Gene 499 (2012) 176–181

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Short communication

Euchromatic and heterochromatic compositional properties emerging from the analysis of *Solanum lycopersicum* BAC sequences

Miriam Di Filippo, Alessandra Traini, Nunzio D'Agostino [1], Luigi Frusciante, Maria Luisa Chiusano *

University of Naples Federico II, Dept. of Soil, Plant, Environmental and Animal Production Sciences, Via Università 100, 80055 Portici, Italy

ARTICLE INFO

Article history:
Accepted 20 February 2012
Available online 26 February 2012

Keywords:
Tomato genome
Heterochromatin
Euchromatin
Repetitive elements
Gene content
Repeat content

ABSTRACT

The consortium responsible for the sequencing of the tomato (*Solanum lycopersicum*) genome initially focused on the sequencing of the euchromatic regions using a BAC-by-BAC strategy. We analyzed the compositional features of the whole collection of BAC sequences publically available. This analysis highlights specific peculiarities of heterochromatic and euchromatic BACs, in particular: the whole BAC collection has i) a large variability in repeat and gene content, ii) a positive and significant correlation of LTR retrotransposons of the Gypsy class with the repeat content and iii) the preferential location of the SINEs (short interspersed nuclear elements) in BAC sequences showing a low repeat content. Our results point out a typical design of the tomato chromosomes and pave the way for further investigations on the relationship between DNA primary structure and chromatin organization in Solanaceae genomes.

1. Introduction

The *Solanum lycopersicum* genome (950 Mb; Arumuganathan and Earle, 1991) is organized into 12 chromosomes with extended heterochromatic regions (Peterson et al., 1996). The majority of genes (~90%) are estimated to be placed in the euchromatic domains, which account for one quarter (~250 Mb) of the whole genome sequence (van der Hoeven et al., 2002). Cytogenetic analysis (De Jong, 1998; De Jong et al., 2000) showed that the euchromatin is generally located in the distal part of the chromosomes and is included between the heterocromatic regions of the telomere and the pericentromere. Since the heterochromatic DNA has been reported to be repeat-rich (Peterson et al., 1996) (hence difficult to sequence and assemble) and gene-poor, the tomato genome sequencing first focused on the euchromatic portion of the genome (Mueller et al., 2009). The effort proceeded using a BAC-by-BAC (bacterial artificial chromosome) approach, which has been successfully applied to the sequencing of other plant genomes such as *Arabidopsis thaliana* (AGI, 2000) and *Oryza sativa* (International Rice Genome Sequencing Project, 2005). This strategy basically consists in anchoring several BACs to the high density F_2-2000 genetic map (Fulton et al., 2002), which contains a set of restriction fragment length polymorphism (RFLP) markers from the Tomato-EXPEN 1992 map (Tanksley et al., 1992). As a BAC is anchored and sequenced, it represents a reference for extending the sequence by BAC walking so as to generate a tiling path along the chromosomes (Mueller et al., 2009). Tiling Path Format (TPF) files indicate the order of the sequenced BACs along each chromosome. TPF files for tomato are publicly accessible at the SOL Genomics Network (SGN) web site (Bombarely et al., 2001, http://solgenomics.net/genomes/Solanum_lycopersicum/index.pl; see Methods).

Although the BAC-by-BAC approach takes into consideration only one quarter of the total genome, it supports the design of a chromosome-scale BAC scaffold and provides a reference for preliminary investigations on specific features such as their assignment to eu- or heterochromatin (Asamizu, 2007; Mueller et al., 2009; Peters et al., 2009; Szinay et al., 2008; Tang et al., 2008; Wang et al., 2006; Yang et al., 2005). Obviously, they also provide a framework for the reliable completion of the whole genome which is now full-finished by a whole genome shotgun effort (http://solgenomics.net/organism/Solanum_lycopersicum/genome).

Molecular markers that define the eu/heterochromatin boundaries for each tomato chromosome and the identifiers of the BACs associated to these markers have been made available by the SGN website (see Methods). In addition, fluorescence in situ hybridization (FISH) supports the assignment of BACs to the corresponding chromosomes and specifically to euchromatic or to heterochromatic regions (De Jong, 1998; De Jong et al., 2000; Peters et al., 2009; Szinay et al., 2008; Tang et al., 2008; Wang et al., 2006). Extensive information on gene and repeat content emerged from the sequenced BACs. Gene-rich BACs

Abbreviations: BAC, Bacterial artificial chromosome; TPF, tiling path format; FISH, Fluorescence in situ hybridization; ESTs, Expressed sequence tags; LTR, Long terminal repeats; SINEs, short interspersed nuclear elements; SGN, SOL Genomics Network; RFLP, restriction fragment length polymorphism; TGR, Tomato genomic repeats; HTGS, High Throughput Genomic Sequences; RT, Reverse transcriptase; TE, Transposable elements; RR, repeat-richer; GR, gene-richer.

* Corresponding author.
E-mail addresses: miriam.difilippo@gmail.com (M. Di Filippo), traini.alessandra@gmail.com (A. Traini), nunzio.dagostino@gmail.com (N. D'Agostino), fruscian@unina.it (L. Frusciante), chiusano@unina.it (M.L. Chiusano).
[1] Present address: Research Centre for Vegetable Crops, via Cavalleggeri, 25 84098 Pontecagnano (SA).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.02.044

showed lower repeat content, supporting the initial assumption that genes were predominantly located in repeat-poor euchromatic regions, whereas heterochromatic BACs proved to be repeat-rich, though not completely depleted of genes (Mueller et al., 2009; Yasuhara and Wakimoto, 2006). Different repeat classes were also identified. As an example, the TGR IV was classified as a typical centromeric satellite, while the LTR (long terminal repeat) retroelements Ty3–Gypsy and Ty1–Copia proved to be abundant in the pericentromeric heterochromatin (Asamizu, 2007; Chang et al., 2008; Peters et al., 2009; Szinay et al., 2008; Tang et al., 2008; Wang et al., 2006; Yasuhara and Wakimoto, 2006).

In addition, genetic and physical information on chromosome 6 has also become available thanks to previous studies on genetic maps and on the organization of the pericentromere (Liharska et al., 1997; Ouyang and Buell, 2004; Van Wordragen et al., 1994, 1996; Weide et al., 1998; http://plantrepeats.plantbiology.msu.edu/). Finally, the most recent investigations on the structure of the chromosome 6 (Peters et al., 2009; Szinay et al., 2008; Tang et al., 2008), based on cytogenetic analysis, attempted to define eu/heterochromatin boundaries.

We investigated the gene and the repeat content of the sequenced BACs according to their association to eu- and heterochromatic region in tomato. Our results highlighted specific compositional properties as well as a typical design of the tomato chromosome organization.

2. Methods

2.1. Data availability

1095 BAC sequences (a total of 119.15 Mb) from *S. lycopersicum* were retrieved from GenBank using the keyword "TOMGEN" (December 2010).

Tiling Path Format (TPF) files for chromosomes 4, 5, 6, 9 and 12 (chr04.v19.tpf, chr05.v8.tpf, chr06.v3.tpf, chr09.v3.tpf and chr12.v5.tpf) were downloaded from the SGN FTP site, ftp://ftp.sgn.cornell.edu/tomato_genome/tpf/.

The list of the markers and the associated BACs at the boundaries of the pericentromeric heterochromatin for each tomato chromosome are available at ftp://ftp.sgn.cornell.edu/tomato_genome/seedbacs/. Eu- and heterochromatic BACs associated to the chromosome 6, as assigned through BAC-FISH by Peters et al. (2009) and Tang et al. (2008), were also considered.

2.2. EST/TC sequences to genome mapping

Both EST and TC sequences, the latter automatically generated by the ParPEST pipeline (D'Agostino et al., 2005, 2009; Gremme et al., 2005) and collected in a dedicated database called SolEST (D'Agostino et al., 2009), were aligned along BAC sequences using the GenomeThreader software (Gremme et al., 2005). GenomeThreader is used to generate splice-alignments of each EST along genomic sequences. Alignments with a minimum identity score of 90% and a minimum sequence coverage of 80% was filtered out. Data are accessible in ISOL@ at http://biosrv.cab.unina.it/isola/ (Chiusano et al., 2008).

2.3. Repeat and gene content definition

Both gene and repeat content for each BAC were calculated as percentage of nucleotides covered by *S. lycopersicum* ESTs and by repeats from the Plant Repeat Database, respectively. The same approach was used to calculate the content of each single class of repeats per BAC. For the gene content the extension of each EST to genome alignment, i.e. exon + intron regions, was considered.

Identification of interspersed repeats was performed by the RepeatMasker program (RepeatMasker at http://repeatmasker.org) using the Plant Repeat Database at Michigan State University (http://plantrepeats.plantbiology.msu.edu/) (Ouyang and Buell, 2004), RepBase.13.06 (http://www.girinst.org/server/archive/) (Jurka et al., 2005), and the SGN tomato UniRepeats (ftp://ftp.sgn.cornell.edu/tomato_genome/repeats/) as filtering databases.

The Pearson correlation coefficient between the content of a single class of repetitive elements and the overall repeat content for each BAC as well as the corresponding p-values (Student's *T* test) were calculated by the MatLab software (The Mathworks, Natick, MA http://www.mathworks.it/products/matlab/).

2.4. Tomato gene models

The gene models (GME) was obtained by the software GeneModelEST (D'Agostino et al., 2007), which aimed to define a data-set of candidate gene models using solely indication derived from cDNA/EST sequences. The genome coordinates of the splice-alignments of ESTs and TCs from different tomato and potato species: *S. lycopersicum*, *S. pennellii*, *S. habrochaites*, *S. lycopersicum X S. pimpinellifolium*, *S. tuberosum*, *S. chacoense* were fed into GeneModelEST. The in silico derived coordinates of candidate gene models are available through the graphical annotation viewer Gbrowse (http://biosrv.cab.unina.it/GBrowse/).

3. Results

An overview of all the 1095 BACs analyzed herein is given in Table 1, including a synopsis of specific features for each chromosome. We report the HTGS (high throughput genomic sequences) phases of the BACs, the amount of sequenced nucleotides, the percentage of nucleotides matching ESTs (expressed sequence tags) from *S. lycopersicum* or plant repeat sequences, the number of gene models (GME; see Table S1). Similar information is accessible at http://biosrv.cab.unina.it/isola/ (Chiusano et al., 2008) a website for Solanaceae genomics where the BAC sequence annotation can be explored.

We considered all the BACs assigned to the preliminary backbone (Tiling Path Format) of the chromosome 6 and we calculated the gene content per BAC by counting the total number of nucleotides covered by *S. lycopersicum* ESTs, introns included, following the same approach we used in Mueller et al. (2009), while the repeat content was assessed considering all the nucleotides covered by known plant repeats. We also calculated the Δ(RG), i.e. the difference

Table 1
For each chromosome the BAC sequencing status, indicated as High Throughput Genome Sequence (HTGS) phase, the number of all analyzed BACs (TOT), total sequenced nucleotides, percentage of nucleotides covered by ESTs from *S. lycopersicum* and by Plant Repeat Database and the number of gene models (GME).

chr	HTGS1[a]	HTGS2[a]	HTGS3[a]	TOT	BAC length[b] (Mb)	*S. lycopersicum* EST regions (%)	Plant repeat regions (%)	GME[c]
0[d]	2	2	0	4	0.54	9.58	49.29	4
1	5	0	14	19	2.52	24.59	30.65	13
2	0	0	180	180	20.13	30.41	21.07	99
3	0	0	16	16	1.85	25.43	29.72	11
4	17	5	189	211	22.72	17.72	41.99	51
5	0	35	29	64	6.38	31.89	16.10	30
6	116	0	41	157	17.28	26.92	25.94	71
7	33	67	24	124	12.03	27.57	25.78	48
8	0	2	162	164	18.41	21.08	37.59	47
9	10	30	30	70	7.24	26.92	21.26	37
10	0	0	4	4	0.49	19.87	49.23	–
11	0	1	23	24	2.95	31.93	22.84	16
12	21	11	26	58	6.61	30.24	22.21	30

[a] More information about HTGS phases are available at http://www.ncbi.nlm.nih.gov/HTGS/.
[b] Ranging from 1837 to 227,938 Kb with an average length of 108,808 Kb.
[c] Gene models obtained by D'Agostino et al. (2007).
[d] BACs with ambiguous positioning on tomato chromosomes were assigned to the arbitrary-defined chromosome 0.