Theor Appl Genet (2005) 112: 72–84
DOI 10.1007/s00122-005-0107-z

ORIGINAL PAPER

Y. Wang · R. S. van der Hoeven · R. Nielsen
L. A. Mueller · S. D. Tanksley

Characteristics of the tomato nuclear genome as determined by sequencing undermethylated *EcoR*I digested fragments

Received: 29 June 2005 / Accepted: 1 September 2005 / Published online: 6 October 2005
© Springer-Verlag 2005

Abstract A collection of 9,990 single-pass nuclear genomic sequences, corresponding to 5 Mb of tomato DNA, were obtained using methylation filtration (MF) strategy and reduced to 7,053 unique undermethylated genomic islands (UGIs) distributed as follows: (1) 59% non-coding sequences, (2) 28% coding sequences, (3) 12% transposons—96% of which are class I retroelements, and (4) 1% organellar sequences integrated into the nuclear genome over the past approximately 100 million years. A more detailed analysis of coding UGIs indicates that the unmethylated portion of tomato genes extends as far as 676 bp upstream and 766 bp downstream of coding regions with an average of 174 and 171 bp, respectively. Based on the analysis of the UGI copy distribution, the undermethylated portion of the tomato genome is determined to account for the majority of the unmethylated genes in the genome and is estimated to constitute 61 ± 15 Mb of DNA (~5% of the entire genome)—which is significantly less than the 220 Mb estimated for gene-rich euchromatic arms of the tomato genome. This result indicates that, while most genes reside in the euchromatin, a significant portion of euchromatin is methylated in the intergenic spacer regions. Implications of the results for sequencing the genome of tomato and other solanaceous species are discussed.

Communicated by R. Hagemann

Y. Wang · R. S. van der Hoeven · L. A. Mueller
S. D. Tanksley (✉)
Department of Plant Breeding and Genetics,
Department of Plant Biology, Cornell University,
Ithaca, NY , 14853 USA
E-mail: sdt4@cornell.edu
Tel.: +1-607-2551673
Fax: +1-607-2556683

R. Nielsen
Department of Biological Statistics and Computational Biology,
Cornell University, Ithaca, NY, 14853 USA

Introduction

CpG and CpNpG methylation is common in plants and plays a pivotal role in controlling gene expression, especially silencing genes and maintaining epigenetic states during plant developmental processes (Cao and Jacobsen 2002; Fojtova et al. 2003; Rabinowicz et al. 2003; Steimer et al. 2004; Ye and Signer 1996; Zemach and Grafi 2003). Methylcytosine is mostly concentrated in repetitive sequences in plants, such as long tandem arrays located around centromeres, at telomeres, and in nucleolar organizer regions (Bird 2002; Martienssen 1998; Rabinowicz et al. 1999). Furthermore, in the complex genomes of monocots (e.g., grasses such as maize, wheat) small blocks of undermethyltated genic regions (usually 5–20 kb and contain 1–4 genes) are intermixed with 5–200 kb blocks of repetitive and methylated DNA (Bennetzen et al. 1994; Panstruga et al. 1998; Tikhonov et al. 1999; Wicker et al. 2001). In such genomes, unmethylated regions can extend about 1–2 kb upstream and downstream of genes, based on bisulfite sequencing and methylation-sensitive PCR amplification and enzyme digestion (Antequera and Bird 1988; Bennetzen et al. 1994; Martienssen et al. 2004; Nick et al. 1986; Walbot and Warren 1990).

Besides providing a potential mechanism of gene control, differential methylation patterns between genic and non-genic regions have also provided a possible method of selectively cloning and sequencing unmethylated, and hence gene-rich, regions of plant genomes, while avoiding highly methylated, gene-deficient regions (Burr et al. 1988; Rabinowicz et al. 1999; 2003; Yuan et al. 2002; Palmer et al. 2003). Such methods could reduce the cost of sequencing the "gene space" of plant genomes and might be complimentary to expressed sequence tag methods, which are inherently biased towards more highly expressed genes. However, the potential downside of selectively sequencing unmethylated genomic fragments as a surrogate for full genome sequencing are that: (1) signif-

icant portions of repetitive DNA may not be methylated, (2) significant portion of genes are methylated (at least during some periods of the life cycle) (Ashapkin et al. 2002; Fojtova et al. 2003; Jacobsen and Meyerowitz 1997; Walker and Panavas 2001; Lippman et al. 2004), and (3) sequencing genes via unmethylated islands, while potentially enriching for genes, will not likely lead to a contiguous sequence, hence losing the important parameter of gene order—a fundamental parameter for comparative genomics or positional cloning. With regard to the first point, in maize it is estimated that 17% of its genome is undermethylated DNA, approximately 25% of which comprises repetitive elements (Palmer et al. 2003). Concerning the second point, it is known that some expressed genes are methylated, for example, *SUPERMAN* and *DRM2* (domains rearranged methyltransferase) in Arabidopsis, *nptII* (neomycin phosphotransferase II) in tobacco, and *r1* locus in maize (Ashapkin et al. 2002; Fojtova et al. 2003; Jacobsen and Meyerowitz 1997; Walker and Panavas 2001). Finally, with regard to the third point, most of the studies of unmethylated DNA have focused on monocots, such as maize that have an interspersed pattern of coding and non-coding heterochromatic regions (Palmer et al. 2003; Peterson et al. 2002; Rabinowicz et al. 1999; White and Doebley 1998; Yuan et al. 2002). Patterns of methylation in dicot genomes are less well studied than that of monocots. Many dicot genomes, such as Arabidopsis, *Medicago truncatula*, *Cochlearia pyrenaica*, tomato, potato, and eggplant, are organized in long stretches of gene-rich euchromatin and flanked by heterochromatic centromeric and telomeric regions (Gottschalk 1954; Kakes 1973; Kulikova et al. 2001). While it is believed that the heterochromatin is highly methylated, the degree of methylation in the gene-rich euchromatin is currently unknown.

The current study is focused on deciphering the content and distribution of unmethylated portion of the genome of tomato, *Solanum lycopersicum* L. [= *Lycopersicon esculentum* Mill]. While tomato is one of the best-studied dicot species, with a long history of genetic, cytogenetic, and molecular research, relatively little is known about the overall structure and distribution of methylation in this species. Tomato genome contains 950 Mb of DNA, approximately 25% of which is contained in long stretches of gene-rich euchromatin found at the distal ends of most chromosomes, whereas the other 75% is contained in the gene-deficient, pericentrimeric heterochromatin (Arumuganathan et al. 1991; de Jong 1998; Peterson et al. 1998). In addition, tomato has one of the lowest G + C contents (37%) of any plant species; and based on isoschizomer studies, it is estimated that 23% of the cytosine residues are methylated (Messeguer et al. 1991).

In order to shed light on the amount, composition and distribution of unmethylated DNA in tomato, 10,370 undermethylated clones were sequenced from one end with the M13 reverse primer. Through annotation of these undermethylated sequences and comparisons with high-density EST databases and previously sequenced genomic BAC clones, one can provide an overall description of the distribution and informational content of the unmethylated portion of this important dicot genome.

Materials and methods

Undermethylated tomato sequence library construction

The genomic sequencing survey in this report employed the methylation filtration (MF) strategy to enrich the unmethylated tomato sequences (Rabinowicz et al. 1999). Young leaves were taken from growing tips of ~4-week-old tomato plants, *S. lycopersicum cv*. TA496. Tomato genomic DNA was extracted from pelleted nuclei according to Bernatzky and Tanksley (1986), completely digested with *EcoR*I and size-fractioned by agarose electrophoresis to enrich the 0.5–4 kb fragments. Short *EcoR*I digested fragments were ligated into pBluescript SK(−) (Stratagene) and cloned into the methylation restrictive *mcrBC*$^{+}$ *E. coli*. strain JM109 (Stratagene) with the purpose of enriching the library for non-methylated DNA fragments. The enzyme McrBC in this host cell can degrade DNA containing methylation (Raleigh et al. 1988).

Estimation of percentage tomato genomic DNA fragments in size of 0.5–4 kb

After *EcoR*I digestion, tomato genomic DNA from TA496 was end-labeled with $\gamma-^{33}$P and electrophoresised on agarose gel. Radioactive signals were quantitated using the ImageQuant program (Amersham). One kilobase plus ladder (Invitrogen) was used as size standard to indicate the sizes of tomato fragments. The percentage of tomato fragments in the range of 0.5–4 kb was determined based on the ratio of radioactive signals in this size range.

Sequencing, assembly, and annotation of undermethylated DNA

A single-end sequence was generated for each clone using the M13R primer at the Institute for Genomic Research (TIGR). In total, 1,0370 sequence reads were obtained and submitted to Genbank (accession numbers BH011826 to BH146113). Trace files were analyzed with phred (Ewing and Green 1998) for base calling and then passed through the CAP3 program (Huang and Madan 1999) for assembly. The major parameters used for CAP3 were 90% identity and 20 bp for an overlap, which had been successfully used for tomato EST assembling projects (www.sgn.cornell.edu).

For computational prediction of the coding regions, unique assembled sequences (hereafter called undermethylated genomic islands—UGIs) were subjected to gene prediction using Genscan+ (Burge and Karlin 1997), Glimmer (Salzberg et al. 1998), Unveil (Majoros et al. 2003), and GeneMark (Borodovsky and McIninch 1993). For functional annotation, Blast was employed to compare UGIs with 62,447 tomato/potato/pepper EST-derived unigenes (BLASTN cutoff score 100 and *E* value $< 10^{-10}$) and the entire Arabidopsis proteome (tBLASTX cutoff score 50 or *E* value $< 10^{-10}$) (van der Hoeven et al. 2002; www.sgn.cornell.edu; www.Arabidopsis.org). The cutoff score 50 and 100 for tBLASTX and BLASTN program were employed to avoid the detection of short virtually identical sequences that might have relatively low *E* values. As a control, each undermethylated sequence was randomized with respect to nucleotide order and the entire set was subjected to the same assembly and annotation processes as described previously. The randomization maintained the nucleotide composition while scrambling the nucleotide order (Bedell et al. 2005; Korf et al. 2003).

Southern hybridization for genetic mapping

For genetic mapping of BAC clones, DNA probes were amplified from BAC DNA using primers designed from annotated exons. These clones were then mapped onto the high-density tomato map based on a population of 80 F2 individuals from the cross *S. lycopersicum* LA925×*S. pennellii* LA716 (Fulton et al. 2002; http://www.sgn.cornell.edu/cgi-bin/mapviewer). RFLP analysis and Southern blot analysis were performed as described in Fulton et al. (2002).

Detection of organellar DNA-like sequences in tomato nuclear genome

UGIs with significant homology to tobacco chloroplast (Z00044) and Arabidopsis mitochondrial (NC_001284) sequences were identified with BLASTN program (*E* value $< 10^{-10}$ and bit score > 100). As a point of reference, two tomato chloroplast genes, *RBCL* (L14403) and *NADH* (U08921), have 97–98% sequence identity to their tobacco chloroplast counterparts (Shinozaki et al. 1986). Likewise, BLASTN results of five bonafide tomato mitochondrial genes (D84426, X54738, X54409, X53397, and AF362735) showed 93–97% identity as the Arabidopsis mitochondrial sequences (Unseld et al. 1997). UGIs containing only chloroplast or mitochondrial homologous DNAs and have sequence identity of more than 97% for chloroplast and 93% for mitochondria sequences were assumed to be organellar sequences and were eliminated for the further analysis. UGIs with homology to organellar DNA, yet containing flanking non-organellar (and presumably nuclear) sequences were classified as organellar sequences transferred into the tomato nuclear genome. In order to compare the amount of organellar sequence insertions in the tomato nuclear genome and the organellar sequences in the Arabidopsis and rice nuclear genome, the pairwise comparison of organellar sequences and nuclear sequences for each species was performed by the BLASTN program for Arabidopsis and rice. The following sequences were used: complete chloroplast genome sequences of *A. thaliana* (NC_000932), and *Oryza sativa* (NC_001320), complete mitochondrial genome sequences of *A. thaliana* (NC_001284) and *O. sativa* (NC_001751, NC_001776), all five assembled Arabidopsis chromosomes (NC_003071, NC_003071, NC_003074, NC_003075, NC_003076), and 12 rice pseudomolecules of *O. sativa* ssp. *japonica* (358,546,961 bp) (www.tigr.org).

Comparison of UGIs with sequenced/annotated BAC clones

Tomato UGIs were compared with ten previously sequenced tomato BACs using BLASTN program (Mao et al. 2001; van der Hoeven et al. 2002; plus AY881150, AY881151, and AY881152). Seven of these BACs were isolated from a tomato BAC library by the virtue of a screen with a known gene and three were randomly selected from the same library (Budiman et al. 2000; Mao et al. 2001; van der Hoeven et al. 2002). In order to be counted as a match with a BAC clone, an UGI had to share $> 98\%$ identity and the aligned had to extend $\geq 90\%$ of the total length of the UGI (Palmer et al. 2003). BACs were re-annotated based on the guideline of rice genome annotation (Table 1, Goff et al. 2002; Yu et al. 2002). The gene content of each BAC was predicted with four computational gene finder programs: FGENESH (Salamov and Solovyev 2000), GenemarkHMM (Borodovsky and McIninch 1993), Genscan+ (Burge and Karlin 1997, Arabidopsis matrix; http://genes.mit.edu/GENSCAN.html), and GlimmerM (Salzberg et al. 1999). BACs were further annotated for gene content via comparisons with both the predicted Arabidopsis proteome (www.Arabidospsis.org) and the tomato EST-derived unigene set (www.sgn.cornell.edu). For this analysis, a BAC region had to match a member of the Arabidopsis proteome (*E* value $< 10^{-10}$ for tBLASTX) and/or tomato EST-derived unigenes (95% identity over 80% sequence length of each unigene) to be considered a coding region.

Estimating the extent to which UGIs cover 5′ non-coding, coding and 3′ non-coding regions

After comparing UGIs with Arabidopsis gene models using BLASTX, high-scoring segment pairs (HSPs) were mapped to the corresponding non-transposon Arabidopsis gene model for searching the boundaries of coding/non-coding regions. The non-aligned UGI sequences

Table 1 Ten tomato BAC clones used for comparing with UGIs

BAC name	GenBank accession number	Number of predicted genes	Non-TE genes confirmed by Arabidopsis proteome or tomato ESTs	Number of transposon-related genes	Location in tomato genome	References
181O9[a]	AY881150	10	0	10	Heterochromatin	Unpublished data
181C9[a]	AY881151	6	0	6	Heterochromatin	Unpublished data
181K1[a]	AY881152	10	0	10	Heterochromatin	unpublished data
47I13	AF411804	6	0	6	Heterochromatin	van der Hoeven et al. 2002
2o7[a]	AF411805AF411806	6	2	4	Heterochromatin	van der Hoeven et al. 2002
62O11	AF411808	7	7	0	Euchromatin	van der Hoeven et al. 2002
127E11	AF411807	19	18	0	Euchromatin	van der Hoeven et al. 2002
FW2.2	AF411809	20	20	0	Euchromatin	van der Hoeven et al. 2002
BAC19	AF27333	18	16	0	Euchromatin	Ku et al. 2000
240K04	AF275345	15	13	2	Euchromatin	Mao et al. 2001
Total		117	76	38		

[a]These BACs have unknown size sequence gaps

were assumed to be non-coding sequences. The UGIs were then categorized as 5′ upstream non-coding, coding, 3′ downstream non-coding based on the HSPs.

Use of maximum likelihood statistics to estimate the total unmethylated portion of the tomato genome

Maximum likelihood was used to estimate the total unmethylated portion of the tomato genome based on five assumptions: (1) the tomato genome is fully digested at *EcoR*I sites regardless of methylation status; (2) only clones containing *EcoR*I fragments free of methylation can survive the cloning process and pass through the methylation restrictive *mcrBC*$^+$ *E. coli* (Sutherland et al. 1992); (3) tomato genome is comprised of a mixture of sequences that occur both as single copy as well as repeated sequences, which exist in two or more copies; (4) a total of n sequence copies have been sampled with equal probability; (5) the total length of the unmethylated genome is mN, where m is the average distance between two restriction sites, N is the number of distinct sequences with *EcoR*I ends in the tomato genome. Due to the size selection for constructing the methylation filtration library, only a total of N' distinct sequences with *EcoR*I ends were cloned into the library. Thus, the percentage of *Eco*RI fragments in the tomato genome, which fall within the size constraints of the *EcoRI* library used in this project is $b = N'/N$. Therefore, if the sampling fraction is defined as $f = n/N' = n/Nb$, the total length of the unmethylated genome should be estimated as $mN = mn/fb$. The probability of sampling a single copy sequence Y times is

$$\Pr(Y=y)=\binom{n}{y}(1/N)^y(1-1/N)^{n-y}\approx \mathrm{e}^{-f}f^y/y! \qquad (1)$$

where the last equality holds as $n \to \infty$, $N \to \infty$, and $n/N \to f$. Similarly, the probability of sampling Z copies of a sequence that exists in K copies in the genome is approximately

$$\Pr(Z=z|K=k)=\mathrm{e}^{-fk}(fk)^z/z! \qquad (2)$$

Assuming the probability that any particular sequence exists in the genome as a single copy with probability p, and belongs to a family of repeated sequences with probability $1-p$, the total probability of observing X copies of the sequence in the sample is

$$\Pr(X=x)=p\Pr(Y=x)+(1-p)\sum_{k=2}^{\infty}\Pr(Z=x|K=k)\Pr(K=k), \qquad (3)$$

where $\Pr(K=k)$ is a proper probability mass function on $k\in\{2,3,...,\}$. It is assumed that $\Pr(K=k)$ is given by a truncated geometric distribution

$$\Pr(K=k)=q(1-q)^{k-2},\quad 0<q<1, \qquad (4)$$

because this distribution provided a better fit to data than any other single parameter distribution examined, such as the Poisson distribution. The resulting probability mass function for X can be expressed in terms of a polylogarithm function as

$$\Pr(X=x)=\frac{\mathrm{e}^{-f}f^x(1-q)p\{p-1+\mathrm{e}^f\mathrm{Li}_{-x}(\mathrm{e}^f(1-p))\}}{(1-p)^2x!}, \qquad (5)$$

where the polylogarithm function $\mathrm{Li}_{-x}(a)$ for an integer x is given by

$$\mathrm{Li}_{-x}(a)=(1-a)^{-1-x}\sum_{i=0}^{x}\left(\sum_{j=0}^{i+1}(-1)^j\binom{x+1}{j}(i-j+1)\right)a^{x-i}. \qquad (6)$$

To correct for the fact that sequences occurring in frequency zero in the sample are unobserved, the likelihood

function for the three parameters is defined as

$$L(p,q,f|X=x)=\frac{\Pr(X=x)}{1-\Pr(X=0)}. \quad (7)$$

This likelihood function can be maximized with respect to p, q, and f to yield a joint maximum likelihood estimate of these parameters. However, due to the limitation of computational power with only one parameter $(1-p)$ indicating the probability of repetitive sequences in the genome, sequences with very high copy numbers were not well represented in this model (Fig. 1a). Therefore, any sequences occurring in a frequency larger than 40 in UGIs were treated as outliers. The maximum likelihood estimated with and without outliers was −4,540.5 and −4,402.0, respectively. Although the estimate of the parameters of the geometric distribution does change substantially depending on whether outliers are included or not ($q=0.026$ for data including outliers vs. $q=0.068$ for data excluding outliers), the estimate of the sampling fraction is relatively robust to assumptions regarding outliers ($f=0.380$ vs. $f=0.359$). Thus, the outliers did not significantly affect the estimation.

Statistical comparisons of potential methylation sites, GC contents, and the maximum length of ORFs among UGIs, random sequences, and tomato EST-derived unigenes

UGIs, tomato EST unigenes, and random sequences were scanned in a sliding window of size 3 to determine the number of potential methylation sites, CpG/GpC and CpNpG/GpNpC. The paired Student's t test was carried out in Microsoft excel for comparing the number of CpG/GpC and the number of CpNpG/GpNpC in each type of sequences, for example, UGIs, sequences with randomized nucleotide bases, and EST-derived unigenes. The multiple comparisons with Turkey test were carried out using the SAS program for GC contents, the maximum length of ORFs, and the number of potential methylation sites among different types of sequences.

Results and discussion

Contig assembly of MF sequence reads

The average size of clone inserts in this *EcoR*I digested methyl-filtered genomic library was 1.4 kb and ranged from 205 bp to 5.7 kb based on a sample of 56 clones. A total of 10,370 sequence reads were obtained from this library, with an average length of 483 bp ranging from 150–845 bp. After assembly, these reads were constructed into 7,107 unique sequence contigs (UGIs) with a mean length of 508 bp, mode of 699 bp, and range of 103-2,009 bp. The distribution of numbers of singletons and UGIs with ≥2 members is depicted in Fig. 1b.

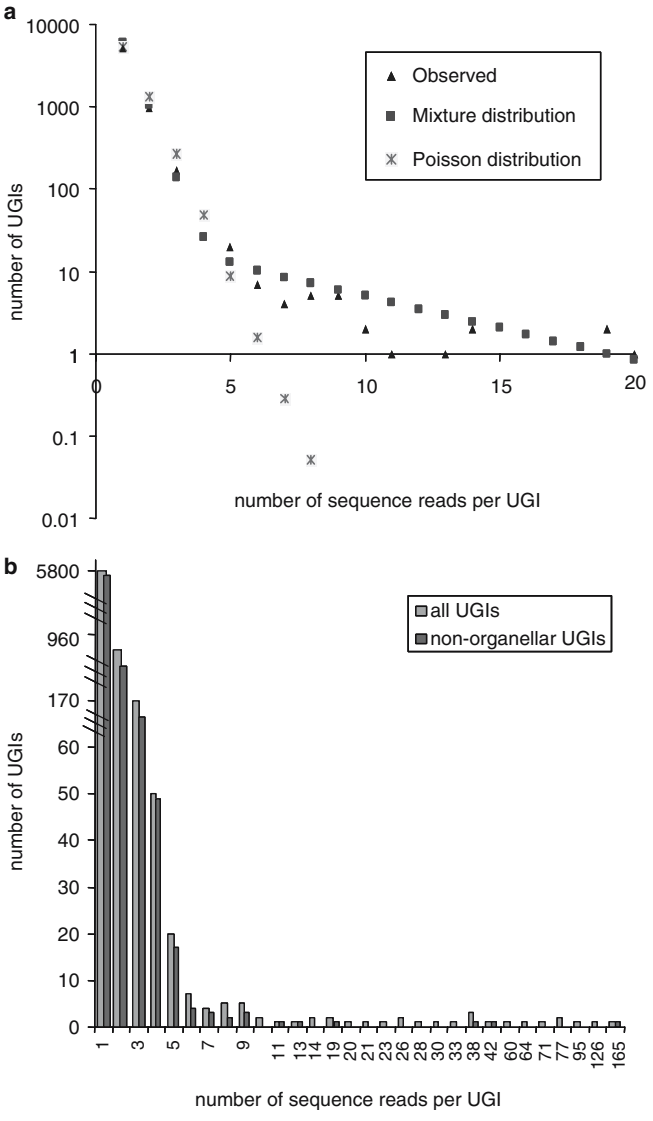

Fig. 1 **a** The distribution of observed and expected number of UGIs assuming that the tomato genome contains only unique sequences (poisson distribution) or both single-copy sequences and repeated sequences (mixture distribution). The mixture distribution was used for maximum likelihood estimation of the undermethylated portion of tomato genome. **b** Comparison of the distributions of all tomato nuclear UGIs and non-organellar UGIs, suggesting that UGIs with large membership are mostly organellar-like sequences in the tomato nuclear genome

Organellar-derived UGIs

UGIs comprised of true organellar DNA

BLASTN searches identified 136 and 28 UGIs having significant homologies with the fully sequenced tobacco chloroplast genome (Z00044) and Arabidopsis mitochondrial genome (NC_001284), with E value $< 10^{-10}$ and bit score > 100. For both cpDNA and mtDNA, the histograms of BLASTN match scores gave bimodal distributions (Fig. 2). The results suggest that the peak

showing highest overall homology to either chloroplast or mitochondrial DNA may actually derive from the organellar versus nuclear genome. As a point of reference, two previously sequenced tomato chloroplast genes, *RBCL* (L14403) and *NADH* (U08921), have 97 to 98% sequence identity with match scores of 2,663 and 3,700 to their tobacco chloroplast counterparts, respectively. Likewise, BLASTN results of five bonafide tomato mitochondrial genes (D84426, X54738, X54409, X53397, and AF362735) showed 93 to 97% identity with match scores from 383 to 2,825 when compared with the sequence of the Arabidopsis mitochondrial genome. Based on these results, it is inferred that the second, higher homology peak, corresponds to true organellar clones (Fig. 2). Thus, in total, there are four UGIs (0.05%) for mitochondrial DNA and 50 UGIs (0.7%) for chloroplast DNA. A more detailed examination of these UGIs revealed that they are homologous throughout their length with their organellar counterparts (thus containing no contiguous nuclear DNA sequences) further bolstering the conclusion that these UGIs correspond to tomato chloroplast or mitochondrial genomes. Therefore, these 54 UGIs corresponding to 380 sequence reads were discarded for further analysis. After eliminating true organellar-derived UGIs, this survey sampled 9,990 undermethylated sequences from tomato nuclear genome, corresponding to 7,053 UGIs with an average size of 507 bp.

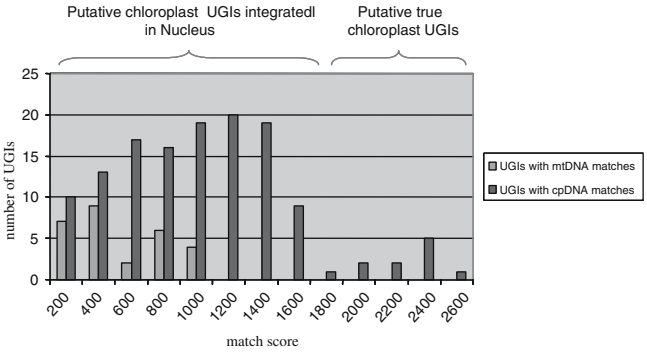

Fig. 2 Histogram of BLASTN scores of UGIs with tobacco chloroplast and Arabidopsis mitochondrial DNA matches. *Two histogram peaks* representing putative chloroplast UGIs integrated in nucleus and putative true chloroplast UGIs are depicted

UGIs corresponding to organellar sequences integrated into the nuclear genome

The remaining 110 UGIs, showed much lower homologies with their organellar counterparts (Fig. 2). Moreover, 84 (76%) of these also contained sequences bearing no homology with organellar genomes. These combined pieces of evidences lead to the conclusion that the majority of these 110 UGIs represent fragments of organellar DNA integrated throughout the nuclear genome, a phenomenon apparently common in all plant genomes (Pichersky et al. 1991; Shahmuradov et al. 2003). Since the majority of these UGIs did not represent full-length organellar genes and/or had premature stop codons, it seems unlikely that they are functional. It is estimated that these sequences represent integrations that have occurred over the time of 4.8–95 million years ago, assuming random decay of the originally integrated sequence and local molecular clock on a phylogenetic tree, as well as 100 million year divergence time between tomato and Arabidopsis (Wikstrom et al. 2001; Yang and Yoder 2003). The level of sequence divergence prior to the estimated 95 million years would likely be too great to allow original detection via homology searches. Likewise, more recent organellar integrations may not have diverged sufficiently to allow differentiation from true organellar sequences with the methodology applied.

A total of 24 UGIs were determined to be mtDNA-like sequences in the tomato nuclear genome, with an average insertion length of 261 bp ranging from 73 to 693 bp. The average sequence identity to Arabidopsis mtDNA of these UGIs was 92% with a range of 80–98%. Similarly, 86 UGIs were identified as cpDNA-like sequences in tomato nuclear genome, with an average integration of 360 bp ranging from 70 to 1,205 bp. The average sequence identity between these cpDNA-like sequences and the tobacco chloroplast counterparts was 96% with a range of 83–100%. The homologous cpDNA and mtDNA segments were uniformly distributed across the tobacco chloroplast genome and Arabidopsis mitochondrial genome (Kolmogorov–Smirnov test of uniform distribution: $P = 0.293$ for tomato homologs distributed on tobacco chloroplast genome and $P = 0.052$ for tomato homologs distributed across the Arabidopsis mitochondrial genome). Moreover, there was also no significant difference between the distribution of UGI counterparts in coding and non-coding chloroplast segments (Kolmogorov–Smirnov test of no difference, $P = 0.1$). Therefore, it was most likely that random portions of the organellar genomes were integrated into the tomato nuclear genome over a long evolutionary period.

Estimate of the proportion of unmethylated tomato DNA with an organellar origin

The entire set of 7,053 UGIs generated in this study comprises approximately 3.6 Mb of undermethylated nuclear genome, 43 kb (1.2%) of which is integrated organellar DNA. If these results are extrapolated to the entire unmethylated portion of the tomato genome (estimated to be approximately 61 Mb—see following section), it is estimated that this portion of the tomato genome contains approximately 1,886 organellar insertions for a total of 734 kb organellar-derived DNA. For comparison, the Arabidopsis and rice genomes are estimated to contain 300 kb (0.2%) and 680 kb (0.2%) of organellar insertions, respectively (www.Arabidopsis.org; www.tigr.org) based on the criteria used for the analysis of UGIs. Since the current study represented

only the unmethylated portion of the tomato genome, the amount and number of organellar insertions in the entire tomato genome is likely to be much greater.

Evidence for the insertion of organellar sequences into nuclear genes

Blast searching the non-organellar portions of these nuclear inserted, organellar-derived UGIs against the tomato EST-derived unigene set and Arabidopsis proteome revealed that only six of these UGIs contain nuclear sequences with homology to putative coding regions. Two of these included sequences with homology to Arabidopsis retroelements (At2g01024 and At2g01028). The other four UGIs contained nuclear sequences with homology to annotated Arabidopsis genes: zinc knuckle (CCHC-type) family protein (At4g00980), proton extrusion protein-related (At4g31040), type II intron maturase (At2g07747), and H^+-transporting two-sector ATPase (At2g07671). Presumably, the insertion of organellar DNA into these genes would render them non-functional. Whether these ancient organellar insertions caused them to become non-functional, or whether they were non-functional prior to insertion, cannot be determined from these data. However, the demonstrations that organellar sequences can insert into coding regions provide further support that organellar genes have periodically been integrated into the nucleus in a manner that they come under the control of nuclear promoters, which may eventually take over the function of the original organellar counterpart. Both functional and non-functional organellar DNA-like sequences were revealed in the Arabidopsis, rice, and tobacco nuclear genomes (Adams et al. 2002; Ayliffe et al. 1998; Huang et al. 2004; Timmis et al. 2004).

Non-organellar UGIs

After elimination of all organellar-derived UGIs, 6,943 UGIs—5751 singletons and 1,192 contigs containing two or more reads were further annotated. The average number of sequence reads per UGI was 1.3, ranging from 1 to 165 (Fig. 1b).

UGIs corresponding to highly repetitive sequences

The distribution of UGI copy number violates a theoretical poisson distribution (likelihood ratio test, $P < 0.01$), based on an assumption that the unmethylated portion of the genome is single copy, due to a large extent by the presence of higher-than-expected UGIs with multiple copy membership (Fig. 1a). If the unmethylated portion of the tomato genome was comprised entirely of single-copy DNA, a draw of 9,990 clones corresponding to 5 Mb (as was done in this project), would be unlikely to generate any contigs with more than eight read members. Nonetheless, nine non-organellar UGIs with >8 read members were observed (Fig. 1b). A manual examination of these nine UGIs revealed that these high-membership UGIs are usually comprised of "stacked" alignments. A stacked alignment refers to each read member sharing high level of identity ($>95\%$) in the aligned nucleotides with the consensus sequences. BLASTN and BLASTX against GenBank sequences showed that these large contigs correspond to retroelements, sequences from large gene families (e.g., serine protease), or ribosomal DNAs. Moreover, retroelements and rDNAs have been reported to be highly repetitive in plant genome, which further indicates that UGIs with large membership corresponding to the repeat sequences in tomato genome.

The UGI (2,009 bp) derived from the largest cluster with 165 MF reads matched tomato 18S rRNA sequence (X51576) and had high sequence identities among MF reads ($>97\%$) (Kiss et al. 1989). Assuming random sampling from the unmethylated portion of tomato genome, 165 copies of 18S rDNA from 3.6 Mb UGIs is not proportional to the 2,300 copies of 45S RNA in the tomato genome (950 Mb) (Ganal et al. 1988). These UGIs contained 497 GpC/CpG and GpNpC/CpNpG sites for potential methylation, implying preferential methylation of a significant portion of rDNA in tomato—a phenomenon already demonstrated in Arabidopsis (Jasencakova et al. 2003; Johnson et al. 2002; Qu et al. 2001).

GC percent content of UGIs

The average GC percent of non-organellar UGIs was 37%, which is the same value as previously reported for the non-coding portion of the tomato genome and among the lowest GC content reported for any plant species (Messeguer et al. 1991; Wagner and Capesius 1981; Yu et al. 2002). The portion of UGIs determined to correspond to coding regions by virtue of homology to the Arabidopsis proteome or tomato EST-derived unigene set (see next section) showed a GC content of 41%—significantly higher than the overall UGI dataset ($P < 0.0001$) and is closer to the 46% GC content previously determined for known tomato genes (Table 2, Messeguer et al. 1991).

Potential methylation sites in non-organellar UGIs

In plant genomes, DNA methylation is maintained preferentially at cytosines in the CpG and CpNpG islands. DNA methylation is one of the important mechanisms to maintain the heterochromatin state of DNA and thus to control the gene expression. Examination of the potential methylation sites in UGIs revealed a significant difference ($P < 0.0001$, t test) between coding and non-coding UGIs (33 vs. 27 CpNpG sites per 500 bp and 27 vs. 23 CpG sites per 500 bp) (Table 2). This result is consistent with the notion that non-coding,

Table 2 The GC content and the distribution of potential methylation sites of CpG or CpNpG in tomato non-organellar UGIs and EST-derived unigenes

	GC%	Number of potential methylation sites per 500 bp					
		GpC	CpG	GpNpC	CpNpG	CpG/GpC	CpNpG/GpNpC
All UGIs	36.7^{c}	15^{cz}	10^{c}	16^{b}	13^{c}	25^{d}	29^{cz}
UGIs with Arabidopsis hits	41.2^{a}	17^{bz}	10^{c}	17^{a}	16^{b}	27^{c}	33^{bz}
UGIs without Arabidopsis hits	36.0^{d}	14^{dz}	10^{c}	16^{b}	11^{d}	24^{e}	27^{dz}
Random sequence	36.9^{c}	17^{b}	17^{a}	17^{a}	17^{b}	34^{a}	34^{b}
Tomato EST-derived unigenes	39.3^{b}	20^{az}	10^{b}	17^{a}	19^{a}	30^{b}	36^{az}

Small superscript letters indicate the significant difference at $P \leq 0.05$ from the multiple comparison results
[z]Indicates that the number of CpG sites is significantly different from the number of GpC sites and the number of CpNpG/GpNpC sites is significantly different from the number of CpG/GpC sites for each type of sequences at $P \leq 0.05$ of paired Student's t test

methylated DNA through deamination and transitions from '5-mC' to 'T' reduces the methylation sites and GC content of non-coding portions of the genome (McClelland 1983; Messeguer et al. 1991; Rabinowicz et al. 2003). It is also consistent with the difference in GC content observed between coding and non-coding UGIs (previous paragraph). Assuming randomized sequences don't possess any biological significance of coding versus non-coding information, a negative control of randomized sequences was created by randomizing the order of nucleotides in each UGI. Randomized sequences had equal numbers of GpC, CpG, GpNpC, and CpNpG sites. Thus, CpNpG/GpNpC occurs much more often than CpG/GpC in the tomato UGIs and EST-derived unigenes, and the unmethylated CpG and CpNpG islands in UGIs are more likely enriched for genic sequences (Table 2). The number of potential methylation sites were also significantly different between EST-derived unigenes and UGIs (36 vs. 33 CpNpG sites per 500 bp and 30 vs. 27 CpG sites per 500 bp), which might indicate the low level of methylation in both coding and UTR regions of EST-derived unigenes.

Annotation of non-organellar UGIs

Of the 6,943 non-organellar nuclear UGIs, 32% (2,210) were classified as putative functional genes based on the gene prediction results from GeneMark and GenScan. As a control, the nucleotides in each UGI were randomized (see Materials and methods for details) and subjected to the same analysis. Of these randomized UGIs, 1,405 were also classified as putative genes using the same criteria. These results suggested that at least half of the putative 2,210 UGIs classified as coding regions might be false positive. If this is indeed the case, it is estimated that only 12% (805 out of 6,943 UGIs) of the UGIs correspond to coding regions based on gene prediction programs.

As an independent test of the percentage of UGIs that correspond to coding regions, the same 6,943 UGIs were blasted against the Arabidopsis proteome and tomato/potato/pepper EST-derived unigene sets (Fig. 3, Table 3). About 33% (2,649) of the UGIs had significant matches (E value $< 10^{-10}$) with either the Solanaceae EST-derived unigene set (1,683 matches), Arabidopsis proteome (1,418 matches), or both (452) (Fig. 3). By comparison, none of the randomized sequences had significant matches to Arabidopsis proteome and tomato EST-derived unigenes (Table 3).

Of the 966 UGIs with only Arabidopsis matches, 704 were transposons-related genes. Thus a total of 1,945 (28%) non-organellar UGIs are predicted to be non-transposon genes. As much as 50% of the UGI detected non-transposons genes have no clear match in the Arabidopsis genome.

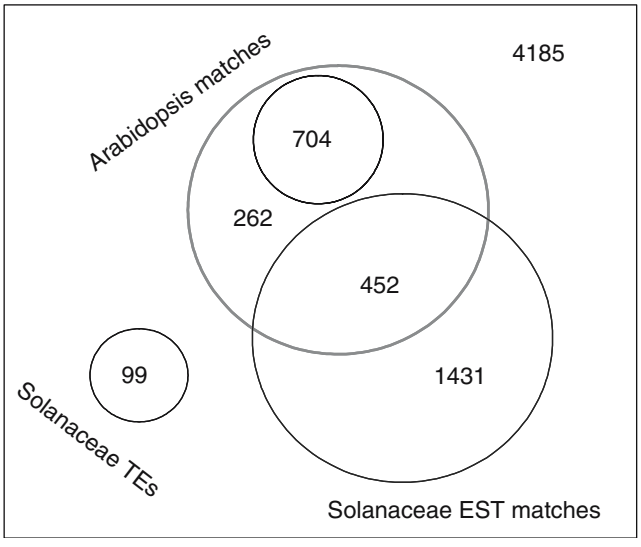

Fig. 3 The number of genic regions identified in UGIs. Two approaches, BLASTN against tomato/potato/pepper EST-derived unigene set and BLASTX against Arabidopsis proteome, touched 1,883 and 1,418 genes, respectively, with an overlap of 452 UGIs potential genes. Since there was 704 TEs in the 966 Arabidopsis-specific matches, 1,945 (28%) sequences were discovered as non-transposon genes with a significant match to either an Arabidopsis predicted protein or a tomato/potato/pepper EST-derived unigene. In total, 4,185 (59%) sequences out of 7,053 UGIs were not identified as genic regions by any of these approaches and 803 (12%) (704 + 99) UGIs corresponded to TEs

Significant portion of UGIs correspond to transposon-related genes

Of the 4,294 UGIs without any match to EST-derived unigenes or the Arabidopsis proteome, 99 had significant sequence similarity with tomato and potato retroelements, for example, TLC1, TONT1, TOPIE1, TORTL1, and TOTO1 (Fig. 3). Of the 1,418 UGIs with matches to the Arabidopsis proteome, 704 were classified as transposable elements (TE). Of these, 652 showed significant homology with class I retroelements in Arabidopsis, such as Hopscotch, gag-pol, Ty1/copia-type and so on. Thus, a total of 751 UGIs (29% of all the putative UGI genes or 11% of all the UGIs in this survey) were putative retroelements. The remainder, 52 (3% of all the putative UGI genes or 0.7% of all the UGIs in this survey) were homologous to known class II DNA transposable elements, for example, TNP2, Ac element, En/Spm, Mutator.

In an effort to shed more light on the transposon content of the unmethylated portion of the tomato genome, the number and categorization of TEs was compared among three independent sequence databases: (1) the UGIs reported herein, (2) ten previously sequenced tomato BACs (five in heterochromatin and five in euchromatin, Zhukuan Cheng, unpublished data), and (3) the EST-derived tomato unigene set (Mao et al. 2001; van der Hoeven et al. 2002). Of 117 predicted genes in ten BACs (Table 1), 36 transposon-related genes were found in five heterochromatic BACs, and two in one euchromatic BAC. Thus, there was similar percentage of TEs found in putative genic UGIs as that in gene models in BACs (29% for UGIs vs. 32% for BACs). Those 38 TEs from BACs were categorized as 37 (97%) class I elements and 1 (3%) as class II element. In contrast, 128 TEs, 97 (76%) class I, and 31 (24%) class II TEs, were identified respectively in the 30,100 tomato EST-derived unigene set. The proportion of class I to class II TEs in BACs (97 vs. 3%) and UGIs (96 vs. 4%) were similar to each other, but significantly higher than that in the tomato EST-derived unigene set (76 vs. 24%) ($P < 0.01$). This finding suggests that DNA methylation is not the only characteristic that differentiates the euchromatin and heterochromatin. Other modifications, such as lysine and arginine methylation, acetylation, and phosphorylation of histones, co-operate to regulate chromatin structure (see review by Lee et al. 2005). In addition, lack of DNA methylation does not assure transposon transcription/activity, as only 0.4% of the tomato EST-derived unigenes versus 29% of the UGI genes correspond to TE. Hence, mechanisms beyond DNA methylation must be in play to repress non-methylated TEs.

Distribution of UGIs across genic regions

The 714 UGIs with non-transposon Arabidopsis matches were mapped against the Arabidopsis gene models based on the high-scoring segment pairs (HSPs) from BLAST results. All these UGIs matched to coding regions of Arabidopsis genes, and contiguous, non-matched sequences were assumed to be 5′ upstream or 3′ downstream non-coding regions. UGIs extended upto 676 bp upstream, with an average of 174 bp, and 766 bp downstream with an average of 171 bp. Former studies also indicated that methylation around the transcription start positions could severely affect the gene expression, while methylated cytosines were found near the 3′ end of the genes (Ashapkin et al. 2002; Rabinowicz et al. 2003; Steimer et al. 2004). Thus, part of the intergenic sequences, especially sequences close to the start and the end sites of tomato genes, in euchromatic regions are likely to be unmethylated. The BLASTN result of tomato UGIs against the 5′ and 3′ UTRs of Arabidopsis genes revealed only short HSPs (< 50 bp) with high identity ($> 90\%$), which are mostly short-sequence repeats. Therefore, it is implausible that the UTRs are conserved between tomato and Arabidopsis. Due to the fact that only UGIs with Arabidopsis hits (37% of putative genes in UGIs) were used for the estimation of 5′ upstream and 3′ downstream extensions, one cannot rule out the possibility that some UGIs, classified as non-coding, may correspond to unmethylated regions extending even further upstream or downstream of functional genes.

Non-coding UGIs

The remaining 4,185 UGIs showed no significant homology with tomato/potato/pepper EST-derived unigenes and Arabidopsis proteome, thus, are presumable non-coding sequences. Moreover, none of these UGIs showed sequence homology to the three most

Table 3 Annotation of UGIs using computational prediction, BLAST search against tomato/potato/pepper EST-derived unigenes and Arabidopsis proteome

	Number of computationally predicted genes	EST matches	Arabidopsis matches	Average length of maximum ORF (bp)
UGIs with Arabidopsis match	821	452	1,418	234[a]
UGIs without Arabidopsis match	1,389	1,431	0	159[c]
All UGIs	2,210	1,883	1,418	177[b]
Randomized sequence	1,405	0	0	128[d]

[z]Small superscript letters indicate the significant difference at $P \leq 0.05$ from the multiple comparison results

highly repeated sequence families in tomato: TGRI (163 bp, X87233), which constitutes the subtelomeric repeat; TGRII (1.8 kp, AY880062), which is found throughout the genome, but more intensely in the centromeric heterochromatin; TGRIII (504 bp, AY880063), which consists of interspersed repeats clustering at or near the centromeres (Ganal et al. 1988). Although the *Eco*RI recognition site (GAATTC) was not found in TGRI, TGRII, and TGRIII, potential methylation sites of GpC and GpNpC exist in these repeats. Therefore, these results suggest that TGRI, TGRII, and TGRIII may be eliminated during cloning due to lack of restriction sites or because they are highly methylated and hence would not survive the methylation screen employed in developing the library for sequencing in this project.

Functional annotation of non-transposon coding UGIs

Of 1,945 UGIs with putative coding regions, functional annotation was focused on 714 (10%) UGIs with significant homologies to non-transposon genes in the Arabidopsis proteome, due to the lack of information for the UGIs that only matched to tomato/potato/pepper EST-derived unigenes. This percentage is similar to studies of methyl-filtered sequences in maize for which 8% of such sequences had significant matches to known genes (Palmer et al. 2003).

Out of 714 UGIs with significant homologies to non-TE Arabidopsis genes, 262 UGIs did not match the tomato/potato/pepper EST-derived unigene sets and thus likely represent tomato genes that were missed during the EST sequencing process. Based on these results, it is estimated that the tomato/potato/pepper EST-derived unigene sets cover approximately 63% of the total genes in the tomato genome.

Comparison of UGIs and genes identified on sequenced tomato BACs

Ten previously sequenced tomato genomic BACs were used for this comparison (see Material and methods for details). In situ hybridization FISH experiments indicate that five BACs are localized in pericentrimeric heterochromatin and the other five BACs in euchromatic regions (Zhukuan Cheng, unpublished data, Table 1). Together, these BACs contain 117 predicted genes—38 of which correspond to transposons. The average length of the 79 non-transposon genes, including coding regions and introns, is 3.2 Kb. A comparison of the 1,945 non-transposon, coding UGIs with these BACs revealed that three of the UGIs had a perfect match to one of the 79 non-transposons genes on the BACs. All three of these genic matches occurred in BACs localized in the euchromatin (FW2.2, 127E11, and 240K04) (Table 1). Three more UGIs mapped to non-genic regions in heterochromatic BACs (181O9, 181K1). The fact that three non-transposon, coding UGIs matched only genes in euchromatic BACs, supports the assertion that the majority of the tomato genes reside in the euchromatin (van der Hoeven et al. 2002).

Estimating the size of the unmethylated portion of the tomato genome

The size (in Mb) of unmethylated DNA in the tomato genome was estimated with the maximum likelihood method using the following pieces of information: (1) m, the average distance between the two *Eco*RI digestion sites in the unmethylated portion of the genome and the number of unique fragments sampled from the tomato genome. This was estimated by determining the insert size of 56 randomly selected clones from the undermethylated *Eco*RI library used in this study. The average was 1.4 kb ± 346 bp with a range of 205–5.7 kb. (2) b, the percentage of *Eco*RI fragments in the tomato genome, which falls within the size constraints of the *Eco*RI library used in this project. This was estimated to be 64% based on the quantification of end-labeled *Eco*RI genomic fragments in the range of 500–4 kb separated via gel electrophoresis. (3) n, the total number of undermethylated clones sequenced, which is 9,990. (4) f, the frequency with which the unmethylated genomic regions were sampled during sequencing. The maximum likelihood estimate for f was 0.354. Thus, the total length of the unmethylated tomato genome should be estimated as mn/fb (see details in Materials and methods), 61 ± 15 Mb.

The previous calculations lead to the prediction that 61 ± 15 Mb of the tomato genome is unmethylated. Based on the annotation of ten sequenced tomato BACs, it is estimated that the average size of a tomato gene, including coding regions and introns, is 3.2 kb (see previous section). If the average length of unmethylated 5′ upstream and 3′ downstream was 345 bp (see previous section), each gene resides in an island of approximate 3.5 kb unmethylated sequence. Assuming all coding regions and introns are fully unmethylated, the unmethylated DNA genome could accommodate 17,500 ± 4,300 (95% confidence interval) genes. This number is substantially less than the 35,000 genes estimated based on analysis of the tomato EST database (van der Hoeven et al. 2002). The difference in these two estimates may be the result of several factors: (1) some tomato genes and promoters may be entirely or partially methylated. As mentioned earlier, there is good evidence for this from Arabidopsis (Ashapkin et al. 2002; Jacobsen and Meyerowitz 1997; Lippman et al. 2004), and CG methylation clusters were also found in the genic regions (Tran et al. 2005); (2) some portion of the tomato genes reside in heterochromatin. Based on the BAC analysis described earlier, while the heterochromatin is very poor in genes, there are clearly some genes that reside in the heterochromatin and insulated from the heterochromatin-related gene silencing with low level of methylation

for example, *Pto* gene and At4g04020 (Lippman et al. 2004; Martin et al. 1993). It was estimated that about 17% of the maize genome (2,600 Mb) and 34% of the sorghum genome (735 Mb) are unmethylated based on the MF sequencing of randomly sheared library (Bedell et al. 2005; Palmer et al. 2003). In addition, the nuclear genomes of Sorghum and maize have an interspersed pattern of coding euchromatin and non-coding heterochromatic regions, which is different from the tomato genome containing centromeric heterochromatin flanked by long stretch of euchromatin (de Jong 1998; Rabinowicz et al. 1999; White and Doebley 1998). Thus, the unmethylated portion of the plant nuclear genome might vary considerably due to the size and structure of the genome.

It is estimated that the euchromatic arms of the tomato genome account for approximately 220 Mb of the total 950 Mb in the tomato genome (Arumuganathan et al. 1991; de Jong 1998). Hence, if only 61 Mb of the genome is unmethylated, methylation must also occur in euchromatic portions of the genome—most likely in the spacer regions between genes. This conjecture is consistent with Messeguer et al. (1991) who reported that tomato nuclear DNA contains islands of unmethylated CpG and CpNpG sites. These combined results support the notion that some intergenic sequences, even in euchromatin, are methylated. If this is correct, then shot-gun sequencing of methyl-filtered clones would not allow assembly of the euchromatic arms. Thus, providing an ordered sequence of the euchromatin would depend on use of other strategies, such as map-based BAC by BAC sequencing as was applied to Arabidopsis (The Arabidopsis genome initiative 2000).

Amount of shot-gun sequences required to cover unmethylated portion of tomato genome

The number of sequence reads required to recover the unmethylated regions, estimated to be from 61 ± 15 Mb, were calculated based on the Lander-Waterman model and the assumption of an average of 600 bp per sequence read and a minimum of 20 bp overlap for sequence assembling (Lander-Waterman 1988). Considering the estimation of 61 Mb unmethylated sequences, 310,000 sequence reads would be required to cover 95% of the unmethylated tomato gene space. Most closely related solanaceous species (e.g., potato, eggplant, pepper) have the same basic chromosome number as tomato ($x = 12$) and similar chromosome structure (pericentrimeric heterochromatin and euchromatic arms) (Gottschalk 1954; Valarik et al. 2004). Moreover, current evidence suggests that most solanaceous species have similar gene content as tomato (Doganlar et al. 2002; Paran et al. 2004). Assuming methylation patterns are also similar in these species, the same calculation for sequencing the unmethylated portion of the genome is also likely to be valid. Currently, the tomato genome is being sequenced on a BAC-by-BAC basis, which will yield, not only gene content, but also gene order (http://www.sgn.cornell.edu/help/about/tomato_sequencing.html). Based on the high degree of gene conservation and synteny among solanaceous species, it should be possible to apply MF sequencing to the genomes of other solanaceous species (e.g., potato, pepper, eggplant, petunia, tobacco), and then use the order of tomato sequence and existing synteny maps to determine the order of the derived UGI genes. Such approach has been used in rat and Drosophila, and would be rapid, cost-effective and capitalize on the more expensive, but necessary BAC-by-BAC sequencing of the tomato genome (Celniker and Rubin 2003; Chen et al. 2001; Pop et al. 2004).

Acknowledgements This work was partially supported by the National Science Foundation Grant DBI-0116076 "Exploitation of tomato as a model for comparative and functional genomics" and 0421634 "Sequence and annotation of the euchromatin of tomato". Sequencing of MF tomato clones was accomplished at the Institute for Genomic Research (TIGR, www.tigr.org). Thanks to Dr. Valentina Vysotskaia from Exelixis, Inc. for sharing sequences of three BACs (181O9, 181C9, and 181K1) with us, and Dr. Rick Durrett and Dr. Pablo Rabinowicz for helpful discussion.

References

Adams KL, Qiu YL, Stoutemyer M, Palmer JD (2002) Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci USA 99:9905–9912

Antequera F, Bird AP (1988) Unmethylated CpG islands associated with genes in higher-plant DNA. Embo J 7:2295–2299

Arumuganathan K, Slattery JP, Tanksley SD, Earle ED (1991) Preparation and flow cytometric analysis of metaphase chromosomes of tomato. Theor Appl Genet 82:101–111

Ashapkin VV, Kutueva LI, Vanyushin BF (2002) The gene for domains rearranged methyltransferase (DRM2) in *Arabidopsis thaliana* plants is methylated at both cytosine and adenine residues. Febs Lett 532:367–372

Ayliffe MA, Scott NS, Timmis JN (1998) Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Mol Biol Evol 15:738–745

Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, Jones J, Flick E, Rohlfing T, Fries J, Bradford K, McMenamy J, Smith M, Holeman H, Roe BA, Wiley G, Korf IF, Rabinowicz PD, Lakey N, McCombie WR, Jeddeloh JA, Martienssen RA (2005) Sorghum genome sequencing by methylation filtration. Plos Biology 3:103–115

Bennetzen JL, Schrick K, Springer PS, Brown WE, Sanmiguel P (1994) Active maize genes are unmodified and flanked by diverse classes of modified highly repetitive DNA. Genome 37:565–576

Bernatzky R, Tanksley SD (1986) Toward a saturated linkage map in tomato based on isozymes and random cDNA Sequences. Genetics 112:887–898

Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16:6–21

Borodovsky M, McIninch J (1993) GeneMark: parallel gene recognition for both DNA strands. Comput Chem 17:123–133

Budiman MA, Mao L, Wood TC, Wing RA (2000) A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome Res 10:129–136

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Burr B, Burr FA, Thompson KH, Albertson MC, Stuber CW (1988) Gene-mapping with recombinant inbreds in maize. Genetics 118:519–526

Cao XF, Jacobsen SE (2002) Locus-specific control of asymmetric and CpNpG methylation by the *DRM* and *CMT3* methyltransferase genes. Proc Natl Acad Sci USA 99:16491–16498

Celniker SE, Rubin GM (2003) The *Drosophila melanogaster* genome. Annu Rev Genomics Hum Genet 4:89–117

Chen R, Bouck JB, Weinstock GM, Gibbs RA (2001) Comparing vertebrate whole-genome shotgun reads to the human genome. Genome Res 11:1807–1816

de Jong JH (1998) High resolution FISH reveals the molecular and chromosomal organisation of repetitive sequences in tomato. Cytogenet Cell Genet 81:104–104

Doganlar S, Frary A, Daunay MC, Lester RN, Tanksley SD (2002) A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the Solanaceae. Genetics 161:1697–1711

Ewing B, Green P (1998) Base-calling of automated sequencer traces using *Phred* I.accuracy assessment. Genome Res 8:186–194

Fojtova M, Van Houdt H, Depicker A, Kovarik A (2003) Epigenetic switch from posttranscriptional to transcriptional silencing is correlated with promoter hypermethylation. Plant Physiol 133:1240–1250

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification analysis and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467

Ganal MW, Lapitan NLV, Tanksley SD (1988) A molecular and cytogenetic survey of major repeated DNA-sequences in tomato (*Lycopersicon esculentum*). Mol Gen Genet 213:262–268

Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchinson D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong JP, Miguel T, Paszkowski U, Zhang SP, Colbert M, Sun WL, Chen LL, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu YS, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science 296:92–100

Gottschalk W (1954) Die chromosomenstruktur der Solanaceen unter Berucksichtigung phylogenetischer Fragestellungen. Chromosoma 6:539–626

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Huang CY, Ayliffe MA, Timmis JN (2004) Simple and complex nuclear loci created by newly transferred chloroplast DNA in tobacco. Proc Natl Acad Sci USA 101:9710–9715

Jacobsen SE, Meyerowitz EM (1997) Hypermethylated *SUPERMAN* epigenetic alleles in Arabidopsis. Science 277:1100–1103

Jasencakova Z, Soppe WJJ, Meister A, Gernand D, Turner BM, Schubert I (2003) Histone modifications in Arabidopsis—high methylation of H3 lysine 9 is dispensable for constitutive heterochromatin. Plant J 33:471–480

Johnson LM, Cao XF, Jacobsen SE (2002) Interplay between two epigenetic marks: DNA methylation and histone H3 lysine 9 methylation. Curr Biol 12:1360–1367

Kakes P (1973) Chromosome number of *Cochlearia pyrenaica* DC near Moresnet (Belgium). Acta Botanica Neerlandica 22:206–208

Kiss T, Szkukalek A, Solymosy F (1989) Nucleotide sequence of a 17S (18S) ribosomal RNA gene from tomato. Nucleic Acids Res 17:2127–2127

Korf I, Yandell M, Bedell J (2003) Blast. O'Reilly & Associates, Inc. Sebastopol, pp 357

Kulikova O, Gualtieri G, Geurts R, Kim D-J, Cook D, Huguet T, de Jong JH, Fransz PF, Bisseling T (2001) Integration of the FISH pachytene and genetic maps of *Medicago truncatula*. Plant J 27:49–58

Lee DY, Teyssier C, Strahl BD, Stallcup MR (2005) Role of protein methylation in regulation of transcription. Endocr Rev 26:147–170

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430:471–476

Majoros WH, Pertea M, Antonescu C, Salzberg SL (2003) GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic genefinders. Nucleic Acids Res 31:3601–3604

Mao L, Begum D, Goff SA, Wing RA (2001) Sequence and analysis of the tomato JOINTLESS locus. Plant Physiol 126:1331–1340

Martienssen R (1998) Transposons DNA methylation and gene control. Trends Genet 14:263–264

Martienssen RA, Rabinowicz PD, O'Shaughnessy A, McCombie WR (2004) Sequencing the maize genome. Curr Opin Plant Biol 7:102–107

Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW, Spivey R, Wu TY, Earle ED, Tanksley SD (1993) Map-based cloning of a protein-kinase gene conferring disease resistance in tomato. Science 262:1432–1436

McClelland M (1983) The frequency and distribution of methylatable DNA sequences in leguminous plant protein coding genes. J Mol Evol 19:346–354

Messeguer R, Ganal MW, Steffens JC, Tanksley SD (1991) Characterization of the level, target sites and inheritance of cytosine methylation in tomato nuclear-DNA. Plant Mol Biol 16:753–770

Nick H, Bowen B, Ferl RJ, Gilnert W (1986) Detection of cytosine methylation in the maize *alcohol dehydrogenase* gene by genomic sequencing. Nature 319:243–246

Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtrations. Science 302:2115–2117

Panstruga R, Buschges R, Piffanelli P, Schulze-Lefert P (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res 26:1056–1062

Paran I, van der Voort JR, Lefebvre V, Jahn M, Landry L, van Schriek M, Tanyolac B, Caranta C, Ben Chaim A, Livingstone K, Palloix A, Peleman J (2004) An integrated genetic linkage map of pepper (*Capsicum spp.*). Mol Breed 13:251–261

Peterson DG, Pearson WR, Stack SM (1998) Characterization of the tomato (*Lycopersicon esculentum*) genome using *in vitro* and *in situ* DNA reassociation. Genome 41:346–356

Peterson DG, Wessler SR, Paterson AH (2002) Efficient capture of unique sequences from eukaryotic genomes. Trends Genet 18:547–550

Pichersky E, Logsdon JM, McGrath JM, Stasys RA (1991) Fragments of plastid DNA in the nuclear genome of tomato—prevalence, chromosomal location, and possible mechanism of integration. Mol Gen Genet 225:453–458

Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. Brief Bioinform 5:237–248

Qu LH, Meng Q, Zhou H, Chen YQ (2001) Identification of 10 novel snoRNA gene clusters from *Arabidopsis thaliana*. Nucleic Acids Res 29:1623–1630

Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nature Genet 23:305–308

Rabinowicz PD, Palmer LE, May BP, Hemann MT, Lowe SW, McCombie WR, Martienssen RA (2003) Genes and transposons are differentially methylated in plants, but not in mammals. Genome Res 13:2658–2664

Raleigh EA, Murray NE, Revel H, Blumenthal RM, Westaway D, Reith AD, Rigby PW, Elhai J, Hanahan D (1988) McrA and McrB restriction phenotypes of some *E coli* strains and implications for gene cloning. Nucleic Acids Res 16:1563–1575

Salamov AA, Solovyev VV (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res 10:516–522

Salzberg S, Delcher A, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26:544–548

Shahmuradov IA, Akbarova YY, Solovyev VV, Aliyev JA (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and Arabidopsis. Plant Mol Biol 52:923–934

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchishinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide-sequence of the tobacco chloroplast genome - its gene organization and expression. Embo J 5:2043–2049

Sutherland E, Coe L, Raleigh EA (1992) McrBC: a multisubunit GTP-dependent restriction endonuclease. J Mol Biol 225:327–348

Steimer A, Schob H, Grossniklaus U (2004) Epigenetic control of plant development: new layers of complexity. Curr Opin Plant Biol 7:11–19

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

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous *Adh* regions of maize and sorghum. Proc Natl Acad Sci USA 96:7409–7414

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123–135

Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S (2005) DNA methylation profiling identifies CG methylation clusters Arabidopsis genes. Curr Biol 15:154–159

Unseld M, Marienfeld JR, Brandt P, Brennicke A (1997) The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nature Genet 15:57–61

Valarik M, Bartos J, Kovarova P, Kubalakova M, de Jong JH, Dolezel J (2004) High-resolution FISH on super-stretched flow-sorted plant chromosomes. Plant J 37:940–950

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 a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14:1441–1456

Wagner I, Capesius I (1981) Determination of 5-methycytosine from plant DNA by high-performance liquid chromatograph. Biochem Biophys Acta 654:52–56

Walker EL, Panavas T (2001) Structural features and methylation patterns associated with paramutation at the *r1* locus of *Zea mays*. Genetics 159:1201–1215

Walbot V, Warren C (1990) DNA methylation in the *Alcohol-dehydrogenase-1* gene of maize. Plant Mol Biol 15:121–125

White S, Doebley J (1998) Of genes and genomes and the origin of maize. Trends Genet 14:327–332

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26:307–316

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond Ser B Biol Sci 268:2211–2220

Ye F, Signer ER (1996) RIGS (repeat-induced gene silencing) in Arabidopsis is transcriptional and alters chromatin configuration. Proc Natl Acad Sci USA 93:10881–10886

Yang ZH, Yoder AD (2003) Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol 52:705–716

Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, Cao ML, Liu J, Sun JD, Tang JB, Chen YJ, Huang XB, Lin W, Ye C, Tong W, Cong LJ, Geng JN, Han YJ, Li L, Li W, Hu GQ, Huang XG, Li WJ, Li J, Liu ZW, Liu JP, Qi QH, Liu JS, Li T, Wang XG, Lu H, Wu TT, Zhu M, Ni PX, Han H, Dong W, Ren XY, Feng XL, Cui P, Li XR, Wang H, Xu X, Zhai WX, Xu Z, Zhang JS, He SJ, Zhang JG, Xu JC, Zhang KL, Zheng XW, Dong JH, Zeng WY, Tao L, Ye J, Tan J, Ren XD, Chen XW, He J, Liu DF, Tian W, Tian CG, Xia HG, Bao QY, Li G, Gao H, Cao T, Zhao WM, Li P, Chen W, Wang XD, Zhang Y, Hu JF, Liu S, Yang J, Zhang GY, Xiong YQ, Li ZJ, Mao L, Zhou CS, Zhu Z, Chen RS, Hao BL, Zheng WM, Chen SY, Guo W, Li GJ, Liu SQ, Tao M, Zhu LH, Yuan LP, Yang HM (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science 296:79–92

Yuan YN, SanMiguel PJ, Bennetzen JL (2002) Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. Genome Res 12:1345–1349

Zemach A, Grafi G (2003) Characterization of *Arabidopsis thaliana* methyl-CpG-binding domain (MBD) proteins. Plant J 34:565–572