Justification of Size Estimates for Tomato Genome Sequencing

A critical aspect of successful sequencing of the tomato genome is accurate size estimates of the targeted space to be covered with high-quality sequence. Accurate sequencing target estimations are obviously important for determining final objectives, project pace, and budget, in addition to providing milestones which can be tested during the sequencing process to adjust strategy and budgetary requirements for the entire project if necessary.

Definition of tomato euchromatin and goals of sequencing effort.

Tomato euchromatin. Euchromatin is defined cytologically as less condensed and thus lightly staining DNA of chromatin spreads, in contrast to the more condensed heterochromatin. While heterochromatin can be observed through most stages of the cell cycle, the pachytene stage is most commonly used in staining and for situ hybridization of tomato chromosomes (Peterson et al., 1996). Heterochromatin is thought to be largely transcriptionally inactive though recent cloning and characterization of rice centromeres for chromosomes 4 and 8 indicates the presence of active genes in centromeric and flanking heterochromatin sequences (Wu et al., 2004, Yan et al., 2005).

The tomato genome is comprised of a majority of paracentric heterochromatin typically flanked by large euchromatin islands that comprise the majority of the chromosome "arms" (see below). *For the purpose of the international tomato genome sequencing project our objective is sequencing of the gene-space residing in the euchromatin arms flanking each centromere.* Approximately one quarter of the tomato genome is comprised of lightly staining euchromatin, the vast majority of which is in these arms (Peterson et al., 1996 and data below).

Little genome sequence data is available from tomato heterochromatin as initial sequencing efforts to date have targeted euchromatin. However, using as a guide data available from rice, the only complete/nearly-complete crop genome sequence available, (and which was also sequenced using a similar map-based BAC-by-BAC approach), we can make a number of relevant estimations and define several guiding principals for the tomato genome sequencing effort.

Goals for the International Tomato Genome Sequencing effort. We shall use as our targeted sequencing goals two guiding principals: 1) complete sequencing of the major euchromatin "arms" flanking each of the 12 tomato chromosomes 2) to a degree of completion comparable to the standards of completion used to guide the international rice genome sequencing project (IRGSP, 2005). Specifically, these standards include:

An error rate of less than 1 error in 10,000 bases.

An average of 8-fold redundancy in sequencing coverage with a minimum of one high quality read in both directions at any specific sequence.

Use of all reasonable state of the art approaches available at the time for gap filling.

Acceptance of a small number of gaps that are recalcitrant to filling but are physically defined by in situ hybridization. Based on the degree of completion of the rice genome and excluding gaps defined by centromeres, this would mean

approximately 4 gaps per tomato chromosome.

We further define our objectives to include sequencing to at least the closest mapped marker to the visible euchromatin heterochromatin borders of each chromosome arm. In situ hybridization will be used to determine if these borders define the true euchromatin/heterochromatin borders or a gap that will be at minimum physically defined.

Estimation of gene space missed in this approach. Extrapolating from data obtained in rice we can calculate the number of genes that we might expect to miss in an approach that focuses on just the gene dense tomato euchromatin. For example, sequencing of rice chromosome 8 revealed 86 active genes in the centromere and distal non-recombinant regions (Yan et al., 2005). 86 genes/centromere X 12 tomato chromosomes = 1032 centromeric genes. Prior to initiation of the international tomato sequencing effort, Exelexsis Biosciences sequenced and deposited two random and highly repetitive heterochromatin BACs which together covered greater than 200 kb and harbored one gene. While this is clearly limited data, we can make a further rough estimate that refraining from sequencing the heterochromatin of tomato we might estimate the loss of an additional 705,000 kb / 200 kb per gene = 3525 euchromatin genes or a total of approximately 4500 genes that could be missed by focusing solely on the euchromatin arms (see below for the 705,000 kb estimate of the heterochromatin). The estimated gene content of tomato is 35,000 genes (van der Hoeven et al., 2002) suggesting that approximately 35,000 – 4,500 = 30,500 genes (87%) might be anticipated to be recovered through the euchromatin-only approach. Correcting further for the fact that non-centromere gaps represented approximately 3% of the targeted sequence space in rice, we would estimate recovery of 85% of the tomato gene space (apx. 30,000 genes) under the efforts of the international tomato sequencing effort. *In summary, the target of the international genome sequencing effort is sequencing of the euchromatin arms of all twelve tomato chromosomes which we estimate will represent approximately 85% of the tomato gene space.*

<u>Estimates of amount of sequencing required to meet project goals</u>.

Success of a large-scale sequencing project requires accuracy in determining the amount of sequencing required to meet project objectives. We have developed estimates of the physical distance to be covered in sequencing the euchromatin gene space of tomato centromeric arms. While more accurate estimates will develop as the project proceeds and more sequence is generated, we note that the current estimates are similar.

Cytologically Based Measurement of Euchromatin Content.

We took a straightforward approach to determine the amount of DNA in euchromatin and heterochromatin of tomato chromosomes (Peterson et al. 1996). First, tomato pachytene chromosomes were spread on glass slides using a technique that did not stretch (deform) the chromosomes. We stained the chromosomes by the Feulgen technique that has been proven to be a reliable, quantitative stain for DNA (see Price 1988). Whether stained by the Feulgen technique, aceto-carmine, aceto-orcein, or fluorescent dyes there is a consistent, clear distinction between distal euchromatin and proximal heterochromatin in tomato pachytene chromosomes (Fig. 1; *e.g.*, Ramanna and Prakken1967). Relative density (absorbance) of Feulgen stained euchromatin and

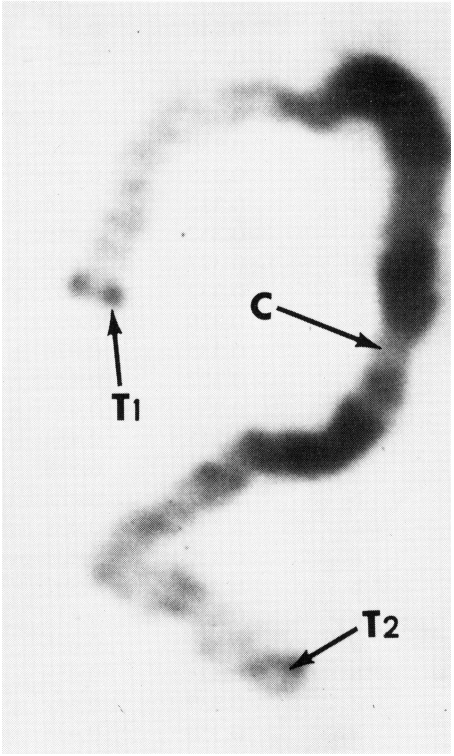

Figure 1. Tomato pachytene chromosome 11 stained with DAPI. The fluorescent image has been reversed so that chromatin appears dark on a light background . The centromere (C) is a constriction within the block of thick, darkly stained pericentric heterochromatin. More distally, the arms are thinner and consist of more lightly stained euchromatin. Telomeres (T) appear as darkly stained spots at the ends of the arms.

heterochromatin was determined in ten different spreads of pachytene chromosomes. Using twenty unstretched tomato pachytene chromosomes, the average width of the chromosomes in euchromatin was determined from fifty separate measurements, and the average width of the chromosomes in heterochromatin was determined from fifty additional measurements. Transverse measurements for diameter were made only in relatively straight parts of chromosomes. Lengths of pachytene chromosomes were taken from Sherman and Stack (1992) who carefully measured tomato pachytene chromosome lengths, arm ratios, and fractions of arms in euchromatin and heterochromatin on electron micrographs. Their results were found to be generally in agreement with other published measurements of tomato pachytene chromosomes. This information was used to calculate the total fraction of the genome in euchromatin and heterochromatin.

	Heterochromatin	**Euchromatin**
Relative chromosome length	0.36	0.64
Relative bivalent diameter	X 1.23	X 1.00
Relative area	0.44	0.64
Relative optical density	X 4.78	X 1.00
Relative OD X relative area	2.10	0.64

Total OD X area	2.74	2.74
Fraction of genome	**0.77**	**0.23**

Thus, we calculate that 77%of the Feulgen-stained DNA is in heterochromatin, and 23% of the Feulgen-stained DNA is in euchromatin. If there is significant error in this calculation, it might be due to observed tapering of pericentric heterochromatin in the transition from heterochromatin to distal euchromatin (Fig. 1). This could cause the average width of heterochromatin to be underestimated. If so, this would result in a slight underestimate of the amount of DNA in heterochromatin and a slight overestimate of the amount of DNA in euchromatin. However, we think this is unlikely because many measurements of the width of heterochromatin were made, so the mean heterochromatin diameter should take into account any tapering of heterochromatin width near euchromatin-heterochromatin transitions. Because of this, we believe 23% of the genome is an accurate measurement of the amount of DNA in euchromatin.

Estimates of the absolute size (1C amount) of the tomato genome are in general agreement at approximately 95 pg of DNA, e.g., Michaelson et al. (1991). Thus, the amount of DNA in euchromatin in one tomato genome is (0.23 x 0.95 pg =) 0.22 pg, a value that is only (0.22 pg/0.16 pg =) 1.4 times as large as the *Arabidopsis thaliana* genome. Converting the DNA amount in euchromatin to base pairs (Bennett and Smith 1976) there are [0.22 pg x (965 $x10^6$ pb/pg) =] 2.12 x 10^8 bp (**212 Mb**) of DNA in the euchromatin of one tomato genome. [0.73 pg x (965 $x10^6$ pb/pg) =] 7.05 x 10^8 bp (705 Mb) of DNA in the heterochromatin of one tomato genome.

Estimating Euchromatin Arm Size Based on Available Genome and EST Sequence.
To date a total of 15.5 Mb of non-overlapping tomato genomic sequence has been submitted to SGN by the US team and our international sequencing partners. A test set of high quality tomato gene sequences was created by combining 1) all published tomato gene sequences in GENBANK, 2) 2898 redundantly sequenced full-length tomato cDNAs available through TIGR, and 3) 6742 tomato contigs containing five or more overlapping EST sequences. 8,097 high quality unigene sequences remained after correcting for redundancy. This set of tomato unigenes was then searched against the available tomato genome sequence with stringency criteria of 90% or greater identify and 80% coverage. 456 of 8,097 unigenes were identified in the genome sequence. Assuming this gene set is representative of the gene space in terms of localization throughout the tomato genome, we estimate that 456/8,097 = 5.6% of the gene space has been covered. Correcting for the percentage of gene space present in the euchromatin arms (85%) we can calculate that 5.6/0.85 = 6.6% of the target gene space has been covered. If 15.5 Mb represents 6.6% of the euchromatin arms then 15.5/0.066 = **234 Mb** of genomic DNA would be calculated to represent the target non-overlapping genome space for the international genome sequencing project.

In a separate analysis, the 15.5 Mb of available tomato genomic DNA was searched for homologies to gene sequences and 2100 non-redundant gene models were identified following removal or transposon, viral and other repetitive sequences. 2100 genes out of 35,000 corresponds to 6% of the predicted gene space. Correcting for the percentage of gene space present in the euchromatin arms (85%) we can calculate that 6.0/0.85 = 7.05% of the target gene space has been covered. If 15.5 Mb represents 7.05% of the

euchromatin arms then 15.5/0.0705 = **220 Mb** of genomic DNA would be calculated to represent the target genome space for the international genome sequencing project.

Method	**Sequencing Target**
Cytology	212 Mb
Available Sequence and percent high quality gene models	234 Mb
Available sequence and total gene models	220 Mb

Additional Information. When the sequencing project is advanced to the stage where BAC contigs can be assayed for both total non-redundant sequence length and physical distance based on in situ hybridization, we will be able to develop an additional estimate of euchromatin physical size through validation of the cytological measurements with actual sequence data. At present there is no data available to make such estimations though the UK group has developed large BAC contigs covering most of chromosome 4 that will move into their sequencing pipeline in coming months. Based on BAC FPC data alone they have reported that their physical size estimate for chromosome 4 is consistent with the original cytological estimates used in planning the international sequencing effort (C. Nicholson, personal communication). In addition, the Korean group has completed more BAC sequencing than any other group in the consortium to date with 49 finished BACs representing approximately 20% of their projected total for chromosome 2. In line with project plans they have started from BACs anchored to the genetic map and spaced along chromosome 2. As such, they still have few and short contigs. Nevertheless, based on the physical distances between mapped marker sequences found in their sequenced BACs, they have estimated that the BACs sequenced to date represent approximately 20% of the genetic map for chromosome 2. While genetic to physical distance ratios can vary widely, and these numbers could change dramatically (for example in an area of suppressed recombination), at present their available data is consistent with the original cytological results on which the project was based.

In summary, the data described above is consistent with a sequencing target of 212 – 234 Mb for completion of the objectives of the international tomato genome sequencing project. The use of total as opposed to high quality gene models could result in an underestimation of target gene sequencing space as some predictions are likely not representative of true genes. *At present we propose use of the larger estimate, 234 Mb, as it is likely more accurate and more conservative* (in terms of justifying budget and activity for completion of project goals).

References

Bennett, M.D. and J.B. Smith. 1976. Nuclear DNA amounts in angiosperms. Phil. Trans. Roy. Soc. (Lond.) B 274:227-274

IRGSP 2005. The map-based sequence of the rice genome. Nature. 436: 793-800

Michaelson, M.J., H.J. Price, J.R. Ellison, and J.S. Johnston.1991. Comparison of plant DNA contents determined by Feulgen microspectrophotometry and laser flow cytometry. Am. J. Bot. 78:183-188

D.G. Peterson, H.J. Price, J.S. Johnston, and S.M. Stack. 1996. DNA content of heterochromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes. Genome 39: 77-82.

Price, H.J. 1988. DNA content variation among higher plants. Ann. Mo. Bot. Gard. 75:1248-1257

Ramanna, M.S. and R. Prakken. 1967. Structure and homology between pachytene and somatic metaphase chromosomes of the tomato. Genetica 38:115-133.

Sherman,J.D. and S.M. Stack. 1992. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. V. Tomato (*Lycopersicum esculentum*) karyotype and idiogram. Genome 35:354-359

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

Wu et al. 2004. Composition and structure of the centromeric region of rice chromosome 8. The Plant Cell 16: 967-976

Yan, H., Jin, W., Nagaki, K., Tian, S., Ouyanf, S., Buell, R., Talbert, P., Henikoff, S. and Jiang, J. 2005. Transcription and histone modification n the recombination-free region spanning a rice centromere. The Plant Cell. 17:3227-3238