BMC Plant Biology

RESEARCH ARTICLE   Open Access

Intergenomic gene transfer in diploid and allopolyploid *Gossypium*

Nan Zhao[1], Corrinne E. Grover[2], Zhiwen Chen[1], Jonathan F. Wendel[2] and Jinping Hua[1*]

Abstract

Background: Intergenomic gene transfer (IGT) between nuclear and organellar genomes is a common phenomenon during plant evolution. *Gossypium* is a useful model to evaluate the genomic consequences of IGT for both diploid and polyploid species. Here, we explore IGT among nuclear, mitochondrial, and plastid genomes of four cotton species, including two allopolyploids and their model diploid progenitors (genome donors, *G. arboreum*: A_2 and *G. raimondii*: D_5).

Results: Extensive IGT events exist for both diploid and allotetraploid cotton (*Gossypium*) species, with the nuclear genome being the predominant recipient of transferred DNA followed by the mitochondrial genome. The nuclear genome has integrated 100 times more foreign sequences than the mitochondrial genome has in total length. In the nucleus, the integrated length of chloroplast DNA (cpDNA) was between 1.87 times (in diploids) to nearly four times (in allopolyploids) greater than that of mitochondrial DNA (mtDNA). In the mitochondrion, the length of nuclear DNA (nuDNA) was typically three times than that of cpDNA. *Gossypium* mitochondrial genomes integrated three nuclear retrotransposons and eight chloroplast tRNA genes, and incorporated chloroplast DNA prior to divergence between the diploids and allopolyploid formation. For mitochondrial chloroplast-tRNA genes, there were 2-6 bp conserved microhomologies flanking their insertion sites across distantly related genera, which increased to 10 bp microhomologies for the four cotton species studied. For organellar DNA sequences, there are source hotspots, e.g., the *atp6-trnW* intergenic region in the mitochondrion and the inverted repeat region in the chloroplast. Organellar DNAs in the nucleus were rarely expressed, and at low levels. Surprisingly, there was asymmetry in the survivorship of ancestral insertions following allopolyploidy, with most *numts* (nuclear mitochondrial insertions) decaying or being lost whereas most *nupts* (nuclear plastidial insertions) were retained.

Conclusions: This study characterized and compared intracellular transfer among nuclear and organellar genomes within two cultivated allopolyploids and their ancestral diploid cotton species. A striking asymmetry in the fate of IGTs in allopolyploid cotton was discovered, with *numts* being preferentially lost relative to *nupts*. Our results connect intergenomic gene transfer with allotetraploidy and provide new insight into intracellular genome evolution.

Keywords: Intergenomic gene transfer, Allopolyploidization, *Gossypium*, Mitochondrial genome, Chloroplast genome, *Numt*, *Nupt*

* Correspondence: jinping_hua@cau.edu.cn
[1]Laboratory of Cotton Genetics, Genomics and Breeding /Joint Laboratory for International Cooperation in Crop Molecular Breeding, Ministry of Education / Key Laboratory of Crop Heterosis and Utilization of Ministry of Education, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China
Full list of author information is available at the end of the article

Background

Prokaryotic α-proteobacteria and cyanobacteria are known to be the forerunners of modern eukaryotic mitochondria [1] and chloroplasts [2, 3], as described by the endosymbiont theory. The transformation from endosymbionts to organelles was accompanied by massive DNA transfer among intracellular genomes, or intergenomic gene transfers (IGT). Although the pace of IGT has slowed considerably since eukaryote formation, it remains a common process that is characteristic of nuclear and organellar genome evolution in plants [4]. Among the three types of genomes in a plant cell, there are six possible directions of gene transfer. The most prominent directions of IGT are from either organellar genome into the nuclear genome [2, 4–25], then from the nuclear and plastid genomes into the mitochondrial genomes [2, 26–38]. Interorganellar transfer to the highly compact plastid genome appears to be quite rare [39–43] .

Recent research has revealed that plant mitochondrial genomes frequently integrate DNA from the other two cellular compartments. As the number of sequenced plant mitochondrial genomes has increased [27, 29, 39, 42, 44–61], the extent of integration from both the chloroplast [28, 30, 32, 62, 63] and nuclear genomes [33, 60, 62] has become more apparent. In general, plant mitochondrial genomes have between 0.56% (*Marchantia polymorpha*) – 10.85% (*Phoenix dactylifera*) plastid-derived sequences [9, 10, 34]. Nuclear sequence integration tends to be more abundant and more difficult to identify, as these commonly include retrotransposon and other repetitive fragments [28, 31, 33, 55].

Nuclear integrants derived from the mitochondrial and plastic genomes are termed *numts* [19] and *nupts* [4], respectively, with these collectively referred to as norgDNAs [14], or nuclear organellar DNAs. Environmental stresses have been shown to increase entry of organellar DNA into the nucleus in plants [8], with insertions commonly occurring in open chromatin regions [12]. While norgDNAs are commonly thought to be inactive, there is some evidence of norgDNA transcription in plant species, including in rice [11] and cotton [64].

While IGT has relevance to broad questions of genome evolution, IGT takes on additional importance due to its possible relationship to plant fertility [65–68]. Repeated IGT transfers can create regions of homology within the mitochondrial genome, which can provide hotspots for intra-organellar recombination. Mitogenomic recombination is a common phenomenon that may generate novel chimeric sequences [69–72]. These novel chimeric sequences may co-transcribe with adjacent functional genes [73–75], subsequently affecting or interfering with mitochondrial electron transfer chain pathways [76, 77]. Furthermore, the phenomenon known as cytoplasmic male sterility is influenced both by the nuclear and mitochondrial compartments, as well as by their interactions [65]. Thus, improving our understanding of IGT may help inform breeding strategies that utilize plant fertility differences, e.g., in male sterility to develop hybrids.

The genus *Gossypium* contains approximately 50 species [78–80], including four that have been domesticated and presently are cultivated for their seed trichomes, or cotton fiber. Two of these domesticated species belong to the clade of seven extant allopolyploid species (AD genome), which formed about 1–2 million years ago (Mya) when an A-genome diploid (resembling modern *G. arboreum*, A_2 or *G. herbaceum,* A_1) hybridized with a D-genome species (resembling modern *G. raimondii*, D_5) and subsequently doubled in chromosome number [79, 81]. High-quality nuclear genome assemblies have recently become available for the diploids *G. raimondii* [64, 82] and *G. arboreum* [83], and for the allopolyploids *G. hirsutum* [84–86] and *G. barbadense* [86–88]. There also exist multiple organellar genome sequences for both diploid and allopolyploid cotton species [60, 61, 63, 65, 89–93]. These genomic data provide the foundation for discovery and description of IGT events in *Gossypium*.
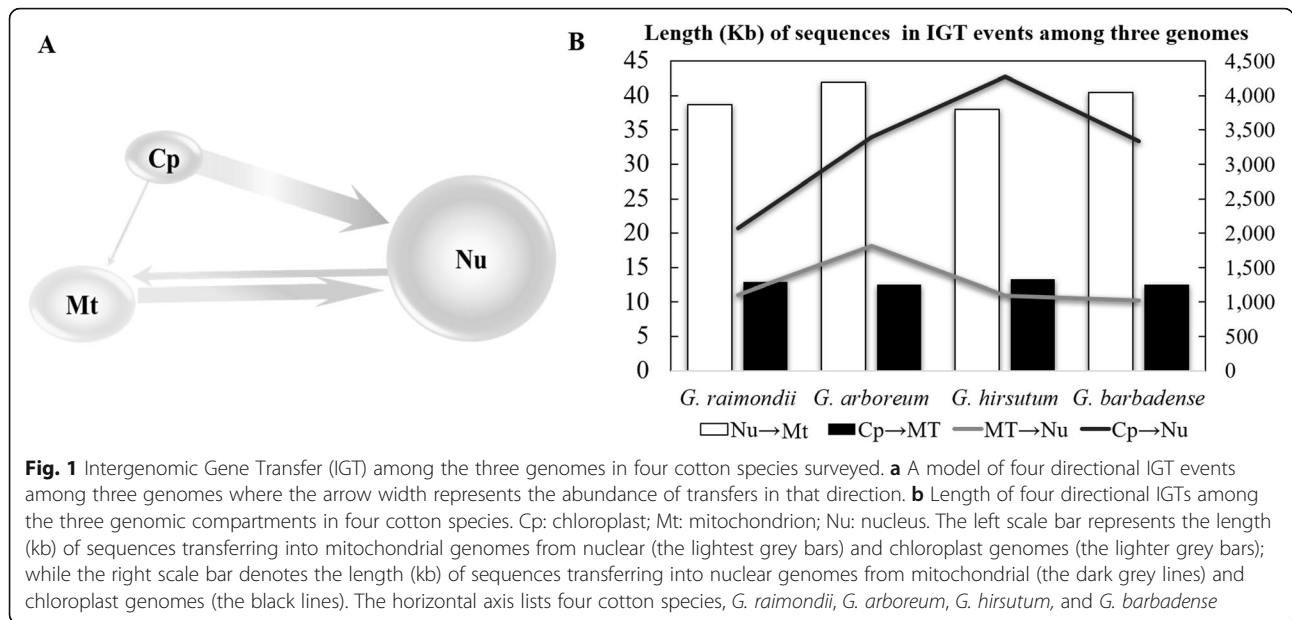
Here we analyze intracellular transfer among nuclear and organellar genomes within four cotton species, including the two cultivated allopolyploids (AD genome) and models of their ancestral diploid (A, D) genome donors, to explore the prevalence of IGT in cotton (*Gossypium*). We characterize and compare the frequencies of the six possible classes of IGT events, as well as the sources and sizes of the inter-organellar sequences. We report a striking asymmetry in the fate of IGTs in allopolyploid cotton, with *numts* being preferentially lost relative to *nupts*. Finally, we explored the expression of norgDNAs.

Results

General profiles of intergenomic gene transfer in *Gossypium*

We screened the nuclear, mitochondrial, and chloroplast genomes of four cotton species, two polyploids and their model diploid progenitors (Additional file 1), for evidence of IGT events. As expected, most of the detected IGTs involve four of the six possible classes of IGT events: nucleus-to-mitochondrion, chloroplast-to-mitochondrion, mitochondrion-to-nucleus, and chloroplast-to-nucleus (Fig. 1a). We did not detect nuclear or mitochondrial insertions into the chloroplast in any of four cotton species.

In the four cotton species surveyed, the length of the integrated fragments was typically 100 times longer for nuclear versus mitochondrial integrants. The size of nuclear integrants ranged from 1028 kb to 4276 kb, whereas mitochondrial integrants varied in size from only 13 kb to 42 kb (Fig. 1b, Table 1). Nuclear insertions

Fig. 1 Intergenomic Gene Transfer (IGT) among the three genomes in four cotton species surveyed. **a** A model of four directional IGT events among three genomes where the arrow width represents the abundance of transfers in that direction. **b** Length of four directional IGTs among the three genomic compartments in four cotton species. Cp: chloroplast; Mt: mitochondrion; Nu: nucleus. The left scale bar represents the length (kb) of sequences transferring into mitochondrial genomes from nuclear (the lightest grey bars) and chloroplast genomes (the lighter grey bars); while the right scale bar denotes the length (kb) of sequences transferring into nuclear genomes from mitochondrial (the dark grey lines) and chloroplast genomes (the black lines). The horizontal axis lists four cotton species, *G. raimondii*, *G. arboreum*, *G. hirsutum,* and *G. barbadense*

into the mitochondrion were typically three times the length of chloroplast insertions (40 kb vs 13 kb; Table 1), and chloroplast insertions into the nucleus (*nupts*) were two to four times greater than mitochondrial insertions (*numts*). With the exception of *G. arboreum*, *numt* length was approximately equivalent among species (Fig. 1b), whereas *nupt* length varied about twofold (from 2072 kb in *G. raimondii* to 4276 kb in *G. hirsutum*; Table 1). Interestingly, the ratio of *nupt* to *numt* was 1.87 for both diploid cotton species and nearly double that (average 3.58) for the two allotetraploid cotton species (Table 1).
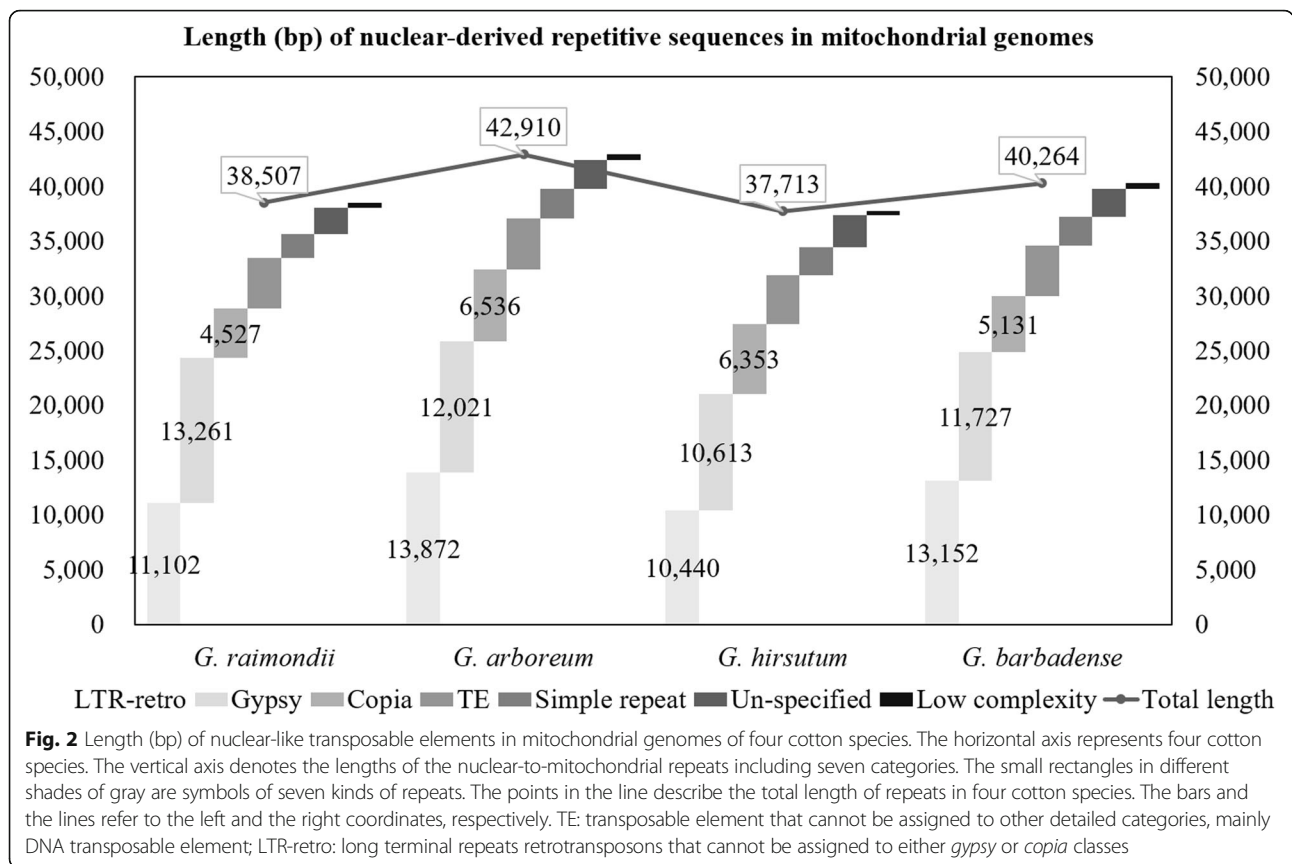
Mitochondrial integrations are variable for nuclear repeats and conserved for chloroplast tRNA genes

Repetitive sequences are common to the mitogenomes of various seed plants, including the monocot *Oryza sativa* [55] and the eudicots *Arabidopsis thaliana* [33], *Cucumis melo* [31], *Cucumis sativus* [28] and *Gossypium* species [60, 91]. Long terminal repeat (LTR) retrotransposons (LTR-retro) typically comprise the biggest component of plant nuclear repeats [94, 95], and they often are a predominant influence on nuclear and mitochondrial genome

size [96–99]. Here, we identified nuclear-derived repeats in the mitochondrial genomes of all four cotton species. The total length of nuclear-derived repeats (for all repetitive classes) ranged from 37.7 kb in the *G. hirsutum* mitochondrial genome to over 42.9 kb in *G. arboreum* (Fig. 2), suggesting that nuclear-derived repeats comprise between 5.64 – 6.24% of a given cotton mitochondrial genome. These repeats were partitioned into seven classes: *copia*, *gypsy*, low complexity, unclassified long terminal repeat retrotransposon (LTR-retro), simple repeat, transposable element (TE) and unspecified (Fig. 2). Here, *copia* and *gypsy* represent a further partitioning of the general LTR-retro (Class 1) transposable elements into their two main classes [100]. The rank of each class (based on element abundance) was nearly identical for each species, with unclassified LTR-retro and *gypsy* elements contributing the most sequence in all four species (Fig. 2); however, the total sequence length for each class varied somewhat among species, leading to the approximately 5 kb difference in total nuclear-derived repeat length noted above. Neither the total amount of repetitive sequence nor the individual length of sequence per repeat class exhibited evidence of bias with respect to ploidy.

Table 1 The total length of IGT events among three genomes in four cotton species

Species	Mitochondrial integrants			Nuclear integrants		
	Nu→Mt (kb)	Cp → Mt (kb)	Nu→Mt/Cp → Mt	Mt → Nu (kb)	Cp → Nu (kb)	Cp → Nu/Mt → Nu
G. raimondii	38.74	12.89	3.01	1106.78	2072.12	1.87
G. arboreum	41.94	12.54	3.34	1813.38	3391.15	1.87
G. hirsutum	37.99	13.36	2.84	1092.08	4275.84	3.92
G. barbadense	40.50	12.54	3.23	1028.21	3331.78	3.24
Average	39.79	12.83	3.11	1260.11	3267.72	2.72

Fig. 2 Length (bp) of nuclear-like transposable elements in mitochondrial genomes of four cotton species. The horizontal axis represents four cotton species. The vertical axis denotes the lengths of the nuclear-to-mitochondrial repeats including seven categories. The small rectangles in different shades of gray are symbols of seven kinds of repeats. The points in the line describe the total length of repeats in four cotton species. The bars and the lines refer to the left and the right coordinates, respectively. TE: transposable element that cannot be assigned to other detailed categories, mainly DNA transposable element; LTR-retro: long terminal repeats retrotransposons that cannot be assigned to either *gypsy* or *copia* classes

In addition to incorporating nuclear repeats, the presence of chloroplast-like tRNA genes in plant mitogenomes has been described [27, 34, 65]. In our previous research, we identified eight chloroplast-derived tRNA genes (*trnD*, *trnH*, *trnM*, *trnN*, *trnP*, *trnS*, *trnV* and *trnW*) in the mitogenomes of all four cotton species [65]. The presence of these genes in species dispersed on the cotton phylogeny likely indicates that their transfer occurred in a common ancestor and that they have been preserved. To evaluate this suggested history, we analyzed nucleotides flanking insertion sites using two chloroplast-derived tRNA genes (*trnH* and *trnD*) as examples. For both genes, we found that the 10 bp upstream and downstream of the insertion were shared among the four cotton species (Fig. 3). When we broadened the number of plant species to include diverse angiosperms, we discovered shared 2-6 bp microhomologies flanking the insertion sites of these two genes (Fig. 3).

Additionally, we evaluated the contribution of both nuclear and chloroplast derived sequences to overall mitochondrial genome size for 26 land plants. Although both nuclear and chloroplast sequences have contributed to mitogenome expansion, the total length of nuclear-like sequences in mitogenomes is more strongly correlated with mitogenome size variation ($R^2 = 0.77$) than is the length of chloroplast sequences ($R^2 = 0.36$)

(Additional file 2A, B). This is partly due to variation in the total amount of repetitive sequence inserted into the mitogenome; however, this correlation is weak ($R^2 = 0.13$) (Additional file 2C), as is the correlation between total repeat length and total nuclear/chloroplast length ($R^2 = 0.23$ and 0.0048; Additional file 2D, E).

Nuclear insertion of mitochondrial DNAs (*numts*) and chloroplast DNAs (*nupts*)

We evaluated the presence-absence pattern of 42 common mitochondrial protein-coding genes for the four cotton species studied, in both the mitochondrion and nucleus, to provide insight into nuclear-mitochondrial coevolution. As expected, most mitochondrial genes were still present in mitogenomes. Only six ribosomal subunit genes (*rpl6*, *rps1*, *rps2*, *rps11*, *rps13* and *rps19*) were absent in all four cotton species (Fig. 4, yellow cells), which likely represent shared loss at some point in their evolutionary history. For mitochondrial genes encoding both complex II (succinate dehydrogenase, *sdh* genes) and ribosomal subunits (*rpl* and *rps* genes) *Gossypium* retains more genes (e.g., *sdh3* and *sdh4*, *rpl10* and *rps10*) in its mitochondrial genome. All remaining mitochondrial genes (except *nad7*) experienced full- or partial-length transfer to the nucleus in at least one cotton species, (Fig. 4, white cells).

Zhao *et al. BMC Plant Biology*   (2019) 19:492

Page 5 of 18

Fig. 3 Flanking nucleotide analysis of chloroplast-derived genes, *trnH* (**a**) and *trnD* (**b**), in mitogenomes. cpDNA and mtDNA are the abbreviations of chloroplast DNA and mitochondrial DNA. The light-dark-light gray strip represents the fusion sequence of mtDNA-cpDNA-mtDNA. The capital English letters in the long rectangular boxes, such as "GG" and "AT", close to cpDNA show the nucleotides of micro-homologies among the different species. Dotted rectangles enclose flanking sequences for the four cotton species studied here

Within *Gossypium*, the number of *numts* in diploids is two to three times that found in tetraploids (33 in *Gossypium raimondii* and 23 in *G. arboreum*, versus 11 in *Gossypium hirsutum* and 13 in *G. barbadense*; Fig. 4). Many of the diploid *numts* (23) are shared between the two diploids, *G. raimondii* (D_5) and *G. arboreum* (A_2), which indicates that these *numts* were incorporated into the nuclear genome prior to the divergence of the A and D clades at the base of *Gossypium*. Subsequently, the lineage leading to *G. raimondii* acquired a further 10 *numts* (*nad5, atp9, ccmB, ccmC, rps3, rps7, nad6, cox2, sdh3* and *rpl2*). In contrast, the two tetraploids (*G. hirsutum* and *G. barbadense*) suffered massive, differential *numt* decay after allotetraploidization. Only three of the shared diploid *numts* (*nad9, ccmFC,* and *rps10*) and one D_5-specific *numt* (*nad6*) are present in both *G. hirsutum* and *G. barbadense*. *Gossypium hirsutum* contains an additional six shared diploid *numt* genes (*nad4, sdh4, cob, cox1, atp1* and *rps14*) and one D_5-specific *numt* gene (*cox2*). On the other hand, *G. barbadense* contains five different diploid-common *numts* (*nad4L, cox3, atp4, atp6* and *mttB*), two different D_5-specific *numts* (*sdh3* and *rpl2*), and two AD_2-specific *numts* (*nad3* and *rps12*).

Interestingly, most *numts* remain intact (full length) after insertion (Fig. 4, red cells), with few that degraded to truncated pseudogenes (Fig. 4, green cells). In a few cases (i.e., *cox2, rpl2, rpl16, rps10* and *rps14*), there is evidence of pseudogenes in diploids and polyploids. The remaining 10 pseudogenes (*nad9, sdh4, cob, cox1, cox3,*

atp8, ccmFC, rpl5, rps4, and *matR*) were intact in diploids but truncated in polyploids. Additionally, four *numt* pseudogenes (*nad5, atp9, ccmB,* and *rps7*) occurred only in *G. raimondii* (Fig. 4, green cells).

We evaluated each genome for nuclear insertions of all 78 chloroplast genes (*nupts*) in the four cotton species. None of the 78 chloroplast genes were lost from the plastids (Fig. 5, yellow cells), and almost all of the 78 genes experienced transfer to nucleus in at least one of the four cotton species. Only six *nupts* (*psaA, ycf3, psbC, ndhB, ndhD, ndhF, rpoC2, rps12* and *matK*) in *G. raimondii* (D_5) and one *nupt* (*rps16*) in *G. arboreum* (A_2) were absent (Fig. 5, white cells).

In contrast to the greater abundance of *numts* in *G. raimondii*, *nupts* experienced a general decay in *G. raimondii* relative to the remaining cotton species (Fig. 5). Nearly all of the 78 *nupts* are present in *G. arboreum*, *G. hirsutum*, and *G. barbadense* (except for *rps16 nupt* loss in *G. arboreum*) as full or partial *nupts*; 69 *nupts* are common to the four cotton species (Fig. 5). In general, *G. arboreum* (A_2) retains the most *nupts* (77 out of 78), whereas there was considerable *nupt* degradation (39) or removal (9) from the *G. raimondii* genome. As in *G. arboreum*, *nupts* in *G. barbadense* are typically retained as intact (75/78) whereas in the other polyploid species, *G. hirsutum,* there has been a modest amount of degradation (12/78 are degraded; Fig. 5). One *nupt* (*rps16*) that is absent in *G. arboreum* is a pseudogene in *G. raimondii* and *G. hirsutum*, only presenting as intact in *G.*
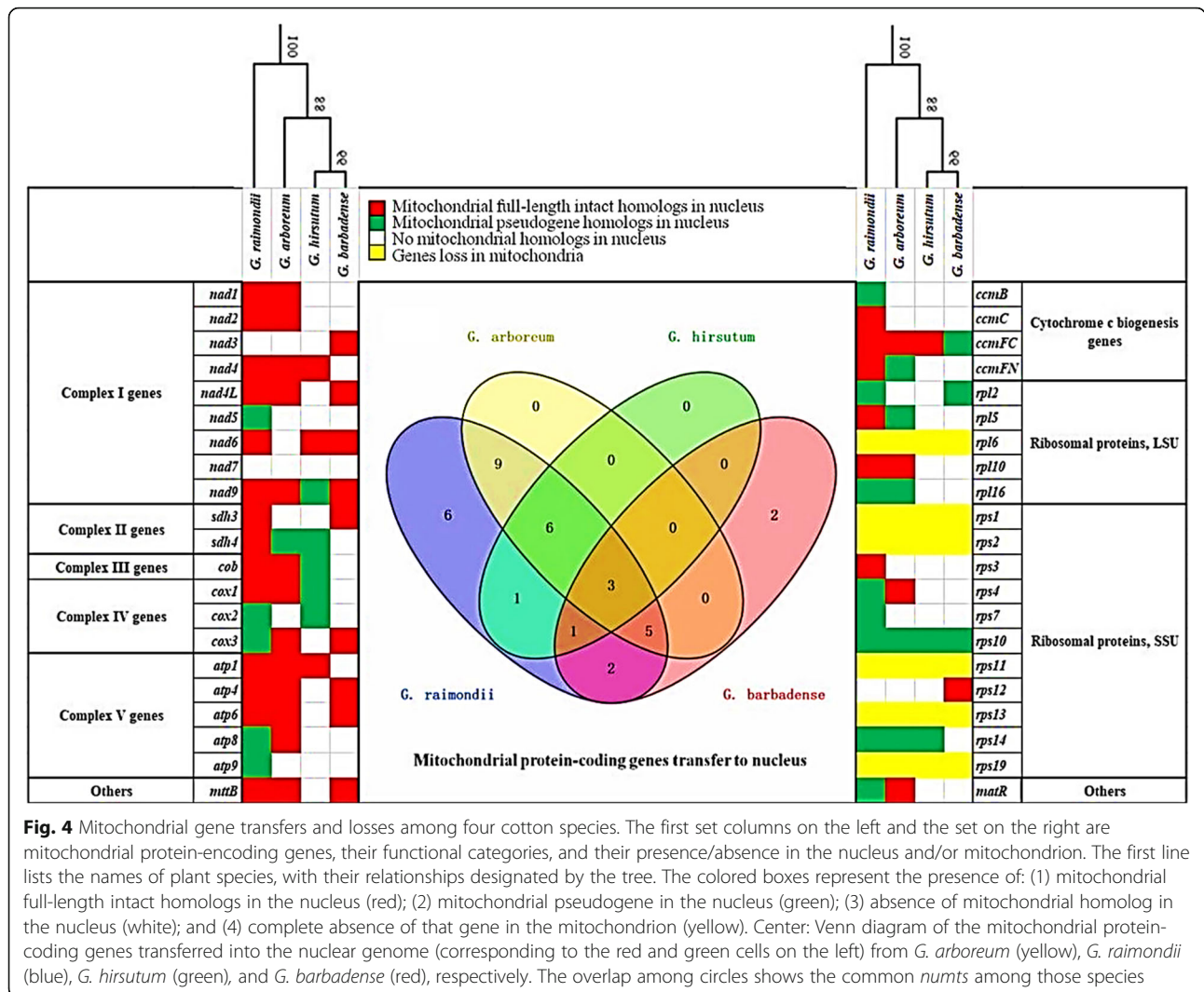
Fig. 4 Mitochondrial gene transfers and losses among four cotton species. The first set columns on the left and the set on the right are mitochondrial protein-encoding genes, their functional categories, and their presence/absence in the nucleus and/or mitochondrion. The first line lists the names of plant species, with their relationships designated by the tree. The colored boxes represent the presence of: (1) mitochondrial full-length intact homologs in the nucleus (red); (2) mitochondrial pseudogene in the nucleus (green); (3) absence of mitochondrial homolog in the nucleus (white); and (4) complete absence of that gene in the mitochondrion (yellow). Center: Venn diagram of the mitochondrial protein-coding genes transferred into the nuclear genome (corresponding to the red and green cells on the left) from *G. arboreum* (yellow), *G. raimondii* (blue), *G. hirsutum* (green), and *G. barbadense* (red), respectively. The overlap among circles shows the common *numts* among those species

barbadense. This suggests that this *rps16 nupt* was transferred to the nucleus sometime during the evolution of *G. raimondii* after divergence from the *G. arboreum* lineage, and that it was subsequently retained in polyploid *Gossypium* until it experienced decay in the lineage leading to *G. hirsutum*. In *G. barbadense*, the three *nupts* (out of 78) that have experienced degradation (i.e., *atpF*, *rpoB*, and *ycf1*) all experienced IGT in the common ancestor of the diploid species (*G. raimondii* and *G. arboreum*), and experienced differential degradation in *G. raimondii* and both allopolyploids (except the intact *rpoB* in *G. hirsutum*). Degradation of *nupts* was more prominent in *G. hirsutum*, where 15% of *nupts* were degraded (12/78; *psaB*, *psbC*, *petA*, *atpF*, *atpI*, *rpoC2*, *rps16*, *matK*, *accD*, *accmA*, *ycf1* and *ycf2*). Almost all of these (nine of 12) are also pseudogenes in *G. raimondii*, while chloroplast *psbC*, *rpoC2* and *matK* were absent in the nucleus. Overall, *G. raimondii* experienced the most degradation of *nupts*, where over half (39 out of 78) were decayed and

nine were not present (i.e., *psaA*, *ycf3*, *psbC*, *ndhB*, *ndhD*, *ndhF*, *rpoC2*, *rps12* and *matK*).

Interestingly, not all regions of the chloroplast and mitochondrial genomes transferred with equal frequency. A hotspot of mitochondrial source material was located in the *atp6-trnW* intergenic region (Fig. 6), and in the chloroplast genome there were three hotspots, i.e., the large single copy region (LSC), inverted repeat region (IR), and small single copy region (SSC), these transferred at a relative rate of 1: 2: 1 (Fig. 7).

Congruent with most angiosperm species, the majority of norgDNAs in cotton are small to medium in size (100 bp – 5 kb) (Additional files 3 and 4), and their distribution patterns vary among species (Additional file 3). Two previously noted complete mitochondrial genome transfers were found on Chr01 of *G. raimondii* and ChrA03 of *G. hirsutum* [64, 93], similar to large-scale norgDNAs found in other plants, e.g., the *numt* on Chr2 of *A. thaliana* [61, 101] and the *nupt* inserted into one scaffold of *S. bicolor* (Additional file 3).
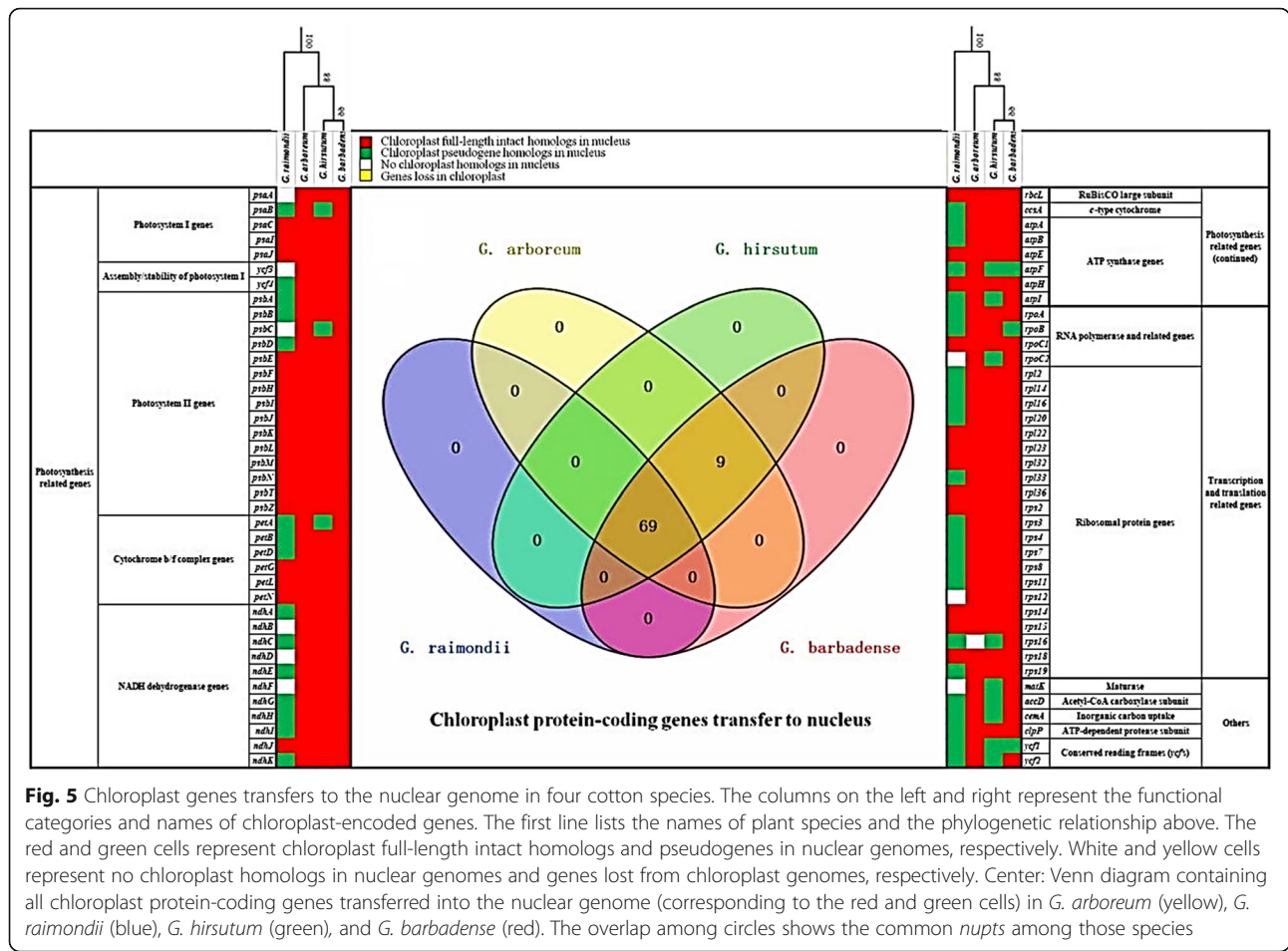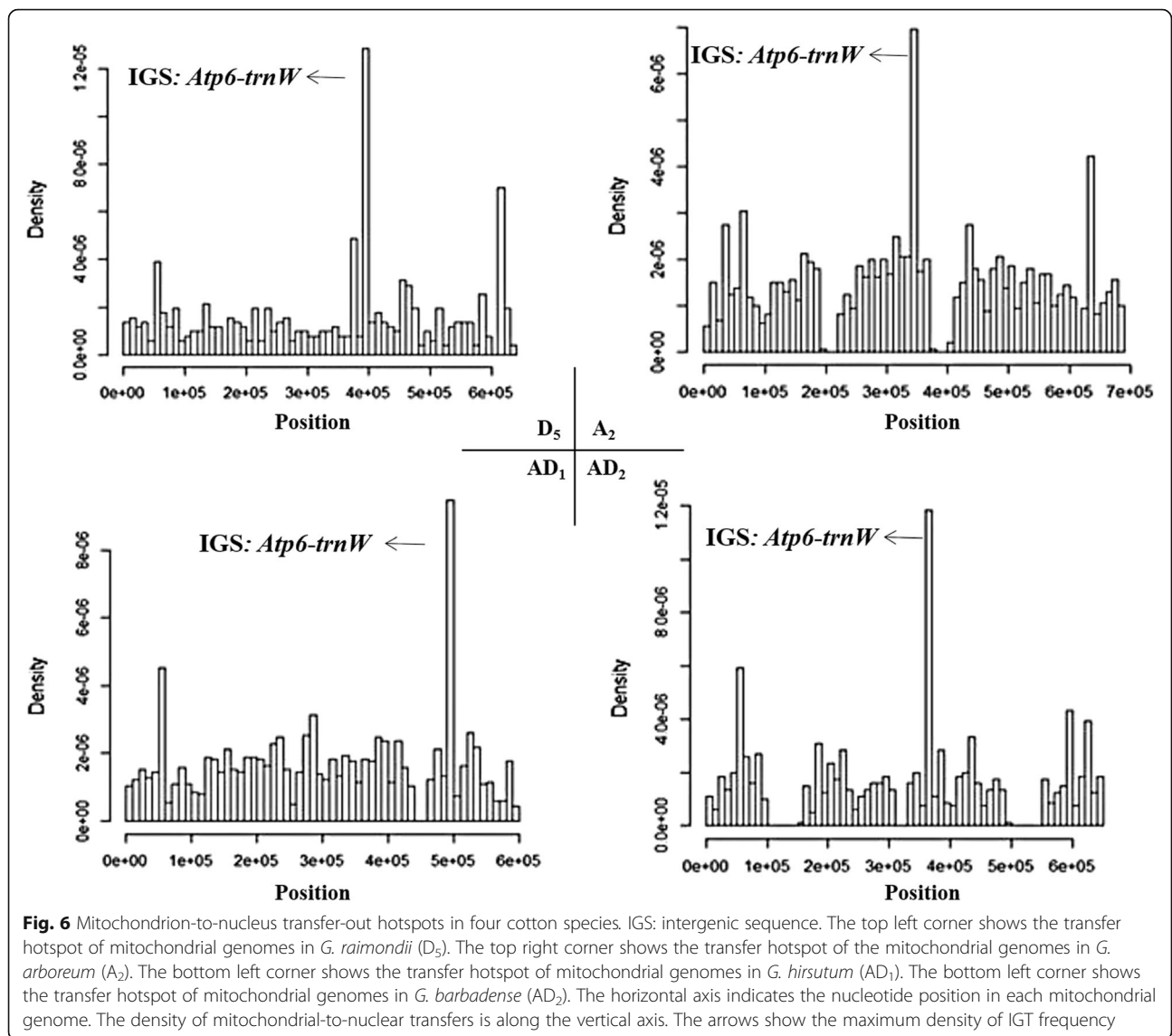
Fig. 5 Chloroplast genes transfers to the nuclear genome in four cotton species. The columns on the left and right represent the functional categories and names of chloroplast-encoded genes. The first line lists the names of plant species and the phylogenetic relationship above. The red and green cells represent chloroplast full-length intact homologs and pseudogenes in nuclear genomes, respectively. White and yellow cells represent no chloroplast homologs in nuclear genomes and genes lost from chloroplast genomes, respectively. Center: Venn diagram containing all chloroplast protein-coding genes transferred into the nuclear genome (corresponding to the red and green cells) in *G. arboreum* (yellow), *G. raimondii* (blue), *G. hirsutum* (green), and *G. barbadense* (red). The overlap among circles shows the common *nupts* among those species

The number of reverse matches (the direction of the sequence intervals along the genome coordinates in donor and receptor genomes are reverse, the start-to-end position of one genome is from small number to large number, while the other genome from large to small) is no fewer than that of the positive matches (both positions in donor and receptor genomes are from small to large, or from large to small). Two kinds of matches coexist in the same chromosome in most studied species. The *nupts* of six chromosomes (chr01, chr03, chr05, chr06, chr08 and chr09) are all reverse matches (blue dots or fragments) and the other seven chromosomes positive matches (red dots or fragments) in *G. raimondii*. Third, norgDNAs in some species frequently transfer into certain chromosomes, but such nuclear hotspots of integration vary from species to species, like *numts* on Chr2 of *A. thaliana*, Chr01 of *G. raimondii*, and ChrA03 of *G. hirsutum*. Fourth, the transfer of norgDNAs is phylogenetically sporadic, as there often is little *norg* similarity between closely related species.

We examined the relationship between IGT into the nucleus and genome size, comparing this to the effects of repetitive content on genome size variation. While nuclear repeat size is most strongly correlated with genome size ($R^2 = 0.9813$; Additional file 5C and Table 2), norgDNA length was moderately correlated with nuclear genome size ($R^2 = 0.5917$ and $R^2 = 0.4675$ for *numts* and *nupts*, respectively; Additional file 5A and B). Additionally, the length of nuclear repeats and norgDNAs was also moderately correlated ($R^2 = 0.6321$ for *numts* and $R^2 = 0.4511$ for *nupts*; Additional file 5D and E).

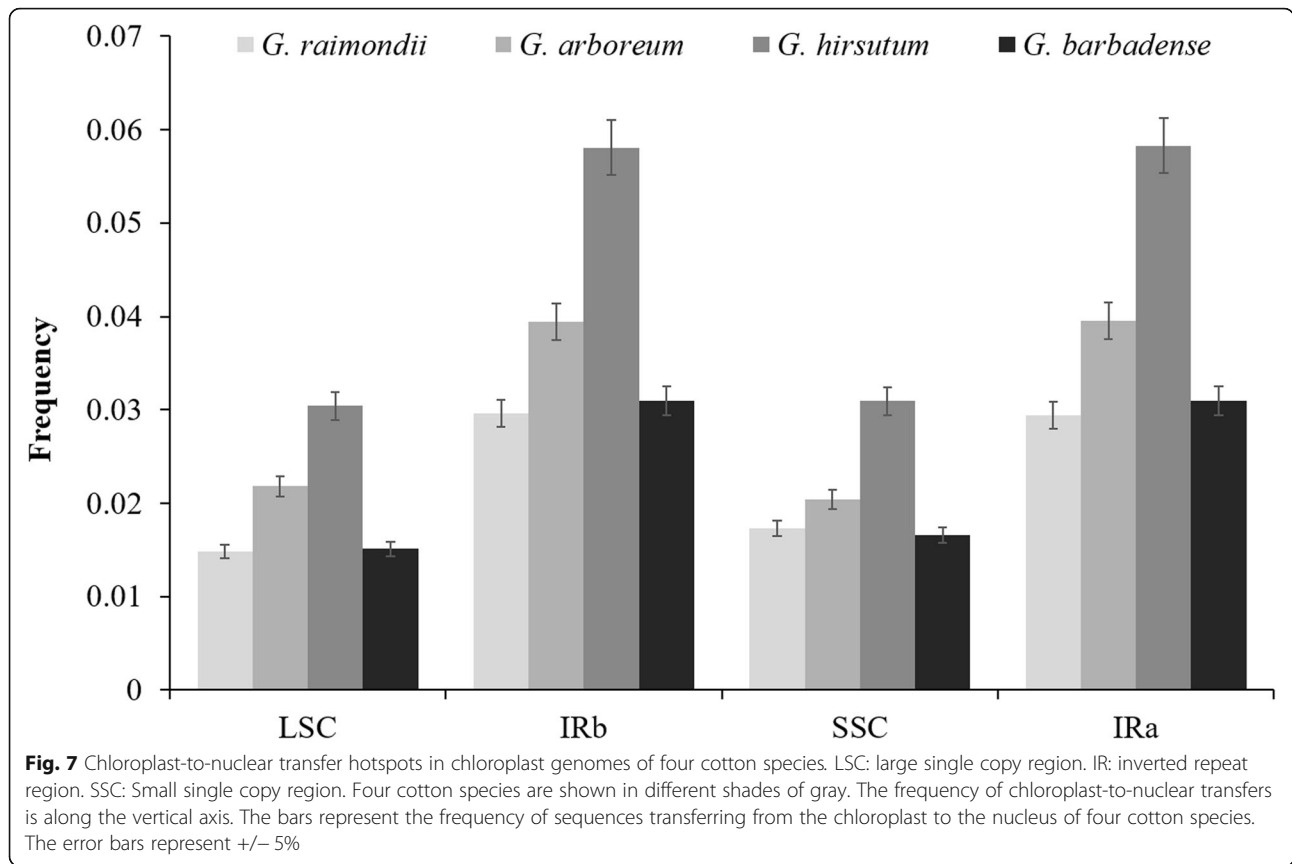## Low expression levels of organellar genes in nucleus (norgDNAs)

To assess if the norgs in cotton are expressed, we analyzed leaf RNA-seq data from two accessions of *G. hirsutum*, i.e., X11 and X42. We found that the relative expression of organellar genes was generally much higher than their intact nuclear counterparts (Table 3), although one norgDNA (the *nupt atpE*_cp) had a relatively high RPKM value compared to other norgDNAs. To explore this further, we evaluated the amount and pattern of sequence variation differentiating the organellar gene from the norgDNA and the resulting expression difference. Only six SNPs differentiate the 402 nucleotides of *atpE*_A09 and *atpE*_cp (~ 1.5% sequence divergence), indicating that this is a relatively recent or highly

Fig. 6 Mitochondrion-to-nucleus transfer-out hotspots in four cotton species. IGS: intergenic sequence. The top left corner shows the transfer hotspot of mitochondrial genomes in *G. raimondii* (D₅). The top right corner shows the transfer hotspot of the mitochondrial genomes in *G. arboreum* (A₂). The bottom left corner shows the transfer hotspot of mitochondrial genomes in *G. hirsutum* (AD₁). The bottom left corner shows the transfer hotspot of mitochondrial genomes in *G. barbadense* (AD₂). The horizontal axis indicates the nucleotide position in each mitochondrial genome. The density of mitochondrial-to-nuclear transfers is along the vertical axis. The arrows show the maximum density of IGT frequency

conserved norgDNA. When we removed all reads that could not be accurately assigned to *atpE*_A09 or *atpE*_cp, we found that no reads actually mapped to the region distinguishing the nuclear and organellar copies of *atpE* (Fig. 8). When we repeat this partitioning for an additional chloroplast gene (*petG*_cp) whose percent divergence between nuclear and chloroplast copies was sufficiently high to distinguish the two copies (~ 6.1% divergence, fourfold greater than *atpE*_A09), relatively few reads were both ambiguously mapped and assigned to a nuclear origin (Additional file 6). While this suggests that our RNA-seq analyses are suitable for capturing the expression differences between organellar and norg genes, we also validated these interpretations via qRT-PCR (Additional file 7) using *atpE* and *petG* as exemplars.

NorgDNAs changes during the process of diploids and allopolyploids evolution

>For IGT into the nucleus, the impact of allopolyploidy is notable and appears different for mitochondrial-to-nuclear IGT and chloroplast-to-nuclear IGT. For *numts*, fewer transfers are inferred in the allopolyploid species relative to their diploid progenitors, and the *numts* present are more frequently partial and/or decayed copies (Fig. 9a). Conversely, almost all *nupts* present in the diploids are retained in the allopolyploid species (Fig. 9b). To evaluate if this pattern is repeated for other polyploid systems, thereby suggesting that this might be a general phenomenon associated with genome doubling, we analyzed IGT for allopolyploid *Brassica napus* (AACC) relative to its model diploid progenitors, *B. oleracea* (CC). *B. rapa* (AA). Unlike *Gossypium*, allopolyploid *Brassica*

Fig. 7 Chloroplast-to-nuclear transfer hotspots in chloroplast genomes of four cotton species. LSC: large single copy region. IR: inverted repeat region. SSC: Small single copy region. Four cotton species are shown in different shades of gray. The frequency of chloroplast-to-nuclear transfers is along the vertical axis. The bars represent the frequency of sequences transferring from the chloroplast to the nucleus of four cotton species. The error bars represent +/− 5%

shows similar rates of integration for both *numts* and *nupts* (Additional file 8).

Discussion

Variable rates of IGT occur between the intracellular genomes of *Gossypium*

In our study, we characterized the rate and direction (i.e., nucleus ↔ chloroplast, nucleus ↔ mitochondrion, chloroplast ↔ mitochondrion) of IGT between the intracellular genomes of four *Gossypium* species. We detected most mitochondrial genes in present mitogenomes, which suggest that most genes (i.e., protein-encoding, rRNA, and tRNA genes) in *Gossypium* mitochondrial genomes are much conserved [29, 102–104]. For example, the mitochondrial genome of *Gossypium*

retains some genes encoding both complex II (succinate dehydrogenase, *sdh* genes) and ribosomal subunits (*rpl* and *rps* genes) that have been massively lost in other angiosperms [9, 17], e.g., *sdh3* and *sdh4*, *rpl10* and *rps10*. Meanwhile, we found all chloroplast genes in any of the four cotton species' plastids, which validates the general conservation of chloroplast genes [105], as in most plant species [9]. While all three intracellular genomes were involved in IGT, the only directions of IGT not detected were nuclear or mitochondrial transfers into the chloroplast. This is consistent with observations from most other plants [9], with two notable exceptions, i.e., *Daucus carota* [39] and *Asclepias syriaca* [40], where intracellular transfers into the chloroplast have been reported. Together, these observations suggest that the

Table 2 Repeats variation of nuclear genomes in four cotton species

Species	Genome sizes (Mb)	Repeat sizes (Mb)	Repeat percentage (%)	Numt length (kb)	Numt percentage (%)	Nupt length (kb)	Nupt percentage (%)	References
G. raimondii	761.57	423.43	55.60	1107	0.15	2072	0.27	[64]
G. arboreum	1694.60	1160.80	68.50	1813	0.11	3391	0.20	[83]
G. hirsutum	2173.00	1460.26	67.20	1092	0.05	4276	0.20	[84]
G. barbadense	2570.00	1776.13	69.11	1028	0.04	3332	0.13	[88]

Zhao *et al. BMC Plant Biology*   (2019) 19:492

Page 10 of 18

Table 3 Transcription levels of 14 organellar genes and nuclear homologs in two *G. hirsutum* varieties

Organellar genes/ Nuclear homologies	DNA lengths (bp)		RPKM value_X42[a]		RPKM value_X11[a]		SNP numbers
	OrgDNA	NuDNA	OrgDNA	NuDNA	OrgDNA	NuDNA	
*atpE*_cp/*atpE*_A09	402	402	661.56	71.19	1219.42	165.73	6
*ndhC*_cp/*ndhC*_D07	363	363	90.29	2.95	177.01	3.45	12
*ndhE*_cp/*ndhE*_A01	306	306	27.38	0.00	52.07	0.65	22
*ndhK*_cp/*ndhK*_A10	678	678	156.19	1.68	257.03	3.50	21
*petG*_cp/*petG*_D12	114	114	129.93	1.18	199.81	2.31	7
*petL*_cp/*petL*_D01	96	96	154.29	7.68	378.55	17.14	6
*petN*_cp/*petN*_D04	90	90	75.96	0.00	112.65	0.00	5
*psaJ*_cp/*psaJ*_D02	129	129	560.61	3.64	1214.12	5.61	6
*psbE*_cp/*psbE*_D06	252	252	184.58	0.53	230.42	0.26	14
*psbH*_cp/*psbH*_A05	222	222	63.70	8.15	105.87	7.41	9
sdh4-mt/*sdh4*_D04	399	399	8.90	1.85	9.07	5.11	1
rps14-mt/*rps14*_A01	303	303	25.44	0.00	21.29	0.22	8
nad6-mt/*nad6*_A08/A03	621	621/621	8.42	0.00/3.78	7.00	0.00/2.54	33/1
cob-mt/*cob*_A01	1179	1179	7.50	0.11	7.15	0.00	27

[a]RPKM values were calculated by RNA-seq data; X11 and X42 represented two *G. hirsutum* varieties, Xinluzao 11 and Xinluzhong 42

plastid genome lacks an active mechanism to integrate exogenous sequences.

Overall, we found substantially more sequences were integrated into the nucleus than the mitochondrion, a phenomenon that may be due to the mechanisms by which sequences get integrated into each genome, limitations on mitochondrial genome size, or both. Nearly all mitochondrial and chloroplast genes experienced transfer to the nucleus in *Gossypium*, which is a pattern seen in most other angiosperm plants [9], and at a transfer rate that is 100-fold greater than into the mitochondrial genome.

Non-additive effects of allopolyploidization on IGT in *Gossypium*

Genome doubling via polyploidy has numerous consequences [106]. Polyploidy can alter the size, content, and complexity of the genome [107], consequently affecting genetic variation, stress adaptation, biological complexity, speciation, biodiversity [108] and evolutionary novelty [109]. Myriad genomic consequences have been documented for polyploidy [110–112]; however, the cyto-nuclear effects of allopolyploidy in particular (which results from hybridization of divergence species) have been underexplored [113] and little is known regarding the influence of polyploidy on IGT. Here, we evaluate IGT transfers for two allopolyploid species, *G. hirsutum* and *G. barbadense,* which arose from a single polyploidization event about 1–2 million years ago involving the ancestors of the two diploid cottons sequenced here [81]. We found one diploid ancestral species (here, *G. raimondii*) experiences more *nupt* truncation and/or loss than

the other diploid ancestral species (here, *G. arboreum*), and the resulting polyploid retains nearly all of the *nupts* found in the diploids, similar to the extensive retention found in *G. arboreum*. These patterns are consistent with the general observation found from a survey of 21 land plants [9]. Conversely, mitochondrial-to-nuclear IGT are massively lost in the allopolyploid species. That is, only a few decayed *numts* are retained in the allopolyploid species, less than retained in either diploid progenitor, and those that were retained were more commonly the older *numts* shared between diploid species.

The total length of retained *numts* and *nupts* did not approach additivity for either polyploid species, *G. hirsutum* (AD$_1$) or *G. barbadense* (AD$_2$), with the possible exception of *nupts* in *G. hirsutum* whose total length was 80% of that found in their representative model diploid parents, *G. raimondii* (D$_5$) and *G. arboreum* (A$_2$). This may reflect insertions in the model diploid parents, either after divergence from the polyploid or the true polyploid progenitors, or it may represent differential decay in the polyploids, both upon formation and over time. Interestingly, the ratio of *nupt* to *numt* in the two allotetraploid cotton species was twice that in both diploid cotton species, indicating a possible shift in relative *nupt* and *numt* incorporation and/or retention (toward *nupt*) in polyploid cotton. This could be partially explained by the challenge in uniquely identifying mitochondrial-derived repeats from the background repeats of the nuclear genome, particularly as these degrade over time. Our preliminary results (above) suggest similar norgDNA integration rates in *Gossypium* and *Brassica*, but more data are needed to understand the
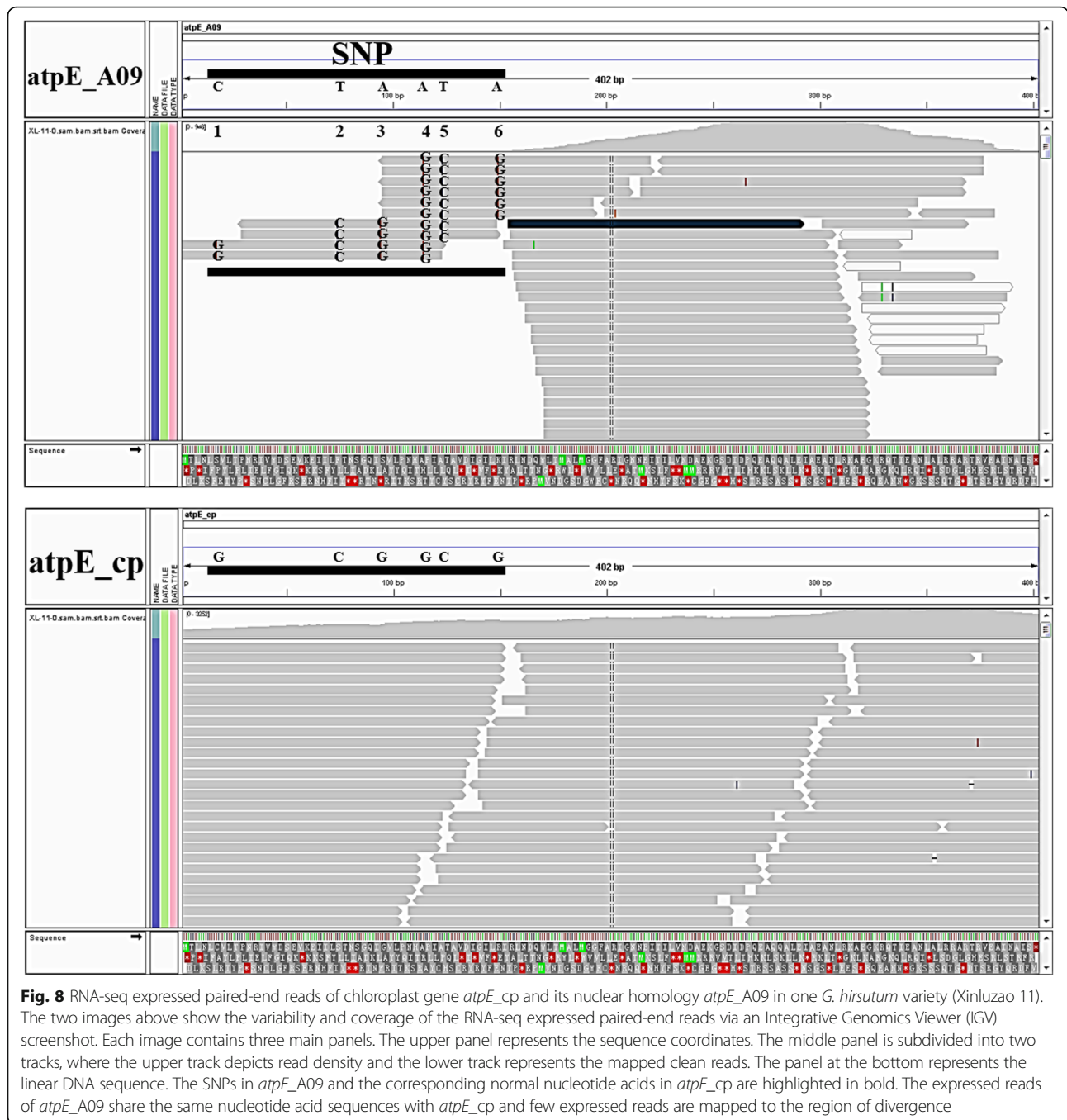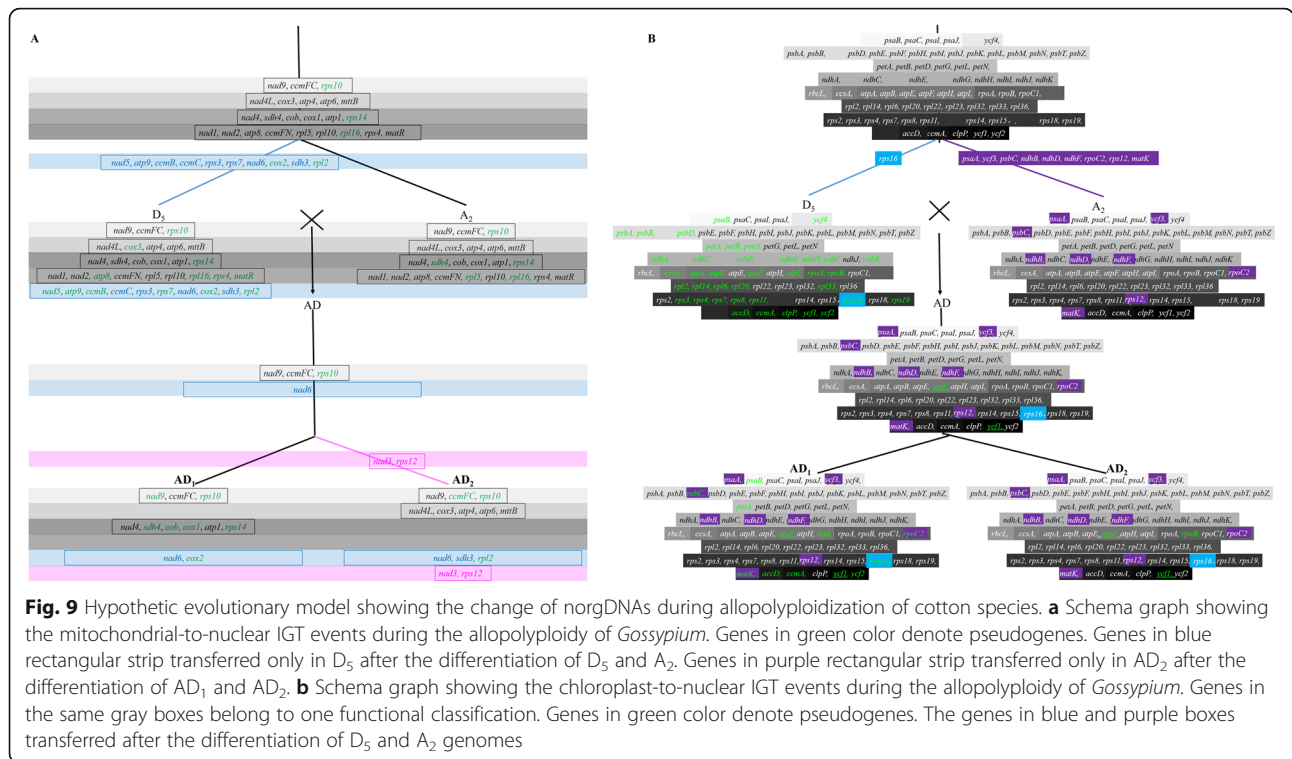
Fig. 8 RNA-seq expressed paired-end reads of chloroplast gene *atpE*_cp and its nuclear homology *atpE*_A09 in one *G. hirsutum* variety (Xinluzao 11). The two images above show the variability and coverage of the RNA-seq expressed paired-end reads via an Integrative Genomics Viewer (IGV) screenshot. Each image contains three main panels. The upper panel represents the sequence coordinates. The middle panel is subdivided into two tracks, where the upper track depicts read density and the lower track represents the mapped clean reads. The panel at the bottom represents the linear DNA sequence. The SNPs in *atpE*_A09 and the corresponding normal nucleotide acids in *atpE*_cp are highlighted in bold. The expressed reads of *atpE*_A09 share the same nucleotide acid sequences with *atpE*_cp and few expressed reads are mapped to the region of divergence

influences of ploidy on *nupt* and *numt* integration and degradation dynamics. While this hints at the differences among allopolyploid systems with respect to IGT, it is premature to draw general, more widely applicable conclusions. Clearly, this area requires further study to understand the evolutionary implications of norgDNAs, the patterns and processes by which they evolve in different biological systems, and the influences of ploidy on integration and degradation.

Patterns of IGT among intracellular genomes

Here we found that the mitochondrial *atp6-trnW* intergenic region and the chloroplast inverted repeat region represent hotspots for IGT source material, i.e., these regions were frequently transferred to other intracellular genomes. In the mitochondrial genome, these (and other) transfers were frequently associated with 10 bp microhomologies, a phenomenon that was observed across distantly related genera (as 2-6 bp microhomologies).

Fig. 9 Hypothetic evolutionary model showing the change of norgDNAs during allopolyploidization of cotton species. **a** Schema graph showing the mitochondrial-to-nuclear IGT events during the allopolyploidy of *Gossypium*. Genes in green color denote pseudogenes. Genes in blue rectangular strip transferred only in D₅ after the differentiation of D₅ and A₂. Genes in purple rectangular strip transferred only in AD₂ after the differentiation of AD₁ and AD₂. **b** Schema graph showing the chloroplast-to-nuclear IGT events during the allopolyploidy of *Gossypium*. Genes in the same gray boxes belong to one functional classification. Genes in green color denote pseudogenes. The genes in blue and purple boxes transferred after the differentiation of D₅ and A₂ genomes

Detected *numts* were found in various states of completeness, either due to the length during transfer or subsequent decay. Here we found that most *numts* remain full-length after insertion, suggesting that the mechanism responsible for *numt* generation may preferentially operate on full length genes; however, a few genes did transfer to the nucleus as partial sequences, indicating that partial genes are not excluded from transfer. Using a phylogenetic approach, we also detected genes that were transferred intact but decayed afterwards, e.g. *nad9* and *atp8*. Because younger *numts* are more readily identified and detection becomes more difficult as the *numt* decays, there is a natural bias toward detection of younger, more intact, and/or conserved numts, decaying *numts/nupts* slowly becoming increasingly difficult to detect as they lose sequence similarity to their source. Here, however, we describe several *numts/nupts* that have survived the basal-most radiation of the genus, approximately 5-10 MYA. Further insight into the potential reasons for these uncommon retentions will require additional functional study.

Contributions of IGT to genome expansion
The underlying causes of genome size variation represent an old question for the nuclear genome and a relatively recent one for the mitochondrial genome. With respect to the latter, we found that both nuclear and chloroplast sequences are correlated with mitogenome expansion, concordant with the view that contamination of plant mitochondrial with nuclear and chloroplast DNA is at the heart of mitogenome expansion in plants [114]. As most plant genomes are composed of massive amounts of repetitive sequence, it is tempting to suggest that nuclear-derived repeats should represent the most frequent transfers; however, correlations between repetitive sequence characteristics (e.g., number, total amount) and the mitogenome size were weak. Therefore, although nuclear repeat-derived transfers do contribute to mitogenome size increase, they cannot fully explain the correlation between nuclear-to-mitochondrial transfer and mitogenome expansion.

With respect to nuclear genome size variation, it is commonly accepted that repetitive content underlies most of the size variations among species; however, the contribution of other sources of genome size expansion are less well characterized. Here we found that while nuclear repeats do contribute the most (more than half) to genomes size differences among species (i.e., 55.60, 68.50, 67.20 and 69.11% in *G. raimondii*, *G. arboreum*, *G. hirsutum* and *G. barbadense*, respectively), the contribution of *numts* and *nupts* to genome size is not insignificant (Table 2). Presence of *numts* and *nupts* were both positively correlated with nuclear genome size, with *nupts* affecting the genome size to a somewhat greater extent (i.e., 0.27, 0.20, 0.20 and 0.13% in *G. raimondii*, *G. arboreum*, *G. hirsutum* and *G. barbadense*, respectively) than *numt* doing (i.e., 0.15,

0.11, 0.05 and 0.04% in *G. raimondii*, *G. arboreum*, *G. hirsutum* and *G. barbadense*, respectively, Table 2). We also found a positive correlation between nuclear repeats and norgDNAs, which may reflect a greater ability for norgDNAs to successfully integrate into genomes with larger gene-free regions.

The consequences of organelle-to-nuclear transfers are largely unknown

Organelles have a history of functional transfers to the nucleus, some of which are conserved among distantly related lineages and others which are lineage specific. These are in addition to non-functional transfers, which may represent sequences varying in size and content from gene fragments to large regions of the organellar genome [4, 14, 18, 19, 46, 101, 115, 116]. While many recent organellar-derived sequences are inactive and/or nonfunctional, some transfers to the nucleus may have function [11, 16, 17, 117], and both functional and non-functional transfers can have consequences for intracellular metabolism and genome evolution [14, 17, 18, 114, 117, 118]. We analyzed the expression of organellar genes and their norgDNAs as a proxy for function, finding that those organelle-derived genes were generally expressed at a far lower level than their organellar counterparts, suggesting limited functional potential. Therefore, while expression of norgDNA does appear to occur, it may reflect leaky transcription rather than function. Expressed and potentially functional norgDNAs, however, do occur in plants (despite the lack of a nuclear promoter) [11], cautioning against a ubiquitous assumption of non-function. Additionally, as we noted, characterization of norgDNA expression needs to acknowledge the high similarity of organelle-derived reads to those from nuclear integrants.

Evolutionary asymmetry of norgDNAs in the allopolyploidization process of *Gossypium*

During polyploid evolution, many of the *numts* present in the diploids have been lost and nearly all *nupts* in diploids are still retained in current polyploids. Given that assembly quality differences among the genomes represents a possible alternative explanation for presence vs. absence of specific genes, we compared our results to a new analysis using the more recently released and even higher quality genome assemblies for *G. hirsutum* and *G. barbadense* [86]. This analysis mostly reiterated our results, but with some differences. In *G. hirsutum*, there are three additional *numts* (*sdh3*, *rpl5* and *rps7*) and one lost *numt* (*nad4*), and one intact *nupt* (*matK*, the previous inference was not intact) and one decayed *nupt* (*rps12*, the previous inference was intact). In *G. barbadense*, there are six new *numts* (*sdh4*, *cob*, *cox1*, *rpl5*, *rps7* and *rps14*) and four lost *numts* (*nad3*, *nad4L*, *rps12* and *mttB*), and four newly decayed *nupts* (*ycf2, ycf3, ycf4* and *rps12*) that were intact using the earlier released assembly. These results emphasize both the potential impact of assembly quality on inference of *numt* gain and loss, but also confirm that our general conclusions about differential loss and gain in the polyploids are valid. In addition, we also detected some previously published large-fragment transfers, such as the complete mitochondrial genome transfers in Chr01 of *G. raimondii* and ChrA03 of *G. hirsutum* [64, 93], the *numt* on Chr2 of *A. thaliana* [61, 101], and additional, newly detected transfers, further validating reliability and robustness of our methods. Therefore, we conclude that the *numt* and *nupt* dynamism reported here is a genuine biological phenomenon.

Conclusions

This study has concluded that among the three intracellular genomes, most IGT was from organelles into the nuclear genome in two cultivated allopolyploids and their ancestral diploid cotton species; however, both nuclear retrotransposons and chloroplast tRNA genes integrated into mitochondrial genomes at a rate sufficient to correlate with mitogenome size increase. We detected hotspot regions for both the source of IGT (e.g., the *atp6-trnW* intergenic region) and the destination, which require further study in diverse plants to determine the patterns and generality of these observations. We also found that following allopolyploidy, there was a striking asymmetry in IGT retention in the nuclear genome, with most *numts* being lost but most *nupts* retained. While it is tempting to attribute parental origin to the loss of these fragments, in that paternally-derived norgDNAs could potentially be interfering, and therefore deleterious, we saw no bias in loss with respect to parental origin. As this is the first report of the relationship between intergenomic gene transfer and allotetraploidy, data from additional polyploid systems is required to understand the evolutionary dynamics of IGT in polyploids.

Methods

Plant materials and genome data

We used two varieties of upland cotton (*G. hirsutum*), Xinluzao 11 (X11) and Xinluzhong 42 (X42) for expression analysis. The seeds of X11 and X42 were provided by our own laboratory. X11 (original name: Yuzao 202) was introduced by Bole Seed Station from the Institute of Cash Crops, Henan Academy of Agricultural Sciences, Henan, China, in 1994. After many years of breeding trial, it was approved by the Crop Variety Approval Committee of Xinjiang Autonomous Region, Xinjiang,

China, in 1999, and named as Xinluzao 11. X42 was bred by the Institute of Cash Crops of Xinjiang Academy of Agricultural Sciences and approved by the Crop Variety Approval Committee of Xinjiang Autonomous Region, Xinjiang, China, in 2009.

The chloroplast, mitochondrial and nuclear genome sequences of two diploids (*G. raimondii* and *G. arboreum*), and two allotetraploids (*G. hirsutum* and *G. barbadense*) were downloaded from the NCBI database (accession numbers listed in Additional file 1).

Identification of intergenomic-transfer gene

For each species in the study, we performed pair-wise comparisons of chloroplast or mitochondrial genes and nuclear chromosomes sequences using BLAST (command code 1 in Additional file 9) [118]. We set *e*-value to $1e^{-5}$ and a 100-bp minimal length for a high match (95%). We also identified the mitochondrial insertions of chloroplast DNA (*mtpts*) using local BLASTN (version 2.2.23) with the 50-bp minimal length of the match (identity > 95%, coverage > 90%). We cataloged those transfers as full-length genes. Pseudogenes are without full-length or existing mutations resulting in premature stop codons.

Detection of nuclear transposable elements and repeats in mitochondrial genomes

We detected nuclear transposable elements from nuclear sources using RepeatMasker (command code 2 in Additional file 9) (http://www.repeatmasker.org) with a custom *Gossypium*-enriched repeat database for the four cotton species studied. We used two-tailed *t*-tests to evaluate the significant levels of the different types. The repeats in mitochondrial genome were identified by repeat-match algorithm (command code 3 in Additional file 9) in MUMmer [116]. Specific parameters include: -f (use the forward strand only), −n (minimum match length; default 20), and -t (only output tandem repeats).

Microhomologies analysis

The analysis was performed as previously described [13, 119]. If there were same nucleotides next to the *mtpt* fusion point shared by the different land species, we identified as microhomologies.

IGT hotspot analyses in *Gossypium*

We performed dot matrix comparisons between the mitochondrial or chloroplast genomes and nuclear chromosomes of four *Gossypium* species using nucmer program of MUMmer. We set 100-bp minimum size for an exact match and 500-bp minimal interval between every two matches [116]. We calculated middle positions of all *Gossypium* organellar insertions into nuclear chromosomes to tabulate transfer

hotspots. Then, we draw the charts of the frequency distribution by R (command code 4 in Additional file 9) (https://www.r-project.org/).

Expression analysis of norgDNAs

Extracted total RNA using improved cetyltrimethylammonium bromide (CTAB) and sodium dodecyl sulfate (SDS) method from leaves of two accessions of upland cotton, X11 and X42, were sequenced on an Illumina HiSeq2500 at Shanghai Hanyu Biotech Co., Ltd. Sequencing libraries were generated using the Illumina TruSeq RNA Sample Preparation Kit (Illumina, USA) following the manufacturer's recommendations, and four index codes were added to diagnose the sample origins (nuclear or organellar) for each sequence. Following experimental confirmation of concentration and purity, poly-(T) oligo-attached magnetic beads were utilized for nuclear mRNA enrichment. Fragments, preferentially 200-300 bp in length, were enriched using the Illumina PCR Primer Cocktail in a 10 cycle PCR amplification to form cDNA libraries. Finally, libraries were the paired-end sequenced in one lane with 4 Gb clean reads/sample of an average length of 125 nt. RNA sequences data quality was checked using FastQC. The reads were mapped to the norgDNAs homologies using bowtie 2 (command code 5 in Additional file 7) [120], then samtools idxstats [121] (command code 6 in Additional file 9) were used to calculate the expression reads counts of each gene. The RPKM values were used to estimate relative expressions. Expressed paired-end reads were mapped onto their respective consensus sequences using BWA 0.7.10-r789 [122]; then the results were transformed into BAM files using SAMtools view [121]; and structural variations (SVs) and InDels were visualized using the Integrative Genomics Viewer [123]. The total RNA of X42 and X11 were reverse-transcribed using two pairs of primers at once, i.e., oligo dT primer and random 6 mers, to capture both nuclear expression (NE) and organellar expression (OE). The cDNA produced by the oligo dT primer represents nuclear expression, whereas cDNA produced by the random 6 mers represents a combination of nuclear and organellar genes (NOE). OE is equal to NOE minus NE. Both kinds of cDNA were used as the templates for qRT-PCR. qRT-PCR experiments were conducted using SYBR Premix Ex Taq™ (Tli RNaseH Plus) RR420A kit (TaKaRa) by Applied Biosystems 7500 Real-Time PCR System. The procedure consisted of three stages: stage 1, 95 °C, 30 s, 1 cycle; stage 2: 95 °C, 5 s, 60 °C, 35 s, 40 cycles; stage 3: 95 °C, 15 s, 60 °C, 1 min, 95 °C, 35 s, 1 cycle. Using the cotton housekeeping gene *UBQ7* as internal control, we analysed the relative expression levels of two organellar genes and their nuclear copies, using the $2^{-\triangle\triangle Ct}$ method. Each sample is repeated three times.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12870-019-2041-2.

Additional file 1. Accession numbers for nuclear, mitochondrial and chloroplast genomes of four cotton species.

Additional file 2. Correlation between the length of nuclear and chloroplast sequences transferring to mitochondrial genome, mitochondrial genome size and repeat sizes in mitochondrial genomes in 26 land plants. (A) Correlation between the length of the nuclear sequences transferring to the mitochondrion and mitochondrial genome size. (B) Correlation between the length of the chloroplast sequences transferring to the mitochondrion and mitochondrial genome size. (C) Correlation between repeat sizes in mitochondrial genomes and the mitochondrial genome. (D) Correlation between length of nuclear sequences transferring to mitochondrial genomes and repeat sizes of mitochondrial genomes. (E) Correlation between length of chloroplast sequences transferring to mitochondrial genomes and repeat sizes of mitochondrial genomes. Each dot represents a two-dimensional value (X, Y) of one species. Back dots denote four cotton species and the gray dots mean the other species. The slash represents the linear regression function of the distribution tendency of the dots. R^2 is the regression coefficient.

Additional file 3. Dot matrix analysis of nuclear insertions of chloroplast DNAs (above) and mitochondrial DNAs (below) in 14 plant species including four cotton species identified by whole-genome alignment. The results were filtered so that only those alignments with the one-to-one mapping between reference and query were selected. The red and blue lines refer to positive and reverse matches, respectively. The upper-most phylogenetic clades are drawn based on the maximum likelihood (ML) method with the model GTR + G + I. The figures related to mitochondrial DNAs insertions into four the nuclear genomes of four cotton species are quoted from a previous paper in our lab [61].

Additional file 4. The identification of nuclear organellar DNA in *Gossypium*. A: *G. raimondii*. B: *G. arboreum*. C: *G. hirsutum* (A_t). D: *G. hirsutum* (D_t). Red bands around the circles indicate the nuclear chromosomes. Orange and green lines represent insertions more than 5 kb from mitogenome and chloroplast genome, respectively. While grey lines represent insertions between 100 bp to 5 kb from both genomes. MT: mitochondrial genome. CP: chloroplast genome.

Additional file 5. Correlation between the length of mitochondrial and chloroplast sequences transferring to the nuclear genome, nuclear genome size and repeat sizes in nuclear genomes in 26 land plants. (A) Correlation between the length of mitochondrial sequences transferring to nuclear and nuclear genome sizes. (B) Correlation between the length of the chloroplast sequences transferring to nuclear and nuclear genome sizes. (C) Correlation between repeat sizes in nuclear genomes and nuclear genome sizes. (D) Correlation between length of mitochondrial sequences transferring to nuclear genomes and repeat sizes of nuclear genomes. (E) Correlation between length of chloroplast sequences transferring to nuclear genomes and repeat sizes of nuclear genomes. Each dot represents a two-dimensional value (X, Y) of one species. Black dots denote four cotton species and the gray dots mean the other species. The slash represents the linear regression function of the distribution tendency of the dots. R^2 is the regression coefficient.

Additional file 6. RNA-seq expressed paired-end reads of chloroplast gene *petG*_cp and its nuclear homology *petG*_D12 in *G. hirsutum* variety (Xinluzao 11). The two images above show the variability and coverage of the RNA-seq expressed paired-end reads via an Integrative Genomics Viewer (IGV) screenshot. Each image contains three main panels. The upper panel represents the sequence coordinates. The middle panel is subdivided into two tracks, where the upper track depicts read density and the lower track shows the mapping of clean reads. The panel at the bottom represents the linear DNA sequence. The SNPs in *petG*_D12 and the corresponding normal nucleotide acids in *petG*_cp are highlighted in

bold. The expressed reads of *petG*_D12 share the same nucleotide acids sequences with *petG*_cp and few expressed reads are mapped to the divergence region.

Additional file 7. Relative expression levels of chloroplast genes *atpE_cp/petG_cp* and their nuclear copies, *atpE_A09/petG_D12*, in two *G. hirsutum* varieties. X42: Xinluzao 42; X11: Xinluzao 11. The relative expression values are calculated with the method of $2^{-\Delta\Delta Ct}$. **, $p < 0.01$. See Methods for details.

Additional file 8. Hypothetic evolutionary model showing the change of norgDNAs during allopolyploidization of *Brassica* species. (A) Schema graph showing the mitochondrial-to-nuclear IGT events during the allopolyploidy of *Brassica*. CC: *B. oleracea*. AA: *B. rapa*. AACC: *B. napus*. Gray rectangular strips represent gene blocks that transferred before the divergence of two diploid species. Blue and purple genes strips represent gene blocks that transferred only in diploid CC and allotetraploid AACC, respectively. Genes in green color denote pseudogenes. (B) Schema graph showing the chloroplast-to-nuclear IGT events during the allopolyploidy of *Brassica*. Genes in the same gray boxes belong to one functional classification. The genes in green color denote pseudogenes. Genes in blue, purple and red boxes transferred only in CC, AA, and AACC, respectively.

Additional file 9. Commands and their codes used in this study.

Abbreviations

B. napus (AACC): *Brassica napus*; *B. oleracea* (CC): *Brassica oleracea*; *B. rapa* (AA): *Brassica rapa*; cpDNA: Chloroplast DNA; *G. arboreum*,(A_2) : *Gossypium arboreum*; *G. barbadense* (AD_2): *barbadense*; *G. herbaceum* (A_1): *Gossypium herbaceum*; *G. hirsutum* (AD_1): *Gossypium hirsutum*; *G. raimondii* (D_5): *Gossypium raimondii*; IGT: Intergenomic gene transfer; LTR-retro: Long terminal repeat retrotransposon; mtDNA: Mitochondrial DNA; norgDNAs: Nuclear organellar DNAs; nuDNA: Nuclear DNA; *numts*: Nuclear mitochondrial insertions; nupts: Nuclear plastidial insertions; TE: Transposable element

Availability of data and materials

The datasets used and/or analysed during the current study are available in the NCBI repository. *G. raimondii* nuclear genome, accession number: PRJNA171262, DOI: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA171262; *G. raimondii* mitochondrial genome, accession number: KR736345, DOI: https://www.ncbi.nlm.nih.gov/nuccore/KR736345.1/; *G. raimondii* chloroplast genome, accession number: HQ325744, DOI: https://www.ncbi.nlm.nih.gov/nuccore/HQ325744.1/; *G. arboreum* nuclear genome, accession number: PRJNA335838, DOI: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA335838; *G. arboreum* mitochondrial genome, accession number: KR736342, DOI: https://www.ncbi.nlm.nih.gov/nuccore/KR736342.1/; *G. arboreum* chloroplast genome, accession number: HQ325740, DOI: https://www.ncbi.nlm.nih.gov/nuccore/HQ325740.1/; *G. hirsutum* nuclear genome, accession number: PRJNA248163, DOI: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA248163; *G. hirsutum* mitochondrial genome, accession number: JX944505, DOI: https://www.ncbi.nlm.nih.gov/nuccore/JX944505.1/; *G. hirsutum* chloroplast genome, accession number: DQ345959, DOI: https://www.ncbi.nlm.nih.gov/nuccore/

Zhao *et al. BMC Plant Biology*   (2019) 19:492

Page 16 of 18

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Laboratory of Cotton Genetics, Genomics and Breeding /Joint Laboratory for International Cooperation in Crop Molecular Breeding, Ministry of Education / Key Laboratory of Crop Heterosis and Utilization of Ministry of Education, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China. [2]Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011, USA.

References

1. Gray MW. Mosaic nature of the mitochondrial proteome: implications for the origin and evolution of mitochondria. Proc Natl Acad Sci U S A. 2015; 112(33):10133–8.
2. Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 1998;393(6681):162–5.
3. Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, et al. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 2005;15(14):1325–30.
4. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 2004; 5(2):123–35.
5. Kleine T, Maier UG, Leister D. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol. 2009;60: 115–38.
6. Bock R. Witnessing genome evolution: experimental reconstruction of endosymbiotic and horizontal gene transfer. Annu Rev Genet. 2017; 51(1):1–22.
7. Michalovova M, Vyskot B, Kejnovsky E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (*nupts* and *numts*) of six plant species: size, relative age and chromosomal localization. Heredity. 2013;111(4):314–20.
8. Wang D, Lloyd AH, Timmis JN. Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants. Proc Natl Acad Sci U S A. 2012;109(7):2444–8.
9. Zhao N, Wang Y, Hua J. The roles of mitochondrion in intergenomic gene transfer in plants: a source and a pool. Int J Mol Sci. 2018;19(2):E547.
10. Cusimano N, Wicke S. Massive intracellular gene transfer during plastid genome reduction in nongreen *Orobanchaceae*. New Phytol. 2016; 210(2):680–93.
11. Wang D, Qu ZP, Adelson DL, Zhu JK, Timmis JN. Transcription of nuclear organellar DNA in a model plant system. Genome Biol Evol. 2014;6(6):1327–34.
12. Wang D, Timmis JN. Cytoplasmic organelle DNA preferentially inserts into open chromatin. Genome Biol Evol. 2013;5(6):1060–4.
13. Hazkani-Covo E, Covo S. *Numt*-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet. 2008; 4(10):e1000237.
14. Leister D. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet. 2005;21(12):655–63.
15. Bergthorsson U, Adams KL, Thomason B, Palmer JD. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature. 2003;424(6945): 197–201.
16. Adams KL, Palmer JD. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol. 2003;29(3):380–95.
17. Adams KL, Qiu Y-L, Stoutemyer M, Palmer JD. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci U S A. 2002;99(15):9905–12.
18. Blanchard JL, Schmidt GW. Pervasive migration of organellar DNA to the nucleus in plants. J Mol Evol. 1995;41(4):397–406.
19. Lopez JV, Yuhki N, Masuda R, Modi W, Obrien SJ. Numt, a recent transfer and tandem amplification of mitochondrial-DNA to the nuclear genome of the domestic cat. J Mol Evol. 1994;39(2):174–90.
20. Park S, Grewe F, Zhu A, Ruhlman T, Sabir J, Mower J, et al. Dynamic evolution of *Geranium* mitochondrial genomes through multiple horizontal and intracellular gene transfers. New Phytol. 2015;208:570–83.
21. Huang CY, Ayliffe MA, Timmis JN. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature. 2003;422:72–6.
22. Richly E, Leister D. *Numts* in sequenced eukaryotic genomes. Mol Biol Evol. 2004;21:1081–4.
23. Sheppard AE, Timmis JN. Instability of plastid DNA in the nuclear genome. PLoS Genet. 2009;5(1):e1000323.
24. Smith DR, Crosby K, Lee RW. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. Genome Biol Evol. 2011;3:365–71.
25. Stegemann S, Hartmann S, Ruf S, Bock R. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A. 2003; 100(15):8828–33.
26. Smith DR. Extending the limited transfer window hypothesis to inter-organelle DNA migration. Genome Biol Evol. 2011;3:743–8.
27. Sloan DB, Alverson AJ, Storchova H, Palmer JD, Taylor DR. Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. BMC Evol Biol. 2010;10:1–15.
28. Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell. 2011;23(7):2499–513.
29. Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, et al. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science. 2013;342(6165):1468–73.
30. Wang D, Rousseau-Gueutin M, Timmis JN. Plastid sequences contribute to some plant mitochondrial genes. Mol Biol Evol. 2012;29(7):1707–11.
31. Rodriguez-Moreno L, Gonzalez VM, Benjak A, Marti MC, Puigdomenech P, Aranda MA, et al. Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. BMC Genomics. 2011;12:424.
32. Wang D, Wu YW, Shih ACC, Wu CS, Wang YN, Chaw SM. Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 mya. Mol Biol Evol. 2007;24(9):2040–8.
33. Knoop V, Unseld M, Marienfeld J, Brandt P, Sunkel S, Ullrich H, et al. Copia-, gypsy- and line-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis thaliana*. Genetics. 1996;142(2):579–85.
34. Dietrich A, Small I, Cosset A, Weil JH, Marechal-Drouard L. Editing and import: strategies for providing plant mitochondria with a complete set of functional transfer RNAs. Biochimie. 1996;78(6):518–29.
35. Gandini CL, Sanchez-Puerta MV. Foreign plastid sequences in plant mitochondria are frequently acquired via mitochondrion-to-mitochondrion horizontal transfer. Sci Rep. 2017;7:43402.
36. Schuster W, Brennicke A. Plastid, nuclear and reverse transcriptase sequences in the mitochondrial genome of *oenothera*: is genetic information transferred between organelles via RNA? EMBO J. 1987;6(10): 2857–63.
37. Wang XC, Chen H, Yang D, Liu C. Diversity of mitochondrial plastid DNAs (*mtpts*) in seed plants. Mitochondrial DNA Part A. 2017;29:635–42.
38. Veronico P, Gallerani R, Ceci LR. Compilation and classification of higher plant mitochondrial trna genes. Nucleic Acids Res. 1996;24(12):2199–203.
39. Goremykin VV, Salamini F, Velasco R, Viola R. Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. Mol Biol Evol. 2009;26(1):99–110.
40. Straub SCK, Cronn RC, Edwards C, Fishbein M, Liston A. Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (*Apocynaceae*). Genome Biol Evol. 2013;5(10):1872–85.

41. Iorizzo M, Grzebelus D, Senalik D, Szklarczyk M, Spooner D, Simon P. Against the traffic: the first evidence for mitochondrial DNA transfer into the plastid genome. Mob Genet Elem. 2012;2(6):261–6.

42. Iorizzo M, Senalik D, Szklarczyk M, Grzebelus D, Spooner D, Simon P. De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. BMC Plant Biol. 2012;12(1):1–17.

43. Smith DR. Mitochondrion-to-plastid DNA transfer: It happens. New Phytol. 2014;202(3):736–8.

44. Chang S, Yang T, Du T, Huang Y, Chen J, Yan J, et al. Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in *Brassica*. BMC Genomics. 2011;12:497.

45. Handa H. The complete nucleotide sequence and rna editing content of the mitochondrial genome of rapeseed (*Brassica napus* l.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. Nucleic Acids Res. 2003;31(20):5907–16.

46. Marienfeld J, Unseld M, Brandt P, Brennicke A. Genomic recombination of the mitochondrial *atp6* gene in *Arabidopsis thaliana* at the protein processing site creates two different presequences. DNA Res. 1996;3(5): 287–90.

47. Rivarola M, Foster JT, Chan AP, Williams AL, Rice DW, Liu X, et al. Castor bean organelle genome sequencing and worldwide genetic diversity analysis. PLoS One. 2011;6(7):e21743.

48. Chang S, Wang Y, Lu J, Gai J, Li J, Chu P, et al. The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. PLoS One. 2013;8(2):e56502.

49. Alverson AJ, Zhuo S, Rice DW, Sloan DB, Palmer JD. The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. PLoS One. 2011;6(1):e16404.

50. Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, et al. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Gen Genomics. 2005;272(6):603–15.

51. Terasawa K, Odahara M, Kabeya Y, Kikugawa T, Sekine Y, Fujiwara M, et al. The mitochondrial genome of the moss *Physcomitrella patens* sheds new light on mitochondrial evolution in land plants. Mol Biol Evol. 2006;24(3): 699–709.

52. Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, et al. Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA: a primitive form of plant mitochondrial genome. J Mol Biol. 1992;223(1):1–7.

53. Chaw S-M, Chun-Chieh Shih A, Wang D, Wu Y-W, Liu S-M, Chou T-Y. The mitochondrial genome of the gymnosperm *cycas taitungensis* contains a novel family of short interspersed elements, bpu sequences, and abundant rna editing sites. Mol Biol Evol. 2008;25(3):603–15.

54. Fang Y, Wu H, Zhang T, Yang M, Yin Y, Pan L, et al. A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera* L.) mitochondrial genome. PLoS One. 2012;7(5):e37164.

55. Notsu YN, Masood SM, Nishikawa TN, Kubo NK, Akiduki GA, Nakazono MN, et al. The complete sequence of the rice (*Oryza sativa* l.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol Gen Genomics. 2002;268(4):434–45.

56. Tian X, Zheng J, Hu S, Yu J. The rice mitochondrial genomes and their variations. Plant Physiol. 2006;140(2):401–10.

57. Ogihara Y, Yamazaki Y, Murai K, Kanno A, Terachi T, Shiina T, et al. Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. Nucleic Acids Res. 2005;33(19):6235–50.

58. Darracq A, Varré J-S, Touzet P. A scenario of mitochondrial genome evolution in maize based on rearrangement events. BMC Genomics. 2010; 11(1):233.

59. Li S, Chen Z, Zhao N, Wang Y, Nie H, Hua J. The comparison of four mitochondrial genomes reveals cytoplasmic male sterility candidate genes in cotton. BMC Genomics. 2018;19(1):775.

60. Tang MY, Chen ZW, Grover CE, Wang YM, Li SS, Liu GZ, et al. Rapid evolutionary divergence of *Gossypium barbadense* and *G hirsutum* mitochondrial genomes. BMC Genomics. 2015;16:770.

61. Chen ZW, Nie HS, Grover CE, Wang YM, Li P, Wang MY, et al. Entire nucleotide sequences of *Gossypium raimondii* and *G. arboreum* mitochondrial genomes revealed a-genome species as cytoplasmic donor of the allotetraploid species. Plant Biol. 2017;19(3):484–93.

62. Stern DB, Lonsdale DM. Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA-sequence in common. Nature. 1982;299(5885):698– 702.

63. Liu GZ, Cao DD, Li SS, Su AG, Geng JN, Grover CE, et al. The complete mitochondrial genome of *Gossypium hirsutum* and evolutionary analysis of higher plant mitochondrial genomes. PLoS One. 2013;8(8):e69476.

64. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin DC, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 2012;492(7429):423–7.

65. Chen ZW, Zhao N, Li SS, Grover CE, Nie HS, Wendel JF, et al. Plant mitochondrial genome evolution and cytoplasmic male sterility. Crit Rev Plant Sci. 2017;36(1):55–69.

66. Rhoads DM, Subbaiah CC. Mitochondrial retrograde regulation in plants. Mitochondrion. 2007;7(3):177–94.

67. Kitazaki K, Arakawa T, Matsunaga M, Yui-Kurino R, Matsuhira H, Mikami T, et al. Post-translational mechanisms are associated with fertility restoration of cytoplasmic male sterility in sugar beet (*Beta vulgaris*). Plant J. 2015;83(2):290–9.

68. Liu Z, Yang Z, Wang X, Li K, An H, Liu J, et al. A mitochondria-targeted ppr protein restores *pol* cytoplasmic male sterility by reducing *orf224* transcript levels in oilseed rape. Mol Plant. 2016;9(7):1082–4.

69. Sloan DB, Muller K, McCauley DE, Taylor DR, Storchova H. Intraspecific variation in mitochondrial genome sequence, structure, and gene content in *Silene vulgaris*, an angiosperm with pervasive cytoplasmic male sterility. New Phytol. 2012;196(4):1228–39.

70. Chen L, Liu YG. Male sterility and fertility restoration in crops. Annu Rev Plant Biol. 2014;65:579–606.

71. Charlesworth D. Origins of rice cytoplasmic male sterility genes. Cell Res. 2017;27(1):3–4.

72. Tang H, Zheng X, Li C, Xie X, Chen Y, Chen L, et al. Multi-step formation, evolution, and functionalization of new cytoplasmic male sterility genes in the plant mitochondrial genomes. Cell Res. 2017;27(1):130–46.

73. Okazaki M, Kazama T, Murata H, Motomura K, Toriyama K. Whole mitochondrial genome sequencing and transcriptional analysis to uncover an RT102-type cytoplasmic male sterility-associated candidate gene derived from *Oryza rufipogon*. Plant Cell Physiol. 2013;54(9):1560–8.

74. Gallagher LJ, Betz SK, Chase CD. Mitochondrial rna editing truncates a chimeric open reading frame associated with s male-sterility in maize. Curr Genet. 2002;42(3):179–84.

75. Wise RP, Fliss AE, Pring DR, Gengenbach BG. *Urf13-t* of T cytoplasm maize mitochondria encodes a 13 kd polypeptide. Plant Mol Biol. 1987;9(2):121–6.

76. Horn R, Gupta KJ, Colombo N. Mitochondrion role in molecular basis of cytoplasmic male sterility. Mitochondrion. 2014;19 Pt B:198–205.

77. Wesolowski W, Szklarczyk M, Szalonek M, Slowinska J. Analysis of the mitochondrial proteome in cytoplasmic male-sterile and male-fertile beets. J Proteome. 2015;119:61–74.

78. Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, Gore MA, et al. Re-evaluating the phylogeny of allopolyploid *Gossypium* I. Mol Phylogenet Evol. 2015;92:45–52.

79. Wendel JF, Grover CE. Taxonomy and evolution of the cotton genus, *Gossypium*. Cotton. 2015;57:25–44.

80. Gallagher J, Grover C, Rex K, Moran M, Wendel J. A new species of cotton from wake atoll, *Gossypium stephensii* (malvaceae). Syst Bot. 2017;42:115–23.

81. Wendel JF. New world tetraploid cottons contain old-world cytoplasm. Proc Natl Acad Sci U S A. 1989;86(11):4132–6.

82. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, et al. The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet. 2012;44(10):1098–103.

83. Li FG, Fan GY, Wang KB, Sun FM, Yuan YL, Song GL, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet. 2014; 46(6):567–72.

84. Li FG, Fan GY, Lu CR, Xiao GH, Zou CS, Kohel RJ, et al. Genome sequence of cultivated upland cotton (*Gossypium hirsutum* tm-1) provides insights into genome evolution. Nat Biotechnol. 2015;33(5):524–30.

85. Zhang TZ, Hu Y, Jiang WK, Fang L, Guan XY, Chen JD, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* l. Acc. Tm-1) provides a resource for fiber improvement. Nat Biotechnol. 2015;33(5):531–7.

86. Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat Genet. 2019;51(2):224–9.

87. Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, Yang CQ, et al. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Sci Rep. 2015;5:14139.

88. Yuan DJ, Tang ZH, Wang MJ, Gao WH, Tu LL, Jin X, et al. The genome sequence of sea-island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. Sci Rep. 2015;5:17662.
89. Lee SB, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, et al. The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. BMC Genomics. 2006;7:61.
90. Xu Q, Xiong GJ, Li PB, He F, Huang Y, Wang KB, et al. Analysis of complete nucleotide sequences of 12 *Gossypium* chloroplast genomes: origin and evolution of allotetraploids. PLoS One. 2012;7(8):e37128.
91. Chen ZW, Feng K, Grover CE, Li PB, Liu F, Wang YM, et al. Chloroplast DNA structural variation, phylogeny, and age of divergence among diploid cotton species. PLoS One. 2016;11(6):e0157183.
92. Chen ZW, Grover CE, Li PB, Wang YM, Nie HS, Zhao YP, et al. Molecular evolution of the plastid genome during diversification of the cotton genus. Mol Phylogenet Evol. 2017;112(Supplement C):268–76.
93. Chen ZW, Nie HS, Wang YM, Pei HL, Li SS, Zhang LD, et al. Rapid evolutionary divergence of diploid and allotetraploid *Gossypium* mitochondrial genomes. BMC Genomics. 2017;18:876.
94. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002;3(5):329–41.
95. Cossu RM, Casola C, Giacomello S, Vidalis A, Scofield DG, Zuccolo A. LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. Genome Biol Evol. 2017;9(12):3449–62.
96. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20(1):43–5.
97. Vitte C, Panaud O, Quesneville H. Ltr retrotransposons in rice (*Oryza sativa*, L.): Recent burst amplifications followed by rapid DNA loss. BMC Genomics. 2007;8:218.
98. Wang H, Liu J-S. LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. BMC Genomics. 2008;9(1):382.
99. Hawkins JS, Proulx SR, Rapp RA, Wendel JF. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc Natl Acad Sci U S A. 2009;106(42):17811–6.
100. Qiu F, Ungerer MC. Genomic abundance and transcriptional activity of diverse *gypsy* and *copia* long terminal repeat retrotransposons in three wild sunflower species. BMC Plant Biol. 2018;18(1):6.
101. Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, et al. Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci U S A. 2001;98(9):5099–103.
102. Kurland CG, Andersson SGE. Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev. 2000;64(4):786–820.
103. Kitazaki K, Kubo T. Cost of having the largest mitochondrial genome: Evolutionary mechanism of plant mitochondrial genome. J Bot. 2010, 2010. https://doi.org/10.1155/2010/620137.
104. Skippington E, Barkman TJ, Rice DW, Palmer JD. Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all *nad* genes. Proc Natl Acad Sci U S A. 2015;112(27):E3515–E24.
105. Huang J, Chen RH, Li XG. Comparative analysis of the complete chloroplast genome of four known *Ziziphus* species. Genes. 2017;8(12):340.
106. Ding MQ, Chen ZJ. Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. Curr Opin Plant Biol. 2018;42:37–48.
107. Jiao Y. Double the genome, double the fun: genome duplications in angiosperms. Mol Plant. 2018;11(3):357–8.
108. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet. 2017;18:411.
109. Soltis PS, Soltis DE. Ancient wgd events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 2016;30:159–65.
110. Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev. 2018;49:1–7.
111. Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, et al. Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc Natl Acad Sci U S A. 2018;115(18):4713–8.
112. Emery M, Willis MMS, Hao Y. Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. PLoS Genet. 2018;14(3):e1007267.
113. Sehrish T, Symonds VV, Soltis DE, Soltis PS, Tate JA. Cytonuclear coordination is not immediate upon allopolyploid formation in *Tragopogon miscellus* (asteraceae) allopolyploids. PLoS One. 2015;10(12):e0144339.
114. Goremykin VV, Lockhart PJ, Viola R, Velasco R. The mitochondrial genome of *Malus domestica* and the import-driven hypothesis of mitochondrial genome expansion in seed plants. Plant J. 2012;71(4):615–26.
115. Kudla J, Albertazzi F, Blazević D, Hermann M, Bock R. Loss of the mitochondrial *cox2* intron 1 in a family of monocotyledonous plants and utilization of mitochondrial intron sequences for the construction of a nuclear intron. Mol Gen Genomics. 2002;267(2):223–30.
116. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30(11):2478–83.
117. Ong HC, Palmer JD. Pervasive survival of expressed mitochondrial *rps14* pseudogenes in grasses and their relatives for 80 million years following three functional transfers to the nucleus. BMC Evol Biol. 2006;6:55.
118. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
119. Ju YS, Tubio JMC, Mifsud W, Fu BY, Davies HR, Ramakrishna M, et al. Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. Genome Res. 2015;25(6):814–24.
120. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
121. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.
122. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010;26(5):589–95.
123. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.