Genomics 97 (2011) 313–320

Contents lists available at ScienceDirect

Genomics

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

Comparative analysis of *Gossypium* and *Vitis* genomes indicates genome duplication specific to the *Gossypium* lineage

Lifeng Lin [a], Haibao Tang [a], Rosana O. Compton [a], Cornelia Lemke [a], Lisa K. Rainville [a], Xiyin Wang [a,b], Junkang Rong [a,1], Mukesh Kumar Rana [a,c], Andrew H. Paterson [a,*]

[a] Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30605, USA
[b] Center for Genomics and Biocomputing, College of Sciences, Hebei Polytechnic University, Tangshan, Hebei, 063000, China
[c] NRC on DNA Fingerprinting, NBPGR, Pusa Campus, New Delhi, 110012, India

ARTICLE INFO

Article history:
Received 17 November 2010
Accepted 15 February 2011
Available online 22 February 2011

Keywords:
Gossypium
Lineage specific genome-wide duplication
Vitis vinifera
Whole genome alignment
Synteny
Collinearity
Gene loss
Gene density

ABSTRACT

Genetic mapping studies have suggested that diploid cotton (*Gossypium*) might be an ancient polyploid. However, further evidence is lacking due to the complexity of the genome and the lack of sequence resources. Here, we used the grape (*Vitis vinifera*) genome as an out-group in two different approaches to further explore evidence regarding ancient genome duplication (WGD) event(s) in the diploid *Gossypium* lineage and its (their) effects: a genome-level alignment analysis and a local-level sequence component analysis. Both studies suggest that at least one round of genome duplication occurred in the *Gossypium* lineage. Also, gene densities in corresponding regions from *Gossypium raimondii*, *V. vinifera*, *Arabidopsis thaliana* and *Carica papaya* genomes are similar, despite the huge difference in their genome sizes and the different number of WGDs each genome has experienced. These observations fit the model that differences in plant genome sizes are largely explained by transposon insertions into heterochromatic regions.

1. Introduction

Whole genome duplication (WGD) events have been more frequent in the lineages of flowering plant species than in most other taxa. With more plant genomes being sequenced and released and the emergence of new tools for genome comparisons, our understanding of the history of genome duplication and its importance in angiosperm evolution is becoming clearer. An ancient genome triplication event is very likely to have been shared by all eudicots [1,2], and different lineages have experienced additional, more recent WGD events [2,3]. For example, *Populus* had one round of tetraploidy in the Salicoid lineage [4] and *Glycine* had two rounds of tetraploidy in the legume lineage [5]. In contrast, *Vitis* and *Carica* have no lineage specific genome duplication events after the common ancestor of all rosids [1,6].

WGD profoundly impacts the genomic landscape in many ways [7]. Synthetic polyploid plants experience abrupt CpG methylation changes after genome doubling [8]. Interchromosomal rearrangements increase after WGD in teleost fish [9]. Duplicated genes created by WGD behave differently from single gene duplications, showing a longer life span before one copy is pseudogenized and/or deleted [10]. In a cross-taxon alignment using a *Gossypium raimondii* (D-genome cotton) physical map [11], more *Gossypium* contigs were aligned to the distantly-related *Vitis vinifera* genome than to the more closely-related *Arabidopsis* genome [11]. It is possible that the two additional WGD events in *Arabidopsis* lineage, along with subsequent gene losses and chromosomal rearrangements, have significantly disrupted the conservation of synteny.

The fact that members of the *Gossypium* genus have a gametic chromosome number of 13 and several related genera have many species with $n=6$ has long hinted that a *Gossypium* ancestor may have experienced a relatively recent WGD [12]. While the history of genome duplication in the *Gossypium* lineage is not yet clear due to the lack of whole genome sequence, classical cytogenetic analysis, *Ks* distributions of duplicated gene pairs, and possible homoeologous relationships among multiple chromosomal segments within the *Gossypium* genome [13–15] all support the hypothesis that *Gossypium* experienced at least one whole genome duplication event since the triplication shared by most if not all eudicots. However, inferred *Gossypium* homology to date is based on genetic mapping, which is dependent on the marker density and might lead to some spurious matchings [14]. Additionally, sequence shuffling between the pericentromeric regions may cause false positives as well [14]. Therefore,

* Corresponding author at: 111 Riverbend Road, Rm 228, Athens, GA 30602-6810, USA. Fax: +1 706 5830160.
E-mail address: paterson@plantbio.uga.edu (A.H. Paterson).
URL: http://www.plantgenome.uga.edu (A.H. Paterson).
[1] Current affiliation: School of Agriculture and Food Sciences, Zhejiang Forestry University, Lin'an, Hangzhou, Zhejiang, 311300, China.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.02.007

although ancient lineage specific WGD in *Gossypium* has been suggested, definitive proof is still lacking.

In sequenced genomes, one common method to search for evidence of ancient WGD is by "all-against-all" dot plot. In this method, ancient homologous genes (or "anchors") are identified using BLAST, with runs of syntenic segments reflected by consecutive strings of homologous genes preserved in a linear pattern parallel to the diagonal or anti-diagonal (the latter indicating segmental inversion). Compared to the *Ks* distribution plot, this approach not only provides structural evidence of ancient duplication events, but also the physical location of the duplicated segment pairs. However, this method is not feasible in species that lack contiguous sequences or information about the relative chromosomal positions of the sequences.

Without whole genome data, local gene loss patterns can also be indicative of the history of WGD [16]. After genome duplication, one homologous gene is thought to be freed from selective pressure, and may adopt new functions (neofunctionalization), share the original gene function with its paralogue (subfunctionalization) or become pseudogenized or removed. Indeed, the majority of duplicated gene copies are lost in just a few million years after polyploidy. If a eudicot genome (such as *G. raimondii*) has experienced WGD with consequent gene loss after its divergence from *V. vinifera*, one would predict that many genes would have been lost from their ancestral locations.

To further our understanding of its evolutionary history, we studied the *Gossypium* genome using two different methods: a whole-genome-level dot plot analysis, and a local-level comparative study of a specific region of *Gossypium–Vitis* synteny on the basis of two sequenced *G. raimondii* BACs with a total base pair of ~184 kb. Both the whole genome dot–plots and local-level sequence comparisons provide new evidence of *Gossypium* lineage-specific genome duplication after the Vitales–Malvales split. Comparison of homologous sequences between the two species also offers new insight into mechanisms of genome size variation.

2. Results

2.1. Gossypium–Vitis whole genome dot plot

Gossypium is a genus consists of 50 allotetraploid species and diploid species. The smallest diploid Gossypium genome, that of *G. raimondii*, has an estimated genome size of around 880 Mb [17]. The construction of a cotton consensus map with 13 chromosomes by merging the most saturated AD tetraploid genetic map (constructed from F2 population of tetraploid cotton *G. hirsutum* and *G. barbadense*) and the D genome (constructed from F2 population of diploid cotton *G. trilobum* and *G. raimondii*) genetic maps, was described in earlier studies [13]. Briefly, there are 333 pairs of loci that were mapped in extensive blocks in different subgenomes. These were used as the basis for merging. The positions of non-sharing markers were interpolated between these common anchor markers based on the relative recombinational distance from the nearest anchor marker. The combined genetic map contains 3016 loci distributed in a reduced number of 13 putative ancestral chromosomes, thus providing a marker density higher than any previously published maps, offering more resolution than using individual maps alone.

In this study, all genetic markers from the *Gossypium* consensus map were compared and plotted against all *Vitis* genes. Among 3016 loci on the cotton consensus genetic map, there were 1865 identified homologies with a total of 3012 genes on the *Vitis* genome. These genes/loci formed 5097 pairs and the positions of these pairs were used in creating the genome-wide dot–plot.

We were able to detect >50 blocks of syntenic regions between *Gossypium* consensus map and *Vitis* chromosomes (Fig. 1). It is clear from the dot plot that there is often more than one region in *Gossypium* that matches the same *Vitis* region. For example, more than half of *Vitis* chromosome 18 matches regions on *Gossypium* consensus chromosome 9 and chromosome 10. Similarly, there are syntenic blocks found between *Vitis* chromosome 3 and *Gossypium* consensus chromosomes 8 and 12, and syntenic blocks found between *Vitis* chromosome 14 and *Gossypium* consensus chromosomes 1 and 6. Across many regions in the *Vitis* genome, two or more blocks of *Gossypium* consensus chromosome fragments are found to be syntenic to the same *Vitis* chromosome region, and we argue that the duplicated *Gossypium* regions are likely derived from a whole genome doubling event not shared with *Vitis*.

The consensus map provided us with improved information about the genome structure in cotton. However, we realize that the syntenic blocks in Fig. 1 appear "fuzzy" because of the uncertainties in the exact order of genetic markers and the construction of consensus map. For example, some areas on the dot plot show a high density of matches, but lack a clear collinear relationship. In places where we could discern significant collinear relationships, there are still fluctuations around the predicted linear order. We should note that the consensus *Gossypium* map was constructed by merging the genetic markers from At-, Dt- and D-genome genetic maps. The interpolation of the positions of unshared markers could be problematic in inferring marker orders on a local scale because both maps are relatively low resolution (ca. 1 cM) and because the genetic/physical distance ratio can fluctuate widely (violating the linear assumption). Nonetheless, 1629 *Vitis* genes and 954 *Gossypium* loci are found in syntenic blocks, among which 263 *Vitis* genes and 314 Gossypium loci are found in blocks that show a 2:1 relationship between *Gossypium* and *Vitis*. These "duplicated blocks" are distributed across many different chromosomes in *Vitis*, which strongly indicates that at least one genome-wide duplication event has occurred in the *Gossypium* lineage since its divergence from *Vitis*.

We further note that the current analysis is feasible because of the high density of markers in our *Gossypium* consensus map. Indeed, we also attempted to detect collinearity using its individual components: the AD tetraploid reference map and the D genome genetic map [13] separately. Although there are isolated cases where homoeologous tetraploid *Gossypium* chromosomes were found to be syntenic to the same *Vitis* chromosome region, they generally fail to reach the same resolution as the analysis with the consensus map. There are many instances where syntenic blocks detected in the plot using the consensus map were missing in the plot using the individual maps due to lack of data points (Supplemental Figure 1).

2.2. Gossypium BAC sequencing and microsynteny detection

We surveyed three BACs from the D-genome *Gossypium* physical map [11] using shotgun sequencing. The BACs selected were GR174O23, GR109E22 and GR163B08, in the order arranged by FPC (Fingerprinted Contigs [11,18]). Two sequence contigs were assembled for GR109E22 with sizes of 30,903 bp (GR109E22contig1) and 78,650 bp (GR109E22contig2) respectively. There is still one sequence gap (<3 kb) between the two contigs but they are ordered and oriented with the mate-pair information from the subclones. The assembled lengths for the other two BACs are: 97,267 bp for GR174O23 and 134,012 bp for GR163B08. Sequence comparison among the three BACs revealed that GR174O23 overlaps with GR109E22contig1, with a merged sequence 104,965 bp long. No overlaps among other BAC sequence fragments were found.

We created putative cotton gene models based on two different methods: *Ab initio* gene predictions were performed using FGENESH, and a similarity-based method was performed by aligning to cotton EST databases (see Methods). A total of 12 genes were identified in GR109E22 contig2 and an additional 12 genes were identified in the combined fragment of GR109E22 contig1 and GR174O23. The BAC GR163B08 has 19 genes identified by FGENESH, but these either failed to show any corresponding EST sequence or are transposon-related