Received: 20 June 2020 | Revised: 17 December 2020 | Accepted: 4 January 2021

DOI: 10.1002/ece3.7222

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

Which factors contribute most to genome size variation within angiosperms?

Dandan Wang | Zeyu Zheng | Ying Li | Hongyin Hu | Zhenyue Wang | Xin Du | Shangzhe Zhang | Mingjia Zhu | Longwei Dong | Guangpeng Ren | Yongzhi Yang

State Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology & School of Life Sciences, Lanzhou University, Lanzhou, China

Correspondence
Guangpeng Ren and Yongzhi Yang, State Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology & School of Life Sciences, Lanzhou University, Lanzhou, China.
Emails: rengp@lzu.edu.cn (GR); yangyongzhi2008@gmail.com (YY)

Funding information
Fundamental Research Funds for the Central Universities, Grant/Award Number: lzujbky-2019-77; National Natural Science Foundation of China, Grant/Award Number: 31900201 and 41901056; Second Tibetan Plateau Scientific Expedition and Research (STEP) program, Grant/Award Number: 2019QZKK0502

Abstract

Genome size varies greatly across the flowering plants and has played an important role in shaping their evolution. It has been reported that many factors correlate with the variation in genome size, but few studies have systematically explored this at the genomic level. Here, we scan genomic information for 74 species from 74 families in 38 orders covering the major groups of angiosperms (the taxonomic information was acquired from the latest Angiosperm Phylogeny Group (APG IV) system) to evaluate the correlation between genome size variation and different genome characteristics: polyploidization, different types of repeat sequence content, and the dynamics of long terminal repeat retrotransposons (LTRs). Surprisingly, we found that polyploidization shows no significant correlation with genome size, while LTR content demonstrates a significantly positive correlation. This may be due to genome instability after polyploidization, and since LTRs occupy most of the genome content, it may directly result in most of the genome variation. We found that the LTR insertion time is significantly negatively correlated with genome size, which may reflect the competition between insertion and deletion of LTRs in each genome, and that the old insertions are usually easy to recognize and eliminate. We also noticed that most of the LTR burst occurred within the last 3 million years, a timeframe consistent with the violent climate fluctuations in the Pleistocene. Our findings enhance our understanding of genome size evolution within angiosperms, and our methods offer immediate implications for corresponding research in other datasets.

KEYWORDS

angiosperm, genome size, long terminal repeat, polyploidization, repeat sequences

1 | INTRODUCTION

Genome size (also known as the C-value) refers to the total amount of DNA contained within one copy of a single complete genome; it is broadly constant within an organism (Greilhuber et al., 2005; Swift, 1950). More and more species' genome sizes have been assessed since early studies in the 1950s, covering more than 12,273 land plants, 6,222 animals, and 2,353 fungi (Gregory, 2015; Kullman

Wang and Zheng are equally contributed to this work.

et al., 2005; Pellicer & Leitch, 2020). Researchers have also discovered that eukaryotic genome size varies greatly over more than a 100,000-fold range and fails to correlate well with apparent complexity; this is the well-known "C-value paradox" (Eddy, 2012; Thomas Jr, 1971). Among the most widely studied land plants, angiosperms (10,770 species searched) exhibit an astonishing diversity of genome size, with a maximum variation by a factor of approximately 2,440 (Leitch et al., 2019; Pellicer et al., 2018), that is, the smallest angiosperm plant genome reported so far is *Genlisea tuberosa* (Lentibulariaceae, 61 Mb/1C) (Fleischmann et al., 2014), a carnivorous angiosperm endemic to Brazil, and the largest is *Paris japonica* (Pellicer et al., 2010), a monocot lily species in the Melanthiaceae family with an astonishingly large genome made up of ca. 149,000 Mb/1C of DNA. Furthermore, the dramatic variation in genome size can occur even among congeners. For example, the variation in genome size can reach ~30-fold in Brassicaceae (0.16–4.63 Gb), ~37-fold in Rosaceae (0.10–3.57 Gb), and ~44-fold in Asteraceae (0.39–25.60 Gb) (Leitch et al., 2019).

Several mechanisms have been proposed to account for the variation in genome size, such as recombination rate, tandem repeats (Tiley & Burleigh, 2015), transposable elements (TEs), and polyploidization, but the relative contribution of these different mechanisms seems to vary between species (Bennetzen et al., 2005). Polyploidization can directly increase the genome size by doubling all the genome contents, and this occurs widely within angiosperms. With the exception of *Amborella*, nearly all the angiosperms have undergone polyploidization events, and the different major lineages (i.e., Ceratophyta, eudicots, monocots, magnolia, and Nymphaeales) have all experienced independent polyploidization events (Initiative, 2019; Van de Peer et al., 2017; Yang et al., 2020). Repeated DNA sequences account for the majority of the genomic DNA in most plant species, occurring in a few to millions of copies. The content of repeated sequences shifts significantly across plant genomes. It can be as low as ~3%, for example, in *Utricularia gibba* (Ibarra-Laclette et al., 2013), and is as high as ~85% in *Zea mays* (Schnable et al., 2009). Among the repetitive sequences, tandem repeats usually occupy a small proportion of the genome and the main repeats fall into four types of transposable elements (TEs): long terminal repeat elements (LTRs), long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and DNA transposon repeat elements (DNA transposons). Among the different types of TEs, LTRs usually occupy the largest proportion of plant genomes and dynamic bursts have acted as major contributors to the genome size differences between plants (Lee & Kim, 2014).

With the development of high-throughput sequencing technologies, more and more angiosperm genomes have been sequenced, assembled, and made publicly available (https://www.plabipd.de/), providing an opportunity to investigate the variation in genome size within angiosperms systematically. Here, we scan the genome sizes of 74 flowering plant genomes from 74 families covering 38 orders (taxonomic foundation sourced from APG IV) and evaluate the correlation between the genome size and three factors: polyploidization, the proportion of repetitive elements, and LTR activity. Based on a series of correlation analyses, we explore which factor is mainly responsible for the genome size variation in angiosperms.

2 | MATERIALS AND METHODS

2.1 | Genome datasets collection

In this study, we sampled 74 species, the genomes of which were derived from previous research, in 74 families representing 38 orders. This dataset included genomes from NCBI, Ensembl Plants, and many other individual genome databases, such as the Herbal Medicine Omics Database, gigaDB datasets, and the *Panax notoginseng* Genome Database. The 74 plant genomes were sampled from 38 diverse orders of five main taxa among the angiosperms. Detailed information about these 74 species and their data sources is presented in Table S1.

2.2 | Repeat sequence identification

To check whether certain types of repeat sequence or whole-genome duplications may have caused the variation in angiosperm genomes, we examined the duplicated genes of 74 species separately and identified whole-genome duplication events from published literature (Table S3). The different kinds of duplicated genes were identified using different pipelines. Tandem repeats, which include minisatellites, microsatellites, and others, divided by nucleotide length were identified using Tandem Repeats Finder v4.90 (Benson, 1999), while transposable elements were identified using RepeatMasker and RepeatModeler. The TEs were identified using a combination of Repbase (Bao et al., 2015) and the de novo prediction results of RepeatModeler. We then used perl script to calculate the proportion of different types of repetitive elements in the genome. The entire pipeline was deposited at Github (https://github.com/dandanWang2019/genome_size_pipeline).

2.3 | Polyploidization fold assessment

The polyploidization fold (PF) was calculated by the formula: $PF = 2^m \times 3^n$, where m refers to the number of times of the whole-genome duplication events and n refers to the number of times of the triplication events, with data sourced from the literature. Because of common duplications, ancient polyploidization events in angiosperms were not taken into account (Soltis et al., 2003).

2.4 | LTR insertion date calculation

Insertion dates were calculated following the methods in the published literature (SanMiguel et al., 1998). The downloaded genomes were scanned using LTR_Finder (Xu & Wang, 2007), and full-length LTRs were extracted by perl scripts. LTR 5′ and 3′ pairs were aligned with MUSCLE

(Edgar, 2004) and ClustalW2 (Larking et al., 2007), and the divergence between LTR pairs was calculated in PHYLIP v3.696. The insertion time of each LTR was estimated in millions of years using the formula: $T = K/2r$ ($r = 1.3 \times 10^{-8}$ per site and per year) (Ma & Bennetzen, 2004), where K refers to nucleotide substitution rates and the arithmetic mean of insertion time was calculated for each species in millions of years.

2.5 | Correlation analysis

Nuclear genome size estimates were determined through scripts from the downloaded genome file and scaled by ancestral haploid genome size of angiosperms (1.73 pg × 978 Mb/pg = 1691.94 Mb) (Carta et al., 2020). The regressions were performed on the proportions of repetitive elements, polyploidization fold, and mean LTR insertion time against genome size fold. The correlation between potentially related factors and genome size fold was calculated using the R lme4 package (Bates et al., 2015). To consider the possible roles of divergence time in the relationship between LTR abundance and insertion time, we conducted a multiple regression with age and insertion time as predictors of abundance (Figure 3h and Table S4). We also analyzed the associations between the factors and genome size fold in a phylogenetic context. The phylogenetic tree was acquired from a recent angiosperm phylogeny study (Li et al., 2019) and pruned with Newick Utilities v1.6.0 (Junier & Zdobnov, 2010). The fitting of a PGLS model in a phylogenetic context with Brownian motion was conducted using the gls function from the nlme package (Pinheiro et al., 2017). All correlation analysis results and phylogenetic trees used for PGLS analysis are presented in the supplementary files (Table S5 and Figure S3). Results were considered significant when $p < .05$.

3 | RESULTS

3.1 | Genome collection and repeat sequences identification

The genome assemblies of 74 flowering plant genomes from 74 families of 38 orders were collected from the NCBI Genome database, GigaDB, and other specific databases (Table S1). Our selected genomes covered the major groups of angiosperms, including two basal angiosperms, twelve monocots, three magnolias, one Ceratophyllale, and 56 eudicots (Table S1). A 30-fold variation in genome size was detected within these 74 species ranging from ~100 Mb (*U. gibba*) to ~ 3,027 Mb (*Helianthus annuus*), with an average of 730.0 Mb and a median value of 566.9 Mb; a genome size of over 1,000 Mb was identified in more than 15 species (Figure 1a and Table S2). A standard method was adopted to annotate the repeat sequences within each genome, and we found that repeats make up a large proportion of the genome in all species ranging from 21.59% in *Spirodela polyrhiza* to 83.23% in *Z. mays* (Table S2). *U. gibba* was estimated to contain 31.81% repeats and the result is higher than in a previous study (Ibarra-Laclette et al., 2013). This may be attributed to the growing number of recognizable repeats and the integration of different types of software in this analysis (Table S2). Transposable elements were the major components of repeats rather than tandem repeats, and the four main types of TEs varied within different species. LTRs were the dominant TE type in the 72 species, ranging from 9.49% to 81.76%, and the two species with the most abundant LTRs were *Z. mays* (81.76%) and *Cephalotus follicularis* (80.61%). Surprisingly, in *Ceratophyllum demersum*, LINEs were the dominant components, accounting for 33.60% of the genome, while LTRs accounted for 30.76%. In *Tripterygium wilfordii*, tandem repeats accounted for 22.54% of the genome, exceeding the 14.88% of LTRs. The two species with the most abundant DNA transposons were *Trichopus zeylanicus* (27.29%) and *Kobresia littledalei* (19.71%), while SINEs contributed very little to the genome size in any of the species (≤0.88%) (Figures 2 and S1; Table S2).

3.2 | LTRs insertion time

As the main component of repeats, we further explored the LTR activity in relation to LTR insertion time. We found a large proportion of the estimated mean LTR insertions occurred recently, with an average insertion time of 2.42 million years ago (Mya) and a median value of 2.40 Mya (Figure 1b). *Elaeis guineensis*, *Argania spinose*, *Carica papaya*, *A. trichopoda*, *Carnegiea gigantean*, and *Populus trichocarpa* had mean LTR insertion times greater than 4 Mya (4.0–4.24 Mya), and all contained a small proportion of LTRs (25.57%–47.41%). In contrast, the younger LTR insertions occurred typically in species that had a relatively high percentage

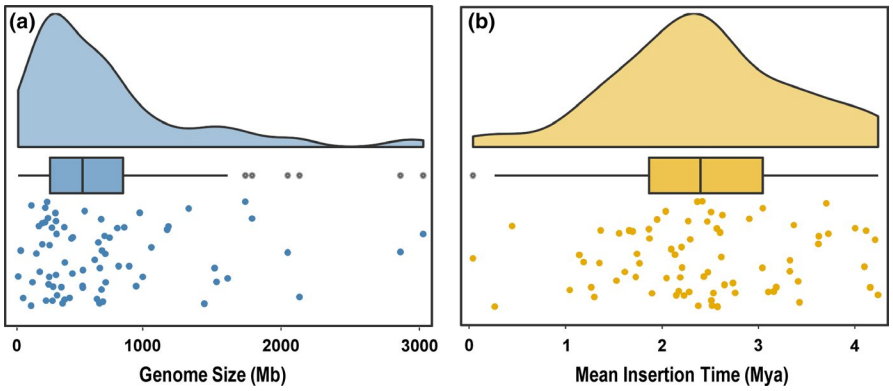

FIGURE 1 The distributions of plant genome sizes (a) and mean LTR insertion times (b). The density curves represent the distribution, while the scatter diagram and the box-plot show the statistics including median, quartile and the outliers

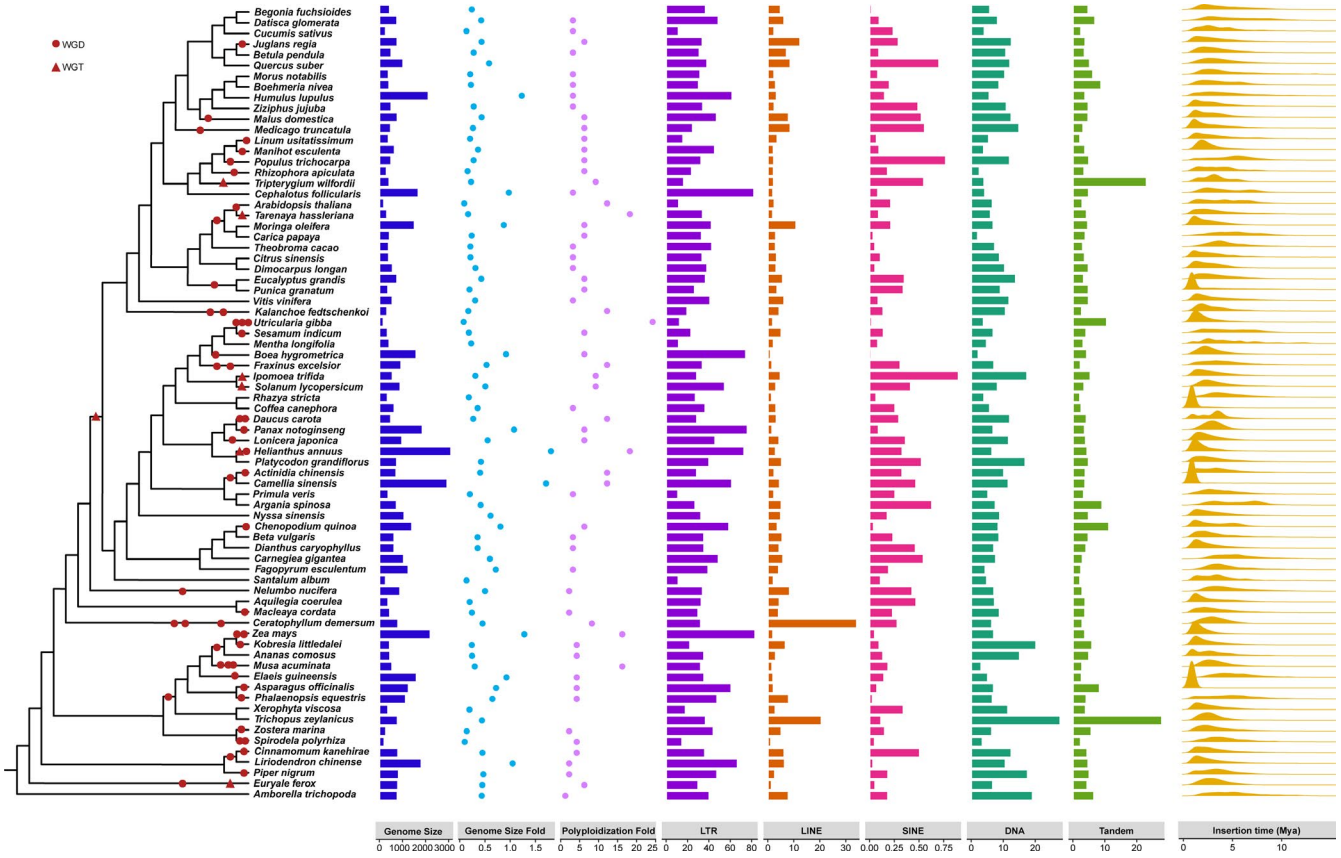

FIGURE 2 Representation of phylogeny and the correlation factors analyzed in 74 genomes. GF: genome size fold, PF: polyploidization fold, tandem: tandem repeats, and LTR insertion dates. GF indicates the genome size fold in plants scaled by the ancestral genome size for angiosperms and PF indicates the value of polyploidization fold which is the number of times that whole-genome duplication and the whole-genome triplication occurred. LTR-tandem indicates different proportions of corresponding repeating elements in genomes as a percentage (%). Mean insertion date indicates the estimated distribution of LTR insertion dates in plants in millions of years. The WGDs and WGTs are labeled in the branches. The topology information cited is from Li et al. (2019)

of LTRs, such as *Asparagus officinalis* (59.26%; mean LTR insertion time: 0.04 Mya) and *Camellia sinensis* (59.87%; mean LTR insertion time: 0.26 Mya) (Figures 2 and S2).

3.3 | Polyploidization event characteristics

As complex and uncertain polyploidization events occurred within ancestral seed plants and angiosperms over a very long time (>200 Mya) (Van de Peer et al., 2017), it is difficult to confirm the real polyploizidization fold (PF) of each species. Here, we assumed the polyploidization fold of *A. trichopoda* to be 1, experiencing only ancient polyploidization events that occurred in the ancestral angiosperms. We collected information on all polyploidization events reported in published papers for each species and scaled the number in relation to *A. trichopoda* (Table S3). Within our dataset, all the species had a value larger than 1, as all the angiosperms except for *A. trichopoda* have experienced additional WGD events. *U. gibba*, with the smallest genome size, had the highest polyploidization value of 24 (Figure 2 and Table S3).

3.4 | Correlation between the factors analyzed and genome size

A series of factors including the proportion of repetitive elements, polyploidization fold, and mean LTR insertion time were examined in the correlation analysis. To ensure phylogenetic independence, we also constructed phylogenetic generalized least-square models (PGLS) to fit the data and the results remained similar, with the exception of the mean LTR insertion time (Figure 3). Across the five repeat elements examined, a strong significant positive correlation was only observed between genome size and the proportion of LTR elements (Figure 3). Species with larger genomes had a relatively larger proportion of LTRs. This was evidenced by the linear regression comparing the proportion of LTR against genome size fold ($R^2 = 0.646$, $y = 0.0169x–0.1628$, $p < 2.0 \times 10^{-16}$; Figure 3a).

With respect to LTR activity, we also considered the effects of divergence times in the relationship between LTR abundance and insertion time. The results of the linear regression showed that LTR insertion time was negatively correlated with genome size fold

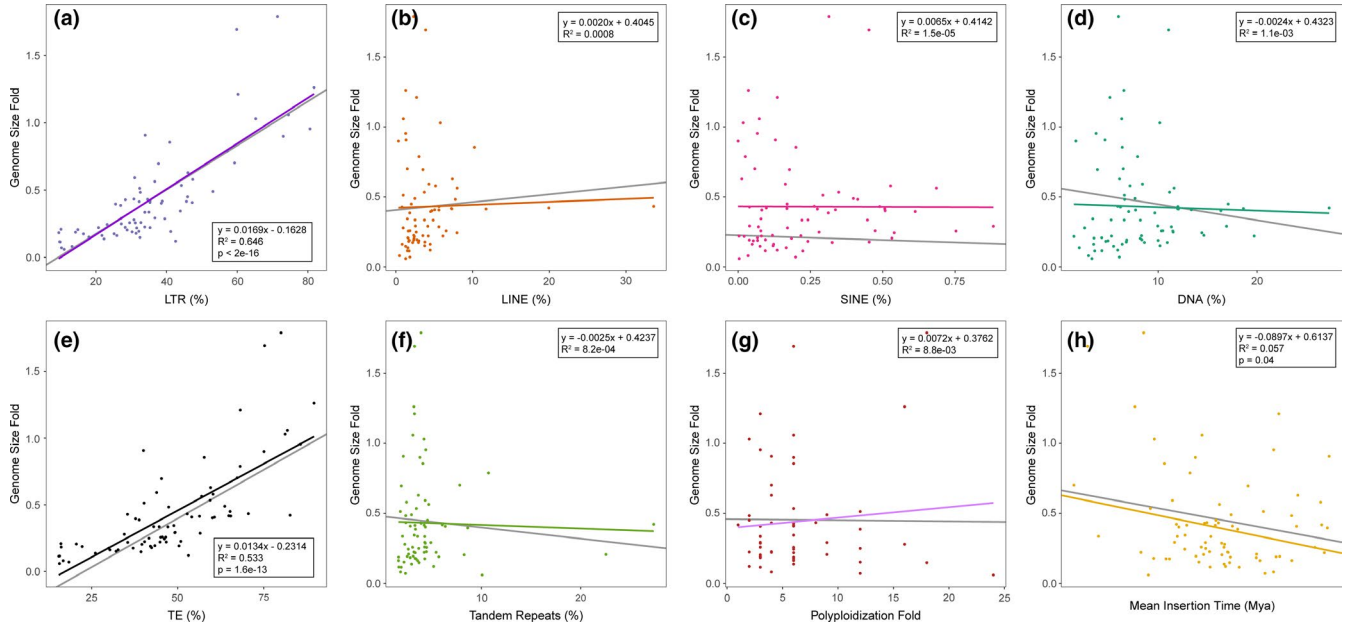

FIGURE 3 Different factors (in colors) as a function of genome size. Different factors fitted against genome size fold (genome size scaled by ancestral genome size for angiosperms). The gray lines represent the estimated result from phylogenetic least squares (PGLS) analysis. (a-f) The relationship between the proportion of different repeated elements against genome size fold in 74 species. (g) The absence of a relationship between polyploidization fold and genome size fold. (h) The mean date of LTR insertions was significantly correlated with genome size fold. Lines are plots of linear regressions

($R^2 = 0.056$, $y = -0.08581x + 0.4080$, $p = .04$; Figure 3h). Species with large genomes consistently have relatively recent mean LTR insertion times; for example, *Asparagus officinalis* and *Camellia sinensis*, with the youngest mean LTR insertion times, had comparatively larger genomes, whereas LTR insertions generally occurred earlier in species with smaller genomes like *Amborella trichopoda*, *C. papaya* and *P. trichocarpa* (Figure 2 and Table S2). Nevertheless, when taking account of phylogenetic nonindependence, the correlation was no longer observed. In addition, to investigate whether polyploidy, as a crucial driving force, also affected the genome size holistically, we calculated the correlation between polyploidization fold and genome size fold. We found that polyploidization fold was not related to the variation in genome size fold ($R^2 = 0.0088$; Figure 3g).

4 | DISCUSSION

4.1 | Absence of a relationship between polyploidization and genome size

From our broad perspective, we were surprised to find that genome size is not significantly correlated to polyploidization, even though the latter is widely known to increase genomes by the inheritance of an additional set (or sets) of chromosomes (Bruggmann et al., 2006; Iorizzo et al., 2016). Multiple ancient polyploidy events occurred in plants around 100 to 120 million years ago and after that relatively recent WGDs occurred in many lineages during the evolution of angiosperms (Fawcett & Van de Peer, 2010; Wu et al., 2020). Those polyploidization helped flowering plants improved acclimatization during severe environmental changes and survival until now (Fawcett & Van de Peer, 2010; Zhang et al., 2020). While, a larger genome comes with the ecological burden of needing more macronutrients to build nucleic acids, particularly nitrogen and phosphorus, with the latter being limited in numerous natural environments (Šmarda et al., 2013; Vitousek et al., 2010). This cost usually varies greatly at different stages as environment changes. For example, CO2 has often been considered to have a dominant role in plant survival as the potentially limiting photosynthetic resource (Boyce & Zwieniecki, 2012), and the atmospheric CO_2 concentrations have fluctuated greatly over the past 400 million years (Rothman, 2002). The CO_2 content in the atmosphere during 100–120 million years ago was much higher, and in the last few million years, it showed a significantly decline (Boyce & Zwieniecki, 2012; Foster et al., 2017), which resulting in an increase in the cost to angiosperms of recent polyploidization (Rothman, 2002). In other words, polyploidization expands genome size in a short period accompanying greater environmental pressure and nutritional needs and maintaining a large genome usually collapsed when external resources in the environment become tense. Thus, diploidization usually follows polyploidization, especially within angiosperms (Dodsworth et al., 2016; Meudt et al., 2015). Diploidization involves the removal of extra DNA (often repetitive DNA) and extraneous gene copies and occurs through recombination-based deletion and other mechanisms, while retaining duplicated genes, some of which may have new or altered functions (Adams & Wendel, 2005; Dodsworth et al., 2016). Diploidization can also downsize the genome by chromosome number reduction,

which potentially involves complex chromosomal rearrangements (including fusions and fissions) (Dodsworth et al., 2016; Franzke et al., 2011; Meudt et al., 2015). Diploidization is considered the key to the evolutionary success of angiosperms, and it has resulted in irregular genome reduction, which explains why polyploidization did not exhibit a significant positive linear correlation with genome size.

4.2 | Effects of TEs especially LTRs on genome size variation

TEs accounted for the most genome content and contributed the most to the genome size variation (Figure 3e). Previous studies have attributed the bigger genome to long-term amplification of TEs, which is associated with a naturally occurring reduction in the efficiency of symmetric DNA methylation in *Arabis alpine* (Willing et al., 2015), and the reduced quantity of small RNAs associated with TE silencing in *Picea abies* (gymnosperms) (Nystedt et al., 2013). In our study, we also found the TEs, especially LTRs exhibited the most significant positive correlation with the genome size variation (Figure 3a, e). So, another reason may be that polyploidization could also induce the activity and burst of TEs, which further diluted the influence of a linear correlation between polyploidization and genome size. Polyploidization usually causes chromatin modifications and epigenetic regulation to accumulate more TEs and produce a bigger genome (McClintock, 1984; Springer et al., 2016; Vicient & Casacuberta, 2017). For example, a widespread DNA methylation variation in TEs was observed in autotetraploid rice and was accompanied by changes in the abundance of 24-nucleotide small interfering RNAs (siRNAs) (Zhang et al., 2015), and demethylation of TEs has been observed in newly formed allopolyploids (Parisod et al., 2009; Yaakov & Kashkush, 2011).

Besides polyploidization, many other variables could also lead to TE bursts and cause changes in genome size; these include abiotic stress, domestication, and the mating system changes (Belyayev, 2014). In a natural population, stress-induced bursts of TEs, especially driven by environmental changes, are important and of special interest because this phenomenon may underlie micro- and macro-evolutionary events and ultimately support the generation and maintenance of biological diversity. We found a burst of LTR insertions mainly in comparatively recent times (<3 Mya, Figure S2 and Table S2), which is likely to have increased plant resistance to the violently fluctuating climate during the Early Pleistocene cooling (Hofreiter & Stewart, 2009; Xu et al., 2018). We also found that the mean insertion time showed a slightly negative correlation with genome size variations (Figure 3h), which is different from previous studies (Nystedt et al., 2013; Willing et al., 2015). This weak negative correlation may be caused by the competition between TE insertion and elimination. Plant genomes have experienced multiple rounds of TE outbreaks in their evolutionary histories, leading abundant TE families to escape from silencing mechanisms (El Baidouri & Panaud, 2013; Fultz et al., 2015; Lisch & Slotkin, 2011). However, as the genome tends to be stable, most TEs are eliminated and only some TEs are able to combat this with silencing, by inactivating the systems that have evolved to recognize them (Fu et al., 2013; McCue et al., 2013). So the ancient TEs usually account for a small proportion of the genome and the recent TEs are mainly responsible for the genome size (Divashuk et al., 2020; Oliver et al., 2013).

4.3 | Adaptation of flowering plants to the environment through genome size variation

Genome size is generally considered to be an evolutionary character, indicating that any change is not a random event, but usually a response to external environmental fluctuations (artificial or natural) (Levin, 2002; Pellicer et al., 2018). Whole-genome duplications and LTR insertions increase the biological complexity and size of the genome, generating novel functions, and altering gene expression patterns. This allows plants to adapt to the environment more easily (Oliver et al., 2013, Van de Peer et al., 2017). Thanks to the polyploidization that was closely associated with complicated climate changes, plants have survived for a long time even in the face of the severe environmental conditions, while retaining certain gene duplicates (Cai et al., 2019; Wu et al., 2020). The insertion of LTRs has been concentrated in the last million years when there have been drastic global environmental changes, indicating their important role in plant survival. This potentially accounts for the extreme diversity in angiosperms compared with the sister clade, gymnosperms, with low LTR activity, but abundant TEs (Kovach et al., 2010; Oliver et al., 2013).

We also found that, when faced with similar environmental conditions, plants may respond in different ways. Within the aquatic plants *Ceratophyllum demersum* and *Euryale ferox*, the proportion of repeats differs greatly: LINE and LTR are the dominant TE types, respectively. In the carnivorous plants *C. follicularis* and *U. gibba*, not only does the proportion of the repetitive elements vary greatly, but this is also the case for the frequency of whole-genome duplication events. *C. follicularis*, with a high proportion of repeated elements, experienced a round of WGT, while *U. gibba* with a low proportion experienced three rounds of whole-genome duplication events and a whole-genome triplication event. Apparently, adaption through different TEs and polyploidization has helped the angiosperms to develop unique modus vivendi, resulting in the survival of a range of taxa. In spite of the fact that diverse strategies may be adopted among species, they still have to confront the same circumstances.

In summary, we systematically scanned 74 species belonging to 74 families from 38 orders, covering the major groups of angiosperms. We performed correlation analysis to compare genome size and polyploidization, different repeat content and LTR insertion times. Our results have enhanced our understanding of genome size variation within angiosperms, and our pipeline will also be of use in future studies examining genome size evolution.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 31900201 and No. 41901056), the

Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2019-77), and the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502). All the computation works were supported by Supercomputing Center of Lanzhou University and the Big Data Computing Platform for Western Ecological Environment and Regional Development.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTION

Dandan Wang: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (lead); Writing-original draft (equal). **Zeyu Zheng:** Data curation (equal); Formal analysis (equal); Methodology (supporting). **Ying Li:** Data curation (equal); Formal analysis (equal); Writing-review & editing (equal). **Hongyin Hu:** Data curation (equal); Formal analysis (equal). **Zhenyue Wang:** Data curation (equal); Investigation (equal); Methodology (equal). **Xin Du:** Data curation (equal); Methodology (equal). **Shangzhe Zhang:** Data curation (equal); Formal analysis (equal); Investigation (equal). **Mingjia Zhu:** Data curation (equal); Methodology (equal). **Longwei Dong:** Conceptualization (equal); Methodology (equal); Software (equal). **Guangpeng Ren:** Conceptualization (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Resources (equal); Writing-review & editing (equal). **Yongzhi Yang:** Conceptualization (equal); Funding acquisition (equal); Investigation (supporting); Methodology (supporting); Project administration (equal); Resources (equal); Writing-original draft (supporting); Writing-review & editing (supporting).

DATA AVAILABILITY STATEMENT

All the repeat elements annotation and LTR sequence insertion times are available from figshare (https://doi.org/10.6084/m9.figshare.12514085.v3).

ORCID

Shangzhe Zhang https://orcid.org/0000-0003-2064-4287
Yongzhi Yang https://orcid.org/0000-0001-6912-6718

REFERENCES

Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, *8*, 135–141.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*, 11.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., & Bolker, M. B. (2015). Package 'lme4'. *Convergence*, *12*, 2.

Belyayev, A. (2014). Bursts of transposable elements as an evolutionary driving force. *Journal of Evolutionary Biology*, *27*, 2573–2584.

Bennetzen, J. L., Ma, J., & Devos, K. M. J. A. O. B. (2005). Mechanisms of recent genome size variation in flowering plants. *Annals of Botany*, *95*, 127–132.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*, 573–580.

Boyce, C. K., & Zwieniecki, M. A. (2012). Leaf fossil record suggests limited influence of atmospheric CO2 on terrestrial productivity prior to angiosperm evolution. *Proceedings of the National Academy of Sciences*, *109*, 10403–10408.

Bruggmann, R., Bharti, A. K., Gundlach, H., Lai, J., Young, S., Pontaroli, A. C., Wei, F., Haberer, G., Fuks, G., & Du, C. (2006). Uneven chromosome contraction and expansion in the maize genome. *Genome Research*, *16*, 1241–1251.

Cai, L., Xi, Z., Amorim, A. M., Sugumaran, M., Rest, J. S., Liu, L., & Davis, C. C. (2019). Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytologist*, *221*, 565–576.

Carta, A., Bedini, G., & Peruzzi, L. (2020). A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytologist*, *228*, 1097-1106.

Divashuk, M. G., Karlov, G. I., & Kroupin, P. Y. (2020). Copy number variation of transposable elements in Thinopyrum intermedium and its diploid relative species. *Plants*, *9*, 15.

Dodsworth, S., Chase, M. W., & Leitch, A. R. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society*, *180*, 1–5.

Eddy, S. R. (2012). The C-value paradox, junk DNA and ENCODE. *Current Biology*, *22*(21), R898–R899. https://doi.org/10.1016/j.cub.2012.10.002

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797.

el Baidouri, M., & Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biology Evolution*, *5*, 954–965.

Fawcett, J. A., & van de Peer, Y. (2010). Angiosperm polyploids and their road to evolutionary success. *Trends in Evolutionary Biology*, *2*, e3–e3.

Fleischmann, A., Michael, T. P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E. M., Greilhuber, J., MüLLER, K. F., & Heubl, G. (2014). Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany*, *114*, 1651–1663. https://doi.org/10.1093/aob/mcu189

Foster, G. L., Royer, D. L., & Lunt, D. J. (2017). Future climate forcing potentially without precedent in the last 420 million years. *Nature Communications*, *8*, 14845.

Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A., & Mummenhoff, K. (2011). Cabbage family affairs: The evolutionary history of Brassicaceae. *Trends in Plant Science*, *16*, 108–116.

Fu, Y., Kawabe, A., Etcheverry, M., Ito, T., Toyoda, A., Fujiyama, A., Colot, V., Tarutani, Y., & Kakutani, T. (2013). Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. *The EMBO Journal*, *32*, 2407–2417.

Fultz, D., Choudury, S. G., & Slotkin, R. K. (2015). Silencing of active transposable elements in plants. *Current Opinion in Plant Biology*, *27*, 67–76.

Gregory, T. R. (2015). *Animal Genome Size Database*. 2015. Retrieved from http://www.genomesize.com

Greilhuber, J., Doležel, J., Lysak, M. A., & Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms 'genome size'and 'C-value'to describe nuclear DNA contents. *Annals of Botany*, *95*, 255–260.

Hofreiter, M., & Stewart, J. (2009). Ecological change, range fluctuations and population dynamics during the Pleistocene. *Current Biology*, *19*, R584–R594.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A. J., Juárez, M. J. A., & Simpson, J. (2013). Architecture and evolution of a minute plant genome. *Nature*, *498*, 94–98.

Initiative, O. T. P. T. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, *574*, 679.

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., Bowman, M., Iovene, M., Sanseverino, W., & Cavagnaro, P. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics*, *48*, 657.

Junier, T., & Zdobnov, E. M. (2010). The Newick utilities: High-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, *26*, 1669–1670.

Kovach, A., Wegrzyn, J. L., Parra, G., Holt, C., Bruening, G. E., Loopstra, C. A., Hartigan, J., Yandell, M., Langley, C. H., & Korf, I. (2010). The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics*, *11*, 420.

Kullman, B., Tamm, H., & Kullman, K. (2005). *Fungal genome size database*.

Larking, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, G., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., & Lopez, R. (2007). ClustalW and ClustalX version 2. *Bioinformatics*, *23*, 2947–2948.

Lee, S.-I., & Kim, N.-S. (2014). Transposable elements and genome size variations in plants. *Genomics Informatics*, *12*, 87.

Leitch, I., Johnston, E., Pellicer, J., Hidalgo, O., & Bennett, M. (2019). *Angiosperm DNA C-values Database, Release 9.0, April 2019*.

Levin, D. A. (2002). *The role of chromosomal change in plant evolution*. Oxford University Press.

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., Gitzendanner, M. A., Fritsch, P. W., Cai, J., & Luo, Y. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants*, *5*, 461–470.

Lisch, D., & Slotkin, R. K. (2011). Strategies for silencing and escape: the ancient struggle between transposable elements and their hosts. *International review of cell and molecular biology*. Elsevier.

Ma, J., & Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences*, *101*, 12404–12410.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, *226*, (4676), 792–801. http://dx.doi.org/10.1126/science.15739260.

McCue, A. D., Nuthikattu, S., & Slotkin, R. K. (2013). Genome-wide identification of genes regulated in trans by transposable element small interfering RNAs. *RNA Biology*, *10*, 1379–1395.

Meudt, H. M., Rojas-Andrés, B. M., Prebble, J. M., Low, E., Garnock-Jones, P. J., & Albach, D. C. (2015). Is genome downsizing associated with diversification in polyploid lineages of Veronica? *Botanical Journal of the Linnean Society*, *178*, 243–266.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., & Alexeyenko, A. (2013). The Norway spruce genome sequence and conifer genome evolution. *Journal of Nature*, *497*, 579–584.

Oliver, K. R., McComb, J. A., & Greene, W. K. (2013). Transposable elements: Powerful contributors to angiosperm evolution and diversity. *Genome Biology Evolution*, *5*, 1886–1901.

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M. A., & Ainouche, M. (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. *New Phytologist*, *184*, 1003–1015.

Pellicer, J., Fay, M. F., & Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, *164*, 10–15.

Pellicer, J., Hidalgo, O., Dodsworth, S., & Leitch, I. J. (2018). Genome size diversity and its impact on the evolution of land plants. *Genes*, *9*, 88.

Pellicer, J., & Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. *New Phytologist*, *226*, 301–305. https://doi.org/10.1111/nph.16261

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., Heisterkamp, S., van Willigen, B., & Maintainer, R. (2017). *Package 'nlme'. Linear and nonlinear mixed effects models, version, 3*.

Rothman, D. H. (2002). Atmospheric carbon dioxide levels for the last 500 million years. *Proceedings of the National Academy of Sciences*, *99*, 4167–4171.

Sanmiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., & Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nature Genetics*, *20*, 43–45.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., & Graves, T. A. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, *326*, 1112–1115.

Šmarda, P., Hejcman, M., Březinová, A., Horová, L., Steigerová, H., Zedek, F., Bureš, P., Hejcmanová, P., & Schellberg, J. J. N. P. (2013). Effect of phosphorus availability on the selection of species with different ploidy levels and genome sizes in a long-term grassland fertilization experiment. *New Phytologist*, *200*, 911–921. https://doi.org/10.1111/nph.12399

Soltis, D. E., Soltis, P. S., Bennett, M. D., & Leitch, I. J. (2003). Evolution of genome size in the angiosperms. *American Journal of Botany*, *90*, 1596–1603.

Springer, N. M., Lisch, D., & Li, Q. (2016). Creating order from chaos: Epigenome dynamics in plants with complex genomes. *The Plant Cell*, *28*, 314–325.

Swift, H. (1950). The constancy of desoxyribose nucleic acid in plant nuclei. *Proceedings of the National Academy of Sciences of the United States of America*, *36*, 643.

Thomas, C. A. Jr (1971). The genetic organization of chromosomes. *Annual Review of Genetics*, *5*, 237–256.

Tiley, G. P., & Burleigh, J. G. (2015). The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. *BMC Evolutionary Biology*, *15*, 194.

van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*, 411.

Vicient, C. M., & Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, *120*, 195–207.

Vitousek, P. M., Porder, S., Houlton, B. Z., & Chadwick, O. A. (2010). Terrestrial phosphorus limitation: Mechanisms, implications, and nitrogen–phosphorus interactions. *Ecological Applications*, *20*, 5–15.

Willing, E.-M., Rawat, V., Mandáková, T., Maumus, F., James, G. V., Nordström, K. J., Becker, C., Warthmann, N., Chica, C., & Szarzynska, B. (2015). Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants*, *1*, 14023.

Wu, S., Han, B., & Jiao, Y. (2020). Genetic contribution of Paleopolyploidy to adaptive evolution in angiosperms. *Molecular Plant*, *13*, 59–71.

Xu, C., Yinan, C., Zhang, L., Jian, W., Liang, J., Cheng, L., Xiaowu, W., & Cheng, F. (2018). Hotspots of independent and multiple rounds of LTR-retrotransposon bursts in Brassica species. *Horticultural Plant Journal*, *4*, 165–174.

Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*, W265–W268.

Yaakov, B., & Kashkush, K. (2011). Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid. *Genome*, *54*, 42–49.

Yang, Y., Sun, P., Lv, L., Wang, D., Ru, D., Li, Y., Ma, T., Zhang, L., Shen, X., & Meng, F. (2020). Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nature Plants*, *6*, 215–222.

Zhang, J., Liu, Y., Xia, E.-H., Yao, Q.-Y., Liu, X.-D., & Gao, L.-Z. (2015). Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. *Proceedings of the National Academy of Sciences*, *112*, E7022–E7029.

Zhang, L., Wu, S., Chang, X., Wang, X., Zhao, Y., Xia, Y., Trigiano, R. N., Jiao, Y., & Chen, F. (2020). The ancient wave of polyploidization

events in flowering plants and their facilitated adaptation to environmental stress. *Plant, Cell & Environment*, *43*, 2847-2856.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Wang D, Zheng Z, Li Y, et al. Which factors contribute most to genome size variation within angiosperms?. *Ecol Evol*. 2021;11:2660–2668. https://doi.org/10.1002/ece3.7222