Repeat Sequences in the Tomato Genome

10

Maria Luisa Chiusano and Chiara Colantuono

Abstract

The sequencing of the tomato genome revealed that, though the moderated size when compared to most of the Solanaceae and other plant species, it comprises more than the 60 % of DNA repeats. This is in contrast with initial estimations assessing that the total genome comprised only about the 10–22 % of repetitive sequences. These preliminary hypotheses were probably biased by the presence of single-copy DNA within the repetitive portion of the genome and by the high sequence divergence of the repeat content. Though the release of the first version of the genome sequences in 2012, the complete view of the repeated regions in tomato at sequence level is still partial, because of difficulties due mainly to DNA repeat sequencing and assembling. However, deeper knowledge on the repeat content of the genome and its distribution was consistently supported by cytogenetics, molecular markers and reassociation kinetics, accompanied by advanced approaches such as *Fluorescence* In Situ *Hybridization* (FISH) and more recently *Optical Mapping*. These techniques helped to clarify many of the principal aspects related to the distribution and the organization of the major repeat classes in tomato, contributing to a consistent overview of this essential part of the genome. The main focus of this chapter is to describe the repeat content of the tomato genome as revealed from the sequencing effort and associated bioinformatics, mainly considering the distribution of highly and moderately repeated DNA sequences. We provide a general overview on plant genome complexity and repeat content, presenting the main repeat categories and their organization. Then we describe the bioinformatics for DNA repeats sequence analysis, focusing on most common approaches for investigations in large genomic sequences, as well as on major repeated sequence collections available to support plant genome annotations. Details on the

M.L. Chiusano (✉) · C. Colantuono
Department of Agraria, University Federico II of Naples, Naples, Italy
e-mail: chiusano@unina.it

M. Causse et al. (eds.), *The Tomato Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-662-53389-5_10

methods employed to analyze the tomato genome sequences (assembly v. 2.40) published in 2012 will be presented. The description of what is known from tomato concerning the major DNA repeat classes is therefore overviewed highlighting the major results or confirmations obtained thanks to the genome sequencing effort. The discussion is mainly focused on the general description of repeat occurrence in the tomato genome, though questions on the specific role and evolution of these extended regions in tomato and in plant genomes, as well as in other eukaryotes, still remain open.

Keywords
Tomato · Repeat · Bioinformatics · Duplication · Cytogenetics

Introduction

The exploitation of evolving experimental techniques, starting from early cytological approaches, molecular markers, *Fluorescence* In Situ *Hybridization (FISH)* and *Optical Mapping,* till the nucleotide sequencing of entire genomes, contributed relevant discoveries on genome organization, also determining relationships among chromosomal peculiarities, in phylogeny, in evolution.

Comparative approaches highlighted that many structure features of plant genomes are remarkably similar among different species, and are also shared with other eukaryotes, animals and fungi (Heslop-Harrison 2000). All eukaryotes have their genomic DNA organized in chromosomes, associated with proteins, showing almost the same organization. Centromeric regions are located in regions that are almost conserved along the chromosome structure, and the terminal regions are organized in telomeres.

Comparative approaches also highlighted the relevance of polyploidy in plants, with chromosome number which varies widely among plant species, such that 2*n* ranges in value from 4 to more than 1000, although the number within any given species is usually constant. Occurrence of polyploidy may be also associated to diploidization events, with rearrangements also implying genome reshuffling, translocations, fusion and fission of chromosomes. These events have been discussed to be some of the consequences why plant genomes are highly duplicated (Lysak et al. 2005; Cui et al. 2006; Tang et al. 2008a, b; Jiao et al. 2011, 2012; Sangiovanni et al. 2013). Beyond the interesting issue of investigating on the mechanisms implied in the occurrence of polyploidy and diploidization events in plants, even in a relatively short time span, tracing plant genome evolution and diversification (Jaillon et al. 2007; Tomato Genome Consortium 2012; Denoeud et al. 2014), it would also be rather intriguing to understand what enabled angiosperms to efficiently manage the presence of homologous chromosomes in comparison to all other eukaryotes, where polyploids are rare. However, in the context of this chapter, it is remarkable to focus on the effects that whole-genome and segmental duplications had on the redundancy of genome regions and of gene copies, with the definition of novel gene families. Though it is not the aim of this chapter to discuss repeats in DNA due to polyploidization events or to retaining of duplicated regions, it is noteworthy, indeed, to underline also here that one of the main outcomes of the tomato genome sequencing effort was the tracing of two consecutive genome triplications in the *Solanum* lineage. The more ancient event was shared with rosids, while, a more recent one appeared specific to the *Solanum* lineage (Tomato Genome Consortium 2012; Denoeud et al. 2014). These events had a relevant impact on diversification

and evolution of novel functionalities in these clade of plants. However, it is discussed that the repeated regions tracing these possible events in the tomato genome were mainly detected only at sequence level (Tomato Genome Consortium 2012), presumably because of the high divergence determined by gene loss or mutations since the last hypothesized polyploidization event (Shearer et al. 2014).

The dynamics of genome evolution in plants offers striking opportunities to have multiple copies of the genome content, i.e. to repeat it, and to keep it duplicated even when diploidization occurred. Furthermore, the transfer of genes or of entire parts of the DNA from organelles to nucleus is now well documented both in plants and animals (Martin and Herrmann 1998; Vaughan et al. 1999).

Worthy to note, though the different occurrences of genome rearrangements in plants, the gene numbers as well as their order are almost conserved over substantial evolutionary distances in plants (Gebhardt et al. 1991; Ahn et al. 1993; Devos and Gale 1993, 1997, 2000).

The tomato genome, as an example, is highly syntenic with those of other economically important Solanaceae (Potato Genome Sequencing Consortium 2011; Tomato Genome Consortium 2012; Hirakawa et al. 2014; Kim et al. 2014; Sierro et al. 2014) as well as other plants (Jaillon et al. 2007). However, plant genome size can strongly vary among different species. Indeed, repetitive sequences contribute significantly to genome size in plants. Understanding the mechanisms and inferring on possible functional reasons favouring these variability and plasticity is still an open challenge.

DNA Content in the Cell

The amount of DNA (in picograms) in an unreplicated haploid cell, which corresponds to the constant value or C-value (Swift 1950; Greilhuber et al. 2005), is relatively homogeneous within a species. However, it is evident that the C-value is particularly variable between species. This variability is not related to the complexity of the organisms in terms of size or developmental mechanisms. The DNA content of the unicellular amoeba was 200 times higher than in human cells, though mammals have evident higher developmental complexity. This initially "unexpected" phenomenon represents the so-called "C-value paradox". The paradox is today explained knowing that the DNA content in a species can be abundant in repetitive sequences, though the numbers of coding genes are of the same order of magnitude in all eukaryotes, which ranges from about 6000 in the unicellular *Saccharomyces cerevisiae* to approximately 20,000 to 25,000 in the human genome (which is 200 times bigger than the genome of the yeast) (Richard et al. 2008).

In general, the term "repetitive sequences" refers to highly similar DNA fragments that are present in multiple copies in a genome. In particular the major contribution to the haploid genome size in eukaryotes is due to highly and moderately repeated sequences, i.e. DNA motifs, ranging in length from a single couple of nucleotides to thousands of nucleotides, repeated many hundreds or thousands of times. These repeated motifs are ubiquitous in eukaryotic genomes (Charlesworth et al. 1994; Kumar and Bennetzen 1999; Bowen and Jordan 2002) and represent a large portion of the chromosome structure (von Sternberg 2002), ranging between 50 and 90 % or more of all the nuclear DNA content. As an example, more than the 50 % of the human genome is composed by repeats (Richard et al. 2008).

In higher plants, the amount of DNA is particularly variable between species (Flavell et al. 1974; Bennett and Smith 1976; Ouyang and Buell 2004; Hawkins et al. 2009). The lowest content reported for *A. thaliana* is one of the main reasons why this genome was the first one to be sequenced among plant species (NSF 1990; Arabidopsis Genome Initiative 2000). Accordingly, mainly thanks to its "modest" genome size, poplar was the first tree to be sequenced (Brunner et al. 2004). Also in the case of plant genomes, the proportion of protein-coding regions is rather similar among

the species (Table 10.1). Indeed, the structural and developmental complexity of plant species with very different amounts of DNA per cell is not fundamentally different from those with the highest amounts (Smyth 1991). It is also evident (Table 10.1) that the contribution of repeats to each genome has a wide range of variability starting from very low percentages, like in *Arabidopsis thaliana*, reaching a very high relative content like in *Capsicum annum* (~82 %) and in several monocots (~85 %).

DNA Repeat Classes

Repetitive DNA was first detected because of its rapid reassociation kinetics when denatured, since the rate at which a particular sequence reassociates is proportional to the number of times it is found in the genome. Based on the renaturation rates, in denaturation–renaturation experiments of genomic DNA after heat exposure, it is possible to identify three major classes of DNA sequence types: the highly repetitive sequences,

Table 10.1 List of plants with sequenced genomes

Scientific name	Monocot/dicot	#Chr (*n*)	Size (Mb)	#Gene	%Repeat	References
Arabidopsis lyrata	Dicot	8	207	32.670	30	Hu et al. (2011)
Arabidopsis thaliana	Dicot	5	125	25.498	14	The Arabidopsis Genome Initiative (2000)
Brassica rapa	Dicot	10	485	41.174	40	The Brassica rapa Genome Sequencing Project Consortium (2011)
Capsicum annum cultivate/wild	Dicot	12	3349/3480	35.336/34.476	81/82	Qin et al. (2014)
Carica papaya	Dicot	9	372	28.629	43	Ming et al. (2008)
Coffee canephora	Dicot	11	710	25.574	50	Denoeud et al. (2014)
Cucumis sativus	Dicot	7	367	26.682	24	Huang et al. (2009)
Fragaria vesca	Dicot	7	240	34.809	23	Shulaev et al. (2011)
Glycine max	Dicot	20	1115	46.430	57	Schmutz et al. (2010)
Hordeum vulgare	Monocot	7	5100	30.400	84	The International Barley Genome Sequencing Consortium (2012)
Lotus japonicus	Dicot	6	472	30.799	56	Sato et al. (2008)
Musa acuminata	Monocot	11	523	36.542	44	D'Hont et al. (2012)

(continued)

Table 10.1 (continued)

Scientific name	Monocot/dicot	#Chr (*n*)	Size (Mb)	#Gene	%Repeat	References
Nelumbo nucifera	Dicot	8	929	26.685	57	Ming et al. (2013)
Nicotiana tabacum K326/TN90/BX	Dicot	24 (2*n*)	4600/4410/4570	91.870/81.404/93.303	73/79/73	Sierro et al. (2014)
Oryza brachyantha	Monocot	12	300	32.038	29	Chen et al. (2013)
Oryza sativa	Monocot	12	389	37.544	26	International Rice Genome Sequencing Project (2005)
Phoenix dactylifera	Monocot	18	658	28.890	40	Al-Mssallem et al. (2013)
Solanum lycopersicum	Dicot	12	900	34.727	63	The Tomato Genome Consortium (2012)
Solanum melongena	Dicot	12	1100	85.446	71	Hirakawa et al. (2014)
Solanum tuberosum	Dicot	12	844	39.031	62	The Potato Genome Sequencing Consortium (2011)
Sorghum bicolor	Monocot	10	818	34.496	62	Paterson et al. (2009)
Theobroma cacao	Dicot	10	430	28.798	24	Argout et al. (2011)
Triticum aestivum	Monocot	42 (6*n*)	17,000	124.201	80	IWGSC (2014)
Triticum urartu	Monocot	7	4940	34.879	67	Ling et al. (2013)
Vitis vinifera	Dicot	19	475	30.434	41	Jaillon et al. (2007)
Zea mays	Monocot	10	2300	32.540	85	Schnable et al. (2009)

Type (monocot or dicot), number of chromosomes (#Chr), size (Mb) and haploid number (*n*), number of annotated genes (#Gene), percentage of repeats and related bibliographic references (author, year) are also reported

representing DNA fragments that reassociate very rapidly; the moderately repetitive ones, i.e. DNA fragments that reassociate at an intermediate rate, the single copy (or very low copy number class) representing fragments that do not repeat at a consistent frequency in DNA sequences. Such approaches to estimate the repetitive content of genomic DNAs in different organisms, though possible underestimations due to diverging repetitive elements, are remarkable since they give out a global accurate picture of genome composition in the absence of sequence information. In parallel to

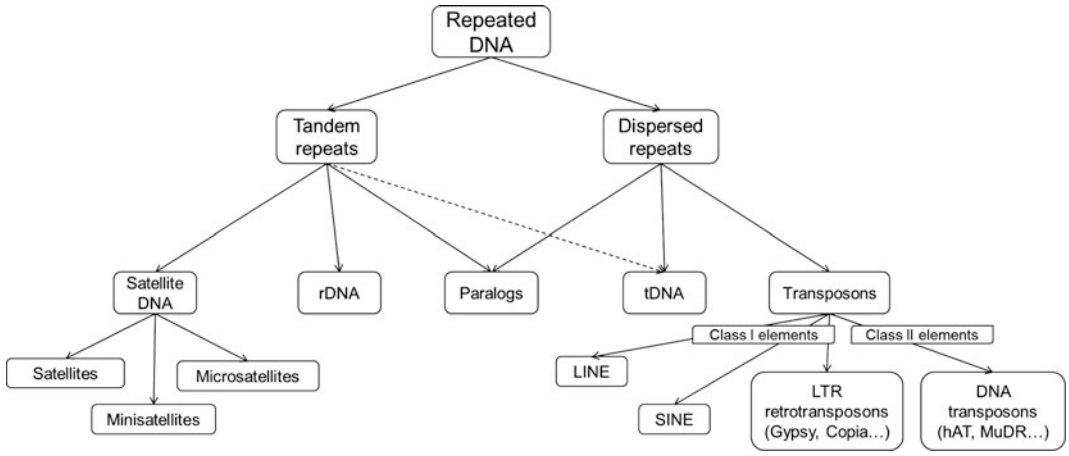

Fig. 10.1 Repeated DNA sequences in eukaryotic genomes. The two main categories of repeated elements (tandem and dispersed repeats) are shown, along with their subcategories

the reassociation kinetics properties, repeated sequences can be also divided in two major categories based on their organization or distribution in a genome: "tandem repeats" and "dispersed repeats" (Fig. 10.1). Tandem repeats are generally corresponding to the highly repetitive sequences. They mostly localize on large conspicuous heterochromatic DNA blocks at the distal ends and interstitial parts of the chromosome (Schmidt and Heslop-Harrison 1998) and include sequences that are repeated in tandem along the genome sequences such as ribosomal DNA repeat arrays (rDNA) and satellite DNA. Among tandem repeats, duplicated protein-coding genes (paralogs) can also be included. Dispersed repeats are usually corresponding to moderately repeated sequences, and include transposons and dispersed gene paralogs. Transfer RNA genes (tDNA) are often distributed in tandem, but they are usually included among the dispersed repeats (Richard et al. 2008).

Tandem Repeats

rDNA

rDNAs represent non protein-coding multigene families usually classified as tandem repeats. rDNAs (Fig. 10.1) are usually head-to-tail arrays of genes encoding the precursor (45S) of the three largest ribosomal RNAs (18S, 5.8S and 25S in plants). The corresponding DNA region generally contains several tandem copies, including active rRNA genes and silent rRNA genes, which are often highly compacted in dense heterochromatin. The rDNA region gives rise to secondary constrictions in metaphase chromosomes that are called the nucleolus organizer regions (NOR), around which the nucleolus forms. rRNA coding genes are usually transcribed by RNA polymerase I. The 5S rRNA genes, highly conserved genes of around 120nts in length, are distributed independently from the 45S rDNA, in multiple copies arranged as tandem arrays separated by a high variable spacer in sequence and in length. The number of copies of the core unit, from 200 to 900 nucleotides, can vary from 1000 to 50,000 copies. The sequences can be adjacent or not to the 45S rDNA region and are usually transcribed by the RNA polymerase III.

Satellite DNA

The name "satellite DNA" refers to a "satellite" band different in density from bulk DNA in a density gradient, due to repetitions of short DNA sequences. It consists of almost large number of repeat units, distributed as tandem arrays of DNA. Satellite DNA is in itself also distinguished in minisatellites or microsatellites. Both subcategories are variable in number of repeats

(Variable Number of Tandem Repeats or VNTR). Minisatellites consist of a core repeat units of 10 to 60–90 nucleotides. Microsatellites (also known as "Simple Sequence Repeats" or SSRs, or "Short Tandem Repeats" or STRs) consist of a core of around 2–6–10 nucleotides. In general satellite DNA can be distributed throughout the chromosomes (King et al. 1997; Richard et al. 2008), both in heterochromatin and euchromatin regions (Cuadrado and Schwarzacher 1998; Cuadrado and Jouve 2007a, b; Chang et al. 2008), in genes, both in the protein-coding regions, in introns, or in their regulatory regions, and within transposable elements.

The tandem satellite DNA sequences exhibit in general characteristic chromosomal locations, with roles depending on their locations. They can be at telomeric, subtelomeric and centromeric regions, with repetitive families that can be shared within a taxonomic family or a genus, or may be specific to the species, genome or even a chromosome (Sharma and Raina 2005). These features have formed the basis of extensive utilization of repetitive sequences for taxonomic and phylogenetic studies. Satellite DNA is the main component of centromeres, with a core units from 9 to 64 bp long, and of telomeric regions, with a conserved core units of around 6 bp, and repetition numbers that can range from hundreds to thousands, depending on the species (Podlevsky et al. 2008), forming the main structural constituent of heterochromatin. Centromeres are essential for chromosome segregation, yet their DNA sequences evolve rapidly in contrast with the high conservation of the core units of telomeres (Henikoff et al. 2001). Centromeres differ greatly in their sequence organization among different species. In *Saccharomyces cerevisiae* a "point centromere" of 125-bp sequence is sufficient to confer centromere function (Meraldi et al. 2006). In most animals and plants, centromeres contain megabase-scale arrays of simple tandem repeats, sometimes interspersed with long terminal repeat transposons (Heslop-Harrison et al. 2003) and, despite their relevant role, very little is known about the degree to which centromere tandem repeats share common properties between different species (Melters et al. 2013). However, the key kinetochore proteins are conserved in both plants and animals, particularly the centromere-specific histone H3-like protein (CENH3) highlighting the importance of epigenetic mechanisms in the establishment and maintenance of centromere identity (Houben and Schubert 2003). Telomere repeats occur predominantly at the ends of eukaryotic chromosomes, arranged in tandem to form large uninterrupted blocks often associated to subtelomeric satellite repeats (Ganal et al. 1991). They appear to protect chromosome ends from degradation and shortening during replication (Mason and Biessmann 1995).

Microsatellites may have high variability in length, due to unequal crossing over, rolling circle amplification and replication slippage, even before meiosis (Tautz and Schlotterer 1994), making these regions endowed of a high rate of mutation per locus per generation (Jarne and Lagoda 1996; Kruglyak et al. 1998). This is why these sequences are important for different approaches (Buschiazzo and Gemmell 2006). Indeed microsatellites can be amplified using unique sequences at the flanking regions to define primers for amplifications, producing variable patterns of fragments lengths which are useful for population studies, fingerprinting, marker assisted selection, and study of breeding patterns of wild or domesticated species (Martinez-Zapater et al. 1986; Maluszynska and Heslop-Harrison 1991; Michelmore et al. 1991; Martin et al. 1992; Maughan et al. 1995; Liu et al. 1996; McCouch et al. 1997; Milbourne et al. 1997; Livingstone et al. 1999).

Dispersed Repeats

tDNA

Genes coding for transfer RNAs represent a non protein-coding multigene family, as rRNA coding genes. Though often distributed in tandem, they are usually classified as dispersed repeats.

In addition to its essential function in protein synthesis, recent studies have shown that tRNAs are multifunctional molecules involved in many

processes of cellular metabolism (Minajigi and Francklyn 2010). Furthermore, tRNA-derived RNAs appear to be used in the RNA silencing pathway, and are a major source of short interspersed nuclear elements (Bermudez-Santana et al. 2010; Phizicky and Hopper 2010).

It is postulated that all tRNA genes (tDNAs) derive from an ancestral molecule (Eigen et al. 1989) that during evolution gave rise to a full set of tRNA genes generated as the result of numerous mutation, duplication and reorganization events. The number of tRNA pseudogenes and organellar-like tRNA genes present in nuclear genomes varies greatly from one plant species to another. Generally, there is no correlation between genome size and tDNA copy number in the nuclear genome (Richard et al. 2008). However, Michaud et al. (2011), in their analysis of tRNA gene distribution in plant genomes, revealed that the tRNA gene content in plants is rather homogenous, and is mostly correlated with genome size.

Transposable Elements

Among dispersed repeats, transposable elements (TEs) are DNA sequences that are capable of "moving" in the cell, integrating into a new site within the genome where they originated from (Craig et al. 2002), creating changes and amplifying and altering the cell's genome size. This is why they were also termed "jumping elements". They were discovered in plants by Barbara McClintock who earned her Nobel Prize for this scientific contribution in 1983 (McClintock 1953). She not only found that genes could move, but also that they could be turned on or off according to the environmental conditions or during different stages of cell development. Transposons consist of two major classes: retrotransposons (class I elements) and DNA transposons (class II elements) (Fig. 10.1), depending on the mechanisms that determine their excision and insertion in the genome.

Retrotransposons replicate by forming RNA intermediates, which are then reverse transcribed to DNA sequences and inserted into new genomic locations. Therefore, retrotransposons need transcription and a reverse transcriptase to move, while DNA transposons are excised from the genome, and the "cut-and-paste" mechanisms for transposition require transposases (Craig et al. 2002). Retrotransposons are commonly grouped in LTR or non-LTR retrotransposons according to the presence or not of long terminal repeats (LTR). In LTR retrotransposons, the terminal repeats range from ~100 bp to over 5 kb in size. They are the most high representative class in plant genomes (Kumar and Bennetzen 1999; Bennetzen 2000) and may be further subclassified into different classes, differing by the degree of sequence similarity and by the order of encoded gene products along their structure. Among these, Ty1-copia-like and Ty3-gypsy-like are commonly found in high copy number in plants genomes, but also in animals, fungi and protista. Retroviruses are often classified separately from the LTR retrotransposons though they share many features with them. A major difference with Ty1-copia and Ty3-gypsy retrotransposons is that Retroviruses have an Envelope protein (ENV) and have domains that enable extracellular mobility (Cotton 2001).

Non-LTR retrotransposons include long interspersed elements (LINEs) and short interspersed elements (SINEs). LINEs encodes for functionalities that are essential for retrotransposition, such as reverse transcriptase and endonucleases activities, and are transcribed by the RNA polymerase II, like mRNAs. Their mechanisms of transposition, however, differ from that of other LTR elements (Bibillo and Eickbush 2004). SINEs are nonautonomous retroelements, with length ranging from 100 to 900 bp, and copy not identical in the genome (Kramerov and Vassetzky 2005). They do not encode reverse transcriptase, and presumably co-opt the LINE machinery to be retrotransposed (Jurka 1997). They are transcribed by RNA polymerase III, being organized at their 5' end like a typical tRNA promoter (Defraia and Slotkin 2014).

Bioinformatics for Repeat Detection

Repeat Sequence Databases

Due to the presence of different types of repeats, there are different dedicated databases that organize repeats, such as *Repbase* (Jurka et al. 2005), the *Tandem Repeats Database* (Gelfand et al. 2007), *RepeatsDB* (Di Domenico et al. 2014). In particular, *RepBase* is a comprehensive repeat collection including prototypes of repetitive DNA sequences derived from the consensus of each of the repeat families from each eukaryotic species. The *Tandem Repeats Database* is specific for repeated regions in tandem, while *RepeatsDB* specifically contains tandem repeats found in protein sequences. In parallel to these resources, *Rfam* (Burge et al. 2013) contains families of non protein-coding RNAs, and is useful to support annotation of the corresponding genes in a genome, rRNA and tRNA coding genes included.

Some available databases are specific for plants, *PGSB Repeat Database* (Nussbaumer et al. 2013) and the *Plant Repeat Database* organized starting from the *TIGR Plant repeat database* (Ouyang and Buell 2004), this last updated till 2008, both designed as comprehensive repeat collections. *PlantSat* (Macas et al. 2002) and *Plant rDNA database* (Garcia et al. 2012) are dedicated to satellite repeats and rDNAs, respectively. Some of these databases have the possibility to allow search for repeated region in specific genera or species, such as the *Plant Repeat Database*, that is made of subsections dedicated to Solanaceae, Gramineae or other plants, or *Plant rDNA database*.

Methodologies

Bioinformatics strategy to identify and annotate repeats in genome sequences is almost similar even in different species. In general, the currently available methods can be based on comparative approaches, which aim to identify and therefore classify the repeated regions aligning a query sequence, the one to be analyzed, with sequences representing repeat classes organized in dedicated databases. Other approaches are based on de novo detections of repeats along a sequence, these methods supporting the identification of novel repeat sequences, i.e. sequences not available in dedicated collections since not yet discovered and classified.

RepeatMasker (Smit et al. 1996) or *Censor* (Kohany et al. 2006) are some of the well-known similarity-based search tools, useful to support the annotation of the repeats detected along a sequence and to provide its masked version, i.e. a sequence in which all the regions identical to repeats are changed to X or Ns, to be ignored in subsequent analyses, like those necessary to detect coding genes.

Similarity methods also may consider comparisons with established genome sequence references find occurrence of similar repeat regions.

Tandem Repeats Finder (Benson 1999) and *mreps* (Kolpakov 2003) are other specific tools helpful to find and annotate tandem repeats in DNA sequences. Like *LTR_STRUC* (McCarthy and McDonald 2003), *Recon* (Bao and Eddy 2002) and *RepeatScout* (Price et al. 2005), they detect repeated DNA sequences by de novo approaches. These approaches are generally based on self-comparisons of repeated similar regions. The exploitation of associated clustering approaches usually permits also to group-related sequences, to classify them into families and or subfamilies.

The identification and the annotation of repeated gene loci, such as those coding for non protein-coding genes (tRNA, rRNA), can be performed by dedicated tools like *Infernal* (Nawrocki et al. 2009), also useful for the identification of other non protein-coding RNAs. Specifically, *Infernal* is used to search RNA families dedicated databases for similar sequences such as *Rfam*. Infernal builds a profile from a structurally annotated multiple sequence alignments of RNA families with a position-specific scoring system. The scoring approach also takes into consideration secondary structure organization of the family being modelledQuery, such as base pairing, combining different levels of structure information to get to the most appropriate result. Other tools, such as *tRNAscan-SE* (Schattner et al. 2005) and *ARAGORN* (Laslett

and Canback 2004) or *SnoReport* (Hertel et al. 2008) are specific for some classes of RNAs, like tRNAs and snoRNAs, respectively.

Repeats in the Tomato Genome

Protein-coding Gene Paralogs

Though the description of protein-coding paralog genes is not the main topic of this chapter, preferred to briefly reported on their distribution in the tomato genome since they represent repeat sequences in a genome and their occurrence contributed to reveal the two consecutive triplications events of the *Solanum* lineage, that moulded the gene set controlling fruit characteristics (Tomato Genome Consortium 2012). The total number of genes with at least one paralog in tomato is 25,992, about 75 % of the total gene content. In Fig. 10.2 we report the distribution of paralog gene numbers per chromosome. This reflects the high duplication level of mRNA coding genes reported in the tomato genome (Tomato Genome Consortium 2012).

Non Protein-coding Repeated Genes

Among paralogs we may also consider large multigene families such as ribosomal RNAs (rDNA) and tRNAs (tDNA) genes.

Non protein-coding RNAs in the tomato genome sequences were annotated by *Infernal* using the *Rfam* database (version 9.1) (specifically, the collection available at ftp://ftp.sanger.ac.uk/pub/databases/Rfam/9.1/infernal-latest.tar.gz and compatible with *Infernal* 1.0) (Tomato Genome Consortium 2012).

Long rDNAs were excluded from the analyses of the tomato assembly released by the consortium, because of a specific option used by the authors when running the software *Infernal,* that excluded the annotation of these specific regions (Tomato Genome Consortium 2012, supplementary materials 2.3.2). Therefore the analysis resulted to be limited to the identification of 1853 non protein-coding RNAs of 90 distinct *Rfam* families in which almost 48 % of all the targets represented tRNA coding genes (RF00005) (Tomato Genome Consortium 2012).

Table 10.2 summarizes the results included in the *iTAG2.4_infernal.gff3* file made available by the tomato genome sequencing consortium at the ftp section of the Sol Genomics Network (http://solgenomics.net/). Moreover, in order to complete the annotation of the non protein-coding rDNAs, we performed a *BLASTn* of the tomato chromosomes versus the Large Subunit sequences (LSU, RF02543), which include the 25S RNA, and the Small Subunit (SSU, RF01960) sequences, corresponding to 18S, both collections available in the *Rfam* database (release 12.0). We considered only locus that corresponded to matches with identity and coverage ≥98 %.

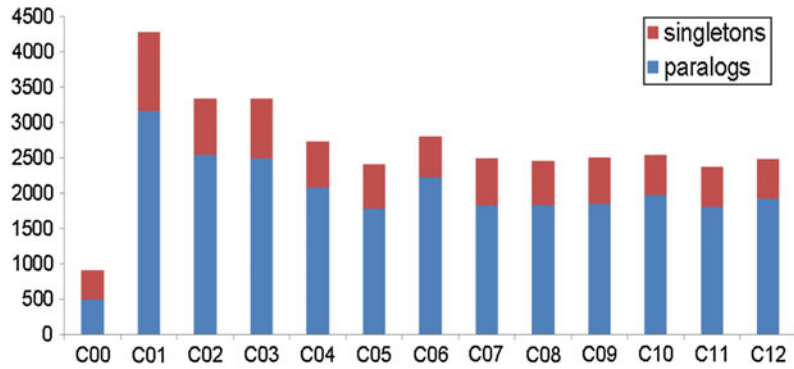

Fig. 10.2 Paralog gene distribution per chromosome. The data source from which we report this summary is obtained from BioMart section of *EnsemblPlants* (http://plants.ensembl.org/)

Table 10.2 Number of 5.8S rRNA, 5S rRNA, tRNA as reported by the Tomato Genome Consortium (2012)

	iTAG v. 2.4			Udated	
	5.8S rRNA	5S rRNA	tRNA	18S rRNA	25S rRNA
chr 00	11	3	16	4	20
chr 01	2	38	109	4	9
chr 02	0	1	76	1	6
chr 03	2	3	83	5	6
chr 04	0	1	71	1	4
chr 05	3	0	60	2	1
chr 06	7	0	102	5	11
chr 07	1	2	52	2	4
chr 08	0	0	70	1	8
chr 09	0	2	44	2	3
chr 10	0	0	90	1	8
chr 11	13	4	48	12	21
chr 12	1	0	64	2	6
Sum	40	54	885	42	107

Updated contents of 25S rRNA and 18S rRNA gene are also shown

5.8S rRNA genes defined by the consortium are listed mainly on chromosomes 11 and 6, while higher figures are reported by our updating corresponding to regions similar to 25S sequences (Table 10.2). It is also evident that there are still matches on the unassigned sequences collected as *unassembled* on "chromosome 0", probably because the difficulties in assigning repeated sequences during the assembly of large and complex genomes.

The table also shows a high number of 5S coding regions on chromosome 1 (Fig. 10.3a), confirming the loci identified as repeated in tandem by FISH on pachythene chromosomes on the short arm of chromosome 1 (1S), close to the centromeric region (Vallejos et al. 1986; Lapitan et al. 1991; Xu and Earle 1996a, b). Though, as explained, the information on the long rDNA regions (45S or at least 18S and 25S families) was not available from the sequencing and annotation effort, we reviewed the information collected from analyses preceding the tomato genome sequencing and exploited our updating based on the *BLASTn* analysis. Indeed, it was known that ribosomal DNA represents the most abundant repetitive DNA family in tomato, comprising approximately 3 % of the genome. From experimental analysis, 5S and 45S rRNA genes were detected as tandemly repeated with 1000 and 2300 copies. Karyotyping in combination with fluorescence in situ hybridization (FISH) on tomato pachytene chromosomes allowed the identification and mapping of the 45S rDNA on the satellite of the short arm of chromosome 2 (2S) and a minor locus on 2L, though these evidence are not confirmed by the tomato genome sequencing, from which no match, neither with the only considered marker 5,8S, was detected (Vallejos et al. 1986; Tanksley et al. 1988; Lapitan et al. 1991; Xu and Earle 1996a, b). However, these results find some confirmation from our updated analysis, with few matches from the 25S confirmed on chromosome 2. Other minor loci were also revealed at 6S, 9S and 11S (Xu and Earle 1996a, b), the first and the last also finding some confirmation by the annotation from the consortium, with stronger support by our update. Indeed, the updated analysis shows regions similar to the 25S (LSU) in all the

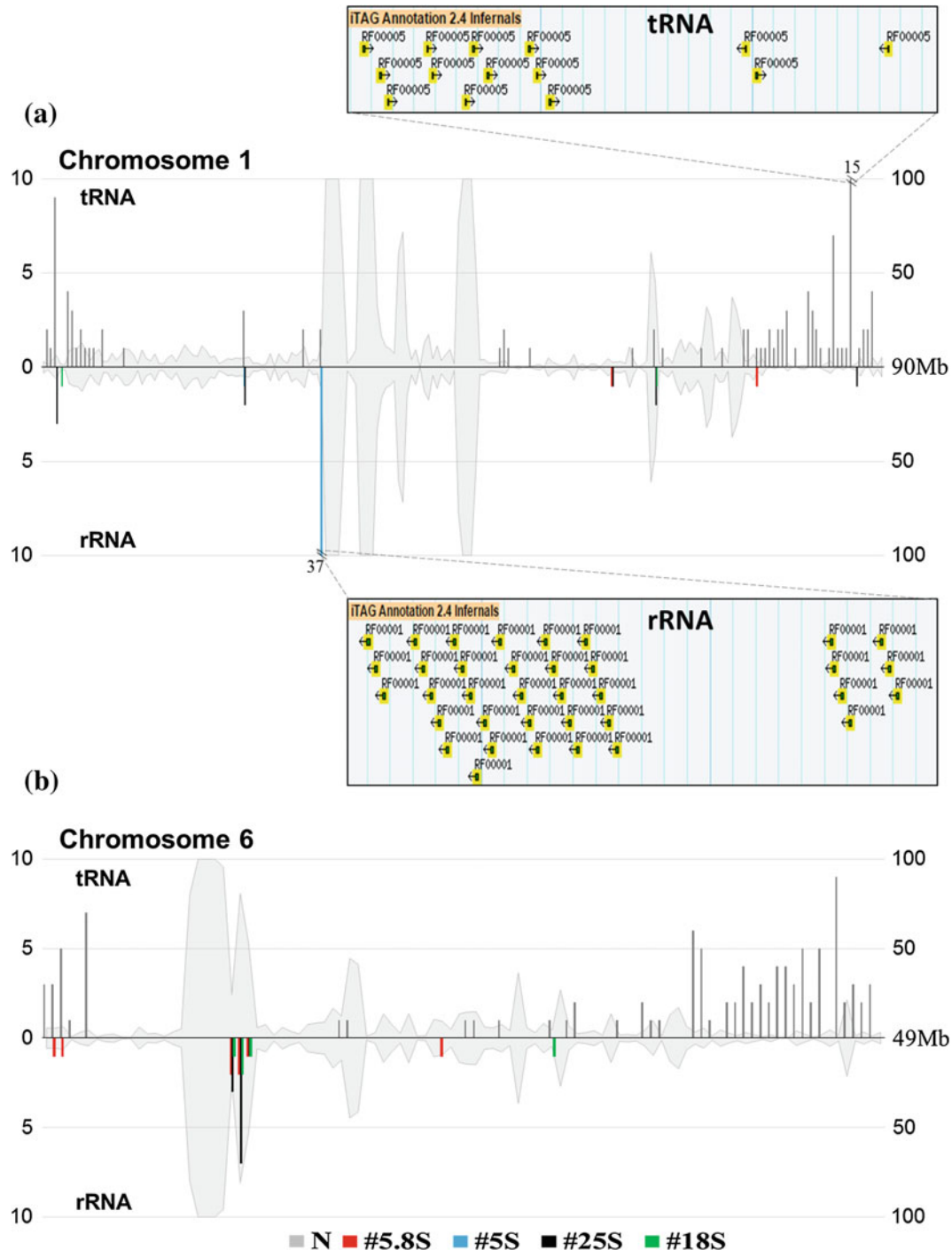

Fig. 10.3 Distribution per chromosome 1 (**a**) and chromosome 6 (**b**) of repeated non protein-coding genes. Percentage of *N* is also reported by a nonoverlapping window analysis of chromosomes divided per 500 Kb, with a total of 197 windows for chromosome 1 and 100 windows for chromosomes 6. Details of regions with 5S rRNA and tRNA in tandem on chromosome 1 are shown

chromosomes, accompanied by a similar distribution by the 18S, though with lower numbers, in contrast with what expected from previous analysis.

In Fig. 10.3a, b the distribution of non protein-coding genes on chromosomes 1 and 6 are shown, respectively. Data are from the *iTAG2.4_infernal.gff* file made available by the tomato genome consortium at ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.4_release/. Moreover, the results from the updated analysis here provided are also shown in the figure.

Our updated analysis also permitted the clear identification of an rDNA locus associated to the occurrence of 45S loci on chromosome 6, since 18S 5.8S and 25S are all located in the region (Fig. 10.3b).

tDNA distribution is shown both in Table 10.2 and in Fig. 10.3. Interestingly to notice, their occurrence is reported in all the chromosomes.

Noncoding Tandem Repeats

Noncoding tandem repeat sequences in tomato chromosomes were detected using the de novo approach of *Tandem Repeats Finder* (Benson 1999), with default parameters. This permitted to classify the sequences by length into microsatellites (2–9 bp), minisatellites (10–99) and satellites (≥100-bp), while overlapping annotations of more than one of the three classes were classified as hybrid type.

The whole collections of tandem repeats resulted to cover 3.2 % of the genome, with the major contribution from minisatellites (1.7 of the entire genome and 53.7 % of the tandem repeats). Microsatellite repeats in tomato genome were also analyzed by Suresh et al. (2014), who detected a total of 68,641 microsatellite repeat motifs. Dinucleotide repeats (60.18 %) resulted much more abundant than tri (19.56 %) and other repeats, of which ∼82.90 and ∼17.10 % were simple and compound repeats, respectively. A total of 5841 and 4773 SSRs were present in the assigned genes and their 5′-upstream sequences, with average frequencies of 0.172 SSRs/gene and 0.14 SSRs/5′-upstream sequences, respectively. Data are accessible at the *Tomato Genomic Resources Database* (http://59.163.192.91/tomato2/).

Telomere

Beyond rDNAs, telomeres are the most ubiquitous tandem repeated arrays in the genome of eukaryotes.

The telomere repeats have been studied extensively in species of the Solanaceae family, which show mostly the Arabidopsis-type telomere (TTTAGGG). The typical tomato telomeric repeat (TR) (TT(T/A)AGGG) is arranged in tandem to form large uninterrupted blocks (Ganal et al. 1991). A block of 162-bp subtelomeric repeats (TGRI) is localized a few hundred kb from the terminal telomere repeats in 20 of the 24 homologous chromosomes (Ganal et al. 1988, 1991; Schweizer et al. 1988; Lapitan et al. 1989). These repeated blocks together accounts for around the 2 % of the total chromosomal DNA and, though the TR repeat is highly conserved, the long range physical structure of these arrays has been shown to be highly variable in different varieties (Broun et al. 1992) and within the genome (Zhong et al. 1998). Zhong et al. (1998) investigated on the relative length and distribution of the TR the spacer and the TGRI blocks in tomato chromosomes. The major evidence from Zhong et al. work was to highlight differences in TR-spacer-TGRI organization in most if not all the chromosome ends in tomato. Concerning the role of the spacer and the TGRI repeats it is assumed that they could represent buffering blocks separating chromosome ends from unique sequences or alternatively, playing a role in favouring or preventing chromosome degradation, fusions and fissions (Meyne et al. 1990). However, they have also been speculated to be regions susceptible to unequal crossing over between homologous and even nonhomologous chromosomes, yielding to high polymorphisms even in conserved genomes (Broun et al. 1992).

Interestingly, interstitial telomeric repeats (ITRs) were also revealed hybridizing the TR repeat on lambda clones of tomato, showing

unexpected telomere homologous sequences on 8 of the 12 tomato centromeres (Ganal et al. 1991; Presting et al. 1996).

ITRs are organized as short tandem arrays and are expected to be evolutionary relics derived from chromosomal rearrangements and DNA repairs (He et al. 2013). However, megabase-sized ITR arrays were reported in *Solanum* species (Tek and Jiang 2004). These results showed that some ITR subfamilies were amplified and invaded the functional centromeres of Solanaceae chromosomes revealing possible other roles than simply being relics of chromosomal rearrangements. The epigenetic landscape and transcription of telomeres and ITRs were also investigated. As an example, in *Nicotiana tabacum* (with no detectable ITRs), and in *Ballantinia antipoda*, (with large blocks of pericentromeric ITRs and relatively short telomeres) Majerová et al. (2014) revealed that genuine telomeres displayed heterochromatic as well as euchromatic marks, while ITRs were just heterochromatic. Methylated cytosines were present at telomeres and ITRs, but showed a bias with more methylation towards distal telomere positions and different blocks of ITRs methylated to different levels (Majerová et al. 2014). Interestingly, the authors also showed that telomeres and ITRs are transcribed, and that the level of telomerase transcripts is tissue dependent, contributing novel insights for the understanding of the specific role and regulation activity of the associated transcripts.

Centromere

The tomato genome sequencing confirmed the presence of a high DNA repeat content in the heterochromatin pericentromeric regions, however no value added information was provided by the sequencing effort to characterize centromeric tandem repeated regions. It is known, however, that both the centromeric satellites and the retroelements are essential for centromere recognition by kinetochore proteins (Zhong et al. 2002; Nagaki and Murata 2005; Nagaki et al. 2011), and previous efforts also revealed the mosaic structure of centromeres in plant species (Nagaki et al. 2012). Interestingly, though it was evident that centromeric repeats evolve rapidly (Melters et al. 2013), Gong et al. (2012) recently reported that six of the 12 potato centromeres contain megabase-sized arrays of satellite repeats different in each centromere. By contrast, five potato centromeres are shown to be composed of single- and low-copy DNA sequences, with no satellite repeats detected. These five potato centromeres structurally resemble neocentromeres. Moreover, they also showed that most of the centromeric satellite repeats in potato were amplified recently from retrotransposon-related sequences and are not present in wild *Solanum* species closely related to potato.

A deeper comparative analysis revealed that different centromeric haplotypes were found to be associated with three potato centromeres, including haplotypes containing megabase-sized satellite repeats and haplotypes that do not contain the same repeats (Wang et al. 2014).

To further understand the evolution of centromeric DNA in *Solanum* species, (Zhang et al. 2014) conducted a genome-wide analysis of DNA sequences associated with the cenH3 nucleosomes in *Solanum verrucosum* ($2n = 2x = 24$), a wild species closely related to potato. They demonstrated a rapid divergence of the centromeric sequences between these two closely related species. Therefore, they hypothesized that centromeric satellite repeats may undergo boom–bust cycles of evolution from which a structurally favourable repeat lengths, maybe favouring the structure ideal for cenH3 nucleosome organization, could take place.

Many existing centromeres are believed to have originated as neocentromeres that activated de novo from noncentromeric regions by acquiring specific histones in the nucleosome (for example, the canonical histone H3 is replaced by cenH3 histone in plants or by CENP-A in animals (Kalitsis and Choo 2012; Rocchi et al. 2012). Newly formed neocentromeres are associated with gene "desert" regions and initially do not contain satellite repeats (Marshall et al. 2008; Wang et al. 2014). The evolutionarily new centromeres presumably accumulate satellite repeats and/or retrotransposons during evolution and eventually evolve

rapidly to become repeat-based centromeres (Yan et al. 2006; Kalitsis and Choo 2012; Sharma et al. 2013).

Transposons

Considering the dispersed repeats, we already reported on tDNA distribution in the tomato genome in the paragraph on non protein-coding repeated gene families.

The other relevant class among dispersed repeats includes the transposons. In Table 10.3, we report the nucleotide coverage in terms of transposon classes of all the chromosomes, as derived from the annotation reported in the *iTAG2.4_repeat.gff3* file released by the tomato genome consortium (Tomato Genome Consortium 2012) and available at http://solgenomics.net.

While the pseudomolecules images in the Nature paper report the general behaviour of repeat content along tomato and potato pseudomolecules, in this chapter we provide, as an example, a more detailed view with a similar approach showing the distribution of all single class of repeats along tomato chromosomes 1 and 6 (Fig. 10.4a, b).

As reported from Nature 2012, full length LTR retrotransposons in the tomato genome sequence, were detected by a curated analysis starting from a de novo approach based on *LTR-STRUC* (McCarthy and McDonald 2003). 1647 intact LTR retrotransposons were detected. These sequences were assigned to the gypsy or copia subgroups which were identified thanks to the order of their inner protein domains.

Additional full length LTR elements were found by sequence similarity, leading to a total of 4052 still intact elements. Moreover, a cluster analyses of these sequences highlighted that tomato and potato (Potato Genome Sequencing Consortium 2011) genome sequences shared common LTR retrotransposons (Tomato Genome Consortium 2012).

The insertion events of LTR retrotransposons were also dated by the sequence divergence between left and right LTRs (Wiley et al. 2009). Interestingly, this analysis showed fewer copies in tomato and potato when compared to sorghum and older insertion age. This appears to be a peculiarity of tomato, and apparently also of potato, among angiosperms (Tomato Genome Consortium 2012).

Transposons along tomato chromosomes were annotated by the *wublast* version of *RepeatMasker* (http://www.repeatmasker.org) against the dicots section of *mipsREdat* (REdat_v8.9_Eudico). This transposon library is connected to a repeat classification scheme (mips_REcat) and contains a collection of known transposons as well as de novo detected LTR retrotransposons from tomato (1647) and potato (1309). The *RepeatMasker* output was subjected to two post-processing filter steps: (a) removal of low confidence hits (length <50 bp, score ≥255) and (b) cleaning of overlapping annotations, considering higher score hits first, and overlapping lower scored hits either shortened or, if the overlap exceeded 80 % of their length, removed.

In Table 10.3 we redefined the nucleotide coverage in terms of repeat classes for all the tomato chromosomes, starting from the available annotation from the consortium (Tomato Genome Consortium 2012).

Moreover, while the pseudomolecule images in the Nature 2012 paper (Tomato Genome Consortium 2012) reports the general behaviour or the global repeat content along tomato pseudomolecules, in this chapter we provide a more detailed view with a similar approach showing the distribution of all single classes of repeats along chromosomes 1 and 6 (Fig. 10.4a, b).

Moreover, in Fig. 10.5 we report the distribution of the transposons by the delta repeat minus gene content in a 500 kb window in chromosome 6. The plots confirmed the high content of LTR retrotransposon in repeat-rich regions, that should correspond to heterochromatin regions (Di Filippo et al. 2012) with higher content of the gypsy-like class and much lower content of the copia-like one. The plots also show that, among non-LTR retrotransposon, the SINE are more frequent in gene richer regions, as also demonstrated at BAC level (Di Filippo et al. 2012), with a similar trend also from LINE.

Table 10.3 Number of nucleotides covered by transposons per chromosomes

Length		*N*	Retro transposon	LTR	LTR copia	LTR gypsy	LINE	SINE	DNA transposon	DNA En-Spm	DNA Harbinger	DNA hAT	DNA MuDR	Other
C00	21,805,821	3,139,100	5	11,455	1762	4741	10	58	135	18	7	29	26	14
C01	98,543,444	12,423,381	3962	32,077,426	5,750,265	19,161,095	72,439	129,011	290,884	43,507	14,911	96,182	16,837	14,521
C02	55,340,444	8,083,432	1605	14,898,384	2,445,478	8,553,327	66,702	75,569	189,419	18,401	11,143	33,888	9377	11,392
C03	70,787,664	9,925,850	3167	22,389,432	3,411,165	13,988,572	66,267	99,920	220,311	29,141	10,310	85,887	9375	5984
C04	66,470,942	6,021,919	4998	23,441,720	4,051,988	13,641,676	93,403	113,876	234,454	30,573	9137	50883	14,996	12,163
C05	65,875,088	4,634,458	4760	24,682,363	4,329,908	16,365,009	66,477	96,407	179,533	27,364	6125	84,475	11,744	9455
C06	49,751,636	6,178,685	3216	14,940,362	2,694,369	8,510,709	64,352	79,036	176,359	15,585	9356	49,379	6312	7342
C07	68,045,021	6,084,209	3232	25,568,639	4,245,047	15,897,705	60,272	103,023	172,764	27,979	7901	62,597	10,107	6780
C08	65,866,657	6,081,969	3236	24,947,566	3,749,174	15,354,025	55,707	90,328	180,081	38,447	7729	72,172	10,124	8276
C09	72,482,091	7,614,308	4151	26,933,831	4,358,334	16,786,921	41,483	109,364	179,824	31,557	8050	36,901	12,588	4313
C10	65,527,505	4,736,321	3387	25,077,433	4,499,404	15,169,358	40,894	94,067	196,755	64,172	5481	99,554	12,679	6774
C11	56,302,525	6,045,240	4510	19,645,202	3,058,258	11,949,620	75,219	93,800	163,561	22,893	8539	44,329	8894	10,355
C12	67,145,203	5,338,821	5274	26,165,887	4,402,233	16,376,894	49,614	119,045	210,762	18,246	7514	44,773	9091	8398
Sum	823,944,041	86,307,693	45,503	280,779,700	46,997,385	171,759,652	752,839	1,203,504	2,394,842	367,883	106,203	761,049	132,150	105,767

The LTR column includes annotations without further classification. Chromosome lengths (length) and number of Ns per chromosome are also specified

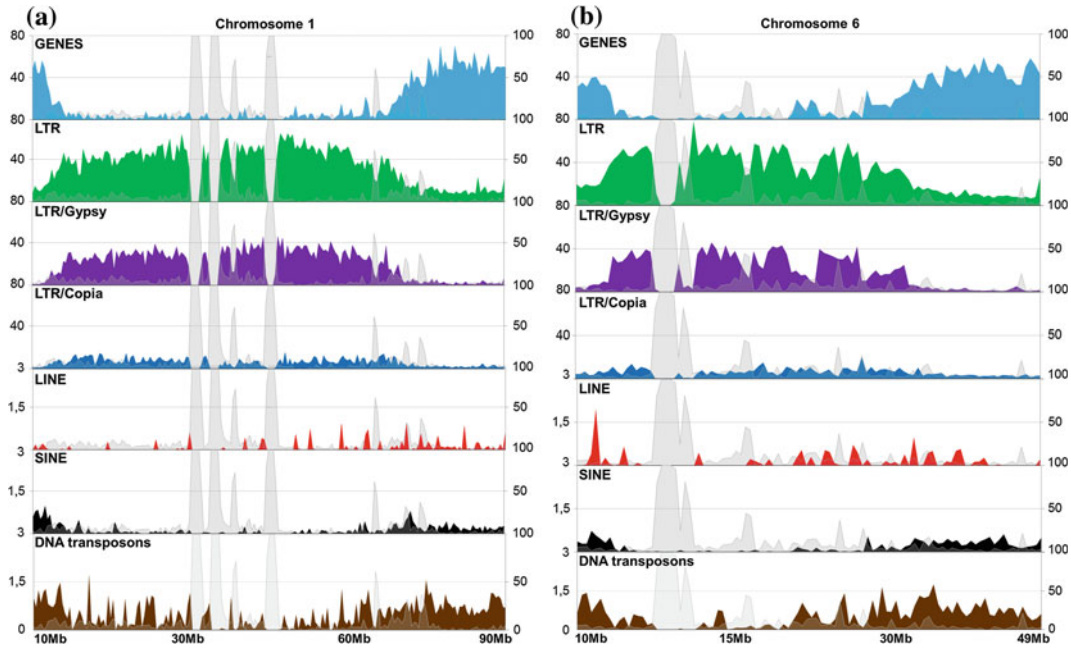

Fig. 10.4 Distribution of gene and repeat content along chromosomes 1 and 6. *Annotation of line*, LTR, Gypsy, Copia, Sine and DNA transposons were obtained from *ITAG2.4_repeats.gff3*; gene annotations were from *ITAG2.4_gene_models.gff3*, both available at http://solgenomics.net/. Data are reported by a 500 Kb nonoverlapping window. *Left* and *right* *y*-axes represent different percentages. The *right* *y*-axes represent the number of undefined nucleotide (*N*) per window

The *iTAG2.4_repeats.gff3* file used to perform this analysis was downloaded from the ftp section at http://solgenomics.net/.

Discussion

Solanaceae is an unusually divergent family consisting of approximately 90 genera and 3000–4000 species (Knapp et al. 2004) and almost all members share the same chromosome number ($x = 12$) (Wikstrom et al. 2001). Though the genomes appeared to have undergone relatively small numbers of chromosomal rearrangements (Park et al. 2011), they maintained a conserved gene content and order (Bonierbale et al. 1988; Tanksley et al. 1988; Prince et al. 1993; Livingstone et al. 1999; Wang et al. 2008; Wu et al. 2009). Though, the sequencing of different genotypes of the same species revealed microscale heterogeneity between cultivated and wild species (Traini et al. 2013; Ercolano et al. 2014; Qin et al. 2014), the overall conservation of the Solanaceae gene regions was generally described as conserved, even at the level of syntenic segments (Wang et al. 2011). The level of conservation revealed at gene level, however, is not confirmed when considering genome size, repetitive sequence content and composition. Within the Solanaceae family, *Solanum lycopersicum* (tomato) has a genome size of ~950 Mb, the genome size of *Solanum tuberosum* (potato) is 840 Mb and *Capsicum annuum* (pepper) genomes is of 3349 Mb, though the estimated gene content is comparable, suggesting a possible significant role of repeats in the speciation of these clade of plants (Zhu et al. 2008).

The 12 tomato chromosomes consist of an extended heterochromatic region (>60 % genome), mostly representing the telomeres and extended pericentromeric regions. The euchromatin regions locate in the distal part of the chromosome (Peterson et al. 1996, 1998), composed of most single-copy sequences with fewer

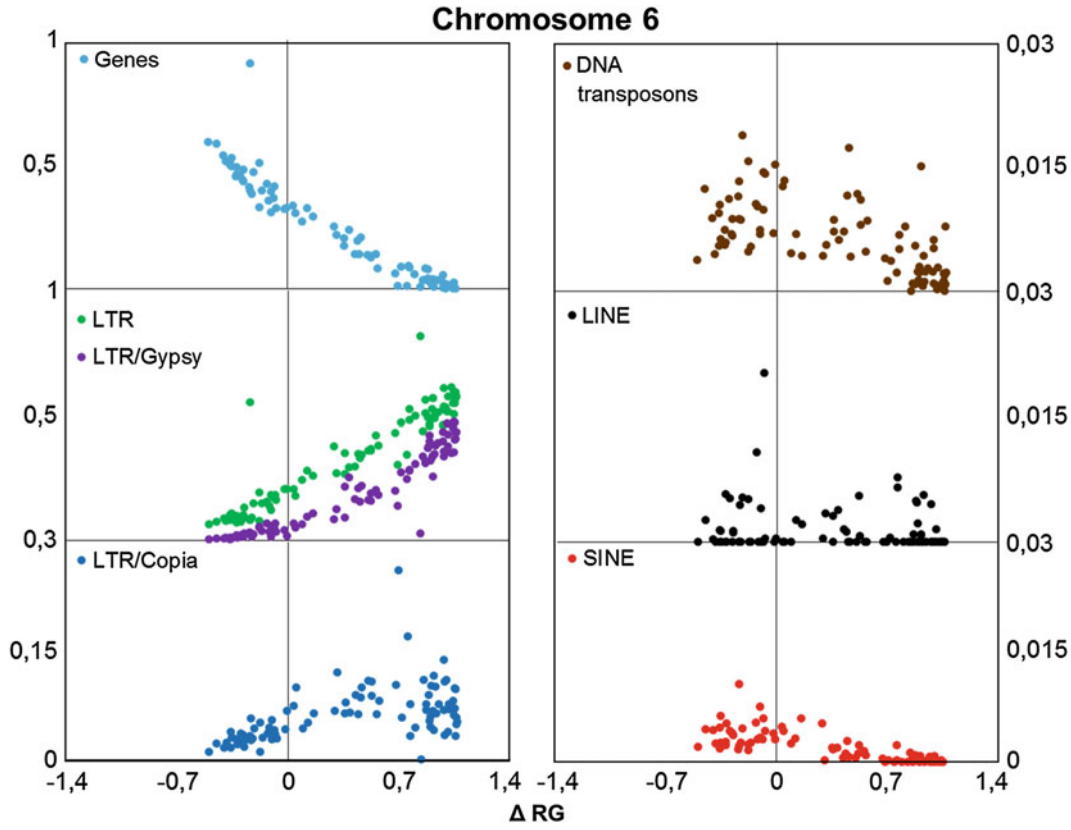

Fig. 10.5 Distribution of main repeat classes by windows of 500 kb along chromosome 6. The data are reported as frequency in the window versus the difference between repeat and gene content frequency (ΔRG). Annotations were obtained as for Fig. 10.3

retrotransposon and the 90 % of the genes (Chang et al. 2008).

Pericentromeric heterochromatin is generally assumed to be gene poor and repeat-rich, where crossing over is severely repressed (Sherman and Stack 1995). The pericentromeric heterochromatic segments contain a large portion of retrotransposons, other types of repeated sequences and some single-copy sequences, which also include a lower but representative gene content (Di Filippo et al. 2012).

Among tandem repeats, ribosomal DNA represents one of the most abundant repetitive DNA family. The repeat unit, estimated to be 9.1 Kb, was expected of 2300 copies and at the end of chromosome 2 by Ganal et al. (1988). rDNA should represent the 3 % of the tomato genome and its distribution was described also by several other efforts (Vallejos et al. 1986; Lapitan et al. 1991). As reported in this chapter, the rDNA regions appear not to be exhaustively covered by the tomato genome sequencing and by the associated annotation, and this is presumably the reason why they are not broadly discussed in the effort (Tomato Genome Consortium 2012). However, the presence of satellite DNA joint to the intergenic spacer of rDNA units also reveals the strong association of these two types of repeats and a possible initiation of satellite repeats from these loci (Jo et al. 2009).

Previous analysis also confirmed a 162 bp satellite repeat, named TGRI, with 77,000 copies in the genome as localized within a few hundred kb of the terminal 7 bp telomeric repeat TT(T/A)AGGG in tomato, at 20 of 24 chromosome ends (Ganal et al. 1988). In addition, internal

telomeric repeats (ITR) were also found at a few centromeric and interstitial sites (Lapitan et al. 1989; Ganal et al. 1992; Presting et al. 1996), opening interesting questions on the reasons of this organization, as also highlighted in this chapter.

Two other tomato genomic repeats, TGRII and TGRIII, are less abundant, and were estimated with 4200 and 2100 copies, respectively. TGRII is apparently randomly distributed with quite a regular spacing of 133 kb (Ganal et al. 1988), while TGRIII is predominantly clustered in the pericentromeric region. The TGRIV repeat was discovered later and it was found mainly associated to satellite repeats in the centromere (Chang et al. 2008).

Microsatellite polymorphism and genomic distribution were studied in tomato by fingerprinting using labelled oligonucleotide probes complementary to GATA or GACA microsatellites (Vosman et al. 1992; Grandillo and Tanksley 1996). The mapping of individual fingerprint bands showed main association to centromeres (Arens et al. 1995). The copy number and the size of microsatellite containing restriction fragments were proved to be highly variable between tomato cultivars (Arens et al. 1995). Structure, abundance, variability and location were also evaluated (Broun and Tanksley 1996) and successfully used for genotyping tomato cultivars and accessions (Smulders et al. 1997; Bredemeijer et al. 2002). Interestingly, what is evident in tomato is the presence of compound satellite repeats, highly variable in length and strongly specific to the species. Ganal et al. (1988), underlined that the distribution of the major classes of tandem repeats described in tomato is limited to this species. This is probably due to high evolving rate of these regions. Zamir and Tanksley (1988) also reported a positive correlation between copy number and rate of divergence of repeats among DNA sequences from related Solanaceae species. This means that highly repeated regions are less conserved when compared to single-copy regions, coherently also with a different selective pressure on the two types of regions. Further analyses revealed rapid evolution of centromere-proximal sequences (Presting et al. 1996) which is also confirmed from analysis in other Solanaceae (Gong et al. 2012; Melters et al. 2013; Wang et al. 2014; Zhang et al. 2014).

Among all classes of repeats, transposons comprise a large proportion of the tomato genome. In general, the highest contribution to dispersed repeats in plant genomes is mainly due to LTR retrotransposons (Piegu et al. 2006; Richard et al. 2008; Lee and Kim 2014). Plants show more C-value variation than other taxa (http://data.kew.org) (Bennett and Leitch 2005), which appears to be correlated with LTR retrotransposon abundance (Michael 2014). In animals non-LTR elements appear to be more abundant (Sakowicz et al. 2009). DNA transposons have minor impact on genome size because of the way they expand (Lee and Kim 2014). In particular, repeat-rich regions of the tomato genome revealed abundance of the LTR retroelements Ty3–gypsy and Ty1–copia (Yasuhara and Wakimoto 2006; Chang et al. 2008; Szinay et al. 2008; Tang et al. 2008a, b; Peters et al. 2009; Di Filippo et al. 2012), though the second class is present at a less extent, as also confirmed by the tomato genome annotation (Table 10.3; Fig. 10.5).

In Di Filippo et al. (2012), tomato genome sequences obtained by the preliminary BAC sequencing that preceded the whole-genome shotgun approach were analyzed to correlate heterochromatin and euchromatin regions with the relative gene and repeat content. Moreover, in the same effort, molecular markers, available to define the eu/heterochromatin boundaries along each tomato chromosome (data from the Solanaceae Genome Network website), and all the BACs associated to the chromosome structure by fluorescence in situ hybridization (FISH) (de Jong 1998; de Jong et al. 2000; Wang et al. 2006; Szinay et al. 2008; Tang et al. 2008a, b; Peters et al. 2009) were used to analyze the associated sequences. This gave out a preliminary confirmation based on sequence analysis that BACs associated to euchromatin in the tomato genome were indeed richer in gene and lower in repeat content when compared to BACs associated to heterocromatin regions. The analyses presented in Di Filippo et al. (2012), while confirming the

initial assumption that genes were predominantly located in repeat-poor euchromatin regions, proved that the repeat-rich heterochromatic BACs were not completely depleted of genes (Yasuhara and Wakimoto 2006; Mueller et al. 2009). Interestingly, Di Filippo et al. (2012) also proposed an immediate approach to show the specific content of repeat classes in tomato gene or repeat richer BACs, corresponding to euchromatic and heterochromatic BACs, respectively. We also exploited the same approach here to confirm, at chromosome level, the distribution of different repeat classes in compositionally different genome regions (Fig. 10.5).

Today it is well known that transposons play various relevant roles in genome evolution, gene expression regulation and genetic instability. They can change position within the genome, contributing to genome reorganizations and altering the genome size, since transposition often results in duplication of the transposable elements, contributing with their movement to changes in cell function and organisms development (Nowacki et al. 2009) as well as to genome reorganization. Interestingly, in most cases transposable elements are silenced through epigenetics mechanism like methylation and chromatin remodelling. As a consequence, no phenotypic effects nor the movement of transposons occur when, in the wild type plant, they are silenced (Martienssen and Colot 2001; Reik et al. 2001). It is important to note, however, that DNA methylation is not conceived as a factor provoking heterochromatin formation (some species may lack methylation) but rather as a factor stabilizing heterochromatin structures (for review, see Wolffe and Matzke 1999).

Type, number and size of repeat domains in a genome can vary among species, but even differ between close genotypes or accessions, being useful as genome markers in karyotype analysis and chromosome markers in a segregating population. However, based on the assumption that a portion that comprises such a large extent of higher eukaryotes genome sequence cannot be without specific reasons, more interesting could be the understanding of the role and, possibly, advantages, if any, in repeat expansion or reduction, as well as association of these phenomena with heterochromatin formation. A prerequisite for heterochromatin formation appears to be the structural organization of the repeats rather than the nature of the particular sequences, or their repetitive character. It is evident that DNA repeats have specific structure role in constitutive heterochromatin, essential in multicellular organisms at chromosomal and nuclear level. At the chromosomal level, constitutive heterochromatin is present around vital areas such as telomeres and centromeres. The centromeric satellite DNA and retrotransposons are known to be essential in the recognition of the kinetochore (Zhong et al. 2002; Nagaki et al. 2003). The pericentromeric repeats are considered important in the recruitment of histone modification enzymes promoting the formation and maintenance of heterochromatin (Hall et al. 2002; Volpe et al. 2002; Zhong et al. 2002; Bender 2004; Lippman et al. 2004) and conferring protection and strength to the centromere. Around secondary constrictions, heterochromatic blocks may ensure against evolutionary change of ribosomal DNA by decreasing the frequency of crossing over in these regions during meiosis, also absorbing the effects of mutagenesis. Indeed, repetitive sequences in the form of constitutive heterochromatin appeared concomitant with the localization of the portion of the genome that was concerned with synthesis of ribosomal RNA, and with the need to protect chromosome structure and function by telomeres and centromeres, when the mitotic spindle developed in evolution. During meiosis heterochromatin may also aid in the initial alignment of chromosomes, facilitating speciation by allowing chromosomal rearrangement but also providing, through the species specificity of its DNA, barriers against cross-fertilization. At the nuclear level, constitutive heterochromatin may help to maintain the spatial relationships through all the steps of cell cycle. The repetitive DNA was therefore kept through natural selection and, because of its innate attitude to amplify and expand, it favoured eukaryotes genome expansion and evolution (Yunis and Yasmineh 1971; Bennetzen and Kellogg 1997). This occurred in the limit of an

efficient management of other cellular activities (Knight et al. 2005). In principle, repeats are prone to expand but there exist also mechanisms to decrease dramatically their content, if necessary, including illegitimate or unequal recombination and other type of deletions (Grover and Wendel 2010). However, beyond the relevance here discussed, and the impact DNA repeats can have on genome evolution and expansion, it would also be rather important to investigate on further possible roles of species specific repeats in structuring and protecting the genome though the energy requirements that genome expansion can take from cell functionality.

References

Ahn S, Anderson JA, Sorrells ME, Tanksley SD (1993) Homoeologous relationships of rice, wheat and maize chromosomes. Mol Gen Genet 241(5–6):483–490

Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W et al (2013) Genome sequence of the date palm Phoenix dactylifera L. Nat Commun 4:2274

Arabidopsis Genome Initiative T (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796–815

Arens P, Odinot P, Heusden AV, Lindhout P, Vosman B (1995) GATA-and GACA-repeats are not evenly distributed throughout the tomato genome. Genome 38(1):84–90

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G et al (2011) The genome of *Theobroma cacao*. Nat Genet 43(2):101–108

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12(8):1269–1276

Bender J (2004) Chromatin-based silencing mechanisms. Curr Opin Plant Biol 7(5):521–526

Bennett M, Leitch I (2005) Plant DNA C-values database. Royal Botanic Gardens, Kew

Bennett MD, Smith JB (1976) Nuclear dna amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci 274 (933):227–274

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12(7):1021–1030

Bennetzen JL, Kellogg EA (1997) Do plants have a one-way ticket to genomic obesity? Plant Cell 9 (9):1509–1514

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27 (2):573–580

Bermudez-Santana C, Attolini CS, Kirsten T, Engelhardt J, Prohaska SJ et al (2010) Genomic organization of eukaryotic tRNAs. BMC Genomics 11(1):270

Bibillo A, Eickbush TH (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. J Biol Chem 279(15):14945–14953

Bonierbale MW, Plaisted RL, Tanksley SD (1988) RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics 120(4):1095–1103

Bowen NJ, Jordan IK (2002) Transposable elements and the evolution of eukaryotic complexity. Curr Issues Mol Biol 4(3):65–76

Bredemeijer G, Cooke R, Ganal M, Peeters R, Isaac P et al (2002) Construction and testing of a microsatellite database containing more than 500 tomato varieties. Theor Appl Genet 105(6–7):1019–1026

Broun P, Ganal MW, Tanksley SD (1992) Telomeric arrays display high levels of heritable polymorphism among closely related plant varieties. Proc Natl Acad Sci 89(4):1354–1357

Broun P, Tanksley S (1996) Characterization and genetic mapping of simple repeat sequences in the tomato genome. Mol Gen Genet MGG 250(1):39–49

Brunner AM, Busov VB, Strauss SH (2004) Poplar genome sequence: functional genomics in an ecologically dominant plant species. Trends Plant Sci 9 (1):49–56

Burge SW, Daub J, Eberhardt R, Tate J, Barquist L et al (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 41:D226–D232

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28(10):1040–1050

Chang SB, Yang TJ, Datema E, van Vugt J, Vosman B et al (2008) FISH mapping and molecular organization of the major repetitive sequences of tomato. Chromosome Res 16(7):919–933

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371(6494):215–220

Chen J, Huang Q, Gao D, Wang J, Lang Y et al (2013) Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. Nat Commun 4:1595

Cotton J (2001) Retroviruses from retrotransposons. Genome Biol 2(2):1

Craig NL, Craigie R, Gellert M, Lambowitz AM (2002) Mobile DNA II. ASM Press, Washington, DC

Cuadrado A, Jouve N (2007a) The nonrandom distribution of long clusters of all possible classes of trinucleotide repeats in barley chromosomes. Chromosome Res 15(6):711–720

Cuadrado A, Jouve N (2007b) Similarities in the chromosomal distribution of AG and AC repeats within and between Drosophila, human and barley chromosomes. Cytogenet Genome Res 119(1–2):91–99

Cuadrado A, Schwarzacher T (1998) The chromosomal organization of simple sequence repeats in wheat and rye genomes. Chromosoma 107(8):587–594

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE et al (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16(6):738–749

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F et al (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488(7410):213–217

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

de Jong JH, Zhong XB, Fransz PF, Wennekes-van Eden J, Jacobsen E et al (2000) High resolution FISH reveals the molecular and chromosomal organisation of repetitive sequences of individual tomato chromosomes. In: Olmo E, Redi C (eds) Chromosomes today. Birkhäuser, Basel, pp 267–275

Defraia C, Slotkin RK (2014) Analysis of retrotransposon activity in plants. Methods Mol Biol 1112:195–210

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345(6201):1181–1184

Devos KM, Gale MD (1993) Extended genetic maps of the homoeologous group 3 chromosomes of wheat, rye and barley. Theor Appl Genet 85(6–7):649–652

Devos KM, Gale MD (1997) Comparative genetics in the grasses. Plant Mol Biol 35(1–2):3–15

Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. Plant Cell 12 (5):637–646

Di Domenico T, Potenza E, Walsh I, Parra RG, Giollo M et al (2014) RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Res 42:D352–D357

Di Filippo M, Traini A, D'Agostino N, Frusciante L, Chiusano ML (2012) Euchromatic and heterochromatic compositional properties emerging from the analysis of *Solanum lycopersicum* BAC sequences. Gene 499(1):176–181

Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress A et al (1989) How old is the genetic code? Statistical geometry of tRNA provides an answer. Science 244(4905):673–679

Ercolano MR, Sacco A, Ferriello F, D'Alessandro R, Tononi P et al (2014) Patchwork sequencing of tomato San Marzano and Vesuviano varieties highlights genome-wide variations. BMC Genomics 15:138

Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12(4):257–269

Ganal MW, Broun P, Tanksley SD (1992) Genetic mapping of tandemly repeated telomeric DNA sequences in tomato (Lycopersicon esculentum). Genomics 14(2):444–448

Ganal MW, Lapitan NL, Tanksley SD (1988) A molecular and cytogenetic survey of major repeated DNA sequences in tomato (Lycopersicon esculentum). Mol Gen Genet MGG 213(2–3):262–268

Ganal MW, Lapitan NL, Tanksley SD (1991) Macrostructure of the tomato telomeres. Plant Cell 3(1):87–94

Garcia S, Garnatje T, Kovarik A (2012) Plant rDNA database: ribosomal DNA loci information goes online. Chromosoma 121(4):389–394

Gebhardt C, Ritter E, Barone A, Debener T, Walkemeier B et al (1991) RFLP maps of potato and their alignment with the homoeologous tomato genome. Theor Appl Genet 83(1):49–57

Gelfand Y, Rodriguez A, Benson G (2007) TRDB—the tandem repeats database. Nucleic Acids Res 35(suppl 1):D80–D87

Gong Z, Wu Y, Koblížková A, Torres GA, Wang K et al (2012) Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell Online 24(9):3559–3574

Grandillo S, Tanksley S (1996) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. Theor Appl Genet 92(8):935–951

Greilhuber J, Doležel J, Lysák MA, Bennett MD (2005) The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. Ann Bot 95(1):255–260

Grover CE, Wendel JF (2010) Recent insights into mechanisms of genome size change in plants. J Bot 2010:8

Hall IM, Shankaranarayana GD, Noma K-I, Ayoub N, Cohen A et al (2002) Establishment and maintenance of a heterochromatin domain. Science 297 (5590):2232–2237

Hawkins JS, Proulx SR, Rapp RA, Wendel JF (2009) Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc Natl Acad Sci U S A 106 (42):17811–17816

He L, Liu J, Torres GA, Zhang H, Jiang J et al (2013) Interstitial telomeric repeats are enriched in the centromeres of chromosomes in Solanum species. Chromosome Res 21(1):5–13

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293(5532):1098–1102

Hertel J, Hofacker IL, Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics 24(2):158–164

Heslop-Harrison JS (2000) Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. Plant Cell 12 (5):617–636

Heslop-Harrison JS, Brandes A, Schwarzacher T (2003) Tandemly repeated DNA sequences and centromeric chromosomal regions of Arabidopsis species. Chromosome Res 11(3):241–253

Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S et al (2014) Draft genome sequence of eggplant (Solanum melongena L.): the representative

solanum species indigenous to the old world. DNA Res 21(6):649–660
Houben A, Schubert I (2003) DNA and proteins of plant centromeres. Curr Opin Plant Biol 6(6):554–560
Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF et al (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet 43(5):476–481
Huang S, Li R, Zhang Z, Li L, Gu X et al (2009) The genome of the cucumber, Cucumis sativus L. Nat Genet 41(12):1275–1281
International Barley Genome Sequencing, Mayer CKF, Waugh R, Brown JW, Schulman A et al (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491(7426):711–716
International Rice Genome Sequencing, P (2005) The map-based sequence of the rice genome. Nature 436 (7052):793–800
IWGSC (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345(6194):1251788
Jaillon O, Aury JM, Noel B, Policriti A, Clepet C et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449(7161):463–467
Jarne P, Lagoda PJ (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11 (10):424–429
Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR et al (2012) A genome triplication associated with early diversification of the core eudicots. Genome Biol 13(1):R3
Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473(7345):97–100
Jo S-H, Koo D-H, Kim J, Hur C-G, Lee S et al (2009) Evolution of ribosomal DNA-derived satellite repeat in tomato genome. BMC Plant Biol 9(1):42
Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci 94(5):1872–1877
Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O et al (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1–4):462–467
Kalitsis P, Choo KH (2012) The evolutionary life cycle of the resilient centromere. Chromosoma 121(4):327–340
Kim S, Park M, Yeom SI, Kim YM, Lee JM et al (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat Genet 46(3):270–278
King DG, Soller M, Kashi Y (1997) Evolutionary tuning knobs. Endeavour 21(1):36–40
Knapp S et al (2004) Solanaceae—a model for linking genomics with biodiversity. Comp Funct Genomics 5(3):285–291
Knight CA, Molinari NA, Petrov DA (2005) The large genome constraint hypothesis: evolution, ecology and phenotype. Ann Bot 95(1):177–190
Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in repbase: repbasesubmitter and censor. BMC Bioinform 7:474
Kolpakov R (2003) mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 31 (13):3672–3678
Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. Int Rev Cytol 247:165–221
Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95(18):10774–10778
Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532
Lapitan NL, Ganal MW, Tanksley SD (1989) Somatic chromosome karyotype of tomato based on in situ hybridization of the TGRI satellite repeat. Genome 32 (6):992–998
Lapitan NLV, Ganal MW, Tanksley SD (1991) Organization of the 5S ribosomal RNA genes in the genome of tomato. Genome 34(4):509–514
Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32(1):11–16
Lee S-I, Kim N-S (2014) Transposable elements and genome size variations in plants. Genom Inform 12 (3):87–97
Ling HQ, Zhao S, Liu D, Wang J, Sun H et al (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. Nature 496(7443):87–90
Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430 (6998):471–476
Liu Z-W, Biyashev R, Maroof MS (1996) Development of simple sequence repeat DNA markers and their integration into a barley linkage map. Theor Appl Genet 93(5–6):869–876
Livingstone KD, Lackney VK, Blauth JR, Van Wijk R, Jahn MK (1999) Genome mapping in Capsicum and the evolution of genome structure in the Solanaceae. Genetics 152(3):1183–1202
Lysak MA, Koch MA, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe Brassiceae. Genome Res 15(4):516–525
Macas J, Meszaros T, Nouzova M (2002) PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18(1):28–35
Majerová E, Mandáková T, Vu GTH, Fajkus J, Lysak MA et al (2014) Chromatin features of plant telomeric sequences at terminal vs. internal positions. Front Plant Sci 5:593
Maluszynska J, Heslop-Harrison J (1991) Localization of tandemly repeated DMA sequences in Arabidopsis thaliana. Plant J 1(2):159–166
Marshall OJ et al (2008) Neocentromeres: new insights into centromere structure, disease

development, and karyotype evolution. Am J Hum Genet 82(2):261–282
Martienssen RA, Colot V (2001) DNA methylation and epigenetic inheritance in plants and filamentous fungi. Science 293(5532):1070–1074
Martin GB, Ganal MW, Tanksley SD (1992) Construction of a yeast artificial chromosome library of tomato and identification of cloned segments linked to two disease resistance loci. Mol Gen Genet MGG 233(1–2):25–32
Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol 118(1):9–17
Martinez-Zapater JM, Estelle MA, Somerville CR (1986) A highly repeated DNA sequence in *Arabidopsis thaliana*. Mol Gen Genet MGG 204(3):417–423
Mason JM, Biessmann H (1995) The unusual telomeres of Drosophila. Trends Genet 11(2):58–62
Maughan P, Maroof MS, Buss G (1995) Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. Genome 38(4):715–723
McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19(3):362–367
McClintock B (1953) Induction of instability at selected loci in maize. Genetics 38(6):579–599
McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y et al (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol 35(1–2):89–99
Melters DP, Bradnam KR, Young HA, Telis N, May MR et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14(1):R10
Meraldi P, McAinsh AD, Rheinbay E, Sorger PK (2006) Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. Genome Biol 7(3):R23
Meyne J, Baker RJ, Hobart HH, Hsu T, Ryder OA et al (1990) Distribution of non-telomeric sites of the (TTAGGG) n telomeric sequence in vertebrate chromosomes. Chromosoma 99(1):3–10
Michael TP (2014) Plant genome size variation: bloating and purging DNA. Brief Funct Genomics 13(4):308–317
Michaud M, Cognat V, Duchêne A-M, Maréchal-Drouard L (2011) A global picture of tRNA genes in plant genomes. Plant J 66(1):80–93
Michelmore RW, Paran I, Kesseli R (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci 88(21):9828–9832
Milbourne D, Meyer R, Bradshaw JE, Baird E, Bonar N et al (1997) Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. Mol Breed 3(2):127–136
Minajigi A, Francklyn CS (2010) Aminoacyl transfer rate dictates choice of editing pathway in threonyl-tRNA synthetase. J Biol Chem 285(31):23810–23817
Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452(7190):991–996
Ming R, VanBuren R, Liu Y, Yang M, Han Y et al (2013) Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biol 14(5):R41
Mueller LA, Lankhorst RK, Tanksley SD, Giovannoni JJ, White R et al (2009) A snapshot of the emerging tomato genome sequence. Plant Gen 2(1):78–92
Nagaki K, Murata M (2005) Characterization of CENH3 and centromere-associated DNA sequences in sugarcane. Chromosome Res 13(2):195–203
Nagaki K, Shibata F, Kanatani A, Kashihara K, Murata M (2012) Isolation of centromeric-tandem repetitive DNA sequences by chromatin affinity purification using a HaloTag7-fused centromere-specific histone H3 in tobacco. Plant Cell Rep 31(4):771–779
Nagaki K, Shibata F, Suzuki G, Kanatani A, Ozaki S et al (2011) Coexistence of NtCENH3 and two retrotransposons in tobacco centromeres. Chromosome Res 19 (5):591–605
Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q et al (2003) Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics 163(2):759–770
Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25(10):1335–1337
Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG et al (2009) A functional role for transposases in a large eukaryotic genome. Science 324(5929):935–938
NSF (1990) Document 90–80. A long-range plan for the multinational coordinated Arabidopsis thaliana genome research project. National Science Foundation, Washington, DC
Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC et al (2013) MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res 41:D1144–D1151
Ouyang S, Buell CR (2004) The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:D360–D363
Park M, Jo S, Kwon J-K, Park J, Ahn JH et al (2011) Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. BMC Genomics 12(1):85
Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457(7229):551–556
Peters SA, Datema E, Szinay D, van Staveren MJ, Schijlen EG et al (2009) *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. Plant J 58(5):857–869
Peterson DG, Pearson WR, Stack SM (1998) Characterization of the tomato (*Lycopersicon esculentum*)

genome using in vitro and in situ DNA reassociation. Genome 41(3):346–356
Peterson DG, Stack SM, Price HJ, Johnston JS (1996) DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes. Genome 39(1):77–82
Phizicky EM, Hopper AK (2010) tRNA biology charges to the front. Genes Dev 24(17):1832–1860
Piegu B, Guyot R, Picault N, Roulin A, Saniyal A et al (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res 16(10):1262–1269
Podlevsky JD, Bley CJ, Omana RV, Qi X, Chen JJL (2008) The telomerase database. Nucleic Acids Res 36:D339–D343
Potato Genome Sequencing Consortium, T (2011) Genome sequence and analysis of the tuber crop potato. Nature 475(7355):189–195
Presting GG, Frary A, Pillen K, Tanksley SD (1996) Telomere-homologous sequences occur near the centromeres of many tomato chromosomes. Mol Gen Genet MGG 251(5):526–531
Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):i351–i358
Prince JP, Pochard E, Tanksley SD (1993) Construction of a molecular linkage map of pepper and a comparison of synteny with tomato. Genome 36 (3):404–417
Qin C, Yu C, Shen Y, Fang X, Chen L et al (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc Natl Acad Sci USA 111 (14):5135–5140
Reik W, Dean W, Walter J (2001) Epigenetic reprogramming in mammalian development. Science 293 (5532):1089–1093
Richard G-F, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72(4):686–727
Rocchi M, Archidiacono N, Schempp W, Capozzi O, Stanyon R (2012) Centromere repositioning in mammals. Heredity 108(1):59–67
Sakowicz T, Gadzalski M, Pszczółkowski W (2009) Short interspersed elements (SINEs) in plant genomes. Adv Cell Biol 1:1–12
Sangiovanni M, Vigilante A, Chiusano ML (2013) Exploiting a reference genome in terms of duplications: the network of paralogs and single copy genes in *Arabidopsis thaliana*. Biol (Basel) 2(4):1465–1487
Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T et al (2008) Genome structure of the legume, Lotus japonicus. DNA Res 15(4):227–239
Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686–W689
Schmidt T, Heslop-Harrison J (1998) Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends Plant Sci 3(5):195–199
Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463(7278):178–183
Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112–1115
Schweizer G, Ganal M, Ninnemann H, Hemleben V (1988) Species-specific DNA sequences for identification of somatic hybrids between Lycopersicon esculentum and Solanum acaule. Theor Appl Genet 75(5):679–684
Sharma S, Raina SN (2005) Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. Cytogenet Genome Res 109(1–3):15–26
Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W et al (2013) Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 Genes Genomes Genet 3(11):2031–2047
Shearer LA, Anderson LK, de Jong H, Smit S, Goicoechea JL, et al (2014) Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome. G3 Genes Genomes Genet 4(8):1395–1405
Sherman JD, Stack SM (1995) Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). Genetics 141 (2):683–708
Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O et al (2011) The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43 (2):109–116
Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L et al (2014) The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun 5:3833
Smit AFA, Hubley R, Green P (1996) RepeatMasker Open-3.0. http://www.repeatmasker.org
Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other Lycopersicon species. Theor Appl Genet 94(2):264–272
Smyth DR (1991) Dispersed repeats in plant genomes. Chromosoma 100(6):355–359
Suresh BV, Roy R, Sahu K, Misra G, Chattopadhyay D (2014) Tomato genomic resources database: an integrated repository of useful tomato genomic information for basic and applied research. PLoS One 9(1):e86387
Swift H (1950) The constancy of desoxyribose nucleic acid in plant nuclei. Proc Natl Acad Sci USA 36 (11):643–654

Szinay D, Chang SB, Khrustaleva L, Peters S, Schijlen E et al (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(4):627–637

Tang H, Bowers JE, Wang X, Ming R, Alam M et al (2008a) Synteny and collinearity in plant genomes. Science 320(5875):486–488

Tang X, Szinay D, Lang C, Ramanna MS, van der Vossen EA et al (2008b) Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. Genetics 180(3):1319–1328

Tanksley SD, Bernatzky R, Lapitan NL, Prince JP (1988) Conservation of gene repertoire but not gene order in pepper and tomato. Proc Natl Acad Sci USA 85 (17):6419–6423

Tautz D, Schlotterer (1994) Simple sequences. Curr Opin Genet Dev 4(6):832–837

Tek AL, Jiang J (2004) The centromeric regions of potato chromosomes contain megabase-sized tandem arrays of telomere-similar sequence. Chromosoma 113 (2):77–83

Tomato Genome Consortium, T (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485(7400):635–641

Traini A, Iorizzo M, Mann H, Bradeen JM, Carputo D, Frusciante L, Chiusano ML (2013) Genome microscale heterogeneity among wild potatoes revealed by diversity arrays technology marker sequences. Int Journal Genomics 2013:9

Vallejos CE, Tanksley SD, Bernatzky R (1986) Localization in the tomato genome of DNA restriction fragments containing sequences homologous to the rRNA (45s), the major chlorophyll a/b binding polypeptide and the ribulose bisphosphate carboxylase genes. Genetics 112(1):93–105

Vaughan H, Heslop-Harrison J, Hewitt G (1999) The localization of mitochondrial sequences to chromosomal DNA in orthopterans. Genome 42(5):874–880

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI et al (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 297(5588):1833–1837

von Sternberg R (2002) On the roles of repetitive DNA elements in the context of a unified genomic-epigenetic system. Ann NY Acad Sci 981:154–188

Vosman B, Arens P, Rus-Kortekaas W, Smulders M (1992) Identification of highly polymorphic DNA regions in tomato. Theor Appl Genet 85(2–3):239–244

Wang L, Zeng Z, Zhang W, Jiang J (2014) Three potato centromeres are associated with distinct haplotypes with or without megabase-sized satellite repeat arrays. Genetics 196(2):397–401

Wang X, Wang H, Wang J, Sun R, Wu J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43(10):1035–1039

Wang Y, Diehl A, Wu F, Vrebalov J, Giovannoni J et al (2008) Sequencing and comparative analysis of a conserved syntenic segment in the Solanaceae. Genetics 180(1):391–408

Wang Y, Tang X, Cheng Z, Mueller L, Giovannoni J et al (2006) Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. Genetics 172(4):2529–2540

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc Biol Sci 268(1482):2211–2220

Wiley G, Macmil S, Qu C, Wang P, Xing Y et al (2009) Methods for generating shotgun and mixed shotgun/paired-end libraries for the 454 DNA sequencer. Curr Prot Hum Genet 18.11:11-18–11-21

Wolffe AP, Matzke MA (1999) Epigenetics: regulation through repression. Science 286(5439):481–486

Wu F, Eannetta NT, Xu Y, Durrett R, Mazourek M et al (2009) A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus Capsicum. Theor Appl Genet 118(7):1279–1293

Xu J, Earle ED (1996a) Direct FISH of 5S rDNA on tomato pachytene chromosomes places the gene at the heterochromatic knob immediately adjacent to the centromere of chromosome 1. Genome 39(1):216–221

Xu J, Earle ED (1996b) High resolution physical mapping of 45S (5.8S, 18S and 25S) rDNA gene loci in the tomato genome using a combination of karyotyping and FISH of pachytene chromosomes. Chromosoma 104(8):545–550

Yan H, Ito H, Nobuta K, Ouyang S, Jin W et al (2006) Genomic and genetic characterization of rice Cen3 reveals extensive transcription and evolutionary implications of a complex centromere. Plant Cell 18 (9):2123–2133

Yasuhara JC, Wakimoto BT (2006) Oxymoron no more: the expanding world of heterochromatic genes. Trends Genet 22(6):330–338

Yunis JJ, Yasmineh WG (1971) Heterochromatin, satellite DNA, and cell function. Science 174(4015):1200–1209

Zamir D, Tanksley S (1988) Tomato genome is comprised largely of fast-evolving, low copy-number sequences. Mol Gen Genet MGG 213(2–3):254–261

Zhang H, Koblizkova A, Wang K, Gong Z, Oliveira L et al (2014) Boom-bust turnovers of megabase-sized centromeric DNA in solanum species: rapid evolution

of DNA sequences associated with centromeres. Plant Cell 26(4):1436–1447

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A et al (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell Online 14(11):2825–2836

Zhong XB, Fransz PF, Wennekes-van Eden J, Kammen AV, Zabel P et al (1998) FISH studies reveal the molecular and chromosomal organization of individual telomere domains in tomato. Plant J 13(4):507–517

Zhu W, Ouyang S, Iovene M, O'Brien K, Vuong H et al (2008) Analysis of 90 Mb of the potato genome reveals conservation of gene structures and order with tomato but divergence in repetitive sequence composition. BMC Genomics 9:286