An Overlooked Paleotetraploidization in Cucurbitaceae

Jinpeng Wang,[1,2] Pengchuan Sun,[2] Yuxian Li,[1,2] Yinzhe Liu,[1,2] Nanshan Yang,[1,2] Jigao Yu,[1,2] Xuelian Ma,[1] Sangrong Sun,[1,2] Ruiyan Xia,[1] Xiaojian Liu,[1] Dongcen Ge,[1] Sainan Luo,[1] Yinmeng Liu,[1] Youting Kong,[1] Xiaobo Cui,[1] Tianyu Lei,[1,2] Li Wang,[1,2] Zhenyi Wang,[1,2] Weina Ge,[1,2] Lan Zhang,[1,2] Xiaoming Song,[1,2] Min Yuan,[1,2] Di Guo,[1,2] Dianchuan Jin,[2] Wei Chen,[2] Yuxin Pan,[1,2] Tao Liu,[2] Guixian Yang,[1] Yue Xiao,[1] Jinshuai Sun,[1] Cong Zhang,[1] Zhibo Li,[1] Haiqing Xu,[1] Xueqian Duan,[1] Shaoqi Shen,[1] Zhonghua Zhang,[3] Sanwen Huang,[3] and Xiyin Wang*,[1,2]

[1]School of Life Sciences, North China University of Science and Technology, Tangshan, Hebei, China

[2]Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, Hebei, China

[3]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Beijing, China

***Corresponding author:** E-mail: wangxiyin@vip.sina.com.

Associate editor: Hideki Innan

Abstract

Cucurbitaceae plants are of considerable biological and economic importance, and genomes of cucumber, watermelon, and melon have been sequenced. However, a comparative genomics exploration of their genome structures and evolution has not been available. Here, we aimed at performing a hierarchical inference of genomic homology resulted from recursive paleopolyploidizations. Unexpectedly, we found that, shortly after a core-eudicot-common hexaploidy, a cucurbit-common tetraploidization (CCT) occurred, overlooked by previous reports. Moreover, we characterized gene loss (and retention) after these respective events, which were significantly unbalanced between inferred subgenomes, and between plants after their split. The inference of a dominant subgenome and a sensitive one suggested an allotetraploid nature of the CCT. Besides, we found divergent evolutionary rates among cucurbits, and after doing rate correction, we dated the CCT to be 90–102 Ma, likely common to all Cucurbitaceae plants, showing its important role in the establishment of the plant family.

***Key words:* Cucurbitaceae, genomics, polyploidy, homology, gene colinearity.**

Article

Introduction

As the fourth most important economic plant family, Cucurbitaceae, consist of 115 proposed genera with 960 species, growing mainly in tropical and subtropical regions (Schaefer et al. 2009). Their edible fruits are among the earliest cultivated ones in the world. Most of the cucurbits are annual vines, and the others are woody lianas, thorny shrubs, or trees. Their leaves are estipulate alternate simple palmately lobed or palmately compound, and their stems are hairy and pentangular. They have large yellow or white unisexual flowers, being monoecious or dioecious. The female flowers have inferior ovaries. The fruit is often a kind of modified ovary. Therefore, they are important model plants to study sex determination and the development of proto-sex-chromosomes. Besides, they are also models for vascular biology. Annually, they are grown on ~9 million hectares of land, and yield ~184 million tons of vegetables, fruits, and seeds (http://faostat.fao.org).

Because of their biological and economic importance, the whole-genome sequences of three cucurbit plants, including cucumber (*Cucumis sativus* L., $2n = 14$; Huang et al. 2009), melon (*Cucumis melo* L., $2n = 24$; Garcia-Mas et al. 2012), and watermelon (*Citrullus lanatus*, $2n = 22$; Guo et al. 2013), have been deciphered so far. These works have provided valuable opportunities to understand their biological functions at divergent levels, from the functions of disease-related genes and regulatory pathways, to the inferred genomic regions contributing to domestication, sex determination, and genetic diversity, and also helped find the evolutionary trajectories of chromosome number evolution after their split from other eudicots, and during the divergence of them.

Recursive polyploidizations are proposed to answer for the fast divergence and success of seed and flowering plants on this planet (Jiao et al. 2011). Similar to many other plant genomes, the deciphered cucurbit genomes revealed complex genome structures (Huang et al. 2009). By characterizing the DNA sequence divergence between duplicated genes, it was shown that with exception of the core-eudicot common hexaploidization (ECH), the cucurbits avoided more recent whole-genome duplication events (Huang et al. 2009; Garcia-Mas et al. 2012; Guo et al. 2013). The ECH event was initially inferred with the Arabidopsis genome, and later much better characterized with the grape genome, to find a triplication of seven ancestral haploid chromosomes (Bowers et al. 2003; Jaillon et al. 2007). Grape genome was often taken as a good reference to understand the genomic changes in other eudicots.

The availability of multiple genomes from a plant family provides an opportunity to us to perform a hierarchical alignment of genomic homology. By deconvoluting the intra and intergenomic homology, and relating it to recursive polyploidizations and speciations, we aligned eight grass genomes (Wang, Wang, Jin et al. 2015), and those of cotton, cacao, and grape (Wang et al. 2016). These efforts contributed to understand their genomic evolution at divergent levels. In grasses, we produced a list of orthologous and paralogous genes, and using them redated the major evolutionary events. The cotton genomics analysis revealed its decaploid ancestor after split from cacao and other eudicots. Especially, these efforts exploited the high-cost and hard-won genomic data sets, and laid a solid foundation for future comparative genomics efforts from the plant research community.

By using the previously developed methods, algorithms, and software to perform hierarchical genomic homology reconstruction, here we aim at performing a comprehensive analysis of the cucurbit genomes with grape genome as an outgroup reference, and explore the evolutionary patterns and rules of gene duplications, gene losses, and genomic fractionation; check the likelihood of divergent evolutionary rates among cucurbits; reconstruct ancestral genome contents at major evolutionary nodes.

Results

Gene Colinearity within and among Genomes

Gene colinearity, reflecting shared ancestral gene order, is crucial to understand the genomic changes, especially in deconvoluting the evolution of complex plant genomes. By using ColinearScan, we detected colinear genes within each cucurbit genome and between each pair of them (supplementary table S1, Supplementary Material online). Grape genome was used as an outgroup reference to decipher the cucurbit genomes. Therefore, we also detected colinear genes within grape and between grape and cucurbit genomes. Homologous blocks with >4, 10, 20, and 50 colinear genes were checked (supplementary tables S2 and S3, Supplementary Material online).

Intragenomic Homology

According to the present assemblies, cucumber, and watermelon genomes have preserved better intragenomic homology. In watermelon, we revealed 499 homologous blocks with four or more colinear genes, containing 3,463 pairs of colinear genes in total (supplementary tables S2 and S3, Supplementary Material online). At the same criteria, we found only 384 and 338 homoeologous blocks in cucumber and melon, containing 2,891 and 2,395 colinear gene pairs in cucumber and melon, respectively. For the blocks with ten or more colinear genes, cucumber and watermelon have similar numbers of them (68 and 69), which contains similar numbers of colinear genes (1,157 and 1,151). The longest colinear block in three plants is located in watermelon, between chromosomes 4 and 10, containing 70 colinear gene pairs. Notably, for blocks with four or more colinear genes, much more (1.3–1.8 times) colinear genes were found in cucurbit genomes than in grape genome. This finding is something weird in that previous publications showed that cucurbits and grape share the same major-eudicot-common polyploidies and avoided further events. Besides, grape genome seems to preserve more longer homoeologous blocks (supplementary tables S2 and S3, Supplementary Material online), showing that more genomic fractionation occurred in cucurbits. These findings show that cucurbits have more complicated genomes than grape does.

Intergenomic Homology

Between three cucurbit plants, we revealed well-preserved intergenomic homology (supplementary tables S2 and S3, Supplementary Material online). For homologous blocks with four or more colinear genes, we found >1,561–3,152 homologous regions, containing 22,809–29,893 colinear gene pairs or 13,814–15,001 genes in colinearity. A small fraction (3%) of long blocks (with 50 or more colinear genes) contains ~60% colinear genes. Between three cucurbit plants and grape, we also found fair intergenomic homology. The numbers of blocks and genes in them are often much fewer than between cucurbits (supplementary tables S2 and S3, Supplementary Material online).

Two-Way Comparison to Distinguish Orthology and Outparalogy

Homologous gene dotplot is helpful to locate homologous correspondence between chromosomes and were produced by our custom software and used to distinguish orthologous regions, established due to grape-cucurbit split, and outparalogous regions, established due to the ECH (fig. 1 and supplementary figs. S1–S3, Supplementary Material online). The 19 chromosomes of grape were denoted with blocks in seven colors in the dotplots, corresponding to seven ancestral eudicot chromosomes before the ECH (Jaillon et al. 2007; Jiao et al. 2012). If no extra polyploidization in cucurbits, they would have a similar genomic structure as grape. Without considering further gene losses or duplications, we would expect a grape gene (or chromosomal region) had one best matched or orthologous cucurbit genes (chromosomal regions), and two secondary or outparalogous cucurbit genes (chromosomal regions). If there had been an extra tetraploidization in cucurbits, we would expect that a grape gene (or chromosomal region) had two best matched or orthologous cucurbit genes (chromosomal regions), and four outparalogous genes (chromosomal regions). In the grape-cucurbit homologous gene dotplots, orthologous blocks were inferred to be those for which a grape chromosomal region was much more similar to a specific cucurbit region than its grape paralogous regions are to the same cucurbit region. This permitted ready discrimination between orthologous regions, and outparalogous regions. In the dotplot figures (fig. 1 and supplementary figs. S1–S3, Supplementary Material online), we circled out orthologous regions using solid rectangles, and outparalogous ones using dash-line rectangles. In certain outparalogous regions with little trace of colinear genes, due to widespread

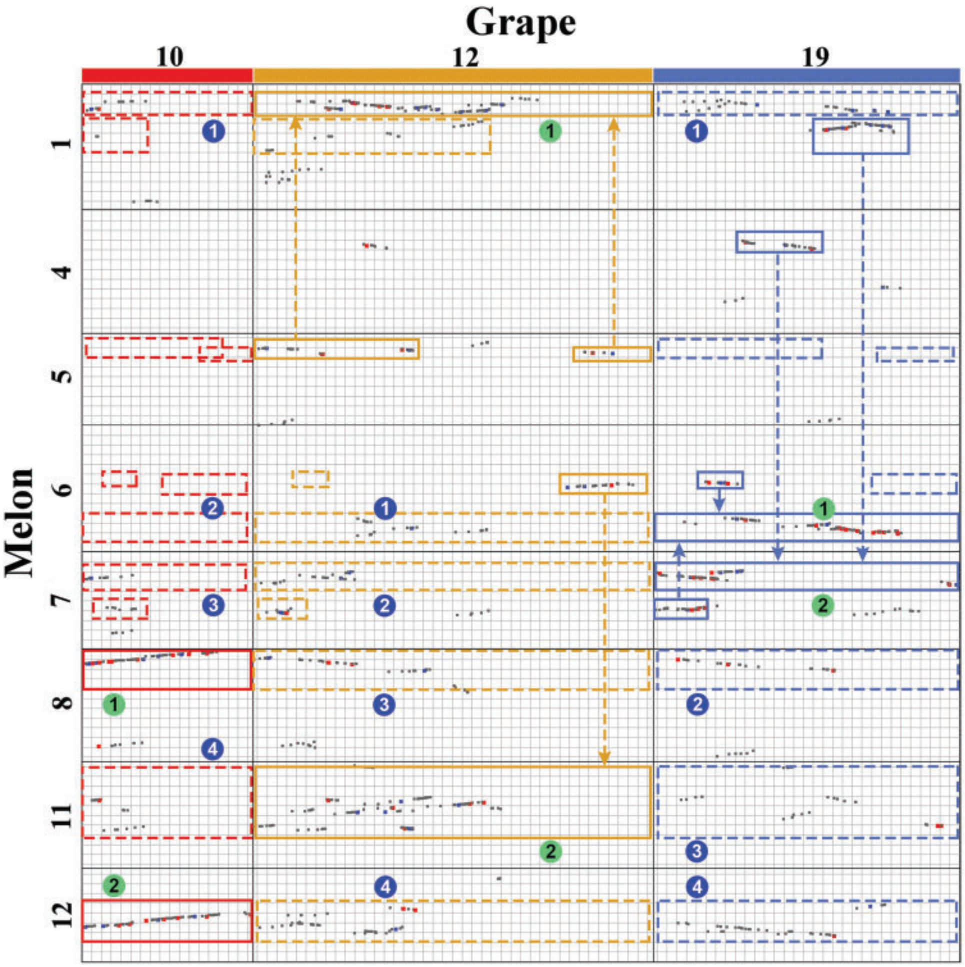

Fig. 1. Homologous dotplot between selected grape and melon chromosomes. The grape chromosomes 10, 12, and 19, being homoeologous triplets produced by the eudicot-common hexaploidy, and their matched melon chromosomes were aligned in horizontal and vertical directions, respectively. Red, blue, and gray dots were used to show the best, secondary, and other matched homologous genes respectively. Best-matched or orthologous regions were marked out by solid-line rectangles numbered by 1 and 2 in lime circles; outparalogous regions or secondary-matched were marked out by broken-lined rectangles numbered by 1–4 in blue circles. Arrows show complement correspondence produced by chromosome breakages during evolution.

and complementary gene losses (Maere et al. 2005), homology between grape chromosomes, and/or between grape and cucurbits can transitively indicate their actual homology.

Here, let us show an example using the ECH produced grape chromosome triplets, being homoeologous (paralogous) chromosomes to one another: Vv10, Vv12, and Vv19. We found that Vv10 has two best-matched or orthologous copies in melon, with one in chromosome Cm8 and the other in Cm12 (fig. 1), containing 95 and 90 colinear genes, respectively. The Vv10's orthologous regions in Cm8 and Cm12 are each outparalogous to Vv12 and Vv19, and the expected regions were circled out by dash-line rectangles. Much fewer dots could be found in the outparalogous blocks (Vv12-Cm8: 53 colinear genes; Vv19-Cm8: 21; Vv12-Cm12: 74; Vv19-Cm12: 42). Often, the homoeology between Vv10, Vv12, and Vv19 provide transitive information to help identify outparalogy between grape and melon chromosomes. In similar strategy, for Vv12 and Vv19, we identified their respective orthologous regions in melon, and these grape homoeologous chromosomes have different orthologous regions and each having two copies (fig. 1). The corresponding orthologous regions in melon were often broken into pieces due to chromosomal rearrangements. Fortunately, a complement pattern of broken segments helps infer that they are derived from the same ancestral chromosome. For Vv12, its two orthologous counterparts can be identified in Cm1, Cm5, Cm6, and Cm11, and the pieces in Cm1 and Cm5 are complement to one another, to form one orthologous copy of Vv12, and the pieces in Cm6 and Cm11 are complement to one another, to form the other orthologous copy (fig. 1).

The above grape homoeologous chromosome triplets have similar homologous patterns in cucumber and watermelon (supplementary fig. S4*A* and *B*, Supplementary Material online). Interactive analysis of their dotplots with the above description in melon helps to consolidate our inference of orthology and paralogy.

Besides the above example of tripled grape chromosomes, each of the other grape chromosomes have the similar finding of two sets of best-matched/orthologous cucurbit chromosomal regions, which is solid evidence of an extra and

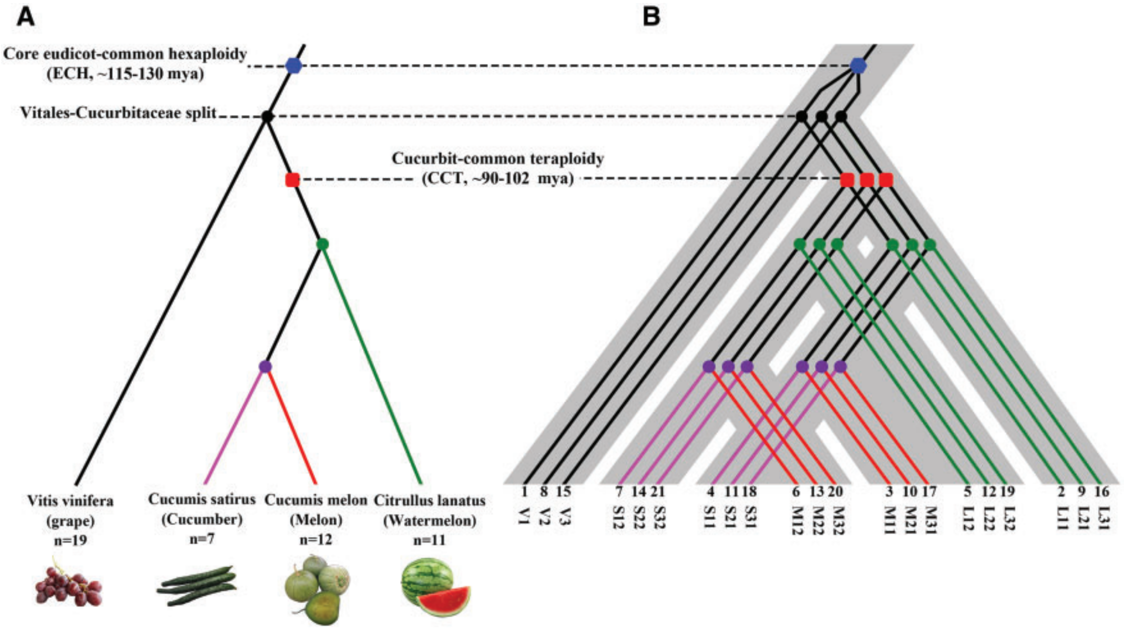

Fig. 2. Species and gene phylogenetic trees for three cucurbits. (*A*) Phylogenetic tree of melon (M), watermelon (L), cucumber (S), and grape (V): Eudicot-common hexaploidy (ECH) denoted by blue hexagon, cucurbit-common teraploidy (CCT) denoted by red square; (*B*) Gene phylogeny: three paralogous genes in the grape genome are denoted by V1, V2, and V3 produced by the ECH, and each has two orthologs and four outparalogs in a Cucurbit genome (e.g., V1 has two orthologs M11 and M12, and four outparalogs M21, M22, M31, and M32 in melon). The species tree is produced based on our present analysis of homologous genes.

cucurbit-common tetraploidy (CCT) after the split from grape and other eudicots (fig. 2). To be more careful, we performed gene phylogeny analysis to find additional evidence to support the CCT. For 1,809 groups of grape gene homoeologs at least one of which have one pair of orthologs in one cucurbit, we constructed gene trees for 175 homologous gene groups that are well aligned with ≥ 10 homologs. Eventually, we found that a fraction of 38.9% (68) trees have a topology supporting the CCT (supplementary fig. S5, Supplementary Material online). This is a prominent support if considering a similar analysis in grasses, with 31–37% trees of duplicated genes in different species supporting a grass-common tetraploidization (Paterson et al. 2004). The other trees have inconsistent topologies likely caused by divergent evolutionary rates of recursively duplicated genes. A nonnegligible groups of trees have all the cucurbit genes often with long branches clustered together, and the grape genes clustered together, showing elevated evolutionary rates in cucurbits.

An extra tetraploidy shared by cucurbits, which must have resulted in another rounds of genomic repatterning, explain the above findings of more colinear genes and smaller homoeologous fragments in cucurbit genomes compared with grape.

Event-Related Genomic Alignment

Homologous genes, orthologs or paralogs, were therefore linked to each event of speciations and polyploidies. The information of orthologous regions and outparalogous regions between genomes were retrieved from the dotplots (supplementary tables S4–S6, Supplementary Material online) to locate the orthologous and outparalogous genes (supplementary tables S7–S9, Supplementary Material online). The analysis actually helped divide duplicated genes from a genome into the ECH-produced paralogs and the CCT-produced paralogs. The ECH event produced 2,432 paralogous pairs, containing 3,866 genes in 87 colinear regions in grape; 2,246–2,902 paralogous genes were found, containing 3,359–4,109 genes in 156–186 colinear regions in cucurbits (table 1). The CCT event produced more paralogous regions in cucurbits, which are about twice of those in grape. Notably, the number of genes does not show significant increase. While for the ECH-produced cucurbit genes, the numbers decrease in about half (1,748–1,976), surely due to gene loss after the extra CCT, which produced 2,236–2,716 new paralogs in extant genomes, respectively.

Table 1. Number of Duplicated Genes within a Genome Related to the ECH and the CCT.

Species	ECH[a]-Related	CCT[b]-Related
Grape	87/2,432/3,866	—
Watermelon	95/1,116/1,748	79/1,137/2,274
Cucumber	92/1,278/1,976	69/1,358/2,716
Melon	108/1,155/1,791	82/1,118/2,236

Note.—As to the note, slashes are used to separate numbers of blocks, gene pairs, and gene numbers.
[a] Core eudicot-common hexaploidy.
[b] Cucurbit-common tetraploidy.

Gene colinearity revealed better intergenomic than intragenomic homology. For example, there are 7,860 (31.0%) cucumber genes having grape orthologs, and 3,075 (12.1%) having grape out-paralogs, whereas there are 7,167 (28.8%)

grape genes having cucumber orthologs, and 3,320 (13.3%) having cucumber out-paralogous genes. With grape as reference, 1,144 (4.6%) and 5,620 (22.6%) genes have one or two corresponding orthologs in melon. Similar findings are for cucumber and watermelon. More information can be found in supplementary tables S7–S9, Supplementary Material online.

With grape genome as reference and by filling colinear gene IDs into a table, we constructed hierarchical and event-related multiple-genome alignment**,** producing a table of homologous genes (supplementary fig. S6 and table S10, Supplementary Material online). To accommodate genes specific to cucurbits but being not available in the grape genome and not represented by the above alignment table, we also constructed the genomic homology table with watermelon as reference (supplementary fig. S7 and table S11, Supplementary Material online), which would better represent pan-cucurbit homology.

Evolutionary Rate Divergence and Dating

By checking molecular distance, we managed to estimate the times of the CCT and other evolutionary events. Here, we characterized synonymous substitutions on synonymous nucleotide sites (Ks) between colinear homoeologs within a genome and between different genomes. The ECH- and CCT-produced paralogs in different cucurbits have overlapping distributions, being not normal for having long tails esp. in the large value site and showing clear different peaks. We developed an approach to find the major normal distributions in each observed Ks distribution. Therefore, the locations of the peaks and their variances were determined statistically (supplementary table S12, Supplementary Material online). Interestingly, cucurbits evolve at considerably divergent rates. We found that the CCT-produced paralogs have different peak locations in three cucurbits (fig. 3*A* and *B*). This phenomenon may be explained by divergent evolutionary rates among different cucurbit plants. Among them, melon evolves slowest, with watermelon and cucumber faster by 23.6% and 27.4%, respectively. Similarly, according to different locations of the ECH-produced paralogs' Ks distributions, we found that the evolution rate of grape is the slowest among these four plants, with melon, watermelon, and cucumber faster than by 29.5%, 57.1%, and 59.0%, respectively.

Significant difference in evolutionary rates leads to distortion when inferring occurrence times of the evolutionary events. Here, based on a modified version of an approach that we previously developed (Wang, Wang, Jin et al. 2015), we performed evolutionary rate correction by aligning the peaks to the CCT event to the same location (see Materials and Methods for details; fig. 3*C* and *D*, and supplementary table S13, Supplementary Material online). This correction at the meantime aligned the ECH peaks to the same location, showing that it could correct the rate differences having accumulated after the ECH event between cucurbits and grape. Supposing that the ECH occurred at ∼115–130 Ma (Jiao et al. 2012; Vekemans et al. 2012), adopted by previous publications (Jaillon et al. 2007; Xu et al. 2011; Paterson et al. 2012), we inferred that the CCT event occurs at ∼90–102 Ma, ∼25 Ma after the ECH event and ∼7 Ma after Cucurbits split from Fagales (Magallon et al. 2015). In addition, we inferred that cucurbit plants' split from grape ∼107–121 Ma. The corrected Ks distributions of the cucurbit-grape orthologs and the cucurbit CCT paralogs are much overlapped with one another. This implies that the CCT event might have directly contributed to the split of cucurbit from grape and other eudicots and the establishment of the cucurbit family. Besides, we inferred that watermelon split from cucumber and melon ∼22–25 Ma, and melon and cucumber split ∼12–14 Ma.

Genomic Fractionation

Vast gene losses and relocations might have occurred after polyploidies in cucurbit duplicated regions. Intragenomic gene colinearity analysis indicated that in cucumber, melon, and watermelon, there was a tiny fraction of 0.8% (190), 0.3% (71), and 0.2% (43) genes preserved all six copies of duplicates produced in two recursive polyploidy events (supplementary table S14, Supplementary Material online), showing extensive gene deletion. Intergenomic analysis with grape as reference came to a little more but similar estimation (0.6–1.5%; supplementary table S15, Supplementary Material online). A fraction of 43–54% of cucurbit genes have no colinear paralogs; 15–17% grape genes have no corresponding colinear orthologs in cucurbits and the similar fraction of grape genes have one colinear orthologs in cucurbits. We found unbalance gene deletions between the CCT duplicated regions. A display of alignment of homologous regions from grape and cucurbits characterized the unbalance pattern (fig. 4). According to grape chromosome 1,87% and 93% of its genes do not have colinear genes in one of the two sets of watermelon orthologous regions, and 76% of its genes do not have correspondence in both sets (supplementary table S16, Supplementary Material online). For the same grape chromosome, similar fractions of missing correspondence are observed in the other two cucurbits, showing shared gene deletions in the cucurbit' orthologous regions, likely occurring before their split. A comparison between cucumber and watermelon shows that 38% of watermelon genes do not have colinear correspondence in cucumber; and 47% of cucumber genes do not have colinear correspondence in watermelon (supplementary tables S17 and S18, Supplementary Material online).

We checked temporal gene deletion by reconstructing the ancestral genome before grape-cucurbit split, that is, an ancestral gene was inferred if a grape gene has an cucurbit ortholog or outparalog at anticipated colinear locations. There were 9,990 genes in the reconstructed ancestral genome. A fraction of 11.3% (1,131/9,990) of the ancestral genes have neither of the CCT orthologs in all cucurbits, showing likely gene deletion after the grape-cucurbit split and before the CCT; 33.9% (6,766/(9,990 × 2)) misses one of the CCT ortholog in all cucurbits, showing likely gene deletion after the CCT and before the cucurbit split; after the cucurbit split,

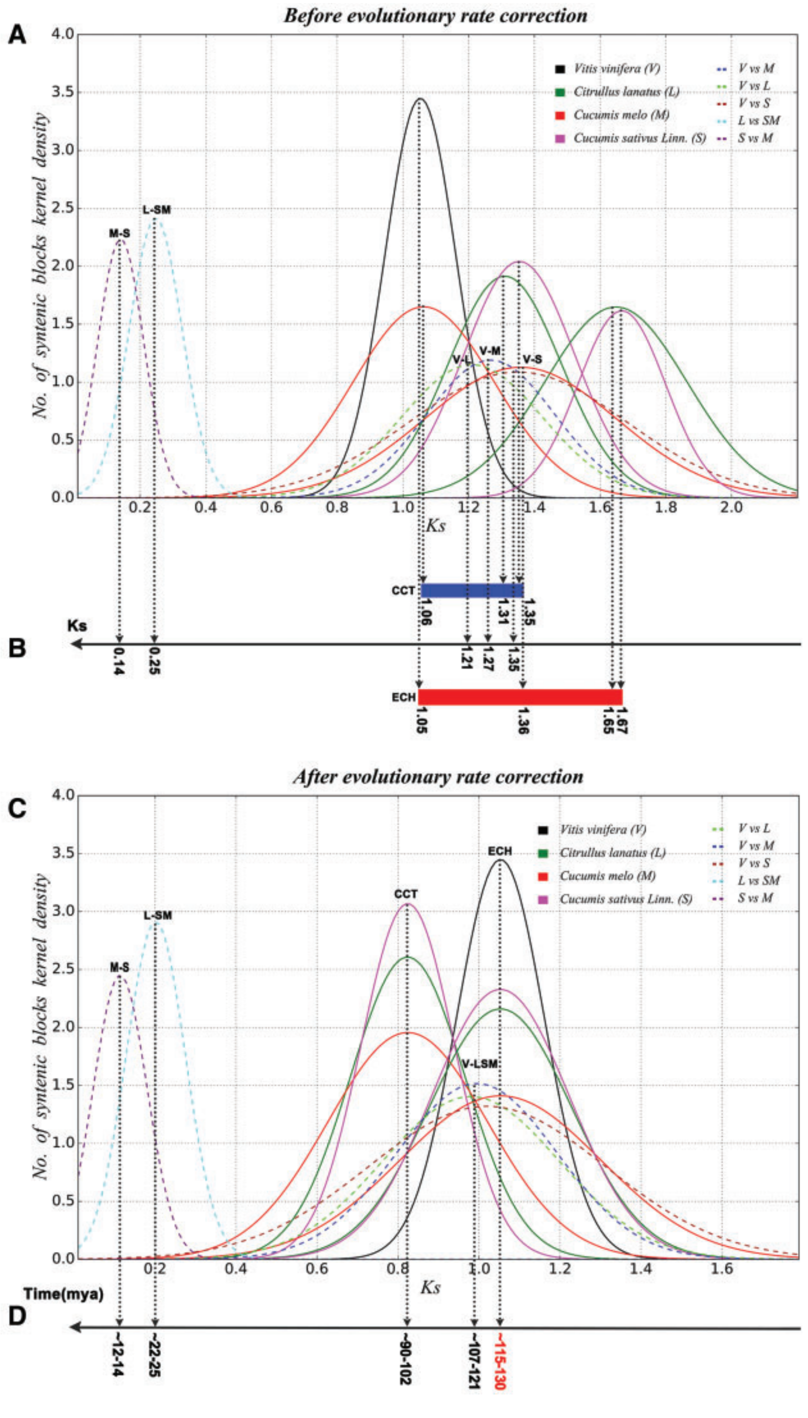

Fig. 3. Dating evolutionary events within and among three cucurbit and grape genomes. Grape (V), Melon (M), Watermelon (L), Cucumber (S). (*A*) Distribution of average synonymous substitution levels (Ks) between colinear gene pairs in intergenomic blocks (solid curves) and intragenomic blocks (dashed curves); (*B*) distribution of average synonymous substitution levels after correction to account for the slower evolution of melon or grape genes; (*C*) correction to the Ks distribution and occurrence of key evolutionary events; (*D*) inferred times.

there occurred species-specific deletion of 41.6% (8,310/(9,990 × 2)) in watermelon, 38.6% (7,715/(9,990 × 2)), and 44.3% (8,859/(9,990 × 2)) in cucumber. As to unbalanced gene deletion between the CCT duplicated regions, the reconstructed ancestral genome inferred lowered gene deletion rates but consistent observation of unbalance (supplementary table S19, Supplementary Material online).

As reported in maize and Brassica plants, subgenomes are often divergent in gene retention or genome fractionation (Schnable et al. 2011; Wang et al. 2011). For the referenced grape chromosome, we found that two sets of orthologous regions often show divergence in gene retention, suggesting that the existence of a dominant subgenome and a sensitive one (supplementary figs. S8–S10, Supplementary Material online). Genes in a local region of the reference chromosome

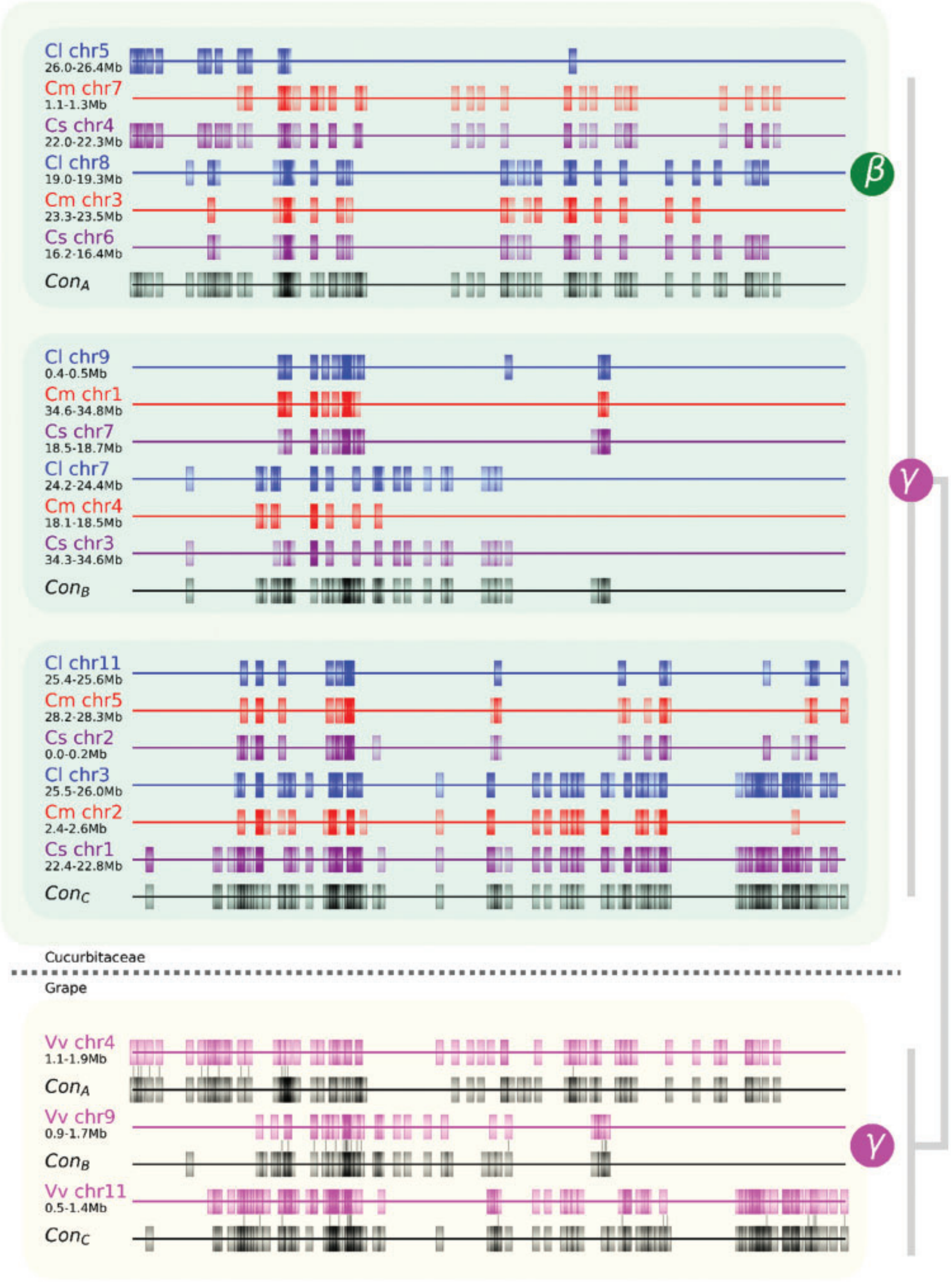

Fig. 4. Local alignment of cucurbit genomes with grape as reference genome. The graph shows details of a short segment of alignment shown in global alignment supplementary fig. S6, Supplementary Material online. Homologous block phylogeny (left): three paralogy chromosome segments in the grape genome, Grape-8, Grape-06, and Grape-13, from which ancestral chromosome affected by ECH, and each of them has two orthology cucurbit chromosome segments. Chromosome numbers are shown after the names of plants, and locations on chromosomes are also shown. A gene is shown by a rectangle. Homologous genes between neighboring chromosomal regions are linked with lines. Reconstructed ancestral chromosome segments, named with *Con-A, B, and C*, are displayed accordingly.

showed a varying retention rate 0–40%. Moreover, we found that the duplicated homoeologous regions could also be derived from two unbalanced fractionated subgenomes (supplementary fig. S8, Supplementary Material online). Taken watermelon as an example, along certain grape chromosomes, for example, chromosomes 1, gene retention shows a complement pattern, whereas, each of grape chromosomes 5 and 6 has two sets of watermelon chromosomal regions showing coordinating gene retention rates. The lower-fractionated watermelon subgenome preserved 3,121 genes, ∼60% than that (2,478) in the other or the higher-fractionated subgenome (*P*-value = 8.46*E*-18). Similar findings were observed in the other two cucurbits (supplementary tables S20–S23, Supplementary Material online). This supports the CCT harbored two divergent subgenomes, with a dominant one to preserve more genes and the other sensitive, showing an allopolyploidy nature. Besides, the same subgenome seems to have been divergently fractionated after the divergence of the three cucurbits, in that, for example, the same subgenome in watermelon has more genes removed

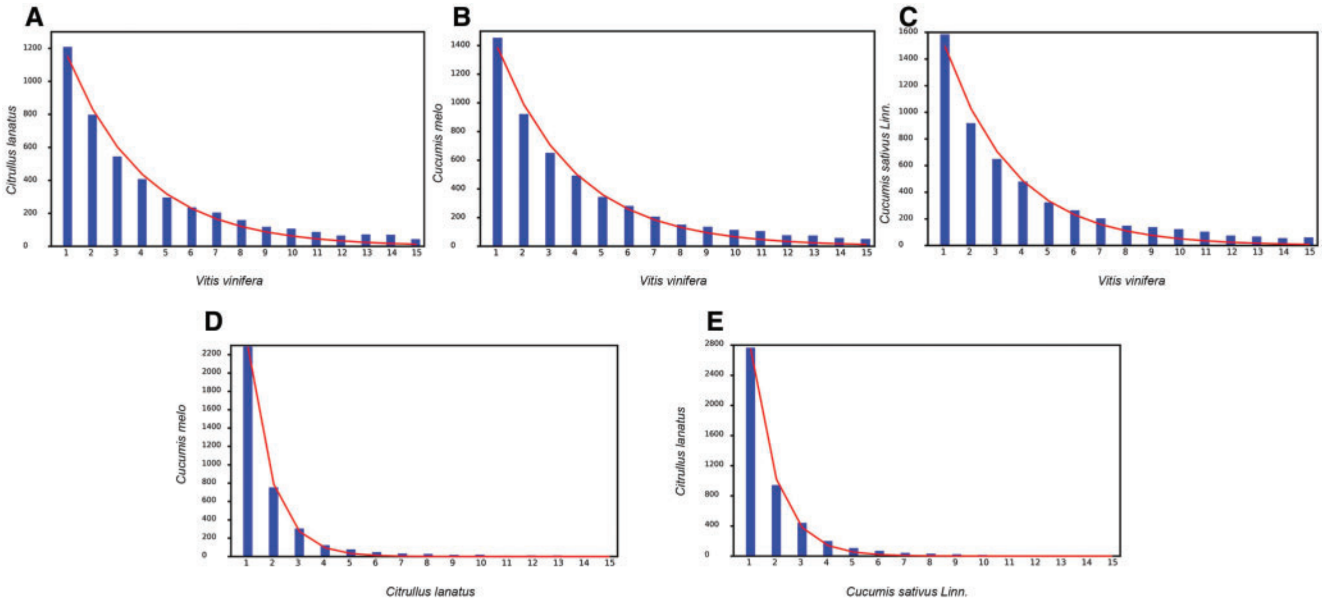

Fig. 5. Fitting a geometric distribution and gene loss rates in cucurbits as to the grape, and among cucurbit genomes. (*A–C*) Cucubitaceae with grape as reference genome; (*D*) Melon with watermelon as reference genome; (*E*) Watermelon with Cucumber as reference genome.

from it than in the other genomes (*P*-values = 4.68*E*-86 and 9.82*E*-17, respectively; supplementary tables S24 and S25, Supplementary Material online), consistent to the above finding of species-specific gene deletion.

To explore the mechanism underlying genomic fractionation, we characterized the runs of continual gene deletions in cucurbits as to the referenced genomes (Methods detailed in Wang et al. 2015). Though there are large patches of chromosomal segmental losses (supplementary figs. S8–S11, Supplementary Material online), most of the runs of gene deletions are of 15 continual genes or fewer. A statistical fitness regression shows that runs of deletions follow a near geometric distribution (fig. 5A–C; supplementary table S26, Supplementary Material online). With grape as a reference, three cucurbits have runs of gene deletion patterns following similar distributions (geometric parameter *P* = 0.276–0.313, the probability deleting one gene a time, and goodness of fitting *F*-test *P*-value 0.89–0.90 to accept the fitness). This shows that 42–44% of genes were deleted in runs containing 1 or 2 genes, indicating a mechanism of fractionation of short DNA segment removal, or ~5–10 kb DNA in length. A comparison within cucurbits themselves showed that although following geometric distribution, even more gene deletions were in short runs (fig. 5*D* and *E*), including 83–85% of deleted genes (supplementary figs. S10–S12, Supplementary Material online). Integrating the observation with grape and cucurbits as references, it seems that short deleting runs accounted the majority initially and then recursive deleting runs overlapping previous ones elongated the observed length of runs revealed with much diverged referenced genome.

Discussion

Plants harbor genomes that are much more complex than animals with regard to gene and genome duplication. This may be a result of natural selection that plants have to stand harsh environmental factors in a niche without any shelter. Therefore, adapting to changing environment, they have to develop relatively fast functional innovations within their genomes (Murat et al. 2010). Polyploidizations often produce thousands of duplicated genes, and numerous chromosomal rearrangements, even whole-genome repatterning, and subsequent DNA mutations, providing enormous genomic opportunities to suffice functional innovations (Soltis et al. 2015). The complexity of their genomes is often difficult to deconvolute, holding back efforts to understand the functional evolution of gene families, pathways, and chromosomal and genomic structures. The precious tools developed by scientists in other domains, such as animals, are often not applicable in plant genome analysis. A problematic explanation of a hard-won genome sequence would be a great pity in consideration of huge money and time invested.

The present effort established a gold-standard to analyze complex genomes. First, an outgroup reference genome, often relatively simple in structure, must be well selected. For eudicots, avoiding further polyploidization, grape is often taken as a reference, in that its genome structure is more similar to the common ancestor prior to the ECH event than many other sequenced dicots. For monocots, especially, grasses, we often take rice as a reference. Second, a two-way comparison through homologous gene dotplots between different genomes is indispensible. Only with two-way comparison can we distinguish orthologous and outparalogous homology and therefore infer paralogy depth in a considered or newly sequenced genome. The paralogy depth tells whether there is one or more polyploidizations after its split with the reference. The production of the homologous gene dotplots can use inferred colinear genes, or use BLAST output between two gene sets, or use that between the reference gene set and the newly sequenced genome. Third, sequence divergence between different paralogous and/or orthologous colinear regions, integratively shown in dotplots, can be used to help distinguish orthologous and outparalogous correspondence between the reference and the newly sequenced genome. Finally, integrating the information of orthology and paralogy, and information of colinear genes, whole-genome alignment of multiple genomes can be constructed, and the alignment lays a solid foundation of gene functional analysis in that it tells how duplicated genes were produced and what genes are colinearity-supported orthologous genes.

Here, we showed that cucurbit genomes were affected by a common tetraploidization but repeatedly overlooked by previous analysis. However, it is not the sole case that a plant genome was not well interpreted. Previously, a potato sequencing effort failed to show that it was affected by a hexaploidization rather than tetraploidization (Xu et al. 2011; International Tomato Genome Sequencing Consortium 2012), and a cotton sequencing effort failed to show that it was affected by a decaploidization rather than a tetraploidization (Paterson et al. 2012; Wang et al. 2012, 2016). These failures might be caused by insufficient usage of reference genomes and problematic analytical approaches. In certain cases, sequence divergence was often estimated with homologous genes, lacking of gene colinearity support. This would mix homologous genes produced by large-scale genomic duplications and small-scale ones together. Often homologous gene dotplotting was not involved to understand the complexity of genomes. As noted above, homologous gene dotplots are very important to distinguish homologous genes produced by recursive polyploidizations. Using homologous gene dotplots, one can compare the genome under study with a well-characterized reference genome, and ratio of best-matched homology between the genomes would tell how many events occurred, whether they were shared or not by the genomes. Besides, in the case of cucurbits, quite low rates (2.5–5.5%) of preserved CCT colinear genes in their genomes might also have led to the ignorance of the CCT event. Comparatively, both maize and soybean were affected by an extra tetraploidization, respectively, and much higher fractions (19.6% and 15.4%) of collinear genes were preserved in extant genomes (Wang, Wang, Jin, et al. 2015; Wang et al. 2017).

The split of Cucurtales and Fagales were inferred to have occurred ~109 Ma (Magallon et al. 2015). The CCT event was inferred to have occurred ~90–102 Ma, therefore it is much likely shared by most cucurbits, if not common to all Cucurtales plants. Actually, it has been realized that polyploidizations had contributed to originations and divergences of seed and flowering plants. Several major plant families were proposed to have their specific ancestral polyploidizations, including Poaceaes, Brassiceaes, Fabaceaes, Solanaceaee, etc., which much likely contributed to their originations and divergences (Young et al. 2011; International Tomato Genome Sequencing Consortium 2012; Chalhoub et al. 2014; Jiao et al. 2014; Wang, Wang, Guo, et al. 2015). The proposed CCT event, being very ancient as compared with polyploidizations in other plant families, might have also contributed to the origination and divergence of Cucurbitaceae plants, eventually establishing one of the largest plant families. Besides, the CCT and the above mentioned polyploidizations in other plant families, were most likely amphidiploid ones or allopolyploids, a result of hybridization between appreciably diverged species, indicating that hybrid vigor might have played a key role in the originations and establishments of plant families. Though we still do not know exactly what is the source of the vigor, fused genomes from two progenitors would transit and combine their adaptive capability to the hybrid offsprings, which harbor thousands of duplicated genes, even after severe gene losses and chromosomal rearrangements, likely rewiring the genetic networks to acquire biological vigor (Chen 2010; Ng et al. 2012; Sattler et al. 2016).

Above we characterized possible genomic fractionation, gene deletions and related modes breaking gene colinearity. We have to note here that missing correspondence (15–17%) in inferred colinearity regions between grape and cucurbit genomes might be caused by gene deletion or insertion reciprocally occurring in them. That is, the occurrence of gene insertion in grape genome may result in an overestimation of gene deletions in cucurbits. Supposing that the grape gene insertion occurred at an equal rate as to cucurbit gene deletion, which is not very likely due to high cucurbit gene loss level inferred above, the grape gene insertion should be much <7%. As to the reconstructed the ancestral genome before grape-cucurbit split, using grape-cucurbit gene colinearity, the inferred gene deletion rates were much lowered, which can lead to an underestimate of gene deletion in that some cucurbit-common ancestral genes not represented by the grape genome could be both deleted after the CCT but not reflected. Besides, the above findings show that cucurbit-specific gene deletion might have occurred in each plant after their splits. However, we have to note that genome assemblies and gene annotations are imperfect representations of the actual genomes, and problematic local ordering and orientation of scaffolds and contigs into chromosomes, chromosomal rearrangements can also disrupt the detection of synteny.

Materials and Methods

Materials

We downloaded genomic sequences and annotations from respective websites for each genome projects, for which complete information can be found in supplementary table S1, Supplementary Material online.

Gene Colinearity

With annotated genes as input, chromosomes from within a genome or between different genomes were compared. Firstly, by performing BLASTP (Altschul et al. 1990), protein sequences were searched against one another to find potentially homologous genes (*E*-value < 1e-5). A less stringent *E*-value may involve more diverged homologous genes. Gene colinearity, describing a batch of genes preserving ancestral gene order, would then complement this loose requirement of gene similarity to help identify very old evolutionary events without jeopardizing the effort here.

Secondly, the information about homologous genes was used as input for the software ColinearScan (Wang et al. 2006) to locate homologous gene pairs in colinearity Maximum gap between colinear genes were set to be 50 intervening genes. Large gene families with 30 or more copies in a genome were removed before inferring gene colinearity, as previously implemented in other angiosperm genomes (Wang, Wang, Guo, et al. 2015; Wang et al. 2016).

Construction of the Event-Related Colinear Gene Table

To construct the table with grape genome as reference (supplementary table S10, Supplementary Material online), all grape genes were listed in the first column. Each grape gene may have two extra colinear genes due to the hexaploidy, and we assign two other columns in the table to contain this information. For a grape gene, when there is a corresponding colinear gene in an expected location, a gene ID was filled in a cell of the corresponding column in the table. When it is missing, often due to gene loss or translocation in the genome, we fill in the cell a dot. For each of the three cucurbits, with the extra tetraploidy, we assign 3 × 2 columns. Therefore, the table has 21 columns, reflecting layers and layers of tripled and then doubled homology due to recursive polyploidies across four plants. The watermelon-referenced table was constructed similarly (supplementary table S11, Supplementary Material online).

Synonymous Substitutions

Synonymous nucleotide substitutions on synonymous sites (Ks) were estimated by using the Nei–Gojobori approach (Nei and Gojobori 1986) implemented by using the Bioperl Statistical module.

Kernel Function Analysis of Ks

Distributions of synonymous nucleotide substitutions on synonymous sites (Ks) of homologous genes could reflect multiple and overlapping genomic duplications, and speciations if the homologs are from different genomes. We adopted a kernel function analysis of Ks distribution of colinear homologs from within a genome, or between different genomes. A Ks distribution was viewed as a mix of multiple normal distributions. We used the kernel smoothing density function **ksdensity** (width is generally set to 0.05) in Matlab to estimate the probability density of each Ks list and obtain the density distribution curve. Then we performed the Gaussian multipeak fitting of the curve by using the gaussian approximation function **Gaussian** in the fitting toolbox **cftool**. We set *R*-squared, a parameter to evaluate the fitting goodness, to be at least 95%, used the smallest number of normal distributions to represent the complex Ks distribution, and the principle one was used to represent the corresponding evolutionary event.

Evolutionary Dating Correction

By aligning the peaks of the CCT from different Ks distributions to the corresponding location in the melon Ks distribution, we performed evolutionary rate correction. We suppose that the melon peak appears at $\mathrm{k_M}$ and for the other cucurbits, supposing that the peak appears at k_i the relative evolutionary rate of cucurbit i can be described with

$$r = (k_i - k_M)/k_M.$$

Then we perform rate correction to find the corrected rate $\mathrm{k}_{i-\mathrm{correction}}$ of the cucurbit *i* relative to k_M:

(1) For a specific cucurbit *i*, for the Ks between its duplicates, we can define a correction coefficient C_i as,

$$\frac{k_{i-\mathrm{correction}}}{k_i} = \frac{k_\mathrm{M}}{k_i} = C_i,$$

therefore, we get

$$k_{i-\mathrm{correction}} = \frac{k_\mathrm{M}}{k_i} \times k_\mathrm{i} = \frac{1}{1+r} \times k_i,$$

$$C_i = \frac{1}{1+r};$$

(2) For Ks between homologous genes from two nonmelon cucurbit i, j, if the peak was located at k_{ij}, we used the arithmetic mean of two correction coefficients in two genomes: $C_{ij} = \frac{C_i + C_j}{2}$, then we calculated a corrected evolutionary rate $k_{ij-\mathrm{correction}} = C_{ij} \times k_{ij}$;

(3) For Ks between homologous genes from melon and another cucurbit *i*, if the peak was located at k_{iM}, supposing the correction coefficient C_i in the cucurbit *i*, then we calculated a corrected evolutionary rate $k_{iM-\mathrm{correction}} = C_i \times k_{iM}$.

(4) For Ks between homologous genes from watermelon and melon-cucumber, if the peak was located at k_{w-mc}, supposing the correction coefficient C_w and C_c in watermelon and cucumber, then we calculated a corrected evolutionary rate

$$\begin{aligned} k_{w-\mathrm{mc}-\mathrm{correction}} &= k_{w-\mathrm{mc}} \times \left(\frac{C_w + C_c}{2} + C_w \right) \\ &= k_{w-mc} \times \left(\frac{3C_w + C_c}{2} \right). \end{aligned}$$

(5) For Ks between homologous genes from grape and cucurbit *i*, if the peak was located at k_{iV}, considering the effects of evolution from the ECH peak $k_{i-\mathrm{ECH}}$ to the CCT peak $k_{i-\mathrm{CCT}}$, then we calculated a corrected evolutionary rate

$$k_{iV} = k_{iV} \times \frac{k_{i-\mathrm{ECH}}}{k_{i-\mathrm{CCT}}}.$$

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We appreciate financial supported from the Ministry of Science and Technology of the People's Republic of China (2016YFD0101001); China National Science Foundation (31371282 to X.W. and 3151333 to J.W.); Natural Science

Foundation of Hebei Province (C2015209069 to J.W. and C2016209097 to W.G.); and Hebei New Century 100 Creative Talents Project, Hebei 100 Talented Scholars Project, and Tangshan Key Laboratory Project to X.W. We thank the helpful discussion with researchers at the iGeno Co. Ltd., China.

Author Contributions

X.W. conceived and led the research. J.W. implemented and coordinated the analysis. P.S., Y.L., N.Y., J.Y., X.M., S.S., Y.L., R.X., X.L., D.G., S.L., X.C., Y.L., Y.K., C.Z., Z.L., T.L., L.W., Z.W., W.G., L.Z., X.S., D.G., D.J., W.C., Y.P., G.Y., Y.X., J.S., C.Z., Z.L., H.X., X.D., S.S. performed the analysis. T.L. and T.L. contributed analyzing tools. Z.Z. and S.H. contributed in manuscript editing. X.W. and J.W. wrote the paper.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215(3):403–410.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930):433–438.

Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. 2014. Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica *napus* oilseed genome. *Science* 345(6199):950–953.

Chen ZJ. 2010. Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* 15(2):57–71.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Gonzalez VM, Henaff E, Camara F, Cozzuto L, Lowy E, et al. 2012. The genome of melon (*Cucumis melo* L.). *Proc Natl Acad Sci U S A.* 109(29):11872–11877.

Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat Genet.* 45(1):51–58.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet.* 41:1275–1281.

International Tomato Genome Sequencing Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–641.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161):463–467.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13(1):R3.

Jiao Y, Li J, Tang H, Paterson AH. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell.* 26(7):2792–2802.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97–100.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* 102(15):5454–5459.

Magallon S, Gomez-Acevedo S, Sanchez-Reyes LL, Hernandez-Hernandez T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* 207(2):437–453.

Murat F, Xu JH, Tannier E, Abrouk M, Guilhot N, Pont C, Messing J, Salse J. 2010. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* 20(11):1545–1557.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3(5):418–426.

Ng DW, Lu J, Chen ZJ. 2012. Big roles for small RNAs in polyploidy, hybrid vigor, and hybrid incompatibility. *Curr Opin Plant Biol.* 15(2):154–161.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A.* 101(26):9903–9908.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. 2012. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492(7429):423–427.

Sattler MC, Carvalho CR, Clarindo WR. 2016. The polyploidy and its key role in plant breeding. *Planta* 243(2):281–296.

Schaefer H, Heibl C, Renner SS. 2009. Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proc Biol Sci.* 276(1658):843–851.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A.* 108(10):4069–4074.

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev.* 35:119–125.

Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, Ruelens P, Maere S, Van de Peer Y, Geuten K. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol Biol Evol.* 29(12):3793–3806.

Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J. 2006. Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. *BMC Bioinformatics.* 7:447.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. *Nat Genet.* 43(10):1035–1039.

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat Genet.* 44(10):1098–1103.

Wang X, Wang J, Guo H, Lee T, Liu T, Jin D, Paterson AH. 2015. Genome alignment spanning major poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Mol Plant.* 8(6):14.

Wang X, Wang J, Jin D, Guo H, Lee TH, Liu T, Paterson AH. 2015. Genome alignment spanning major poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Mol Plant.* 8(6):885–898.

Wang X, Guo H, Wang J, Lei T, Liu T, Wang Z, Li Y, Lee TH, Li J, Tang H, et al. 2016. Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. *New Phytol.* 209(3):1252–1263.

Wang J, Sun P, Li Y, Liu Y, Yu J, Ma X, Sun S, Yang N, Xia R, Lei T, et al. 2017. Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol.* 174(1):284–300.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, et al. 2011. Genome sequence and analysis of the tuber crop potato. *Nature.* 475(7355):189–195.

Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, et al. 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480(7378):520–524.