hoeven, R., C. Ronning, J. Giovannoni, G. Martin, S.D. Tanksley. 2002. Deductions about the number, organization and evolution of genes in the tomato genome based on analysis of a large EST collection and selective genomic sequencing. Plant Cell 14: 1441-1456