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

Lifeng Lin a, Haibo 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,b,c⁎

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

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, Glycine 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]. Intergenic or intron, transposon insertions into heterochromatic regions.

2. Results and Discussion

2.1. Comparative gene density analysis

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.

© 2011 Elsevier Inc. All rights reserved.

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

Abstract

Genic 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.

© 2011 Elsevier Inc. All rights reserved.

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 rearrange-

ments 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 genetic 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 homeologous 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,
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 interpolatated 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
Significant synteny was found between two regions on *Vitis* chromosome 6 and two *Gossypium* BACs GR174O23 and GR109E22. Putative gene sequences from GR163B08 correspond to genes in multiple scattered positions on the *Vitis* genome, and we failed to detect syntenic relationships to any *Vitis* regions using this BAC.

For the ease of analysis, we divided the collinear relationships found between *G. raimondii* BACs and the *V. vinifera* genome into two regions. Region 1 contains the consensus sequence combining GR174O23, GR109E22 contig1 and part of GR109E22 contig2 that is immediately adjacent to contig1 across the sequencing gap in the BAC. This region contains 8 collinear genes that aligned to the region from 21.8 Mb to 22.3 Mb (Vv6g1599–Vv6g1637) on *Vitis* chromosome 6. Region 2 contains the remaining portion of the GR109E22 contig2, which corresponds to 7.5 Mb to 7.8 Mb (Vv6g0801–Vv6g0829) on *Vitis* chromosome 6, with 9 genes in collinear order (Fig. 2).

Region 1 and region 2 are contiguous on the *Gossypium* genome, but are located on separate arms of chromosome 6 in *Vitis* (Fig. 2). To determine if this rearrangement happened in the *Gossypium* genome or the *Vitis* genome, we looked at the corresponding syntenic regions of *Arabidopsis*. Due to the lineage specific genome-wide duplications and rearrangements in the *Arabidopsis* genomes, we have identified several *Arabidopsis* genomic locations that showed synteny to the *Gossypium* BACs. Nonetheless, region 1 and region 2 are found to be syntenic to different *Arabidopsis* genomic locations that are not adjacent to each other. In addition, the synteny between *Vitis* and *Arabidopsis* in these regions does not break at the same point as it does between region 1 and region 2 in the *Gossypium* genome studied here. We conclude that the genome rearrangement that fuses region1 and region 2 happened in the *Gossypium* lineage.

**2.3. Ks value between syntenic gene pairs**

We further calculated the synonymous substitution rate (Ks) between our predicted *Gossypium* gene models and syntenic orthologs in *Arabidopsis* and *Vitis* (Supplemental Table 1). With median value of
1.8 (of 21 gene pairs) Ks between Gossypium–Arabidopsis orthologs are much higher than Gossypium–Vitis orthologs (median value of 1.4 out of 18 gene pairs). This trend is unexpected since Arabidopsis is phylogenetically closer to cotton (both taxa belong to the Eurosid II clade) than to Vitis. However, there are significant variations in the substitution rate among different angiosperm lineages [19]. Indeed, among the four sequenced rosid genomes studied in Ref. [20], Arabidopsis has the fastest substitution rate while Vitis has the slowest substitution rate, which could explain this unexpected Ks trend. Also, the substitution rate of genes from a small region such as the one studied here might not be representative of the whole genome. A genome level comparative analysis, once a cotton genome sequence is available, will clarify the evolutionary history of these species.

2.4. Extensive loss of homologous genes observed in Gossypium sequenced regions

To investigate gene loss in the Gossypium lineage after its split from Vitis, we constructed a putative ancestral gene order and compared it to the homologous regions in Arabidopsis. By comparing genes conserved in collinear arrangements in all four Arabidopsis homologous regions (resulting from the two doublings in the Arabidopsis lineage), we were able to distinguish genes present in ancestral locations; other genes are likely to be lineage specific. The conservation of genes in ancestral locations is still found in the syntenic regions from “lineage-specific” insertion or deletion events. Genes found in collinear blocks across taxa were inferred to be in putative ancestral locations; other genes are likely to be lineage specific gene insertions or deletions. Fig. 3 shows an example using Region 2.

In Region 1, 19 genes were in putative ancestral locations on the Vitis chromosome, of which 8 are still preserved in Gossypium; in Region 2 (Fig. 3), 9 genes are preserved in Gossypium out of 17 in putative ancestral locations in Vitis. In both cases, roughly half the V. vinifera genes in ancestral locations are still found in the syntenic locations in Gossypium.

We further compared the extent of gene-loss in the Gossypium regions with the corresponding regions in Carica (no WGDs after its divergence from Vitis) and Arabidopsis (two WGDs after its divergence from Vitis). Gene numbers are similar in Carica and Vitis regions, each containing approximately twice the number of genes found in collinear positions in Gossypium (Table 2). In the collinear Arabidopsis regions, the preserved gene number is significantly lower than that of Gossypium (Table 2), closer to ¼ of the genes in putative ancestral locations.

Many genes in the collinear regions of these genomes do not fit into putative ancestral gene positions. These are likely to be lineage specific gene insertions or deletions. In particular, in Arabidopsis Region 2 (Fig. 3), seven consecutive genes show no homology in Vitis, Carica or Gossypium, but are found in a collinear block on Vitis chromosome 13, indicating translocation of a large fragment to this region in the Arabidopsis lineage.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Vitis homologous region of Gossypium sequenced BACs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vitis gene number</td>
<td>GR BAC number</td>
</tr>
<tr>
<td>Region 1</td>
<td>Vv6g1599 GR174O23_GR109E22contig1 64</td>
</tr>
<tr>
<td>Vv6g1600 GR174O23_GR109E22contig1 79</td>
<td></td>
</tr>
<tr>
<td>Vv6g1602 GR174O23_GR109E22contig1 84</td>
<td></td>
</tr>
<tr>
<td>Vv6g1615 GR174O23_GR109E22contig1 89</td>
<td></td>
</tr>
<tr>
<td>Vv6g1617 GR174O23_GR109E22contig1 97</td>
<td></td>
</tr>
<tr>
<td>Vv6g1624 GR174O23_GR109E22contig1 105</td>
<td></td>
</tr>
<tr>
<td>Vv6g1625 GR109E22Contig2 2</td>
<td></td>
</tr>
<tr>
<td>Vv6g1637 GR109E22Contig2 16</td>
<td></td>
</tr>
<tr>
<td>Region 2</td>
<td>Vv6g0801 GR109E22Contig2 24</td>
</tr>
<tr>
<td>Vv6g0802 GR109E22Contig2 29</td>
<td></td>
</tr>
<tr>
<td>Vv6g0806 GR109E22Contig2 34</td>
<td></td>
</tr>
<tr>
<td>Vv6g0814 GR109E22Contig2 43</td>
<td></td>
</tr>
<tr>
<td>Vv6g0817 GR109E22Contig2 49</td>
<td></td>
</tr>
<tr>
<td>Vv6g0819 GR109E22Contig2 49</td>
<td></td>
</tr>
<tr>
<td>Vv6g0823 GR109E22Contig2 54</td>
<td></td>
</tr>
<tr>
<td>Vv6g0826 GR109E22Contig2 58</td>
<td></td>
</tr>
<tr>
<td>Vv6g0829 GR109E22Contig2 77</td>
<td></td>
</tr>
</tbody>
</table>

2.5. The Vitis homologous region spans a larger physical distance than the corresponding regions of Gossypium and Arabidopsis

Although the V. vinifera genome is only about 55% of the size of the G. raimondii [1,17], the syntenic region on Vitis is much larger in physical size than the corresponding Gossypium regions in both cases. Region 1 covers a Vitis genomic region of ~446 kb, and a Gossypium region of 58 kb; region 2 covers a Vitis region of 290 kb and a Gossypium region of 35 kb. In both cases, the Vitis region is 5–10 times as large as the corresponding Gossypium region. Gossypium syntenic regions had physical sizes similar to the Gossypium regions (between 22 and 70 kb for both Regions 1 and 2). The size difference between corresponding regions of Gossypium and Vitis could be caused by either expansion in the Vitis genome or condensation in the Gossypium genome, or very likely, both.

Transposons

We analyzed the distribution of transposable elements (Figs. 4 and 5) using RepBase (http://www.girinst.org/) and default parameters. TEs comprise a larger proportion of the sequence in the V. vinifera homologous regions, at 25% and 17% of Region 1 and 2, than the 13% and 7% in G. raimondii. Both DNA transposons and retroelements comprise a larger portion of the Vitis sequences than the Gossypium sequences. The difference in quantity of transposons explains 30% and 18% of the size differences between the compared regions (Fig. 5A).

Fig. 3. Pattern of Gossypium gene loss in Region 2. Genes that showed collinearity across genomes are represented by filled squares; genes not preserved in collinearity (putative lineage specific insertions) are represented by hollow squares. Out of the 17 putative ancestral genes in Vitis, only 9 are still identifiable in Gossypium.
In addition to the size difference explained by transposable elements, there is still a 3× to 4× difference in the size of the corresponding G. raimondii and V. vinifera sequences (Fig. 4). This variation in physical length of syntenic regions is approximately proportional to the number of genes identified. In region 1, Vitis has 38 genes (446 kb) corresponding to 8 genes in Gossypium (15 kb + 43 kb); in region 2, Vitis has 29 genes (290 kb) corresponding to 9 genes in Gossypium (53 kb). In both cases, the size of the genomic region corresponds to the number of genes identified, i.e. gene densities are relatively constant. By plotting the positions of genes and TEs on these regions from the two genomes (Fig. 5), we found that many “extra” gene sequences in the Vitis regions are indeed retained in ancestral positions, suggesting that they may have been lost in this particular region of Gossypium during diploidization following lineage-specific WGD. This suggests that the missing genes in Gossypium are likely to be present in paralogous (homeologous) regions that have yet to be identified and sequenced.

### Table 2

<table>
<thead>
<tr>
<th>Region 1</th>
<th>Region 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size in Vitis</td>
<td>446 kb</td>
</tr>
<tr>
<td>Number of orthologous Vitis genes</td>
<td>19</td>
</tr>
<tr>
<td>Size of the region in Gossypium</td>
<td>~58 kb</td>
</tr>
<tr>
<td>Number of orthologous Gossypium genes</td>
<td>8</td>
</tr>
<tr>
<td>Number of orthologous Arabidopsis genes</td>
<td>6, 4, 9, 5</td>
</tr>
</tbody>
</table>

### Gene loss

In addition to the size difference explained by transposable elements, there is still a 3× to 4× difference in the size of the corresponding G. raimondii and V. vinifera sequences (Fig. 4). This variation in physical length of syntenic regions is approximately proportional to the number of genes identified. In region 1, Vitis has 38 genes (446 kb) corresponding to 8 genes in Gossypium (15 kb + 43 kb); in region 2, Vitis has 29 genes (290 kb) corresponding to 9 genes in Gossypium (53 kb). In both cases, the size of the genomic region corresponds to the number of genes identified, i.e. gene densities are relatively constant. By plotting the positions of genes and TEs on these regions from the two genomes (Fig. 5), we found that many “extra” gene sequences in the Vitis regions are indeed retained in ancestral positions, suggesting that they may have been lost in this particular region of Gossypium during diploidization following lineage-specific WGD. This suggests that the missing genes in Gossypium are likely to be present in paralogous (homeologous) regions that have yet to be identified and sequenced.

**2.6. A non-syntenic Gossypium BAC is enriched for repetitive DNA**

GR163B08, distal to GR109E22 in the same physical BAC contig (Fig. 2), differs markedly from the other two BACs sequenced. Homology searches in Genbank showed that 8 (out of 19) predicted genes on this BAC are retrotransposon related, and the remaining 11 showed either no significant homology to known proteins, or homology to unknown proteins. No collinearity was detected with the Vitis, Carica or Arabidopsis genomes. A total of 11% of the BAC sequence is made up of transposable elements, but unlike the other two BACs, these are almost exclusively (97%) LTR-retrotransposons. The number of tandem repeats found in this BAC is 3 to 8 times higher than in other two BACs.

GR163B08 is closer to the end of the chromosome than the other sequenced BACs (Lin et al. unpublished) and may be in or near a transition zone from gene rich euchromatin to the sub-telomeric region. Common features of sub-telomeric regions include the enrichment of tandem repeats and large transposable element insertions [21], consistent with the sequence composition of GR163B08.

### 3. Discussion

Earlier genome mapping studies suggested that diploid Gossypium might be an ancient polyploid [13]. In this study, we used two different approaches to further investigate this hypothesis. We first showed whole genome dot-plot analysis using genetic markers in Gossypium against all genes in the sequenced V. vinifera genome. Although a significant improvement over prior studies, the resolution of the dot plot is still constrained by the limited number of informative Gossypium markers. Nonetheless, in many cases one V. vinifera chromosome segment corresponded to at least two Gossypium segments, which strongly suggests at least one round of WGD in the diploid Gossypium lineage. Sequencing of one of the collinear regions revealed genome stratification in Gossypium that fits the expected behavior of duplicated gene loss after WGD events. These findings complement and reinforce earlier published findings using different methods that Gossypium species are ancient polyploids.

Despite the smaller genome size of V. vinifera than G. raimondii, the homologous regions in V. vinifera have been analyzed are much larger than the G. raimondii regions. We argue that although TE insertions do play a role in the size differences, diploidization in the Gossypium genome could explain a larger portion of the size difference. The missing genes in the Gossypium regions studied are likely to be retained in paleo-duplicated fragments elsewhere in the Gossypium genome that we have not sampled. Therefore, although gene loss has caused the Gossypium regions studied to be smaller in size than the corresponding Vitis regions, given the similar gene densities it is likely that the overall genome size is not affected much by gene deletion. Earlier studies in other taxa [22] suggest that the genome is composed of two distinctive components, with genes densely packed in euchromatic regions, and the heterochromatic regions being largely repetitive DNA that explains the majority of genome size differences. Therefore, given the similar gene densities in these genomes, the variation of genome sizes is mostly determined by the size of heterochromatin.

### 3.1. New evidence supporting a history of WGD in Gossypium

Both cytogenetic studies [15] and intragenomic comparisons of genetic marker positions and use of the current gene/marker order to deduce the ancestral gene order [13,23] previously suggested that the Gossypium lineage experienced at least one WGD. Two new lines of evidence further support this hypothesis and indicate that the Gossypium WGD was subsequent to the triplication affecting most if not all dicots.
Compared to earlier analyses using software packages CrimeStat2 and FISH in the detection of ancient duplicated segments [14], whole-genome dot–plots of mapped Gossypium genes herein reveal fewer segments. However, our analysis imposed the additional requirement that genes to be ordered in a collinear manner, reducing the likelihood that corresponding segments are explicable by factors other than genome duplication.

Our study also includes a local-level comparative analysis, starting from a putative ancestral gene order prior to the duplication event, in a “top-down” approach [2]. The local analysis, based on BAC sequences, has the advantage of more detailed view of local regions than studies using coarsely-mapped Gossypium markers alone. The fraction of retained collinear genes in all Gossypium regions studied is less than 50% but is appreciably higher than that of any one Arabidopsis segment (the latter having experienced two WGD events after its divergence from Vitis). However, across the four Arabidopsis segments (Fig. 3), a total of 17 genes are preserved in collinear locations in at least one segment, roughly double the 9 in the single G. raimondii segment.

This pattern of gene loss in the G. raimondii sequenced region could be the result of a) individual single gene translocation events; b) gene loss after segmental duplication; or c) gene loss after genome duplication. The first scenario is unlikely because with a similar divergence time from Vitis, the gene content in Carica homologous region is still very well conserved. Although our comparison here includes only four species, the correlation between gene conservation pattern and number of genome duplications is quite significant. However, it is difficult to differentiate the effect of genome duplication versus segmental duplication using the gene loss pattern alone for this particular region, due to the limited sequences that we sampled in this study. Nevertheless, with the evidence provided in our genome-level dot–plot analysis and other published genome-level comparative mapping results, it is likely that the observed gene loss pattern in our Gossypium BACs is representative of a genome-wide event and the gene losses is incurred by WGD rather than segmental duplications. In other words, it is likely that the relatively small number of ancestral genes found in the Gossypium BACs is explained by the existence of one or more paleo-duplicated segments in the G. raimondii genome that have not yet been sequenced. Such a segment would be expected to retain collinearity to the same V. vinifera region but based on a gene set that is largely complementary to what is found on the sequenced BACs, accounting for the missing half of the inferred ancestral gene content.

Whether there has been one round or two rounds of WGD in Gossypium lineage is yet to be determined. Although our local-level gene-loss patterns resemble the effect of one round of WGD, we need to be cautious of our conclusions here for several reasons: first, when inferring the ancestral gene repertoire, we inevitably miss genes that are lost either in V. vinifera or in all Arabidopsis homologous regions, or both. So the real ancestral gene number may be larger than what we infer, and thus the apparent 2:1 ratio of ancestral gene number to G. raimondii preserved gene number may be not significantly different from 4:1 (indicative of two rounds of WGD). Secondly, although on average, Arabidopsis thaliana homologous regions have fewer duplicated genes preserved than G. raimondii, the number of duplicated genes preserved in Gossypium is not significantly larger than what is found in the best preserved Arabidopsis homologous region (Table 1, bold numbers) (Fig. 3). The sequencing of the whole genome of G. raimondii (in progress) will provide us with a relatively complete list of Gossypium genes and their arrangements, clarifying the history of genome duplications in Gossypium species.

3.2. Effect of genome duplication on genome size

There is no obvious correlation between the number of WGDs a genome has experienced, and the size of its genome. Genomes with history of WGD vary greatly in genome size. For example, sorghum and rice share a similar history of WGD, while the sorghum genome is ~72% larger (740 Mb vs 430 Mb) [24]. The Arabidopsis genome, with a history of one triplication and two duplications, has one of the most compact genomes in higher plants.
The loss of duplicate genes (or “diploidization”) is common after whole genome duplication events. Over long periods of time, the diploidization process seems to restore plant genomes to a relatively stable gene number although changing the relative abundance of some gene functional groups. For the (albeit small) genomic region that we studied here, gene density of homologous regions in genomes with and without WGD is similar, consistent with the notion that genome size variation is mostly caused by transposon accumulations in heterochromatic regions. Comparative studies between rice and sorghum showed that in genomes where sizes of gene space are very similar, heterochromatin alone causes huge genome size differences [22,24]. In the regions of our study, however, fewer transposon insertions were detected in the Gossypium sequences. This might be because the Gossypium BACs selected came from a gene-rich region.

Transposon insertions tend to accumulate in heterochromatic regions such as peri-centromeric or sub-telomeric regions [22,25]. In euchromatic regions, duplicated genes in one paralogous region might be removed along with neighboring sequences after WGD, causing the region to be more compact than unduplicated counterparts in outgroup genomes.

Many studies of genome size evolution focus on the effects of transposable elements, particularly the insertion and deletion patterns ofLTR-retrotransposons [25,26]. The rapid expansion of one or a few transposon families could lead to a huge increase in genome size [27–29]. A burst of transposon activity has been described in synthesized polyploids, and retrotransposons alone can cause genome size doubling even without WGD [30]. Our findings here suggest that regardless of the number of WGDs a genome has experienced, the collective size of gene-rich regions in different genomes do not vary much after extensive gene loss, e.g. the sum of the sizes of four Arabidopsis regions homologous to the Vitis region studied are similar in size to the Vitis region. This suggests that most WGDs have little long-term impact on the huge genome size differences between plant species.

3.3. Advantage of using the V. vinifera genome in whole genome dot-plots

The V. vinifera genome is an excellent reference for efforts to determine numbers of WGDs in eudicots. The Vitis genome has experienced no WGD since the ancient hexaploidy event shared by most if not all eudicots [1]. Its slow evolutionary rate [2] and its stable karyotype [1] may have left it closer to the ancestral gene order than most other eudicot lineages. Vitis is also a good phylogenetic outgroup for comparative analysis of many eudicot species. These attributes are very helpful in elucidating the duplication history of a new genome.

Detecting ancient WGD often requires relatively complete information, i.e. sequences and arrangement of most genes in a genome, in order for collinearity and/or synteny to be discernible after extensive gene loss, single gene duplications and translocations. For Gossypium, which is not yet sequenced and has only ~2000 genes genetically mapped, there are relatively few homologous gene pairs available so far to distinguish paleopolyploidy from background noise [13].

We show that the lack of data points to infer paleopolyploidy by intra-genomic comparison can be partially mitigated by using a consensus genetic map (with markers from homeologous chromosomes in tetraploid cotton interleaved into a consensus order) and comparison to an outgroup genome. This approach has two advantages: 1. the consensus genetic map approximately doubled the number of Gossypium data points available; 2. using an outgroup genome such as Vitis helps to detect “ghost duplication” [31] segments that are not detectable in self-plots due to the loss of one homolog. The dot-plot analysis using Gossypium consensus map and the V. vinifera genome in this study has shown patterns of synteny that are not detected using Gossypium–Gossypium dot-plots. This method could be generalized to the study of other genomes with well-developed genetic maps but lacking whole genome sequence.

4. Materials and methods

4.1. Genetic map and genome sequences

Gossypium genetic map and marker sequence data were retrieved from a previously published map [13]. Gene peptide sequences and position information for Vitis, Carica and Arabidopsis were all downloaded from the Plant Genome Duplication Database (PGDD: http://chibba.agtec.uga.edu/duplication/). Gossypium mRNA sequences were downloaded from PlantGDB (http://www.plantgdb.org/prj/ESTCluster/progress.php).

4.2. Gossypium–Vitis whole genome dot-plot

Gossypium genetic marker sequences were aligned against Vitis genes using BLASTx, with an E-value cut-off of 1e-10. The top 5 best hits were retained in the BLAST results. The dot-plot was generated using a Python script (http://github.com/tanghaibao/quota-alignment/blob/master/scripts/blast_plot.py). ColinearScan [32] was used to detect collinear blocks. The maximum gap allowed within a syntenic block on a Vitis chromosome was set to 1 Mb, and the maximum genetic distance allowed on the consensus map was set to 10 cM.

4.3. BAC sequencing

The BACs are sequenced following a shotgun protocol. Each BAC DNA sample was sheared using a HydroShear (GeneMachines) to ensure random fragmentation. Sheared DNA fragments were repaired using End-it DNA End Repair Kit (Epicenter, Madison, Wisconsin, USA). Fragment sizes of ~4–5 kb were selected on a 1% low melting agarose gel, eluting the DNA from the gel using the Qiagen QIAEX II (Qiagen, Valencia, California, USA) gel extraction system. DNA fragments were then ligated into the PCR-Blunt II-TOPO vector and transformed into DH10B Escherichia coli host cells using an electroporator. The transformed cells were spread onto Q-plates and picked by a Q-bot into 96-well plates. Sequencing was performed on an ABI 3730- XL Sequence Analyzer using BigDye Terminator v3.1 Cycle Sequencing Kit. Chromatographs were assembled using PhredPhrap. Quality of sequence assemblies were checked using Sequencher V.4.1.4.

4.4. Gene and repetitive element identification from BAC sequences

Genes were identified from BAC sequences using FGENESH (http://linux1.softberry.com/berry.phtml). In Gossypium, the species parameter was set to “Dicot plants”; for Vitis the parameter was set to V. vinifera. Repetitive elements were identified using RepBase repeat masking service (http://www.girinst.org/), with species set to A. thaliana.

4.5. Local-level collinearity searches

The Vitis peptide sequences were used to BLAST against the BAC sequences using tBLASTn, with a cutoff value of 1e-20. The BLAST results were manually checked for collinearity. For Arabidopsis–Vitis genomes synteny, multiple collinearity search and alignment was performed using MCScan [20].

4.6. Calculation of synonymous substitutions (Ks)

For homologues inferred from syntenic alignments, we aligned the protein sequences using CLUSTALW [33] and used the protein alignments to guide coding sequence alignments by PAL2NAL [34]. To calculate Ks, we used the Nei–Gojobori method implemented in yn00 program in PAML.
package [35]. Python script is used to pipeline all the calculations and available at (http://github.com/tanghaibao/bio-pipeline/tree/master/synonymous_calculation/).

Supplementary materials related to this article can be found online at 10.1016/j.ygeno.2011.02.007.

References