Chapter 13
Papaya Repeat Database

Niranjan Nagarajan and Rafael Navajas-Pérez

Introduction

Thomas (1971) first suggested that the lack of correlation between genome size and structural complexity is mainly due to the accumulation of repetitive sequences by coining the term C-value paradox. Since then, genomes have been proved to actively expand by means of several mechanisms including polyploidization, transposition, and duplication. Today, it is well known that animal, among them mice and humans, and plant genomes, including such agriculturally important plants as rice, corn, or wheat, have acquired a repertoire of repetitive elements accounting for the vast majority of nuclear DNA in many cases (Kubis et al. 1998).

Three main classes of repetitive sequences are considered: transposable elements (TEs), tandem repeats (TRs), and high copy number genes. On the one hand, TEs constitute the most abundant component of many plant genomes, ranging from 40 up to 80 % of total genomic DNA (Bennetzen et al. 2005). TEs can be further divided into DNA-mediated class II transposons and RNA-mediated class I retrotransposons. DNA transposons were first described by Barbara McClintock as genetic elements capable of transposing to different chromosomal locations in maize plants (1950) and today are known to constitute an important family of TEs in plants (Jiang et al. 2003). The most common TEs in plants though are LTR retrotransposons (Novikov et al. 2012); non-LTR retrotransposons, while numerous, remain mostly inactive and under regulation of the host genome (Cheng and Ling 2006). On the other hand, TRs are main constituents of centromeric, telomeric,

N. Nagarajan
Department of Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore

R. Navajas-Pérez (✉)
Facultad de Ciencias, Departamento de Genética, Universidad de Granada, Campus de Fuentenueva s/n, Granada 18071, Spain
e-mail: rnavajas@ugr.es

R. Ming and P.H. Moore (eds.), *Genetics and Genomics of Papaya*, Plant Genetics and Genomics: Crops and Models 10, DOI 10.1007/978-1-4614-8087-7_13,

and subtelomeric regions of many eukaryotes, comprising hundreds to thousands of tandemly arrayed monomeric repeats (Ugarkovic and Plohl 2002). These repeats also appear at interspersed positions and in low-recombining regions, such as sex chromosomes or B chromosomes (Camacho et al. 2000; Navajas-Pérez 2012). This type of sequences can account for a large portion of genomic DNA (Saini et al. 2008). The third class is constituted by high copy number genes. Some molecular data suggest that a great number of plant genes belong to gene families ranging in size from a few members to hundreds (Martienssen and Irish 1999).

Apart from the role of constitutive heterochromatin, traditionally linked to a major architectonic function necessary for cell division (Yunis and Yasmineh 1971), repetitive elements—in the best of cases—have been considered dispensable if not junk or selfish DNA with no function at all (Ohno 1972; Orgel and Crick 1980). This lack of function contrasts with their prevalence in the genomes. In fact, an increasing number of pieces of evidence are changing the whole picture posing that repetitive sequences would play important roles in different biological aspects. It is now evident, for example, that repetitive sequences have been crucial drivers of genome evolution significantly contributing to the expansion of the genomes and consequently shaping contemporary chromosome organization through events of chromosome rearrangements due to interactions between scattered repeats (Fedoroff 2000). Indeed, up to 70 % of flowering plants have evolved through a polyploid ancestor in their lineages so both whole-genome and segmental duplications are common and key events in plant genome evolution (Wang et al. 2012). Also, variations in repeats content are thought to influence the determination of continuous phenotypic characters (Meagher and Vassiliadis 2005; Gemayel et al. 2010) or be related to the response to environmental cues (Schmidt and Anderson 2006). More recently, the implication of repetitive elements in gene regulation has been demonstrated (Thornburg et al. 2006; Lunyak et al. 2007; Román et al. 2011) suggesting they might be fundamental for the creation of new genes and sophisticated regulatory network systems. Thus, the study of repetitive sequence elements is essential to understand the nature and consequences of genome size variation between different species and for studying the large-scale organization and evolution of plant genomes.

Finally, it is worth mentioning the implication of repetitive sequences (mainly satellite DNAs and retrotransposons) in the emergence of sex chromosomes. Many Y chromosomes have more abundant heterochromatin derived from repetitive sequences compared with X chromosomes and autosomes. Accumulation of repetitive sequences contributes to generate gene deserts found in the Y chromosomes, Y-chromosome chromatin expansion, and chromosome breaks and rearrangements and may well be a key factor in the generation of differences in morphology and size observed between X and Y chromosomes that ultimately prevent the recombination in the sex-determining region. This has been proved by cytogenetic analyses and more recent genome-based projects in both plant and animal genomes (Matsunaga 2009; Navajas-Pérez 2012). Also theories on the implication of TEs in the origin of sex reproduction have been put forward (Arkhipova 2005).

Due to their high rate of change, repetitive elements have been used to detect polymorphisms in diverse type of biological analyses. The variation of

minisatellites repeat copy led to the DNA profiling method for general use in human genetic analysis (Jeffreys et al. 1985). Soon, with the advent of PCR, hypervariable and ubiquitous microsatellite markers became widely used in genome mapping and population analysis and genotyping (Ellegren 2004). Also, satellite DNA has helped to clarify phylogenetic relationships among related species by checking the presence/absence status (Navajas-Pérez 2012) or by analyzing the rates of change (Robles et al. 2004) and to understand the dynamics of repetitive elements in the genomes (Navajas-Pérez et al. 2009a). Other repetitive DNA as rRNA sequences has been traditionally used in phylogeny (Pace 2009). TEs have been used to examine genome structure and composition as well as to successfully elucidate evolutionary relationships (Ray 2007).

In these grounds, several databases devoted to store, curate, and classify repetitive DNA have been developed recently: as for satellite repeats (Macas et al. 2002), tandem repeats (Navajas-Pérez and Paterson 2009), or transposable elements (Llorens et al. 2011; Bousios et al. 2012). As the genomic information increases a pleiad of methods for mining, detection and further analysis of repetitive DNA are arising (Benson 1999; Jurka 2003; Navajas-Pérez et al. 2007).

Papaya, because of its position in the tree of life sharing a common ancestor with *Arabidopsis* about 72 million years ago, and with the existence of an incipient pair of sex chromosomes is a promising genomic model. In the past decade many genomic resources have been generated, as a draft whole-genome sequence, an integrated genetic and physical map including sex-determining region, three BAC libraries, and a large collection of ESTs (Ming et al. 2008; Na et al. 2012; Wang et al. 2013; also Chap. 17 in this text). This offers a good opportunity to characterize the papaya repeatome among many other issues. In fact, coupled with this development, a large collection of SSR and AFLP markers comprising sex-specific markers have been characterized (Ma et al. 2004; Chen et al. 2007), and a papaya repeat database has been generated (Nagarajan et al. 2008). In this chapter, we highlight the most relevant information regarding this matter.

Transposable Elements

The papaya repeatome is dominated by TEs, comprising of 52 % of the genome and ~93 % of the repeatome as described in Nagarajan et al. (2008). This is almost certainly a conservative estimate as repeat elements and TEs, in particular, are hard to assemble from whole-genome shotgun sequencing data (Nagarajan and Pop 2009). In addition, the vast majority of identified TEs in the genome are papaya-specific (71 %) and unidentifiable using consensus sequences for other plant repeats, underscoring the rapid divergence of TE families in plant genomes (Nagarajan et al. 2008). Using de novo repeat finders and manual curation, a custom library of TE families was constructed for the papaya genome, providing a curated database of 889 papaya TE families that serve as a resource for annotation of newly sequenced plant genomes (ftp://ftp.cbcb.umd.edu/pub/data/CPR-DB).

Table 13.1 Summary of TE content of papaya WGS and sex chromosomes

Class	Element	Percentage of sequence (%)		
		MSY	X	WGS
(I Retrotransposons)	Ty1/copia	4.6	5.8	5.5
	Ty3/gypsy	47.1	35.1	27.8
	LINE	0.6	1.1	1.1
	SINE	0	0	<0.01
	Other	11.4		8.4
		64.3	49.9	42.8
(II Transposons)	CACTA/En-Sp	<0.01	0	0.01
	MuDR-IS905	0	0	<0.01
	Tc1-IS630-Pogo	0	0	<0.01
	Other	0	0	<0.01
		0.1	0.1	0.01
Unclassified	Unknown	13.4	9.4	8.72
Total		77.8	59.3	51.62

A wide representation of known common types of TEs were found in the papaya genome, with retrotransposons (40 % of the genome) being the dominant class and *Ty3–gypsy* (27.8 %) being the dominant type (71 % of these are papaya specific). *Ty1–copia* (5.5 %) and LINE (1 %) retrotransposons, as well as CACTA-like DNA transposons (0.1 %), were the other major identifiable types. A significant fraction of the TE matches were either unknown retrotransposons (8.4 %) or unannotated families (8.5 %), highlighting the need for further characterization of these repeat families. In particular, the observed lack of known DNA transposons (0.2 % of the genome) compared to other plant genomes could be due to the presence of unannotated papaya-specific DNA transposon families (Table 13.1; Nagarajan et al. 2008).

In agreement with earlier observations, the TE content in the papaya genome is intermediate between the much smaller *Arabidopsis* genome (Arabidopsis Genome Initiative 2001) (14 % TE content) and the much larger maize genome (Messing et al. 2004) (58 % TE content), but as a function of the genome size, it is relatively repeat rich (Nagarajan et al. 2008). The high TE content of the papaya genome serves to explain the observation that it has a smaller gene repertoire than the *Arabidopsis* genome despite having a genome that is three times the size (Ming et al. 2008). The expansion of most TE families in the papaya genome is presumably ancient, with most TE matches being inactive fossils that have diverged substantially from their consensus. However, for several families (papaya-specific, often *Ty3–gypsy* elements), dozens of nearly perfect copies can be found in the papaya genome, some with EST matches, suggesting that some elements may still be active (Nagarajan et al. 2008).

A striking feature of TEs in the papaya genome is the similarity with the rice genome despite their divergence on the species tree. As reported in Nagarajan et al. (2008), 57 % of matches to retrotransposons and 81 % of matches to DNA transposons among TIGR plant repeats (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats) were to rice repeats. Phylogenetic analysis of *Ty1–copia* and *Ty3–gypsy* elements (Fig. 13.1)

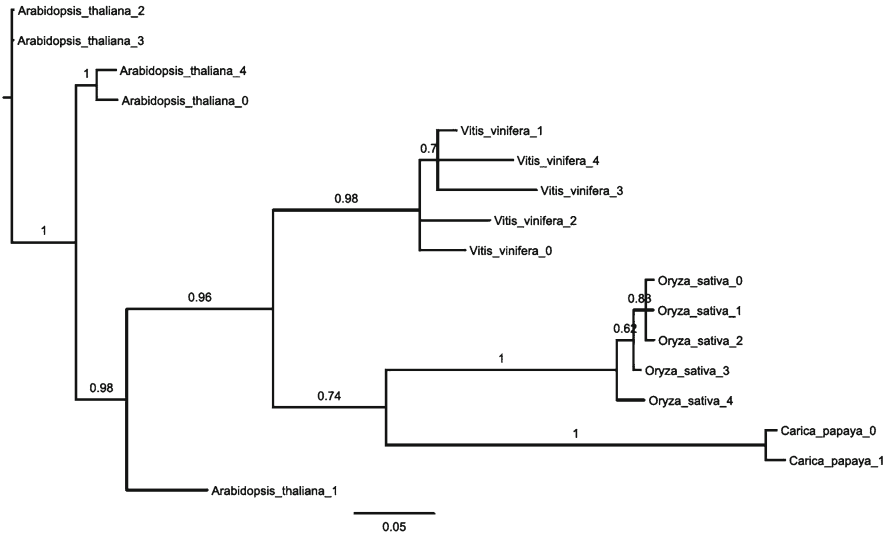

Fig. 13.1 Phylogenetic analysis of plant genome sequences matching the Ty3–gypsy retrotransposon sequence ATGP5A_I in bases 3,700–4,100 (the five best matches for each species were included in the phylogenetic analysis) [modified with kind permission of Springer Science + Business Media from Nagarajan et al. (2008)]

also revealed a similar pattern where papaya sequences tended to cluster with rice sequences. It was also observed that the ratio of *Ty3–gypsy* to *Ty1–copia* elements in the papaya genome was closer to the 2:1 ratio of the rice genome than to the 1:1 of *Arabidopsis* and maize genomes. Taken together these pieces of evidence may suggest a horizontal mode of transfer for introduction of these retrotransposons into the papaya genome.

Tandem Repeats

The existence of 277.4-Mb whole-genome shotgun sequences (WGS) of papaya allows an in silico exploration for TRs. Repeat motifs between 1 and 2,000 bp were analyzed and classified according size into micro- (1–6 bp), mini- (7–100 bp) and satellite (>100 bp) tandemly arrayed sequences, as described in Nagarajan et al. (2008). According to this approach, a total of 414,681 class I (≥20 bp) repeats were characterized in 57,360 loci (spanning a total of 4.8 Mbps, representing 1.3 % of the total genome size). The analysis revealed an average repetitive-unit length of 79 bp and a copy number average of 7.23 (ranging from 1.8 to 969.3 copies). The average AT content was 72 %, slightly higher than the average AT content of the genome (65 %). Tandem repeats are randomly distributed in the papaya genome, and there is no correlation between tandem repeat number and gene density. This supports the observation that papaya genome is mostly euchromatic (Ming et al. 2008).

In terms of physical quantity of DNA, microsatellites represent a 0.19 % of the total papaya genome size, minisatellites a 0.68 %, and satellite DNAs a 0.43 %. Following the same approach, Navajas-Pérez and Paterson (2009) found similar abundance of tandem repeats in angiosperms assemblies, 0.19, 0.83, and 0.5 % on average for micro-, mini-, and satellite DNAs, respectively. Punctual quantifications of TRs in other species reveal that these sequences frequently constitute a large portion of the genomes (Lim et al. 2005; Saini et al. 2008). However, although all papaya TR sequences may not be covered, most known repetitive elements are found to be reasonably well represented.

Despite their low percentage, microsatellites represent the class with the highest number of tandem repeat copies in papaya. Dinucleotides are the best represented with ~180,000 units, the most common being (TA/AT)n and (AG/TC)n along with long A/T stretches. TTC/AAG, AAT/TTA trinucleotides and their multimeric variants (with up to 969.3 repeats in a single locus), and pentanucleotides are also common repeats in papaya. Previous reports based on genomic library screenings and mining of repeated DNA databases have demonstrated the same for a great number of plant species (revised in Navajas-Pérez and Paterson 2009).

Longer TRs are normally specific to a related group of species due to their rapid evolutionary change rate (Miklos 1985). Thus, only a small portion of TRs were annotated. Those sequences fell into DNA binding, pseudogenes, or TE-like categories (Table 13.2) (Navajas-Pérez and Paterson 2009). It can be argued that these sequences could be somehow involved in gene regulation/inactivation or evolved through a TE intermediate. These findings agree with the recent tendency to consider repeat DNAs functional, instead of simply junk or parasitic elements.

There is a general tendency in the distribution of repeat-unit sizes in papaya tandem repeats to sequences between 9 and 50 bp, which account for a high number of copies as well as for the maximum number of variants and loci (Nagarajan et al. 2008). This agrees with data on papaya for perfect SSRs from Wang et al. (2008) who found that the 20-bp repeats were the most common repeats in class I, followed by 24-bp repeats with insignificant variance between EST and WGS or BES sequence data and with data from Navajas-Pérez and Paterson (2009) who found the abundance of repeats in the range 9–30 bp in eight WGS of plants from different sources. This might suggest that structural features such as monomer length could play a role in tandem repeat preservation and evolution (Ugarkovic and Plohl 2002).

Perfect SSR Sampling

Due to the reproducibility of their amplifications and the possibility to better detect polymorphisms among individuals, perfect SSRs are preferred for fine-scale mapping, population analysis, and genotyping. It is important to note that the term perfect repeat is used to denote repeats that do not contain insertions, deletions, and/or mismatches with respect to their basic repetitive motif. In this context, an additional mining has been performed in papaya to detect two types of perfect SSRs: class I or

Table 13.2 Mini- and satellite-DNA BLAST hits summary in *Arabidopsis thaliana* TAIR7 release [modified with kind permission of Springer Science+Business Media from Navajas-Pérez and Paterson (2009)]

Annotation	Arabidopsis	Papaya	Poplar	Grapevine	Rice
Unclassified proteins	4,256	63	71	56	50
Transposable elements, viral, and plasmid proteins	1,047	426	14	9	6
Metabolism	283	14	5	16	12
Cell rescue, defense, and virulence	180	1	6	3	4
Classification not yet clear-cut	179	1	3	6	2
Protein synthesis	88	0	2	8	5
Cellular transport, transport facilitation, and transport routes	80	1	1	2	2
Transcription	75	7	5	6	2
Cellular communication/signal transduction mechanism	58	0	2	1	0
Protein fate	51	2	4	2	3
Subcellular localization	50	1	2	0	3
Biogenesis of cellular components	44	2	0	0	0
Cell cycle and DNA processing	13	2	0	1	1
Energy	12	1	1	1	0
Development (systemic)	10	0	0	1	0
Protein with binding function or cofactor requirement (structural or catalytic)	8	0	0	0	2
Cell fate	8	0	0	0	0
Systemic interaction with environment	3	0	0	0	0
Interaction with environment	2	0	0	0	0
Storage protein	1	0	1	1	0
Regulation of metabolism and protein function	1	0	0	0	0
Total	6,449	521	117	113	92

SSRs ≥20 bp and class II, less variable SSRs between 12 and 20 bp. Following this method, a total of 371,710 perfect SSRs were identified in the papaya genome, of which 32,164 (8.7 %) and 339,546 (91.3 %) belonged to class I and class II SSRs, respectively. The density was of one per 8.6 kb for class I and one per 0.8 kb for class II on average. Thus, according to this approach class II SSRs was substantially more abundant than class I SSRs on a genome-wide scale (Wang et al. 2008).

The same procedure was used to scan 51.2-Mb bacterial artificial chromosome (BAC) end sequences (BES) (Ming et al. 2001) and 13.4-Mb expressed sequence tag (EST) sequences (Ming et al. 2008). A total of 49,738 SSRs were identified from BES, including 3,581 (7.2 %) class I and 61,394 (92.8 %) class II SSRs with densities of one per 14.3 and 1.1 kb, respectively, while 10,688 SSRs with 94.2 % class II and 5.8 % class I were gathered from EST sequences.

The highly mutable nature of SSRs makes them potentially powerful markers for analyzing genetic polymorphisms between closely related genotypes. Around 11,000 primer pairs have been developed by different authors (Santos et al. 2003;

Pérez et al. 2006; Eustice et al. 2008; Wang et al. 2008; Ramos et al. 2011) from different sources (BES, EST, and WGS) for the amplification and polymorphism of class I SSRs in papaya. This batch of primers was tested on four selected genomic DNA samples, including the parents of an F2 mapping population, an "AU9" female and "SunUp" hermaphrodite, and two pooled DNA samples containing either ten female or ten hermaphrodite F2 plants, as described in Wang et al. (2008), and contributed to integrate the WGS data with EST and BES sequences to construct a high-density marker map. This complete set of SSR markers throughout the genome will assist diverse genetic studies in papaya and related species. For example, some of these SSR markers have been used to analyze polymorphisms in tropical accessions of papaya and their cross-amplification with *Vasconcellea* species (Pérez et al. 2006), and others have been used for marker-assisted selection in backcross programs (Ramos et al. 2011).

Gene Families

Despite containing fewer genes overall compared to the *Arabidopsis* genome, the papaya genome has several gene families with increased copy number (Ming et al. 2008). These gene families, identified by a gene "tribe" analysis by comparison with *Arabidopsis*, poplar, grape, and rice genes, highlight the role of gene family expansion in papaya tree and fruit development. In particular, compared to *Arabidopsis*, the papaya genome is marked by an increase in certain families of transcription factors (e.g., RWP-RK), resistance genes (NBS-LRR), lignin synthesis genes, starch-associated genes, and those involved in volatile development (Ming et al. 2008).

While the papaya genome lacks signatures of recent genome duplication, a significant fraction of the genes (>2 %, representing 3 % of the papaya genome) are present in a large number of copies (>20; Nagarajan et al. 2008). Many of the most abundant genes are, not surprisingly, similar to those found in TEs (with matches to integrases and polyproteins). However, a number of them also represent non-TE-associated functions including MADS-box transcription factors, zinc-finger proteins, topoisomerases, and serine/threonine phosphatases (Nagarajan et al. 2008) and could be under strong selection in the papaya genome.

Telomeres

Telomeres are highly conserved structures that maintain chromosome integrity by stabilizing chromosome termini. Telomeric DNA is made up of relatively short arrays of a 7-bp long TG-rich sequence added by a telomerase enzyme. This solves the capping and replication issue at the ends of a DNA double helix (Watson and Riha 2010). The first eukaryotic telomere sequence, TTTAGGG, was cloned for

Arabidopsis thaliana, and found to be present in most higher plants, except for plants of order Asparagales that harbor human-type telomere repeat, TTAGGG (de la Herrán et al. 2005), and plants from several genera of the Solanaceae family (Sykorova et al. 2003). There is no obvious correlation between telomere size and phylogenetic relationships, and the length of telomeric DNA widely varies among plant taxa, ranging from 0.3 kb in green algae (Petracek et al. 1990) to 100 kb in tobacco (Fajkus et al. 1995). Telomeres in *A. thaliana* are 2.5 kb long on average (Richards and Ausubel 1988). Papaya telomeres belong to the *Arabidopsis* type (Nagarajan et al. 2008), and their size ranges from 25 kb to well over 50 kb (Shakirov et al. 2008).

Telomere microsatellite-like repeats are separated from the rest of the genomic DNA by a transitional sequence or subtelomere. Subtelomere does not necessarily participate in telomere function but can facilitate meiotic pairing or protect terminal genes against the loss and gain processes at the chromosome ends (Kipling 1995). Subtelomeric or telomere-associated sequences (TAS), in addition to location and the ineffectiveness for sequence homogenization (Contento et al. 2005), have a similar organization in many plants (Ganal et al. 1991), being frequently constituted by species-specific long tandem repeats, transposons, and degenerate variants of (TTTAGGG)n motifs (Riethman et al. 2005; Navajas-Pérez et al. 2009b). Multiple copia- and gypsy-like retrotransposons and different DNA transposons occupy subtelomeric regions in papaya as well as tracts of microsatellite-like repeats including the vertebrate motif in a small copy number (Ming et al. 2008; Nagarajan et al. 2008). In addition, inspection of subtelomeric regions indicated that nine of them share 0.5–1.5 kb of nearly identical DNA sequence immediately adjacent to terminal telomeric repeats (Ming et al. 2008). Also, Navajas-Pérez and Paterson (2009) found some repeats in papaya WGS showing homology with the telomere-like 500 repeat of *A. thaliana*, all of these, typical features of a TAS. Notably, the organization of subtelomeric DNA in papaya contrasts sharply with *Arabidopsis* subtelomeres, which consist of unique sequence on eight out of ten chromosome arms.

Centromeres

In eukaryotes, centromeres are often composed of cytologically distinctive heterochromatin and are associated with long arrays of satellite DNA (Kipling 1995). This highly repetitive nature makes centromeres difficult for sequencing and fine-scale genetic mapping (Navajas-Pérez and Paterson 2009). Notwithstanding the difficulties, centromere-specific repetitive DNA sequences have been isolated yet from several plant species, including *Brassica napus* (Harrison and Heslop-Harrison 1995), *A. thaliana* (Martinez-Zapater et al. 1986), *Oryza sativa* (Wang et al. 1995), or *Sorghum bicolor* (Miller et al. 1998).

Although the papaya genome is largely euchromatic, highly condensed heterochromatin knobs exist on most chromosomes' centromeric and pericentromeric regions representing an estimated 30–35 % of the genomic DNA. Five BACs that

mapped onto centromeric region have been analyzed up to date. Sequence analysis showed that all of these BACs lack known centromere-specific sequences. BACs contained 19.7 % of known repetitive sequences based on a RepeatMasker search, including 115 gypsy retroelements—which are a typical feature of the pericentromeric region of plant chromosomes—2 copia retroelements, 85 simple repeats, 457 low complexity repeats, one DNA transposon, and one small RNA (Yu et al. 2007). However, a large portion of this heterochromatic DNA was probably not covered neither by the draft genome sequence nor in BAC libraries, and its nature remains understudied (Ming et al. 2008).

It is noteworthy that papaya male-specific region of Y chromosome (MSY) mapped close to the centromere of the Y chromosome (Yu et al. 2007). Fine mapping showed that the centromere of the Y chromosome is either directly associated with knob 4 or is immediately adjacent to either side of this knob, a region showing more divergence between X and Y than the rest of the MSY. It may indicate that the first sex-determining gene of papaya was possibly located within the centromeric region where recombination is severely or completely suppressed. Thus, accumulation of genes related to male functions near the centromere would have favored and triggered the establishment and expansion of the MSY region. Natural selection of such genes may result in a selective advantage to recombination suppression between these genes and the sex-determining region on the proto-sex chromosome (Charlesworth et al. 2005).

Sex-Chromosome Repeatome

Early evolved plant sex chromosomes like those from papaya have given rise to many studies in recent years which have proved chromosomal rearrangements and repetitive DNA accumulation crucial events in sex-chromosome evolution (Sola-Campoy et al. 2012). Sex chromosomes are thought to have evolved from a standard autosomal chromosome pair as a consequence of a rarely recombining region containing genes involved in sex determination (Ming et al. 2011). That progressive suppression of recombination is the ultimate consequence of the accumulation of diverse repetitive sequences, such as mobile elements and satellite DNAs, that consequently gives rise to Y-chromosome degeneration. This may further inhibit recombination between X and Y chromosomes and ensure the maintenance of dimorphic sex chromosomes, while conferring them with exceptional evolutionary features.

In papaya, the lack of recombination might have been caused by the proximity of MSY to the centromere as mentioned before together with two large-scale inversions, followed by numerous additional chromosomal rearrangements (Wang et al. 2012). These data come from a recently constructed physical map of the MSY region and its X counterpart by chromosome walking and sequenced bacterial artificial chromosomes—BACs (Wang et al. 2012; Gschwend et al. 2011). Thus, papaya constitutes the first complete sequencing of a plant Y-specific region together with

its X counterpart. This offers a good opportunity to gain insights into structural organization and composition of plant-sex chromosomes. According to this analysis, papaya MSY encompasses 8.2 Mb, more than twice as large as the corresponding 3.5-Mb female region. As predicted by the model of sex-chromosome evolution, the male-specific region expanded by massively accumulation of repeated DNA, representing 83 %, while the corresponding X region included 70 % of such repeats (Na et al. 2012). In any case, both are much higher than the papaya genome-wide average of 56–58 % (Ming et al. 2008; Nagarajan et al. 2008).

A more detailed analysis revealed that among all interspersed repeats in this region, the retroelements are the most significantly accumulated repeats with 64 % in the MSY and 50 % in the corresponding X region. They are the principal responsible of the larger size of the MSY accounting for nearly 99 % of all identifiable interspersed repeats. DNA transposons were also found but in a minor extension, representing only 0.1 % of the MSY region (Table 13.1). 80.2 % of the younger inversion sequence is already repetitive and at least 80.7 % of second inversion too. This would support the predicted early accumulation of transposable elements in the initial stage of sex-chromosome evolution after recombination stops (Charlesworth et al. 2005). As for the unannotated sequences, a total of 36 new repeats were identified. However, only 21 of them—20 from the MSY and 1 from the X—had no match to papaya genome sequences and then were regarded as potentially sex-specific repeats. Interestingly, all these MSY-specific repeats mapped within two regions where the MSY explosion occurred, suggesting their role in the origin of sex chromosomes (Na et al. 2012). Tandem repeats content has been estimated in 3.1 % for the X region and 3.8 % for the MSY (Na et al. 2012).

Finally, it is remarkable that papaya X counterpart also presented higher repetitive content than the genome-wide average. This has been found in other organisms (Bergero et al. 2007). In papaya, could partly be explained considering the pericentromeric location of sex-determining regions of the X and Y chromosomes (Gschwend et al. 2012). Also, due to the lack of recombination between X and Y chromosomes in males and hermaphrodites, the X region would have a lower effective population size than the autosomes and then a reduced efficacy of purifying selection redounding on a higher accumulation of repeat DNA (Wang et al. 2012).

Methods

Sources

The papaya 277.4-Mb WGS from a "SunUp" female plant (Ming et al. 2008), a total of 51.2-Mb BES from a hermaphrodite BAC library (Ming et al. 2001), and 13.4-Mb EST sequences (Ming et al. 2008) were used for repeats mining.

Annotation of TEs

TEs were annotated using RepeatMasker (http://www.repeatmasker.org) and a non-redundant database combining plant repeats from Repbase (Jurka 2003), CPR-DB (ftp://ftp.cbcb.umd.edu/pub/data/CPR-DB), and TIGR (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats). CPR-DB was constructed by applying the de novo methods RepeatScout (Price et al. 2005) and PILLAR (Edgar and Myers 2005) to the complete set of contigs from the papaya genome. Repeat families were annotated using a combination of manual curation (786 repeat families; N. Jiang, personal communication) and BLAST searches against NR and PTREP (http://wheat.pw.usda.gov/ITMI/Repeats).

Tandem Repeats Detection

Tandem repeats were detected by using the Tandem Repeats Finder software (Benson 1999). Repeat units between 1 and 2,000 bp were analyzed, and only repeats arrayed in tandems >25 bp were considered. Repeats were classified into micro- (1–6 bp), mini- (7–100 bp), and satellite (>100 bp) tandemly arrayed sequences. A nonredundant set of sequences was constructed using the program cd-hit-est, as implemented in the package CD-HIT (Li and Godzik 2006), at the 85 % similarity level. For annotations, the nonredundant sequences were BLASTed with the Arabidopsis TAIR 7 release (Poole 2007) and the hits classified according to the MIPS functional catalogue database (http://mips.gsf.de). Perl scripts were written to automate the process. A perl program Simple Sequence Repeat Identification Tool (SSRIT) available at http://www.gramene.org (Temnykh et al. 2001) was used for perfect SSR automated mining according to Wang et al. (2008).

Analysis of High Copy Number Genes

Annotated genes (DNA and protein sequences) in the papaya genome were BLASTed against the whole-genome sequence to find significant matches (*E*-value < 1e−20), and searches against the NR protein database (NCBI, January 2008) were used to find plant homologs.

Data Access and Retrieval

The sequences and annotations in the papaya repeat database are available via FTP downloads at ftp://ftp.cbcb.umd.edu/pub/data/CPR-DB. The sets of novel TE sequence in papaya (annotated and unannotated) are presented as multi-fasta files in

a format convenient for use with RepeatMasker. For tandem repeats, redundant and nonredundant databases as well as a consensus sequence list are available in multi-fasta files. A file including annotations is also provided. High copy number papaya transcripts and protein sequences are also available as annotated multi-fasta files. Further details can be found in the README file accompanying the database. Also a comprehensive information of 11,000 perfect SSR marker surveyed can be found in Santos et al. (2003), Pérez et al. (2006), Eustice et al. (2008), Wang et al. (2008), and Ramos et al. (2011).

References

Arabidopsis Genome Initiative (2001) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

Arkhipova IR (2005) Mobile genetic elements and sexual reproduction. Cytogenet Genome Res 110(1–4):372–382

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot (Lond) 95:127–132

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

Bergero R, Forrest A, Kamau E, Charlesworth D (2007) Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: evidence from new sex-linked genes. Genetics 175:1945–1954

Bousios A, Minga E, Kalitsou N, Pantermali M, Tsaballa A, Darzentas N (2012) MASiVEdb: the sirevirus plant retrotransposon database. BMC Genomics 13(1):158

Camacho JP, Sharbel TF, Beukeboom LW (2000) B-chromosome evolution. Philos Trans R Soc Lond B Biol Sci 355:163–178

Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity 95(2):118–128

Chen C, Yu Q, Hou S, Li Y, Eustice M, Skelton RL, Veatch O, Herdes RE, Diebold L, Saw J, Feng Y, Qian W, Bynum L, Wang L, Moore PH, Paull RE, Alam M, Ming R (2007) Construction of a sequence-tagged high-density genetic map of papaya for comparative structural and evolutionary genomics in Brassicales. Genetics 177(4):2481–2491

Cheng XD, Ling HQ (2006) Non-LTR retrotransposons: LINEs and SINEs in plant genome. Yichuan 28:731–736

Contento A, Heslop-Harrison JS, Schwarzacher T (2005) Diversity of a major repetitive DNA sequence in diploid and polyploid *Triticeae*. Cytogenet Genome Res 109:34–42

de la Herrán R, Cuñado N, Navajas-Pérez R, Santos JL, Ruiz Rejón C, Garrido-Ramos MA, Ruiz Rejón M (2005) The controversial telomeres of lily plants. Cytogenet Genome Res 109(1–3):144–147

Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21(Suppl 1):i152–i158

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5(6):435–445

Eustice M, Yu Q, Lai C, Hou S, Thimmapuram J, Liu L, Alam M, Moore P, Presting G, Ming R (2008) Development and application of microsatellite markers for genomic analysis of papaya. Tree Genet Genomes 4:333–341

Fajkus J, Kovarik A, Kralovics R, Bezdek M (1995) Organization of telomeric and subtelomeric chromatin in the higher plant *Nicotiana tabacum*. Mol Gen Genet 247:633–638

Fedoroff N (2000) Transposons and genome evolution in plants. Proc Natl Acad Sci USA 97(13):7002–7007

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

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44:445–477

Gschwend AR, Yu Q, Moore P, Saski C, Chen C, Wang J, Na JK, Ming R (2011) Construction of papaya male and female BAC libraries and application in physical mapping of the sex chromosomes. J Biomed Biotechnol 2011:929472

Gschwend AR, Yu Q, Tong EJ, Zeng F, Han J, VanBuren R, Aryal R, Charlesworth D, Moore PH, Paterson AH, Ming R (2012) Rapid divergence and expansion of the X chromosome in papaya. Proc Natl Acad Sci USA. 109(34):13716–13721

Harrison GE, Heslop-Harrison JS (1995) Centromeric repetitive DNA sequences in the genus *Brassica*. Theor Appl Genet 90:157–165

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable "minisatellite" regions in human DNA. Nature 314:67–73

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR (2003) An active DNA transposon family in rice. Nature 421(6919):163–167

Jurka J (2003) Repetitive DNA: detection, annotation, and analysis. In: Krawetz SA, Womble DD (eds) Introduction to bioinformatics: a theoretical and practical approach. Humana Press, Totowa

Kipling D (1995) The telomere. Oxford University Press, Oxford

Kubis SE, Schmidt T, Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Ann Bot (Lond) 82:45–55

Li W, Godzik A (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Lim KB, de Jong H, Yang TJ, Park JY, Kwon SJ, Kim JS, Lim MH, Kim JA, Jin M, Jin YM, Kim SH, Lim YP, Bang JW, Kim HI, Park BS (2005) Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*. Mol Cells 19:436–444

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F, Muñoz-Pomer A, Sempere JM, Latorre A, Moya A (2011) The gypsy database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39 (database issue):D70–D74

Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, García-Díaz A, Zhu X, Yung Y, Montoliu L, Glass CK, Rosenfeld MG (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science 317(5835):248–251

Ma H, Moore PH, Liu Z, Kim MS, Yu Q, Fitch MM, Sekioka T, Paterson AH, Ming R (2004) High-density linkage mapping revealed suppression of recombination at the sex determination locus in papaya. Genetics 166(1):419–436

Macas J, Mészáros T, Nouzová M (2002) PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18:28–35

Martienssen R, Irish V (1999) Copying out our ABCs: the role of gene redundancy in interpreting genetic hierarchies. Trends Genet 15(11):435–437

Martinez-Zapater JM, Estelle MA, Somerville CR (1986) A highly repeated DNA sequence in *Arabidopsis thaliana*. Mol Gen Genet 204:417–423

Matsunaga S (2009) Junk DNA promotes sex chromosome evolution. Heredity 102:525–526

McClintock B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA 36:344–355

Meagher TR, Vassiliadis C (2005) Phenotypic impacts of repetitive DNA in flowering plants. New Phytol 168:71–80

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, Wing RA (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA 101:14349–14354

Miklos GL (1985) Localited highly repetitive DNA sequences in vertebrate and invertebrate genomes. In: McIntryre JR (ed) Molecular evolutionary genetics. Plenum, New York

Miller JT, Jackson SA, Nasuda S, Gill BS, Wing RA, Jiang J (1998) Cloning and characterization of a centromere specific DNA element from Sorghum bicolor. Theor Appl Genet 96:832–839

Ming R, Moore PH, Zee F, Abbey CA, Ma H, Paterson AH (2001) Construction and characterization of a papaya BAC library as a foundation for molecular dissection of a tree-fruit genome. Theor Appl Genet 102:892–899

Ming R, Hou S, Feng Y et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996

Ming R, Bendahmane A, Renner SS (2011) Sex chromosomes in land plants. Annu Rev Plant Biol 62:485–514

Na JK, Wang J, Murray JE, Gschwend AR, Zhang W, Yu Q, Navajas-Pérez R, Feltus FA, Chen C, Kubat Z, Moore PH, Jiang J, Paterson AH, Ming R (2012) Construction of physical maps for the sex-specific regions of papaya sex chromosomes. BMC Genomics 13:176

Nagarajan N, Pop M (2009) Parametric complexity of sequence assembly: theory and applications to next generation sequencing. J Comput Biol 16(7):897–908

Nagarajan N, Navajas-Pérez R, Pop M, Alam M, Ming R, Paterson AH, Salzberg SL (2008) Genome-wide analysis of repetitive elements in papaya. Trop Plant Biol 1(3–4):191–201

Navajas-Pérez R (2012) The genus *Rumex*: a plant model to study sex-chromosomes evolution. In: Navajas-Pérez R (ed) New insights on plant sex chromosomes, 1st edn. Nova, Hauppauge

Navajas-Pérez R, Paterson AH (2009) Patterns of tandem repetition in plant whole genome assemblies. Mol Gen Genomics 281:579–590

Navajas-Pérez R, Rubio-Escudero C, Aznarte JL, Ruiz Rejón M, Garrido-Ramos MA (2007) satDNA Analyzer: a computing tool for satellite-DNA evolutionary analysis. Bioinformatics 23(6):767–768

Navajas-Pérez R, Quesada del Bosque ME, Garrido-Ramos MA (2009a) Effect of location, organization, and repeat copy number in satellite-DNA evolution. Mol Gen Genomics 282: 395–406

Navajas-Pérez R, Schwarzacher T, Ruiz Rejón M, Garrido-Ramos MA (2009b) Characterization of RUSI, a telomere-associated satellite-DNA, in the genus *Rumex* (Polygonaceae). Cytogenet Genome Res 124(1):81–89

Novikov A, Smyshlyaev G, Novikova O (2012) Evolutionary history of LTR retrotransposon chromodomains in plants. Int J Plant Genomics 2012:874743

Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23:366–370

Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607

Pace NR (2009) Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev 73(4):565–576

Pérez OJ, Dambier D, Ollitrault P, Coppens DG et al (2006) Microsatellite markers in *Carica papaya* L.: isolation, characterization and transferability to *Vasconcellea* species. Mol Ecol Notes 6:212–217

Petracek ME, Lefebvre PA, Silflow CD, Berman J (1990) *Chlamydomonas* telomere sequences are A+T rich but contain three consecutive G-C base pairs. Proc Natl Acad Sci USA 87: 8222–8226

Poole RL (2007) The TAIR database. Methods Mol Biol 406:179–212

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:351–358

Ramos HC, Pereira MG, Silva FF, Gonçalves LS, Pinto FO, de Souza Filho GA, Pereira TS (2011) Genetic characterization of papaya plants (*Carica papaya* L.) derived from the first backcross generation. Genet Mol Res 10(1):393–403

Ray DA (2007) SINEs of progress: mobile element applications to molecular ecology. Mol Ecol 16(1):19–33

Richards EJ, Ausubel FM (1988) Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. Cell 53:127–136

Riethman H, Ambrosini A, Paul S (2005) Human subtelomere structure and variation. Chromosome Res 13:505–515

Robles F, De La Herrán R, Ludwig A, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA (2004) Evolution of ancient satellite DNAs in sturgeon genomes. Gene 338:133–142

Román AC, González-Rico FJ, Moltó E, Hernando H, Neto A, Vicente-García C, Ballestar E, Gómez-Skarmeta JL, Vavrova-Anderson J, White RJ, Montoliu L, Fernández-Salguero PM (2011) Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. Genome Res 21(3):422–432

Saini N, Shultz J, Lightfoot DA (2008) Re-annotation of the physical map of Glycine max for polyploid-like regions by BAC end sequence driven whole genome shotgun read assembly. BMC Genomics 9:323

Santos SC, Ruggiero C, Silva CLSP, Lemos GM (2003) A microsatellite library for *Carica papaya* L. cv Sunrise Solo. Rev Bras Frutic 25:263–267

Schmidt AL, Anderson LM (2006) Repetitive DNA elements as mediators of genomic change in response to environmental cues. Biol Rev Camb Philos Soc 81(4):531–543

Shakirov EV, Salzberg SL, Alam M, Shippen DE (2008) Analysis of *Carica papaya* telomeres and telomere-associated proteins: insights into the evolution of telomere maintenance in Brassicales. Trop Plant Biol 1(3–4):202–215

Sola-Campoy PJ, de la Herrán R, Ruiz Rejón C, Navajas-Pérez R (2012) Plant sex-chromosomes evolution. In: Navajas-Pérez R (ed) New insights on plant sex chromosomes, 1st edn. Nova, Hauppauge

Sykorova E, Lim KY, Chase MW, Knapp S, Leitch IJ, Leitch AR, Fajkus J (2003) The absence of *Arabidopsis*-type telomeres in *Cestrum* and closely related genera *Vestia* and *Sessea* (*Solanaceae*): first evidence from eudicots. Plant J 34:283–291

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11: 1441–1452

Thomas CA Jr (1971) The genetic organization of chromosomes. Annu Rev Genet 5:237–256

Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365:104–110

Ugarkovic D, Plohl M (2002) Variation in satellite DNA profiles, causes and effects. EMBO J 21:5955–5959

Wang ZX, Kurata N, Saji S, Katayose Y, Minobe Y (1995) A chromosome 5-specific repetitive DNA sequence in rice (*Oryza sativa* L.). Theor Appl Genet 90:907–913

Wang J, Chen C, Na JK, Yu Q, Hou S, Paull RE, Moore PH, Alam M, Ming R (2008) Genome-wide comparative analyses of microsatellites in papaya. Trop Plant Biol 1(3–4):278–292

Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression divergence: a view from plants. Ann N Y Acad Sci 1256(1):1–14

Wang J, Na JK, Yu Q, Gschwend AR, Han J, Zeng F, Aryal R, VanBuren R, Murray JE, Zhang W, Navajas-Pérez R, Feltus FA, Lemke C, Tong EJ, Chen C, Wai CM, Singh R, Wang ML, Min XJ, Alam M, Charlesworth D, Moore PH, Jiang J, Paterson AH, Ming R (2012) Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci USA. 109(34):13710– 13715

Wang J, Na J-K, Ming R (2013) Physical mapping of papaya sex chromosomes. In: Ming R, Moore P (eds) Genetics and genomics of papaya. Springer Science + Business Media, New York

Watson JM, Riha K (2010) Comparative biology of telomeres: where plants stand. FEBS Lett 584(17):3752–3759

Yu Q, Hou S, Hobza R, Feltus FA, Wang X, Jin W, Skelton RL, Blas A, Lemke C, Saw JH, Moore PH, Alam M, Jiang J, Paterson AH, Vyskot B, Ming R (2007) Chromosomal location and gene paucity of the male specific region on papaya Y chromosome. Mol Genet Genomics 278(2):177–185

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