

Solanum lycopersicum cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements

Sander A. Peters^{1,2,†,*}, Erwin Datema^{1,2,†}, Dóra Szinay³, Marjo J. van Staveren^{1,2}, Elio G.W.M. Schijlen^{1,2}, Jan C. van Haarst^{1,2}, Tamara Hesselink^{1,2}, Marleen H.C. Abma-Henkens^{1,2}, Yuling Bai⁴, Hans de Jong³, Willem J. Stiekema¹, René M. Klein Lankhorst¹ and Roeland C.H.J. van Ham^{1,2}

¹Wageningen University Centre for Biosystems Genomics, Droevendaalsesteeg 1 6708 PB Wageningen, The Netherlands,

²Plant Research International, Business Unit of Bioscience, cluster Applied Bioinformatics, Plant Research International, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands,

³Laboratory of Genetics, Wageningen University and Research Centre, Arboretumlaan 4, 6703 BD Wageningen, The Netherlands, and

⁴Laboratory of Plant Breeding, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands

Received 19 October 2008; revised 14 January 2009; accepted 23 January 2009; published online 6 March 2009.

*For correspondence (fax +31 317 418094; e-mail sander.peters@wur.nl).

†These authors contributed equally.

SUMMARY

We studied the physical and genetic organization of chromosome 6 of tomato (*Solanum lycopersicum*) cv. Heinz 1706 by combining bacterial artificial chromosome (BAC) sequence analysis, high-information-content fingerprinting, genetic analysis, and BAC-fluorescent *in situ* hybridization (FISH) mapping data. The chromosome positions of 81 anchored seed and extension BACs corresponded in most cases with the linear marker order on the high-density EXPEN 2000 linkage map. We assembled 25 BAC contigs and eight singleton BACs spanning 2.0 Mb of the short-arm euchromatin, 1.8 Mb of the pericentromeric heterochromatin and 6.9 Mb of the long-arm euchromatin. Sequence data were combined with their corresponding genetic and pachytene chromosome positions into an integrated map that covers approximately a third of the chromosome 6 euchromatin and a small part of the pericentromeric heterochromatin. We then compared physical length (Mb), genetic (cM) and chromosome distances (μm) for determining gap sizes between contigs, revealing relative hot and cold spots of recombination. Through sequence annotation we identified several clusters of functionally related genes and an uneven distribution of both gene and repeat sequences between heterochromatin and euchromatin domains. Although a greater number of the non-transposon genes were located in the euchromatin, the highly repetitive (22.4%) pericentromeric heterochromatin displayed an unexpectedly high gene content of one gene per 36.7 kb. Surprisingly, the short-arm euchromatin was relatively rich in repeats as well, with a repeat content of 13.4%, yet the ratio of *Ty3/Gypsy* and *Ty1/Copia* retrotransposable elements across the chromosome clearly distinguished euchromatin (2:3) from heterochromatin (3:2).

Keywords: tomato chromosome 6, fluorescent *in situ* hybridization, BAC sequencing, LTR retrotransposons, gene and repeat annotation, integrated mapping.

INTRODUCTION

The SOL Genomics Network (SGN) is an international consortium of groups that aims to develop the family *Solanaceae* as a model for a systems approach for understanding

plant adaptation and diversification. A cornerstone of the SGN is the International Solanaceae Genome Project (SOL), which aims to sequence the genome of tomato (*Solanum*

lycopersicum) cv. Heinz 1706 (<http://www.sgn.cornell.edu/solanaceae-project>). Within the SOL project the Centre for BioSystems Genomics (CBSG) in the Netherlands takes responsibility for sequencing chromosome 6.

Tomato is a diploid species with 12 chromosomes and a genome size of approximately 950 Mb (Arumuganathan and Earle, 1991). Its chromosome morphology has been well studied and cytogenetic analysis of pachytene chromosomes displays long continuous stretches of less condensed euchromatin in both chromosome arms flanked by highly condensed heterochromatin at the telomere ends and the centromeres (Ganal *et al.*, 1991; De Jong *et al.*, 2000; Chang *et al.*, 2008). Based on deletion studies, the pericentromeric heterochromatin contains approximately 75% of the nuclear DNA (Kush *et al.*, 1964). This was subsequently confirmed by recent heterochromatin estimates (Peterson *et al.*, 1996, 1998; Chang *et al.*, 2008). However, approximately 90% of all non-transposon genes were thought to reside in the euchromatin (Van der Hoeven *et al.*, 2002; Wang *et al.*, 2006) and this led to the initial efforts of the SOL partners to concentrate on the euchromatic gene-rich space of the genome (Van der Hoeven *et al.*, 2002). Initial sequencing efforts revealed that the euchromatin was largely devoid of repetitive sequences and had a gene density of 6.7 kb per gene, similar to Arabidopsis and rice (*Oryza sativa*). In contrast, the pericentromeric heterochromatin displayed a 10- to 100-fold lower gene density and was found to be densely packed with transposable elements, of which the *Ty3/Gypsy* class was the most abundant (Wang *et al.*, 2006). The euchromatin/heterochromatin proportion of the tomato genome has been subject of several cytogenetic studies, and recent measurements yielded estimates of 31 Mb of euchromatin and 28 Mb of heterochromatin for chromosome 6, of which short- and long-arm euchromatin measure 4.1 Mb and 26.9 Mb, respectively (Chang *et al.*, 2008; Szinay *et al.*, 2008).

To reconstruct the euchromatic part of the tomato genome, SOL follows a BAC-by-BAC sequencing approach. Physical mapping is an integral part of the reconstruction process as it provides a framework for ordering and joining sequence data, genetically mapped markers and bacterial artificial chromosome (BAC) scaffolds. A classical global mapping approach known as 'map first sequence second' was used, in which a physical map is constructed from a fingerprinted BAC library and from which in turn a minimal tiling path (MTP) of clones is selected for shotgun sequencing. Deep-coverage tomato *HindIII*, *EcoRI* and *Mbol* BAC libraries have been constructed (Budiman *et al.*, 2000; http://www.sgn.cornell.edu/about/tomato_sequencing.pl) and the *HindIII* library has been fingerprinted at the University of Arizona (<http://www.genome.arizona.edu/fpc/tomato>). However, this physical map was built with fingerprinted contigs (FPC) and was based on low-resolution and low-information-content fingerprinting (HCIF), a technique known to

introduce gaps and false overlaps in the MTP (Meyers *et al.*, 2004). Bacterial artificial chromosome contigs have been linked to genetically mapped markers and have provided a framework of clones available for sequencing, positional cloning and comparative analysis. To this end, approximately 500 overgo probes were designed from sequenced markers mapped on the EXPEN 2000 map, which is a high-density genetic map of tomato constructed from an F₂ population (F₂-2000) of 83 individuals derived from the cross *S. lycopersicum* LA925 × *Solanum pennellii* LA716 (Fulton *et al.*, 2002). These probes have been hybridized to BAC filter arrays in order to link overgo sequences to specific BAC clones. Although overgo screening is simple and efficient, spurious hybridization may cause the occurrence of both false-positive and false-negative BACs, as was described for maize (Cone *et al.*, 2002). While two-thirds of the probes could be unequivocally assigned to single BACs, the initial physical map was relatively unsaturated with anchor points. The low level of anchoring and the limited resolution of fingerprinted contigs prompted us to follow a local rather than global physical mapping strategy. In this so-called sequenced tagged connector (STC) approach (Venter *et al.*, 1996; Batzoglou *et al.*, 1999) large-scale BAC end sequencing followed by similarity searches between seed BACs and BAC end sequences (BES) were used to select favorable BACs for extension walking. After sequencing the extension BAC, the process is reiterated, resulting in contigs which are built stepwise as sequencing progresses. This 'map-as-you-go' procedure has already been validated for tomato (Peters *et al.*, 2006). In the course of the project new BAC extension sequence overlaps and fluorescent *in situ* hybridization (FISH) data progressively became available and were used to continuously update the contig construction and improve the scaffolding. Following contig building, identification of marker sequences on BAC inserts and cytogenetic mapping information provided us with the opportunity to tie physical map data to genetic map data. Here we present the integrated mapping results and the physical and genetic organization of 139 BACs on tomato chromosome 6 as revealed by combining high-information-content fingerprinting (HCIF), genetic mapping data, FISH and sequence annotation.

RESULTS

BACs linked to genetic markers and cytogenetic mapping

To obtain seed BACs with mapped anchor positions on chromosome 6, *HindIII* and *Mbol* tomato BAC libraries were previously screened using overgo hybridization (see http://www.sgn.cornell.edu/maps/physical/overgo_process_explained.pl). In addition, we selected a small number of BACs using amplified fragment length polymorphism (AFLP) analysis (data not shown). We verified the

cytogenetic position of 113 candidate BACs on pachytene chromosomes using BAC-FISH to construct a backbone for BAC walking (see Figure 1, Figure S1 in Supporting Information and Szinay *et al.*, 2008). Fifty-one seed BACs and 30 extension BACs were confirmed to be located on chromosome 6. An additional three BACs landed on both chromosome 6 and on other chromosomes. The other 29 BACs did not hybridize to chromosome 6; 24 showed a single focus on one of the other chromosomes and four had multiple foci onto multiple chromosomes. For one BAC we could not detect a clear signal.

With the aim of linking BAC sequences to the tomato EXPEN 2000 genetic map, the BAC sequences were searched against the tomato marker database from SGN (http://www.sgn.cornell.edu/search/direct_search.pl?search=Markers). In total, 154 markers were identified, of which 88 have been mapped on chromosome 6 and 12 have multiple genetic map locations (Figure 1 and Figure S1 and Table S1). The remaining 54 markers had not previously been mapped on tomato chromosome 6. Surprisingly, we found three markers in chromosome 6 BAC sequences that were genetically mapped on chromosomes 2, 4 and 11, respectively. One marker (cLET-5-M3) mapped on both chromosome 7 and 12 but not on chromosome 6. In addition, six markers were mapped onto both chromosome 6 and on another chromosome, whereas another four markers were genetically mapped at multiple positions on chromosome 6. Nevertheless, for all of these cases the corresponding BACs displayed a single clear fluorescent signal on chromosome 6. We then selected sets of seed BACs for multicolor FISH analysis such that each set shared at least one seed BAC included in another set as a reference (Figure S1). From the cytogenetic positions a linear order of seed BACs was determined. Overall, the cytogenetic mapping order of seed BACs was in agreement with the linear mapping order of anchored BACs to the tomato EXPEN 2000 genetic map, although some striking discrepancies were observed. For example, the region between 0 and 5 cM on the short arm contains markers for which genetic positions are clearly inverted compared with the relative physical positions. Likewise, markers and corresponding BACs, which have been genetically mapped to the long-arm euchromatin regions around 47 cM and 97 cM, have discrepant genetic and cytogenetic map orders (Table S1).

BAC-by-BAC walking, physical mapping and chromosome coverage

A bidirectional BAC walk from 64 seed BACs was initiated. For the short-arm euchromatin we sequenced 29 BACs which comprised 2.0 Mb of non-redundant sequence, covering 49% of the short arm. Approximately 6.9 Mb (26%) of non-redundant sequence was recovered for the long-arm euchromatin from 90 BACs and an additional 1.8 Mb (6%) of

pericentromeric heterochromatin was sequenced from 20 BACs (Figure 1).

We assessed the accuracy of 'global' physical mapping for placing the sequenced BACs in contigs on the physical map. To this end, we used 131 SNaPshot fingerprinted BACs, consisting of 12 seed BACs which were cytogenetically mapped to the short-arm euchromatin of chromosome 6 and 119 candidate extension BACs. The fingerprints of these BACs were assembled in a single round with FPC, resulting in seven contigs. After sequencing and assembly, the BAC order of these contigs was compared with the order of the corresponding BES as they were aligned to the 2 Mb of assembled sequence of the short arm. Within six contigs, of which the largest started at SL_Mbo_115P13 and ended with SL_Mbo_134P07, the linear BAC order was identical to the order derived from BES (Figure S2). However in the seventh contig seed BACs clustered into a stacked assembly along with other extension candidates (Figure S3). These BACs localized to non-adjacent positions on the genetic map as was confirmed by cytogenetic mapping (Figure 1 and Figure S1). Assembly analysis indeed indicated that these BACs did not share sufficient sequence overlap to properly assemble, and sequence annotation revealed that these BACs had a high repeat content. Only after exclusion of these repetitive BACs could a mapping order be produced that was in agreement with the contig order as determined from the assembled BES.

In order to avoid the repeat resolution problem associated with 'global' physical mapping we instead focused our efforts on 'local' physical mapping. Using the sequence tagged connector (STC) approach, BES were assembled on seed BAC insert sequences and analyzed with TOPAAS as described previously (Peters *et al.*, 2006). Candidate extension BACs that passed the TOPAAS quality control were subsequently fingerprinted with a HCIF technique using SNaPshot (Luo *et al.*, 2003). Each set of fingerprinted BACs, consisting of a single seed and corresponding candidate extension BACs, was then assembled with FPC into a single contig in multiple rounds.

In total we placed 142 BACs on the physical map, 139 of which were sequenced. The average overlap between BACs in a supercontig was 13.3 kb. The relative order in which supercontigs were placed on the physical map was primarily determined by the FISH map position of the seed BACs (Figure 1 and Figure S1, and Szinay *et al.*, 2008).

During the construction of a MTP for the euchromatic regions of chromosome 6 we identified several domains that were poorly covered by seed BACs, and these areas reflected the physical gaps that were not yet bridged in the BAC walking process. To estimate a global bp/cM relationship, chromosome distances on a micrometer scale were determined for the euchromatic and heterochromatic portion (Table 1). In addition, we estimated gap sizes as a fraction of the total euchromatin size by measuring physical distances

Figure 1. Physical coverage and integrated map for tomato chromosome 6.

Cytogenetic mapping positions of seed and extension bacterial artificial chromosomes (BACs) are displayed in the left panel. Chromosome 6 markers and their corresponding genetic positions on the cM scale are displayed in the middle panel. The right panel displays the reconstructed BAC minimal tiling path of 25 supercontigs and 8 singleton BACs in accordance with the order of cytogenetically mapped seed BACs. Each BAC is shown with its identifier displayed in a colored box: *HindIII* BACs in red, *MboI* BACs in blue and *EcoRI* BACs in green. Solid colors represent seed BACs, and transparent colors display extension BACs. Seed BACs identified by overgo hybridization have the corresponding marker identifier depicted in blue. Markers found by BLASTN analysis have red colored identifiers. The BACs with a fluorescent *in situ* hybridization (FISH) confirmed chromosome 6 position containing a non-chromosome 6 genetic marker, have the corresponding marker identifier depicted in green. Markers located on both chromosome 6 and other chromosomes are depicted in pink. Markers from a genetic map other than the EXPEN 2000 map are in italics.

Table 1 Physical and genetic distances of the short arm (6SE) and long arm (6LE) euchromatin region and the heterochromatin region (6SH + 6LH) of chromosome 6

Domain	Chromosome distance (μm)	Genetic distance (cM)	Size (Mb)	Mb μm^{-1}	cM μm^{-1}	Mb cM^{-1}
6SE	1.80	10	4.1	2.27	5.55	0.41
6SH+6LH	8.00	8.5	50	6.25	1.06	5.94
6LE	29.50	82.5	26.9	0.91	2.80	0.33
Telomere–H147H20	0.70	n.d.	1.59	2.27	n.d.	n.d.
H055E14–H106A20	7.70	20	7.0	0.91	2.60	0.35
H034C13–H098L02	1.29	0.8	1.17	0.91	0.62	1.46
H021K07–telomere	1.11	n.d.	1.44	1.30	n.d.	n.d.

n.d., not determined.

between adjacent BAC-FISH positions in pachytene complements flanking supercontigs (Table 1 and Table S1). As an example, we observed a large gap towards the bottom of the long arm, which was flanked by BACs LE_HBa_055E14 and SL_Mbo_106A20 on the physical map. On the genetic map this gap was flanked by marker T0405 (73 cM) and marker C2_At1g16870 (92.5 cM) (Figure 2). The bp/cM ratio for the corresponding interval was 0.35 Mb/cM and this was comparable to what we observed for the long-arm euchromatin portion of the chromosome, but almost two times more than the 0.2 Mb/cM ratio that was observed for most of chromosome 2 (Koo *et al.*, 2008). This distance makes up approximately 20% of the total linkage group of chromosome 6 and approximately 7 Mb on the physical map. While we could place two seed BACs in this gap, extension of these BACs as well as extension of the BACs bordering the gap was unsuccessful. Proximal to the long arm telomere, in the 97 to 98 cM interval, we observed a gap between LE_HBa_034C13 and LE_HBa_098L02. This region has a relatively high bp/cM ratio of 1.2 Mb/cM and thus this small genetic gap corresponds to a considerable physical gap of 0.96 Mb.

Repeat and gene distribution in euchromatic and heterochromatic domains of chromosome 6

An additional advantage of BAC-FISH on pachytene chromosomes is the heterochromatin differentiation of the distal ends and the pericentromere (Chang *et al.*, 2008). The short arm displayed a relatively clear and distinct border of highly condensed heterochromatin and less condensed euchromatin, whereas the long arm shows a

gradual transition of denser heterochromatin to euchromatin (Figures S4 and S5). We therefore focused on these borders to establish the boundaries of the repeat-rich heterochromatin. The short-arm euchromatin spanned between LE_HBa_016K14 just below the telomere region and LE_HBa_304P16 just north of the pericentromeric region (Figure 2 and Figure S4). On the genetic map these BACs were mapped at 0 cM and 10 cM, respectively. On the long arm SL_Mbo_082G10 landed just south of the pericentromeric domain, mapping at 18.5 cM (Figure 2). LE_HBa_315H13 and LE_HBa_021K07 (Figures 1 and 2, Figure S1, S4, S5) were localized near the long-arm telomeric domain mapping at 98 cM. Using these borders, the assembled sequence data were divided into three domains based on the local chromatin status: the short-arm euchromatin (the first four supercontigs in Figure 1 measuring approximately 2.0 Mb), the pericentromeric heterochromatin (1.8 Mb) and the long-arm euchromatin (the final 15 supercontigs in Figure 1 spanning 6.9 Mb).

Figure 2 displays the gene and repeat content of the assembled sequence contigs divided into bins of 50 kb and clearly identifies the different chromatin domains based on their repeat and gene content. The short-arm euchromatin had an average repeat content of 13.4%. In contrast, the repeat content of the pericentromeric heterochromatin measured 22.4%, yet a clear transition in repeat density between euchromatin and heterochromatin was not detected. The long-arm euchromatin was almost devoid of repetitive sequences; however, the first Mb of the long arm, which contains BAC contigs that map to the transition between the tightly condensed heterochromatin and the

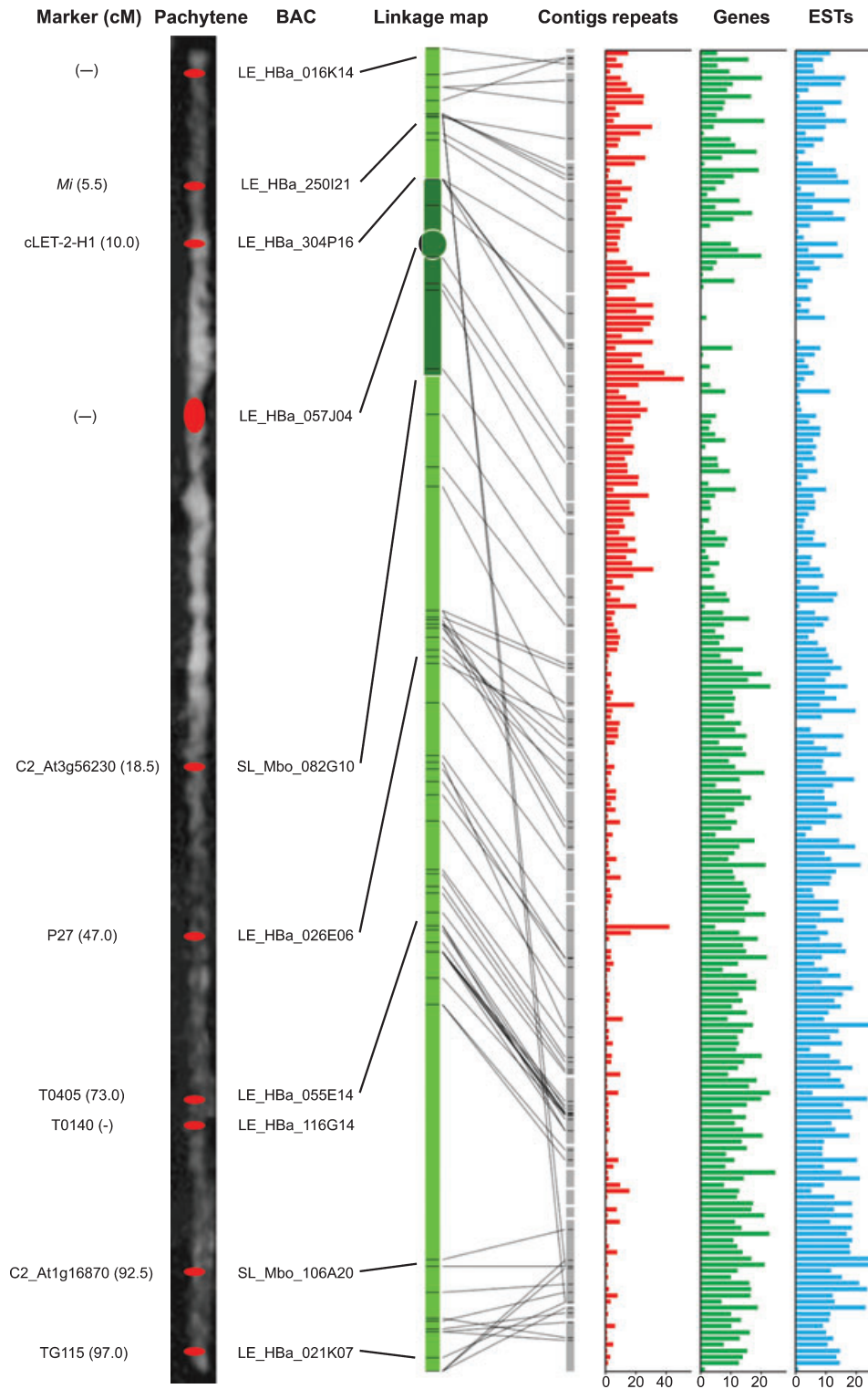


Figure 2. Predicted repeat and gene content of the assembled chromosome 6 sequence. On the left a straightened 4',6-diamidino-2-phenylindole (DAPI) stained chromosome 6 of tomato (*Solanum lycopersicum*) cv. Heinz 1706, with a selected set of bacterial artificial chromosomes (BACs) and corresponding genetic markers as referred to in the text. Intense white and a reduced DAPI fluorescence correspond to condensed heterochromatin and less-condensed euchromatin, respectively. The superimposed colored dots correspond to the BAC-fluorescent *in situ* hybridization (FISH) position taken from Figure 1 and Figure S1. To the right of the chromosome is a cartoon representation of the genetic map on which the assembled supercontigs have been anchored. The three histograms on the right reflect repeats, genes and aligned transcript (expressed sequence tag, EST) contents of the supercontigs. Each bar in the histograms represents a 50-kb interval of the assembled sequence.

more relaxed euchromatin as observed in pachytene chromosomes, clearly had a higher repeat content compared to the rest of the long arm (Figure 2). We also detected a single 20-kb insert of a *Ty3/Gypsy* type retrotransposon similar to *Ogre* (Macas and Neumann, 2007) in an otherwise repeat-poor region (the large peak on the long arm in Figure 2).

The repeat content as a whole, but also *Ty1/Copia* and *Ty3/Gypsy* retrotransposons, was not uniformly distributed across the different chromosome domains. The short-arm euchromatin contained more *Ty1/Copia* than *Ty3/Gypsy* related sequences (7.7% and 5.2%, respectively), whereas *Ty3/Gypsy* retrotransposons were more prevalent in the pericentromeric heterochromatin (14.7% *Ty3/Gypsy* versus 9.0% *Ty1/Copia*). The long-arm euchromatin contained only 2.6% *Ty1/Copia* and 1.5% *Ty3/Gypsy* related sequences. The amount and distribution of DNA transposons also differed between the three domains (Table S2). Approximately one-quarter of the DNA transposons in the short-arm euchromatin were identified as *hAT* and *hAT-Ac* transposons (Rubinl *et al.*, 2001), whereas these only comprised one-sixth and one-ninth of the heterochromatin and long-arm DNA transposon content, respectively. While the overall DNA transposon density of the pericentromeric heterochromatin was similar to that of the short arm, this domain contained a greater number of *En-Spm* and *MuDR* transposons (Bennetzen, 1996; Gierl, 1996). The distribution of several previously identified tomato repeats also differed between the chromosome domains. For example, short interspersed elements (SINEs; the *SoLSINE* family) and terminal repeat retrotransposons in miniature (TRIMs; *Tork2*) were absent from the pericentromeric heterochromatin but prevalent in the euchromatin (Table S3).

In total, 970 putative protein-coding genes were predicted and their distribution of genes varied remarkably between the chromosome domains. The short-arm euchromatin contained one gene per 15.3 kb, whereas the gene density in the long-arm euchromatin was almost twice as high, with one gene per 8.8 kb. Forty-eight genes were predicted in the pericentromeric heterochromatin, yielding an unexpectedly high gene content of one gene per 36.7 kb. With on average 4.0 exons per gene, the genes in this domain differed from those in the short- and long-arm euchromatin, which had 5.0 and 5.2 exons per gene, respectively. This difference was also reflected by a higher relative abundance of single- and two-exon genes and a lower abundance of genes with nine or more exons in the heterochromatin (Figure S6). Several heterochromatic contigs contained 'gene islands' that were separated by long stretches of retrotransposons, whereas others showed a more even distribution of genes. For example, six genes were predicted on the 160 kb supercontig consisting of seed BACs LE_HBa_003K02 and LE_HBa_271L05, which were cytogenetically mapped close to the centromere (Figure S5). For three of these genes ample expressed sequence tag (EST) evidence was found, indicat-

ing that the gene-containing areas extend well into the pericentromeric heterochromatin. On the other hand, no genes were predicted on BAC LE_HBa_057J04, which was located nearest to the centromere. This BAC contained several copies of the centromere-associated TGRIV retrotransposon but we could not identify any large CAA blocks as found near the chromosome 12 centromere, nor any PCRT (pericentromer repeat)-related sequences (Yang *et al.*, 2005; Chang *et al.*, 2008). The BAC contigs that were mapped nearest to the telomeres did not show a decline in gene content, nor did they contain long stretches of subtelomeric TGR1 repeats (Chang *et al.*, 2008), probably because the outermost BAC contigs are not sufficiently close to the telomere to delimit the euchromatin (Table 1). Strikingly, a small cluster of TGR1 repeats (23 copies and several fragments; Table S3) was observed on the long arm, probably corresponding to the interstitial site previously identified through FISH (Zhong *et al.*, 1998).

Six clusters of functionally related genes were identified in the long-arm euchromatin based on GO annotation. The contig harboring the P27 locus held two such clusters: one cluster of four putative GDSL-motif lipases upstream of P27 and another cluster of four predicted Tospovirus resistance genes just downstream of P27. Close to marker TG472 we found a cluster of five (potentially six, when considering *ab initio* gene predictions and BLASTX alignments) cytochrome P450 genes. Four genes resembling ABC2 transporters were found clustered together near marker T0534; a cluster of four (potentially five) single-exon Agenet homologs was predicted in a markerless region on BAC LE_HBa_036J15; and a group of four putative wound-inducible carboxypeptidases was identified between markers T1114 and T1124.

DISCUSSION

Physical mapping accuracy

We used BAC-FISH for cytogenetic confirmation of the chromosomal position of seed and extension BACs that were selected for sequencing the euchromatin parts of chromosome 6. Sequence data progressively became available in the course of this study and 31% of the markers we identified had not been previously mapped on the chromosome. We identified seven EST-derived markers that mapped in addition to chromosome 6 primarily on chromosome 3 and 9, suggesting gene duplications between these chromosomes. Overall, chromosomal positions of the BACs were generally in agreement with the genetically mapped marker order from the EXPEN 2000 map, although we observed several types of inconsistency. These markers may have been erroneously mapped, or the BACs that were picked up with these markers may have been aberrantly identified in the overgo screening process as a result of

spurious hybridization. Alternatively, discrepancies with the genetic map may be the result of different genotypes that have been used for the EXPEN 2000 map construction. This genetic map of tomato is based on a segregating population of *S. lycopersicum* LA 295 × *S. pennellii* LA716 and may be biased by small chromosomal rearrangements between the parents. Furthermore, rearrangements may exist between these lines and Heinz 1706. Nevertheless, the integrated map presents a crosslink between genetic markers, sequences and (cyto)genetic locations of BACs on tomato chromosome 6 and as such is a very valuable resource for the tomato research community.

We assessed the accuracy of the 'local' and 'global' fingerprint mapping approaches. In contrast to a global map-based approach, the STC approach bypasses the need for a global physical map and requires fewer BACs to be fingerprinted. However, the reduced fingerprinting effort comes at the expense of time-saving massive parallel fingerprinting and mapping as it needs successive rounds of fingerprinting for each individual BAC extension. In addition, continuous rounds of BLASTN similarity searches are needed for each successive bidirectional extension. Our 'local' mapping approach involved fingerprinting a single seed BAC together with candidate extension BACs and thereby the seed BAC was isolated from repetitive sequences in other BACs. In this way a drastic reduction of complexity circumvented the adverse effect of repeats that caused repetitive BACs to cluster. Whereas 'global' mapping was less accurate and suffered from repetitive BACs that were displaced in the FPC map, SNaPshot fingerprinting combined with local FPC mapping produced more reliable results and was more robust when mapping high repeat containing BACs. Thus, while the existing 'global' FPC map provides a valuable resource for selecting candidate extension BACs, 'local' FPC maps can resolve repeat-rich contigs, which are abundant in large plant genomes such as tomato.

Chromosome structure and organization

The higher abundance of *Ty3/Gypsy* in chromosome 6 supports earlier observations on the unequal ratio of *Ty3/Gypsy* and *Ty1/Copia* retrotransposons in the tomato genome. A survey of more than 300,000 tomato BESs presented a *Ty3/Gypsy*:*Ty1/Copia* ratio between 2:1 and 3:1 (Datema *et al.*, 2008), and in the current study we found that this ratio varied between the gene-poor heterochromatin (roughly 3:2) and gene-rich euchromatin (roughly 2:3) domains of chromosome 6 (Table S2). An insertion bias for retrotransposons has previously been described for other plant genomes, including *A. thaliana*, *Cestrum* spp., members of the genus *Helianthus*, conifers and maize (Friesen *et al.*, 2001; Pereira, 2004; Natali *et al.*, 2006; Fregonezi *et al.*, 2007; Liu *et al.*, 2007), and several families of retrotransposons have

been identified in maize that preferentially associate with gene-rich or gene-poor regions (Liu *et al.*, 2007). Tam *et al.* (2007) showed that the distribution of *Ty1/Copia* retrotransposons in tomato and related wild species are determined by factors such as genetic drift and mating system, but not recombination rate. We observed an enrichment of *Ty1/Copia* retrotransposons in the short-arm euchromatin and *Ty3/Gypsy* retrotransposons in the pericentromeric heterochromatin, and we know from previous research that the TGRIV *Ty3/Gypsy* retrotransposon is preferentially located in the structural centromeres of tomato (Chang *et al.*, 2008). Consistent with previous genome-wide cytogenetic studies we also observed a preferential localization of TGRII and TGRIII in the pericentromeric heterochromatin (Table S3; Chang *et al.*, 2008). Taken together these data strongly suggest that insertion preferences of different types of retrotransposons also occurred in the tomato genome. It has been suggested that transposable elements play a major role in genome organization, evolution, gene regulation and function (Kazanian, 2004); however, it is not clear how mobile elements have affected the evolution of genes and function in the tomato genome. Additional studies on the insertion distributions are needed in order to understand better the role of transposable elements on tomato genome evolution.

Functional annotation of the predicted genes revealed 15 putative cytochrome P450 genes, thereby reinforcing the observation of a large number of cytochrome P450 domains in the BES data representing 19% of the tomato genome (Datema *et al.*, 2008). Five P450 genes were found clustered together and another six of these genes were present in three pairs. In total we found six clusters of four or more genes that overlapped in their GO annotation. Recently a conserved cluster of four genes was identified in Arabidopsis and oat that plays a role in triterpene synthesis (Field and Osbourn, 2008) and similar metabolic gene clusters have been identified in maize and rice (Gierl and Frey, 2001; Qi *et al.*, 2004). The clusters of lowly transcribed genes we found on chromosome 6 could also indicate a functional relationship between these genes (Hurst *et al.*, 2004). While we only studied one chromosome, these findings may hold true for the complete tomato genome. Using Fisher's exact test we could not identify any significantly over- or under-represented GO terms between the genes annotated on chromosome 6 and the GO terms found in the genome-wide study of the tomato BES (Datema *et al.*, 2008 and data not shown), indicating that the annotation of chromosome 6 can be considered an informative sample of the genome as a whole. As well as exploiting the hierarchical relationships between different GO terms, manual curation of the sequence annotation can also be used to identify more and larger clusters of functionally related genes in tomato and in this way shed light on the molecular evolution of the genome (Yi *et al.*, 2007).

Genetics of tomato chromosome 6

The higher abundance of retrotransposons and repeats in the short-arm euchromatin co-localized with disease resistance gene (*R* gene) loci near markers T1188 and *Mi* in BACs LE_HBa_019E05 and LE_HBa_250I21, respectively. Analysis of 4',6-diamidino-2-phenylindole (DAPI)-stained tomato pachytene chromosomes showed the morphological differentiation into euchromatic and heterochromatic parts. The BAC-FISH analysis showed the repeat-rich regions to be euchromatic in nature. These findings indicate that the distribution of relatively gene-poor and repeat-rich domains is not necessarily confined to the heterochromatin, but also extends to the less condensed short-arm euchromatin. The repeat-rich nature coincides with a suppression of recombination that has been noticed for these regions. The accumulation of retrotransposons as a result of recombinatorial suppression in repeat-rich regions has been observed in the maize genome and is suggested to be a general property of interstitial gene-poor domains intermixed with euchromatin from maize (Liu *et al.*, 2007). Currently, we do not know whether there is a relationship in tomato between repeat content and recombination frequencies.

A previous study revealed a chromosomal inversion in the *Mi-1* region between *S. lycopersicum* and *Solanum peruvianum* (donor of the *Mi-1* gene), which might explain the severe recombination suppression in this region (Seah *et al.*, 2004). Molecular markers flanking two different alleles of the *Mi*-homologs were in the same relative orientation, but markers between the two clusters were in an inverse orientation. Interestingly, a macro-synteny study between tomato (including Cherry VFNT and Heinz 1706) and potato using cross-species BAC-FISH painting revealed a paracentric inversion in the short arm of chromosome 6 that covers the whole euchromatic part. This inversion may have reshuffled the gene order and affected gene function (Tang *et al.*, 2008). If this inversion occurred in wild tomato species, we would expect suppression in recombination frequency for interspecific crosses. Another cluster of *R* genes was found on the long arm near marker P27 (47 cM). For this region a suppression of recombination has been observed (YB, unpublished results). Interestingly, some of the markers that map in this region (including P27 and C2_At4g10030) also appear in reverse orientation compared with the order of the associated BACs in the FISH map.

Distribution of the repeat and gene space

Many genetic markers, such as EST-derived markers, conserved ortholog set (COS) markers and cDNA-derived RFLPs, which anchor BACs to the tomato EXPEN 2000 genetic map, correspond to genes. This has resulted in a small number of anchored BACs in gene-poor and repeat-rich regions and has led, amongst others, to a discontinuous BAC minimal

tiling path for several repeat-rich regions of tomato chromosome 6. Tanksley *et al.* (1992) noted large gaps in the molecular map of tomato and suggested that these gaps might represent areas which were deficient in genes and low-copy sequences. Alternatively, the large genetic distance combined with a relative short physical size, as also observed for the 73 to 93 cM interval on the long arm of chromosome 6, might be explained as a result of a high recombination frequency in this region. In contrast, we observed a relative cold spot of recombination for the 97–98 cM interval, which was reflected by a five-fold increase in the bp/cM ratio. The inconsistent order of markers on the genetic map might be explained by the low rate of recombination in this region.

The mapping and sequence analysis effort presented here is aimed at getting an overview of the gene-rich space of tomato chromosome 6. Since the bulk of all non-transposon-related genes are currently thought to reside in the euchromatin (Van der Hoeven *et al.*, 2002; Wang *et al.*, 2006), we delineated the euchromatin borders by combining BAC-FISH on pachytene chromosomes, sequence annotation and genetic mapping data. While we indeed identified a high repeat content in the pericentromeric heterochromatin, 48 genes were predicted in the 1.8 Mb that represent the various regions of heterochromatic sequence, and transcription for many of these genes was detected from the EST data. The high gene density of these regions corresponds well with that found for six heterochromatic BACs from chromosomes 2, 7, 8, 9 and 10 (Wang *et al.*, 2006) as well as that of the *jointless-2* locus in the pericentromeric region of chromosome 12 (Yang *et al.*, 2005). These regions contain one gene per 56 and 65 kb, respectively; however, they were selected for sequencing using gene-based markers or other single-copy sequences. Similarly, the heterochromatic BACs in this study could also present a biased view, yet the seed BACs of four of these contigs (BACs LE_HBa_003K02, LE_HBa_057J04, LE_HBa_040F08 and LE_HBa_308F14) were selected on the basis of AFLP markers and not on gene-based markers (data not shown). These four contigs span 679 kb and contain 22 genes, implying that the gene content of these contigs is not substantially different from those identified by gene-based markers. Considering the 28 Mb of heterochromatic sequence of chromosome 6, a substantial fraction of tomato genes may in fact reside in the pericentromeric domain.

Recent discoveries have reported on large numbers of heterochromatic genes in plants, mammals and *Drosophila* (Yasuhara and Wakimoto, 2006); in the latter, the expression of such genes has been shown to be dependent on the heterochromatin environment (Weiler and Wakimoto, 1995). The genes we found in the pericentromeric heterochromatin contained on average fewer exons than euchromatic genes and were often grouped into 'gene islands' separated by stretches of retrotransposons, as was previously reported

for the FER locus (Guyot *et al.*, 2005). While we currently do not know whether these findings imply an adaptation of gene structure and organization to this chromosome domain, our observations confirm that tomato heterochromatin cannot merely be regarded as being a functionally inactive region with respect to gene expression.

The predicted gene content and the amount of transcriptional evidence were generally in good agreement with each other. However, several bins had high predicted gene content yet a small amount of EST evidence (Figure 2). One bin in BAC LE_HBa_107H05 contained three putative leucine-rich repeat protein kinases. Two other bins corresponded to the *Mi* locus, which harbored three putative NBS-ARC-LRR disease resistance genes, amongst other predicted genes. We were unable to delineate the proposed topology of the *Mi* locus (Seah *et al.*, 2004), although we predicted one additional disease resistance gene approximately 300 kb upstream and a second one close by based on *ab initio* gene predictions and BLASTX alignments. This chromosomal region is a hot spot of *R* genes. These genes are often clustered in tandem arrays including *Cf-2/Cf-5*, *OI-4/OI-6*, *Mi-1/Mi-9* and *Ty-1* genes conferring resistance to several unrelated pathogens (Bai *et al.*, 2004). Another two bins of high putative gene content and low expression were found in the long arm euchromatin, of which one bin mapped between markers P27 and C2_At4g10030 and contained the cluster of putative Tospovirus resistance genes. Interestingly, *R* genes for virus resistance have been mapped in this region (Ji *et al.*, 2007). Some of the genes in these bins could be non-functional, or were simply not transcribed in detectable amounts under the conditions in which the EST libraries were generated. We also identified a number of gene-poor bins in the pericentromeric heterochromatin, in the short-arm euchromatin and the first Mb of the long-arm euchromatin that contained a large number of aligned transcripts. A considerable number of ESTs in these bins matched to retrotransposon-related genomic sequences, providing further circumstantial evidence for the presence of transcriptionally active retrotransposons in the tomato genome (Manetti *et al.*, 2007).

BAC walking and finishing the tomato genome

We observed chromosomal regions which are not yet targeted with markers, and this has resulted in an assembly containing megabase-sized gaps between BAC supercontigs. Without the identification of new seed BACs the bridging of these gaps by BAC walking will be difficult and time-consuming. To complement the chromosome 6 sequencing project, and indeed the whole tomato sequencing project, we will consider additional approaches. Sequence comparison between tomato and BACs from other *Solanum* species combined with cross-species

multicolor BAC-FISH painting may allow identification of new candidate seed BACs. Whole-genome shotgun sequencing with next generation sequencing (NGS) platforms will undoubtedly speed up and help to complete the sequencing. Yet, while the NGS platforms represent powerful technologies that produce large numbers of sequences, chromosome-based assemblies including the vast amount of repetitive sequences will be a major challenge. Nevertheless, such a whole genome sequencing approach can benefit from the sequence islands that have been produced and which may serve as a backbone for whole chromosome assembly.

EXPERIMENTAL PROCEDURES

Chromosome preparations

All FISH experiments were performed on tomato *S. lycopersicum* cv. Heinz 1706 ($2n = 2x = 24$). The pachytene preparations from young anthers containing pollen mother cells and the spreads of extended DNA fibers from young leaves were made following the protocols of Zhong *et al.* (1996a) and Budiman *et al.* (2004).

Fluorescence in situ hybridization (FISH)

Two-color and multicolor FISH of BAC clones to pachytene chromosomes were performed according to the FISH protocols (Zhong *et al.*, 1996b). Slides were examined under a Zeiss Axioplan 2 Imaging Photomicroscope (<http://www.zeiss.com/>) equipped with epifluorescence illumination, filter sets for DAPI, fluorescein isothiocyanate (FITC), Cy3, Cy5, diethyl aminomethyl coumarin (DEAC) and Cy3.5 fluorescence. To determine distances (in micrometers) between BACs each distance was measured on 10 straightened chromosomes and then averaged. Capturing of selected images, image processing and distance measurements were performed as previously described (Szinay *et al.*, 2008).

The BAC clones

The BAC clones were inoculated in 48-well blocks containing 2.5 ml $2 \times$ LB medium with $12.5 \mu\text{g ml}^{-1}$ chloramphenicol and were grown for 20 h at 37°C and 175 rpm. Overnight cultures were centrifuged at 5796 g for 10 min. Cell pellets were used for high-throughput BAC DNA isolation in a 96-well plate format using a Biorobot 9600 and the R.E.A.L prep kit (Qiagen, <http://www.qiagen.com/>). The BAC DNA was dissolved overnight at 4°C in 50 μl MilliQ. Quantification of the BAC DNA yield was carried out by fluorescence detection (Tecan XFluor4; <http://www.tecan.com/>) in the presence of Quant iT PicoGreen [Molecular Probes, <http://www.invitrogen.com/> (see brands)].

Fingerprinting reaction

On average 1 μg BAC DNA digested by *Bam*HI, *Eco*RI, *Xba*I, *Xho*I and *Hae*III (Invitrogen, <http://www.invitrogen.com/>) according to the protocol described by Luo *et al.* (2003) and labeled with the SNaPshot Multiplex Ready Reaction Mix (ABI, <http://www3.appliedbiosystems.com/>). Sedimented and labeled DNA was washed with 100- μl 70% ethanol and centrifuged again at 3650 g

for 10 min. Air-dried pellets were dissolved in 10 μ l of Hi-Di formamide and 0.5 μ l of internal size standard LIZ-600 (Applied Biosystems). The DNA was denatured at 95°C for 5 min and chilled on ice and analyzed on a ABI 3730 Genetic Analyzer (Applied Biosystems) using POP-7 polymer. Peak detection, intensity and collection were executed with the Any5 Dye-set by the data-collection software version 3.0 (Applied Biosystems).

FPC data processing, BAC selection and assessment of mapping accuracy

The ABI 3730XL Genetic Analyzer data was processed with Genemapper 4.0 (Applied Biosystems). Typically, 115 fragments \pm 52 per fingerprinted BAC with a size range of 100–600 bp were selected with GeneMapper at a peak intensity threshold of 100. The Genemapper sizing results were processed by Genoprofiler 2.1 (You *et al.*, 2007) and converted into a FPC usable format (Soderlund *et al.*, 2000).

The procedure to identify minimal overlapping clones and maximal extending inserts was as previously described (Peters *et al.*, 2006). SNaPshot fingerprinted tomato BACs were mapped with FPC with an optimal probability of coincidence of 10^{-7} and a tolerance level of 0.4 bp, which accounted for the greatest number of true overlapping extension BACs. The mapping result was compared against the contig order within 2.0 Mb of assembled BAC sequence and BAC extension overlaps found with FPC were compared with the overlap positions and linear order of the BES of candidate extensions assembled onto the seed BAC insert sequence, and subsequently the number of displaced BACs was determined.

Sequencing and assembly

Sequencing and assembly were performed as described by Peters *et al.* (2006). The BAC sequences are available at ftp://ftp.sgn.cornell.edu/tomato_genome/.

Sequence annotation

The sequence data were annotated using the Cyrille2 system (Fiers *et al.*, 2008). Interspersed repeats were identified using Repeat-Masker 3.2.5 (<http://www.repeatmasker.org/>) and cross_match 0.990319 (<http://www.phrap.org/>) with the *Magnoliophyta* section of RepBase (release 2008-08-01) (Jurka *et al.*, 2005), the TIGR Lycopersicon repeats v3.1 and Solanaceae repeats v3.1 (<http://www.tigr.org/>) and a tomato-specific retrotransposon library (ED, unpublished results).

Genes were predicted using the linear combiner option of JIGSAW 3.2.8.1 (Allen and Salzberg, 2005) integrating data from *ab initio* gene predictors Augustus 2.0.2 (Stanke and Waack, 2003), Genscan (Burge and Karlin, 1997), Geneid 1.3.7 (Parra *et al.*, 2000) and GlimmerHMM 3.0.1 (Majoros *et al.*, 2004), each using *Arabidopsis thaliana* gene models; alignments of *S. lycopersicum*, *S. tuberosum* and *Nicotiana tabacum* ESTs (obtained from <http://www.kazusa.or.jp/jsol/microtom/index.html>, <http://biosrv.cab.unina.it/potatestdb/> and <http://sgn.cornell.edu/>, respectively) generated using BLASTN 2.2.17 (Altschul *et al.*, 1997) and sim4 (Florea *et al.*, 1998); and alignments of proteins from the plant division of UniProt (Boutet *et al.*, 2007) and *A. thaliana* TAIR8 (<http://www.arabidopsis.org/>) generated using BLASTX 2.2.17 and GeneWise 2.2.0 (Birney *et al.*, 2004). The predicted genes were annotated through BLASTX alignments to the NCBI nr (2008-03-30 release) and RefSeq (2008-05-28 release) databases and InterPro-

Scan 4.3.1 (Finn *et al.*, 2006) domain searches using version 15.0 of the databases. Marker sequences downloaded from SGN were aligned to the genomic sequences using BLASTN and sim4. Sequence and annotation data are available in Files S1 and S2 in the Supporting Information.

To identify clusters of functionally related genes, a list of non-redundant GO terms per gene was produced. Subsequently, overlapping windows of 10 genes were created and tested for overrepresented GO terms using Fisher's exact test. The number of occurrences of each GO term per window was compared with the frequency of that GO term in the full set of 970 genes. A GO term was found to be significantly overrepresented in a window at a *P* value smaller than 0.01 applying the Holm–Bonferroni correction.

ACKNOWLEDGEMENTS

This project was (co)financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research and by the European Commission (EU-SOL project PL 016214).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Cytogenetic mapping of seed and extension bacterial artificial chromosomes (BACs).

Figure S2. Fingerprinted contig (FPC) map of contig 1 bacterial artificial chromosomes (BACs) mapping to the short arm of tomato chromosome 6. The minimal tiling path (MTP) selected for sequencing is highlighted in red.

Figure S3. Physical map of contig 2 bacterial artificial chromosomes (BACs) mapping to the short arm of tomato chromosome 6.

Figure S4. Multicolor fluorescent *in situ* hybridization (FISH) of euchromatin and heterochromatin bordering bacterial artificial chromosomes (BACs) on chromosome 6.

Figure S5. Pericentromeric and subtelomeric bacterial artificial chromosomes (BACs) on tomato chromosome 6.

Figure S6. Number of exons per gene as a percentage of the total gene content per domain.

Table S1. Physical, cytogenetic and genetic distances for tomato chromosome 6 bacterial artificial chromosomes (BACs) and contigs.

Table S2. Classification and distribution of known plant repeats in the chromosome 6 sequence data.

Table S3. Number of (near-) complete BLASTN matches to several known tomato repeats in the chromosome 6 sequence data.

File S1. [tomato_chromosome_6.fasta.gz](#).

File S2. [tomato_chromosome_6.gff.gz](#).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

REFERENCES

- Allen, J.E. and Salzberg, S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

- Arumuganathan, K. and Earle, E. (1991) Estimation of nuclear DNA content of plants by flow cytometry. *Plant Mol. Biol. Rep.* **9**, 229–241.
- Bai, Y., van der Hulster, R., Huang, C.C., Wei, L., Stam, P. and Lindhout, P. (2004) Mapping *Ol-4*, a gene conferring resistance to *Oidium neolycopersici* and originating from *Lycopersicon peruvianum* LA2172, requires multi-allelic, single-locus markers. *Theor. Appl. Genet.* **109**, 1215–1223.
- Batzoglou, S., Berger, B., Mesirov, J. and Lander, E.S. (1999) Sequencing a genome by walking with clone-end sequences: a mathematical analysis. *Genome Res.* **9**, 1163–1174.
- Bennetzen, J.L. (1996) The Mutator transposable element system of maize. *Curr. Topics Microbiol. Immunol.* **204**, 195–229.
- Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. *Genome Res.* **14**, 988–995.
- Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007) UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Methods Mol. Biol.* **406**, 89–112.
- Budiman, M.A., Mao, L., Wood, T.C. and Wing, R.A. (2000) A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res.* **10**, 129–136.
- Budiman, M.A., Chang, S.-B., Lee, S., Yang, T.J., Zhang, H.-B., de Jong, H. and Wing, R.A. (2004) Localization of *jointless-2* gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping. *Theor. Appl. Genet.* **108**, 190–196.
- Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94.
- Chang, S.-B., Yang, T.-J., Datema, E. et al. (2008) FISH mapping and molecular organization of the major repetitive sequences of tomato. *Chromosome Res.* **16**, 919–933.
- Cone, K.C., McMullen, M.D., Vroh Bi, I. et al. (2002) Genetic, physical, and informatics resource for Maize. On the road to an integrated map. *Plant Phys.* **130**, 1598–1605.
- Datema, E., Mueller, L.A., Buels, R., Giovannoni, J.J., Visser, R.G.F., Stiekema, W.J. and van Ham, R.C.H.J. (2008) Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato. *BMC Plant Biol.* **8**, 34.
- De Jong, J.H., Zhong, X.-B., Franz, P.F., Wennekes-van Eden, J., Jacobsen, E. and Zabel, P. (2000) High resolution FISH reveals the molecular and chromosomal organisation of repetitive sequences of individual tomato chromosomes. In *Chromosomes Today 13* (Olmo, E. and Rdi, C.A., eds). Switzerland: Birkhäuser Verlag, pp. 267–275.
- Field, B. and Osbourn, A.E. (2008) Metabolic diversification – independent assembly of operon-like gene clusters in different plants. *Science*, **230**, 543–547.
- Fiers, M.W.E.J., van der Burgt, A., Datema, E., de Groot, J.C.W. and van Ham, R.C.H.J. (2008) High-throughput bioinformatics with the Cyrille2 pipeline system. *BMC Bioinformatics*, **9**, 96.
- Finn, R.D., Mistry, J., Schuster-Bockler, B. et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251.
- Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967–974.
- Fregonezi, J.F., Vilas-Boas, L.A., Peleginelli Fungaro, M.H., Gaeta, M.L. and Vanzela, A.L.L. (2007) Distribution of a Ty3/gypsy-like retroelement on the A and B-chromosomes of *Cestrum strigilatum* Ruiz & Pav. and *Cestrum intermedium* Sendtn. (Solanaceae). *Gen. Mol. Biol.* **30**, 599–604.
- Friesen, N., Brandes, A. and Seymour, J. (2001) Diversity, origin, and distribution of retrotransposons (gypsy and copia) in conifers. *Mol. Biol. Evol.* **18**, 1176–1188.
- Fulton, T.M.R., van der Hoeven, R., Eannetta, N.T. and Tansley, S.D. (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell*, **14**, 1457–1467.
- Ganal, M.W., Lapitan, N.L.V. and Tansley, S.D. (1991) Macrostructure of the tomato telomeres. *Plant Cell*, **3**, 87–94.
- Gierl, A. (1996) The En/Spm transposable element of maize. *Curr. Topics Microbiol. Immunol.* **204**, 145–159.
- Gierl, A. and Frey, M. (2001) Evolution of benzoxazinone biosynthesis and indole production in maize. *Planta*, **213**, 493–498.
- Guyot, R., Cheng, X., Su, Y., Cheng, Z., Schlagenhauf, E., Keller, B. and Ling, H.-Q. (2005) Complex organization and evolution of the tomato pericentromeric region at the *FER* gene locus. *Plant Phys.* **138**, 1205–1215.
- Hurst, L.D., Pál, C. and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **4**, 299–310.
- Ji, Y., Schuster, D.J. and Scott, J.W. (2007) *Ty-3*, a begomovirus resistance locus near the *Tomato yellow leaf curl virus* resistance locus *Ty-1* on chromosome 6 of tomato. *Mol. Breeding*, **20**, 271–284.
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.
- Kazazian, H.H. Jr (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.
- Koo, D.-H., Jo, S.-H., Park, H.-M., Lee, S. and Choi, D. (2008) Integration of cytogenetic and genetic linkage maps unveils the physical architecture of tomato chromosome 2. *Genetics*, **179**, 1211–1220.
- Kush, G.S., Rick, C.M. and Robinson, R.W. (1964) Genetic activity in the heterochromatic segment of the tomato. *Science*, **25**, 1432–1434.
- Liu, R., Vitte, C., Ma, J., Assibi Mahama, A., Dhliwayo, T., Lee, M. and Bennetzen, J.L. (2007) A GeneTrek analysis of the maize genome. *Proc. Natl Acad. Sci. USA*, **104**, 11844–11849.
- Luo, M.-C., Thomas, C., You, F.M., Hsiao, J., Ouyang, S., Buell, C.R., Malandro, M., McGuire, P.E., Anderson, O.D. and Dvorak, J. (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics*, **82**, 378–389.
- Macas, J. and Neumann, P. (2007) Ogre elements - a distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*, **390**, 108–116.
- Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source *ab initio* gene finders. *Bioinformatics*, **20**, 2878–2879.
- Manetti, M.E., Rossi, M., Costa, A.P., Clausen, A.M. and van Sluys, M.A. (2007) Radiation of the Tnt1 retrotransposon superfamily in three Solanaceae genera. *BMC Evol. Biol.* **7**, 34.
- Meyers, B.C., Scalabrin, S. and Morgante, M. (2004) Mapping and Sequencing complex genomes: let's get physical! *Nature*, **5**, 578–589.
- Natali, L., Santini, S., Giordani, T., Minelli, S., Maestrini, P., Cionini, P.G. and Cavallin, A. (2006) Distribution of Ty3-gypsy and Ty1-copia-like DNA sequences in the genus *Helianthus* and other Asteraceae. *Genome*, **49**, 64–72.
- Parra, G., Blanco, E. and Guigó, R. (2000) GenelD in Drosophila. *Genome Res.* **10**, 511–515.
- Pereira, V. (2004) Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. *Gen. Biol.* **5**, R79.
- Peters, S.A., van Haarst, J.C., Jesse, T.P., Woltinge, D., Jansen, K., Hesselink, T., van Staveren, M.J., Abma-Henkens, M.H.C. and

- Klein-Lankhorst, R.M. (2006) TOPAAS, a Tomato and Potato Assembly Assistance System for selection and finishing of bacterial and artificial chromosomes. *Plant Phys.* **140**, 805–817.
- Peterson, D.G., Price, H.J., Johnston, J.S. and Stack, S.M. (1996) DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes. *Genome*, **39**, 77–82.
- Peterson, D.G., Pearson, W.R. and Stack, S.M. (1998) Characterization of the tomato (*Lycopersicon esculentum*) genome using in vitro and in situ DNA reassociation. *Genome*, **41**, 346–356.
- Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R. and Osbourn, A. (2004) A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl Acad. Sci. USA*, **101**, 8233–8238.
- Rubini, E., Lithwick, G. and Levy, A.A. (2001) Structure and evolution of the hAT transposon superfamily. *Genetics*, **158**, 949–957.
- Seah, S.J., Rossi, Y.M., Gleason, C.A. and Williamson, V.M. (2004) The nematode-resistance gene, Mi-1, is associated with an inverted chromosomal segment in susceptible compared to resistant tomato. *Theor. Appl. Genet.* **108**, 1635–1642.
- Soderlund, C., Humphray, S., Dunham, A. and French, L. (2000) Contigs built with fingerprints, markers and FPCV4.7. *Genome Res.* **10**, 1772–1787.
- Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(suppl 2), 215–225.
- Szinay, D., Chang, S.-B., Khrustaleva, L., Peters, S., Schijlen, E., Bai, Y., Stiekema, W.J., van Ham, R.C.H.J., de Jong, H. and Klein Lankhorst, R.M. (2008) High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. *Plant J.* **56**, 627–637.
- Tam, S.M., Causse, M., Garchery, C., Burck, H., Mhiri, C. and Grandbastien, M.-A. (2007) The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species. *J. Evol. Biol.* **20**, 1056–1072.
- Tang, X., Szinay, D., Lang, C. *et al.* (2008) Cross-species BAC FISH painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. *Genetics*, **180**, 1319–1328.
- Tanksley, S.D., Ganai, M.W., Prince, J.P. *et al.* (1992) High density molecular linkage maps of the tomato and potato genomes. *Genetics*, **132**, 1141–1160.
- Van der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G. and 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.
- Venter, J.C., Smith, H.O. and Hood, I. (1996) A new strategy for genome walking. *Nature*, **381**, 364–366.
- Wang, Y., Tang, X., Cheng, Z., Mueller, L., Giovannoni, J. and Tanksley, S.D. (2006) Euchromatin and pericentromeric heterochromatin: comparative composition in the Tomato Genome. *Genetics*, **172**, 2529–2540.
- Weiler, K.S. and Wakimoto, B.T. (1995) Heterochromatin and gene expression in *Drosophila*. *Annu. Rev. Genetics*, **29**, 577–605.
- Yang, T.J., Lee, S., Chang, S.B., Yu, Y., de Jong, H. and Wing, R.A. (2005) In-depth sequence analysis of the tomato chromosome 12 centromeric region: identification of a large CAA block and characterization of pericentromere retrotransposons. *Chromosoma*, **14**, 103–117.
- Yasuhara, J.C. and Wakimoto, B.T. (2006) Oxymoron no more: the expanding world of heterochromatic genes. *Trends Genet.* **22**, 330–338.
- Yi, G., Sze, S.H. and Thon, M.R. (2007) Identifying clusters of functionally related genes in genomes. *Bioinformatics*, **23**, 1053–1060.
- You, F.M., Luo, M.-C., Gu, Y.Q., Lazo, G.R., Deal, K., Dvorak, J. and Anderson, O.D. (2007) Genoprofiler: batch processing of high-throughput capillary fingerprinting data. *Bioinformatics*, **23**, 240–242.
- Zhong, X., Fransz, P.F., Wennekes-van Eden, J., Zabel, P., van Kammen, A. and de Jong, H. (1996a) High-Resolution Mapping on Pachytene chromosomes and extended DNA Fibres by Fluorescence in-situ hybridization. *Plant Mol. Biol. Rep.* **14**, 232–242.
- Zhong, X.B., de Jong, J.H. and Zabel, P. (1996b) Preparation of tomato meiotic pachytene and mitotic metaphase chromosomes suitable for fluorescence in situ hybridization (FISH). *Chromosome Res.* **4**, 24–28.
- Zhong, X.B., Fransz, P.F., Wennekes-van Eden, J., Ramanna, M.S., Van Kammen, A., Zabel, P. and de Jong, J.H. (1998) FISH studies reveal the molecular and chromosomal organization of individual telomere domains in tomato. *Plant J.* **13**, 507–517.