the plant journal

SEB Society for Experimental Biology

The Plant Journal (2018) **94,** 721–734 doi: 10.1111/tpj.13894

RESOURCE

Improving *Nelumbo nucifera* genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements

Songtao Gui[1], Jing Peng[2], Xiaolei Wang[1], Zhihua Wu[1], Rui Cao[1], Jérôme Salse[3], Hongyuan Zhang[1], Zhixuan Zhu[1], Qiuju Xia[4], Zhiwu Quan[4], Liping Shu[5], Wedong Ke[2] and Yi Ding[1,*]

[1]*State Key Laboratory of Hybrid Rice, Department of Genetics, College of Life Sciences, Wuhan University, Wuhan 430072, China,*

[2]*Institute of Vegetable, Wuhan Academy of Agriculture Science and Technology, Wuhan, Hubei, 430065, China,*

[3]*Paleogenomics & Evolution (PaleoEvo) Group, Génétique Diversité & Ecophysiologie des Céréales (GDEC), Institut National de la Recherché Agronomique UMR 1095, Clermont-Ferrand 63100, France,*

[4]*Key Laboratory of Genomics, BGI-Shenzhen, Chinese Ministry of Agriculture, Shenzhen 518083, China, and*

[5]*Wuhan Ice-Harbor Biological Technology Co. Ltd, Wuhan 430040, China*

Received 30 September 2017; revised 31 January 2018; accepted 21 February 2018; published online 25 March 2018.
*For correspondence (e-mail yiding@whu.edu.cn).

SUMMARY

Genetic and physical maps are powerful tools to anchor fragmented draft genome assemblies generated from next-generation sequencing. Currently, two draft assemblies of *Nelumbo nucifera*, the genomes of 'China Antique' and 'Chinese Tai-zi', have been released. However, there is presently no information on how the sequences are assembled into chromosomes in *N. nucifera*. The lack of physical maps and inadequate resolution of available genetic maps hindered the assembly of *N. nucifera* chromosomes. Here, a linkage map of *N. nucifera* containing 2371 bin markers [217 577 single nucleotide polymorphisms (SNPs)] was constructed using restriction-site associated DNA sequencing data of 181 F_2 individuals and validated by adding 197 simple sequence repeat (SSR) markers. Additionally, a BioNano optical map covering 86.20% of the 'Chinese Tai-zi' genome was constructed. The draft assembly of 'Chinese Tai-zi' was improved based on the BioNano optical map, showing an increase of the scaffold N50 from 0.989 to 1.48 Mb. Using a combination of multiple maps, 97.9% of the scaffolds in the 'Chinese Tai-zi' draft assembly and 97.6% of the scaffolds in the 'China Antique' draft assembly were anchored into pseudo-chromosomes, and the centromere regions along the pseudo-chromosomes were identified. An evolutionary scenario was proposed to reach the modern *N. nucifera* karyotype from the seven ancestral eudicot chromosomes. The present study provides the highest-resolution linkage map, the optical map and chromosome level genome assemblies for *N. nucifera*, which are valuable for the breeding and cultivation of *N. nucifera* and future studies of comparative and evolutionary genomics in angiosperms.

Keywords: BioNano optical map, chromosome rearrangement, genetic linkage map, *Nelumbo nucifera*, scaffold anchoring.

INTRODUCTION

Next-generation sequencing (NGS) has fuelled a scientific revolution by facilitating the rapid collection of large amounts of genomic sequence data, enabling whole-genome shotgun (WGS) assemblies in many non-model species (Jaillon *et al.*, 2007; Albert *et al.*, 2013; Olsen *et al.*, 2016). However, because of the formidable challenge of *de novo* sequence assembly relying solely on short NGS reads (Green, 1997; Denton *et al.*, 2014), many draft genomes contain thousands of individual sequences with no information on how these pieces are assembled into chromosomes.

A valid approach to address these shortcomings is to pair WGS with linkage mapping. Linkage maps are

accurate at the large, chromosomal scale, but are imprecise in detailed marker ordering. By contrast, NGS *de novo* assemblies are accurate at a fine scale but lack chromosome-scale information (Fierst, 2015). Integration of *de novo* genome assembly and genetic linkage mapping enables anchoring and ordering of scaffolds along chromosomes (Mascher and Stein, 2014). The advent of NGS has recently enabled the discovery and genotyping of thousands of markers across almost any genome in a single step (Davey *et al.*, 2011), making it feasible to anchor and order the scaffolds along chromosomes using linkage maps. Genetic linkage maps have been used to refine *de novo* assemblies of both plants (Argout *et al.*, 2011; Bartholome *et al.*, 2015) and animals (Kawakami *et al.*, 2014; Nossa *et al.*, 2014).

Another approach is physical mapping. Independent physical maps in combination with sequence assemblies will greatly facilitate the correct ordering of genic and non-genic DNA segments on chromosomes (Lewin *et al.*, 2009). A robust strategy to combine the physical map information with NGS is to use a BAC-by-BAC sequencing approach (Venter *et al.*, 1996). Optical mapping, which is primarily used in bacterial genomes (Anantharaman *et al.*, 1999), has recently become suitable for mapping large genomes (Dong *et al.*, 2013; Stankova *et al.*, 2016; Kaur *et al.*, 2017) because of its high-throughput modifications such as genome mapping in nanochannel arrays (Lam *et al.*, 2012).

Nelumbo nucifera Gaertn. (sacred lotus), an aquatic perennial basal eudicot that has survived since the Late Cretaceous period, belongs to the basal eudicot plants family Nelumbonaceae, with only one genus, *Nelumbo*, and two species: *N. lutea* (restricted to eastern and southern North America) and *N. nucifera* (distributed in Asia, Australia and Russia) (Shen-Miller, 2002; Bremer *et al.*, 2009). *N. nucifera* has been cultivated for thousands of years in Asia for its edible rhizomes, seeds and leaves. This plant has also been used as herbal medicine for the treatment of cancer, depression, diarrhoea, heart problems and insomnia (Sharma *et al.*, 2017). As a basal eudicot species, *N. nucifera* also occupies an important position in understanding the origin of eudicots and ancient polyploidization events (Wu *et al.*, 2014b).

The first genetic linkage maps of Nelumbo were constructed based on the genotype of 171 Simple Sequence Repeat (SSR) markers and 53 sequence-related amplified polymorphism (SRAP) markers in *N. nucifera* 'Chinese Antique', *N. lutea* 'AL1' and their 51 F_1 populations (Yang *et al.*, 2012b). Using the same map population, Zhang *et al.* (2014) refined both the linkage maps of *N. nucifera* and *N. lutea* based on SNPs that were identified using restriction-site associated DNA sequencing (RAD-Seq) data. The resulting linkage map of *N. nucifera* was 656.9 cM with 23 SSR markers and 73 SNPs distributed in eight linkage groups (LGs), and the linkage map of *N. lutea* was 494.3 cM with 562 bins (3894 SNPs) and 136 SSRs distributed in nine LGs. Using the 96 F_2 individuals of two *N. nucifera* cultivars, Liu *et al.* constructed a linkage map of 581.3 cM, with 791 bin markers (8971 SNPs) sorted into 8 LGs (Liu *et al.*, 2016).

Two draft genomes of *N. nucifera* were released in 2013. One genome was generated from the sacred lotus variety 'China Antique', resulting in an assembly of 804 Mb (Ming *et al.*, 2013). The other genome was generated from the wild strain 'Chinese Tai-zi', resulting in an assembly of 792 Mb (Wang *et al.*, 2013). Although several attempts have been made to order the scaffolds of the two draft assemblies (Ming *et al.*, 2013; Zhang *et al.*, 2014), the anchored scaffolds are insufficient to represent the actual gene orders along the *N. nucifera* chromosomes. To anchor more scaffolds of the draft assembly, a linkage map with higher resolution is needed. Recently, the Thailand sacred lotus wild strain 'Thailand Chiang Mai' was re-sequenced (Hu *et al.*, 2015), presented an opportunity to build an eligible mapping population for a linkage map with much higher resolution.

Thus, the main objectives of the present study were to: (1) construct a higher resolution linkage map for *N. nucifera*, (2) improve the assembly of the Chinese 'Tai-zi' draft genome, (3) anchor the scaffolds of *N. nucifera* draft genomes into chromosomes, and (4) reveal the ancient chromosome rearrangements in *N. nucifera*. To achieve these goals, we constructed a F_2 population crossed between two sacred lotus wild strains, 'Chinese Tai-zi' and 'Thailand Chiang Mai', and sequenced the F_2 individuals using RAD-sequencing to generate polymorphic markers for linkage map construction. To validate the initial draft assembly of the Chinese 'Tai-zi' cultivar, an optical map was also generated using genome mapping on nanochannel arrays employing the BioNano Irys System. Using the high-resolution linkage map constructed herein and the linkage map constructed by (Liu *et al.*, 2016), most of the scaffolds in the two *N. nucifera* draft genomes were anchored into pseudo-chromosomes and the centromere regions in the chromosomes were identified. Based on the comparison between the ancestral eudicot karyotype and chromosomes of *N. nucifera*, we proposed an evolutionary scenario for the formation of modern *N. nucifera* chromosomes.

RESULTS

Linkage map construction and validation

To construct a high-resolution genetic linkage map for *N. nucifera*, a F_2 mapping population consisted of 181 individuals was generated. In total, 2146 polymorphic SSRs between the two parents were predicted and 306 of them were verified as polymorphic (Appendix S1, Figure S1 and Table S1). Using RAD-sequencing data of the 181 *N.*

nucifera F_2 individuals, 217 577 polymorphic SNPs were identified and genotyped, and merged into 2371 bin markers (Appendix S1, Figure S2, Figure S3 and Tables S2–S6). Subsequently, we constructed a genetic linkage map (hereafter referred to as 'MapBA') using genotyping data from all of the bin markers by computing a minimal spanning tree of the graph (Wu *et al.*, 2008). The resulting linkage map had a total genetic distance of 789.54 cM and an average interval of 0.33 cM between bins (Figure 1(a) and Table S7).

To evaluate the reliability of MapBA, two additional linkage maps named 'MapBO' and 'MapAll' were reconstructed by changing the mapping strategy to the maximum likelihood approach (Wu *et al.*, 2002) and adding 198 validated polymorphic SSRs (Table S6), respectively. The resulting MapBO contained all 2731 bins, with a total length of 799.59 cM and average interval of 0.337 cM, values only slightly higher than those of MapBA, while the resulting MapAll contained 2731 bins and 195 SSRs, with a total length of 1137.06 cM and an average interval of 0.443 cM (Table S7). The heat maps of pairwise recombination fraction and LOD scores in MapBA, MapBO and MapAll (Figure 1(b)) all showed consistent heat across the markers within LGs, implying that the nearby marker pair along the linkage map had lowest recombination rate and highest LOD scores, indicated that the three linkage maps were all distinctly clustered. Comparing the marker orders among MapBA, MapBO and MapAll showed that the markers were grouped into the same eight LGs in the three maps (Figure 1(a)), 97.26% of MapBO bins and 97.47% of MapAll bins were of the same order with MapBA. Only 65 bins in MapBO and 60 bins in MapAll showed order differences with MapBA, and all the order differences were presented within several adjacent markers (Figure 1(a) and Tables S8 and S9). These results indicated that changing the mapping strategy had little impact on the map construction, the addition of SSRs produced inflated genetic distances but showed little influence on marker clustering and ordering. To further evaluate the map quality, the genetic orders of the 195 SSRs were compared with their physical positions in the bin-based pseudo-molecules (Appendix S1 and Table S3). Approximately 82.56% of the SSRs showed the exact same orders between their genetic and physical positions, while the remaining 34 SSRs clustered into 17 pairs that showed reverse orders between their genetic and physical positions (Figure 1(a) and Table S10). The consistency between the genetic and physical locations of the SSRs further demonstrated that the

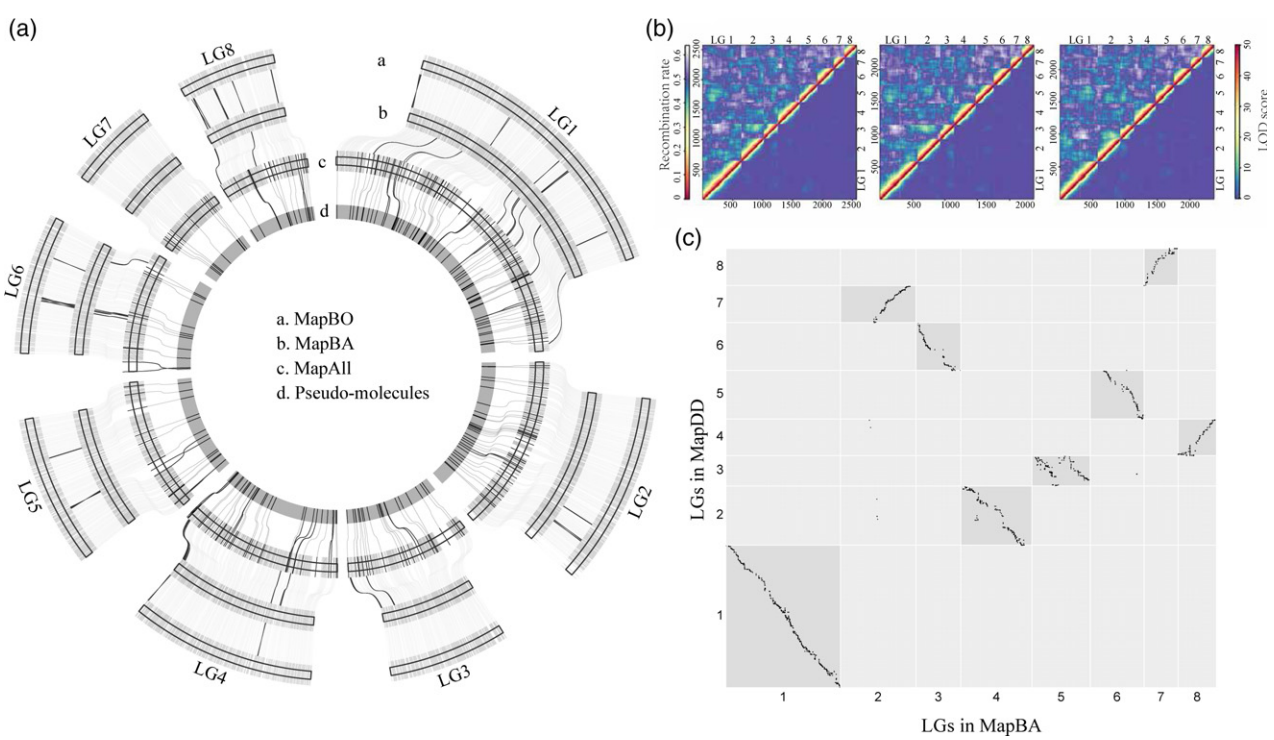

Figure 1. Validations of the linkage maps.
(a) Comparison between MapBA, MapBO, MapAll and the pseudo-molecules of the ordered scaffolds. Tracks a, b, c and d represent MapBA, MapBO, MapAll and the pseudo-molecules of the ordered scaffolds, respectively. SSRs were highlighted as black in tracks c and d. Lines between each track represent the links of same markers between the tracks. The black links represent the differently ordered markers between the adjacent tracks.
(b) The pairwise recombination fraction and the LOD scores in MapBA, MapBO and MapAll. The x and y axes represent the bin markers.
(c) Comparison between MapBA and MapDD. The corresponding blocks were highlighted as dark grey. [Colour figure can be viewed at wileyonlinelibrary.com].

genetic order of the markers were competent to represent their physical distributions along the chromosomes. Eventually, we have compared MapBA with the linkage map of *N. nucifera* reported by Liu *et al.* (2016) (hereafter referred to as 'MapDD'). The results showed the linear relationships were distinct between each matched LG pair of the two maps, except for the two parallel lines resulting from the correspondence between MapDD LG3 and MapBA LG4 (Figure 1(c) and Table S11), further demonstrated the reliability of the linkage map constructed in the present study.

Distorted segregations were identified in 215 bin markers (9.07%) under a *P*-value threshold of 0.05, and 32 bins (1.35%) were identified as significant ($P < 0.01$) segregation distortion (Table S12). As the marker orders were consistent across replications of analysis, the segregation-distorted markers were kept in the linkage maps. Most of the segregation-distorted markers (80.93%) skewed to the maternal genotype ('Chinense Tai-zi'), while 17.21% skewed to genotypes of both parents and 1.86% skewed to paternal genotype ('Thailand Chiang Mai'), no markers skewed to the heterozygous genotype were found (Table S12). Distribution of these segregation-distorted markers on the chromosomes (Figure 2, see sections below for the construction of *N. nucifera* pseudo-chromosomes) showed that markers skewed to the same direction tended to cluster into large segregation distortion regions (SDRs). Markers skewed to 'Chinense Tai-zi' were mainly distributed in two large SDRs spanning more than 30 Mb on TZ-Chr1 and TZ-Chr2 respectively, while markers of 'Thailand Chiang Mai' direction were mainly clustered into a 11 Mb SDRs on TZ-Chr5 (Figure 2).

Improving 'Chinese Tai-zi' assembly using BioNano whole-genome mapping

To conduct whole-genome mapping of 'Chinese Tai-zi' in nanochannel arrays, 241.56 Gb raw data were generated from 10 runs of the BioNano Genomics Irys chips signalling (Table S13). After assessing and filtering, a total of 158.1 Gb clean data was generated and included in the BioNano-based *de novo* physical map assembly (Figure S4 and Table S14). The resulting *de novo* physical map comprised 2099 consensus maps spanning 645.12 Mb (approximately 86.20% of the reference assembly) with a N50 contig size of 0.33 Mb (Table 1). To generate a more contiguous assembly, approximately 72.60% (543.16 Mb) of the reference assembly was aligned to the *de novo* physical map, and 1162 scaffolds coalesced into 694 super-scaffolds (Figure 2), resulting in a noticeable increase in the total length of sequences larger than 1 Mb from 393.68 to 523.01 Mb. The advanced assembly had a N50 of 1.48 Mb (Table 1), approximately 1.5-fold larger than that of the original assembly (0.989 Mb).

To identify possible false joins in the 'Chinese Tai-zi' assembly, the physical locations of all the SNPs and SSRs of the improved assembly of 'Chinese Tai-zi' were compared with their genetic locations to identify 'chimeric scaffolds', which were defined as scaffolds with 'discordant genetic regions' (Bartholome *et al.*, 2015) assigned differently between genetic and physical maps (supported by two or more SNPs). In total, 113 discordant genetic regions among 101 chimeric scaffolds were identified, representing 20.62% of the improved assembly (Figure S5(a)). In order to split the chimeric scaffolds in a more disciplined manner, the discordant genetic regions were subsequently examined to determine whether there were any supports from the BioNano genome map. Only chimeric scaffolds whose candidate breakpoint regions contained one or more gaps not covered by any BioNano genome map were split, with the longest gap selected as the final breakpoint (Figure S6). These operations reduced the number of candidate false joins to 35, representing 7.97% of the improved assembly (33 scaffolds). The 33 chimeric scaffolds were subsequently cut by the largest gap within discordant genetic regions.

N. nucifera pseudo-chromosomes construction

To anchor the scaffolds of 'Chinese Tai-zi' into chromosomes, the refined scaffolds were ordered according to each of the four maps (MapBA, MapBO, MapAll and MapDD) and the consensus of the four maps, respectively. The results showed that both the anchoring rate and orienting rate using each single map were lower than those using the consensus map (Table S15), thus the final pseudo-chromosomes were generated based on the scaffold order using consensus map and renamed according to their physical lengths. Approximately 97.9% of the improved 'Chinese Tai-zi' assembly were ordered and 85.3% of the scaffolds were oriented (Table S15). The total length of the 'Chinese Tai-zi' pseudo-chromosomes (hereafter referred to as 'TZ-Chrs') was approximately 782.76 Mb (Figure 2). The largest chromosome (TZ-Chr1) was approximately 207 Mb, nearly four times the size of the smallest chromosome TZ-Chr8. Each of the pseudo-chromosomes showed a high level of collinearity with the four reference linkage maps (Figure S7). After repeat sequence masking and gene annotating, 30 378 protein coding genes, 527 miRNAs, 890 snRNAs and 1279 tRNAs were identified throughout the TZ-Chrs (Table S16).

Considering that the 'China Antique' draft assembly showed a considerable scaffold N50 of 3.4 Mb (Ming *et al.*, 2013), we also attempted to anchor the scaffolds of the 'China Antique' assembly. Approximately 92.48% (8296) of the SNPs in MapDD, 92.31% (200 859) of the SNPs and all the SSRs in the present study were uniquely mapped to 1464 scaffolds of the 'China Antique' assembly. A total of 1263 SNPs were observed as outliers located on different LGs with their neighbours, and were subsequently excluded (Figure S5(b)). The scaffold anchoring result

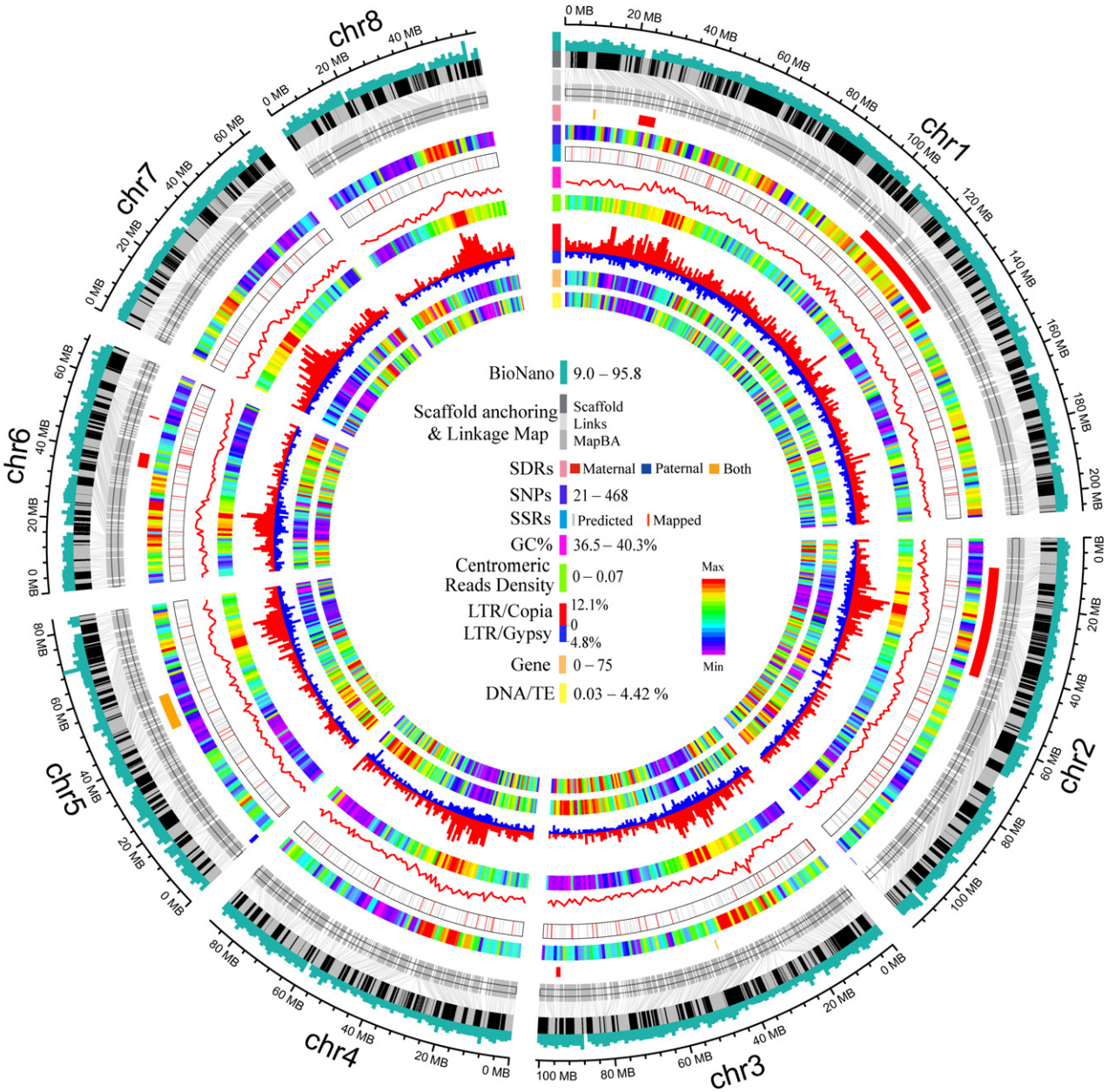

Figure 2. Genomic landscape of the pseudo-chromosomes of the 'Chinese Tai-zi'. Tracks from the outside in represent the coverage of BioNano molecules, the links between the anchored scaffolds and MapBA, distribution of segregation distortion regions (SDRs), distribution of polymorphic SNPs, distribution of polymorphic SSRs, GC content, centromere reads density, density of the Copia class of long terminal repeat retrotransposons (LTR), density of the Gypsy class of LTRs, density of the genes and density of DNA transposons. The window size was 1 Mb. The range of each track is indicated as in the figure. [Colour figure can be viewed at wileyonlinelibrary.com].

using the consensus map was better than that using each individual map, consistent with the results of the 'Chinese Tai-zi' (Table S15). The anchored pseudo-chromosomes of the 'China Antique' (hereafter referred to as 'CA-Chrs') had a total length of approximately 797.68 Mb, ranging from 59.36 to 215.18 Mb (Figure S8). The pseudo-chromosomes represent 97.6% of the 'China Antique' draft assembly and covered 98.8% of the annotations (Table S17). In a previous study, approximately 67.6% of the 'China Antique' genome assembly were anchored into nine megascaffolds according to the genetic map of the American lotus 'AL1'. Comparison the eight pseudo-chromosomes of CA-Chrs in the present study with the nine megascaffolds (67.6% of the 'China Antique' genome assembly) anchored by Zhang *et al.* (2014) showed large blocks of genome collinearity (Figure S9).

Table 1 Statistics of the BioNano *de novo* physical map and the hybrid assembly

Title	Original BNG[a]	Original NGS[b]	BNG in hybrid[c]	NGS in hybrid	Hybrid
Number of contigs[d]	2099	14 895	2099	1162	14 630
Min contig length (Mb)	0.108	0.0001	0.108	0.08	0.0001
Mean contig length (Mb)	0.307	0.053	0.307	0.65	0.0579
Max contig length (Mb)	1.268	4.484	1.268	4.484	5.481
Contig N50 (Mb)	0.33	0.989	0.33	1.043	1.48
Total contig length (Mb)	645.12	790.33	645.12	754.92	847.16

[a]BNG represents the *de novo* BioNano genome mapping.
[b]NGS represents the 'Chinese Tai-zi' *de novo* assembly.
[c]Hybrid represents the hybrid scaffolding result.
[d]Contig represents the genome map for BNG or the scaffold for NGS.

Assessing the completeness of the two pseudo-chromosomes with 1440 plantae lineage-specific Benchmarking Universal Single-Copy Orthologs (BUSCOs) (Simão *et al.*, 2015) showed that 90.1% of the Plantae BUSCOs could be aligned to the pseudo-chromosomes of 'Chinese Tai-zi', slight lower than its draft assembly (91.0%). While in 'China Antique', both the draft assemblies and the pseudo-chromosomes showed 92.2% aligned BUSCOs (Table S18). Synteny analysis between TZ-Chrs and CA-Chrs showed a highly consistent gene order (Figure S10). These results indicated that both TZ-Chrs and CA-Chrs were competent to represent most of the draft assemblies, while CA-Chrs showed better completeness than TZ-Chrs, which may be attributed to the long scaffolds of the 'China Antique' draft assembly generated by sequencing of an additional paired-end 20 kb insert library and an optimized assembly procedure (Ming *et al.*, 2013).

Identification of centromere regions

Centromeres have suppressed levels of crossing over (Talbert and Henikoff, 2010). The centromere regions of *N. nucifera* were initially estimated by identifying the crossover-suppressed regions (CSRs). To identify the CSRs in *N. nucifera*, the genome-wide recombination rates were estimated from the comparison of MapBA and the pseudo-molecules of TZ-Chrs. The spans of CSRs ranged from ~6.7 Mb (TZ-Chr8) to ~13.7 Mb (TZ-Chr1). The CSRs on TZ-Chr1 and TZ-Chr5 were distributed near the ends of the chromosomes while others were positioned near the middle of the chromosomes (Figure 3(a)). This result was consistent with the *N. nucifera* karyotype, showing that the longest chromosome is subtelocentric while other chromosomes are metacentric or submetacentric (Diao *et al.*, 2005). To further determine the centromere regions in the present study, the chromatin immunoprecipitation-based sequencing (ChIP-Seq) reads of the NnCenH3 (sacred lotus centromere-specific histone H3 variant) nucleosome-associated sequences (Zhu *et al.*, 2016) were aligned to the TZ-Chrs and the reads density was calculated in an unbiased manner. Significant sequence enrichment was observed in all eight chromosomes (Figure 3(b)). The sizes of the NnCenH3-enriched centromere cores (from the first to the last CENH3 subdomain in each centromere) in the eight centromeres ranged from ~3.6 Mb (TZ-Chr5) to ~8.7 Mb (TZ-Chr3) (Figure 3(c)). The distributions of NnCenH3-enriched centromere cores were consistent with the distributions of CSRs. The NnCenH3 binding domains of TZ-Chr1, 2, 4 and 7 were wrapped in the CSRs, while the NnCenH3 binding domain extended into the flank of the CSRs in TZ-Chr3, 5, 6 and 8 (Figure 3(c)). The NnCenH3-enriched centromere cores were also identified in the CA-Chrs. The enrichment patterns of ChIP-Seq reads in CA-Chrs were similar to those in TZ-Chrs but with larger spans (Figure S11).

In *N. nucifera*, the genes were primarily distributed in subtelomeric regions while the long terminal repeats (LTRs) were accumulated in subcentromeric or centromeric regions (Figure 2), similar to those in other eukaryotes with large genomes (Jaillon *et al.*, 2007; Argout *et al.*, 2011; Potato Genome Sequencing Consortium *et al.*, 2011). Additionally, the high-GC-content regions were distributed nearby centromeric regions, indicating the GC-rich centromeric DNA of *N. nucifera*. The centromeres in most higher eukaryotic organisms are composed of satellite repeats (Henikoff *et al.*, 2001). The correlation analysis between NnCenH3 binding sequences and the repeat sequences in the present study showed that the LTRs were positively correlated with the NnCenH3-binding sequences, while the DNA TEs formed a negative correlation with the NnCenH3 binding sequences (Figure 2 and Table S19). The *Ty3/Copia* class of LTRs, predominantly observed in *N. nucifera* (Ming *et al.*, 2013; Wang *et al.*, 2013), also had the strongest correlation with the NnCenH3 binding sequences (Table S19). This result was consistent with that of Zhu *et al.* (2016), who identified seven major centromere-associated repeat clusters belonging to the *Ty3/Copia* class of LTRs.

N. nucifera chromosome paleohistory

Eudicots have been proposed to derive from an ancestral eudicot karyotype (AEK) structured with seven protochromosomes (Salse, 2016). AEK experienced a known whole-

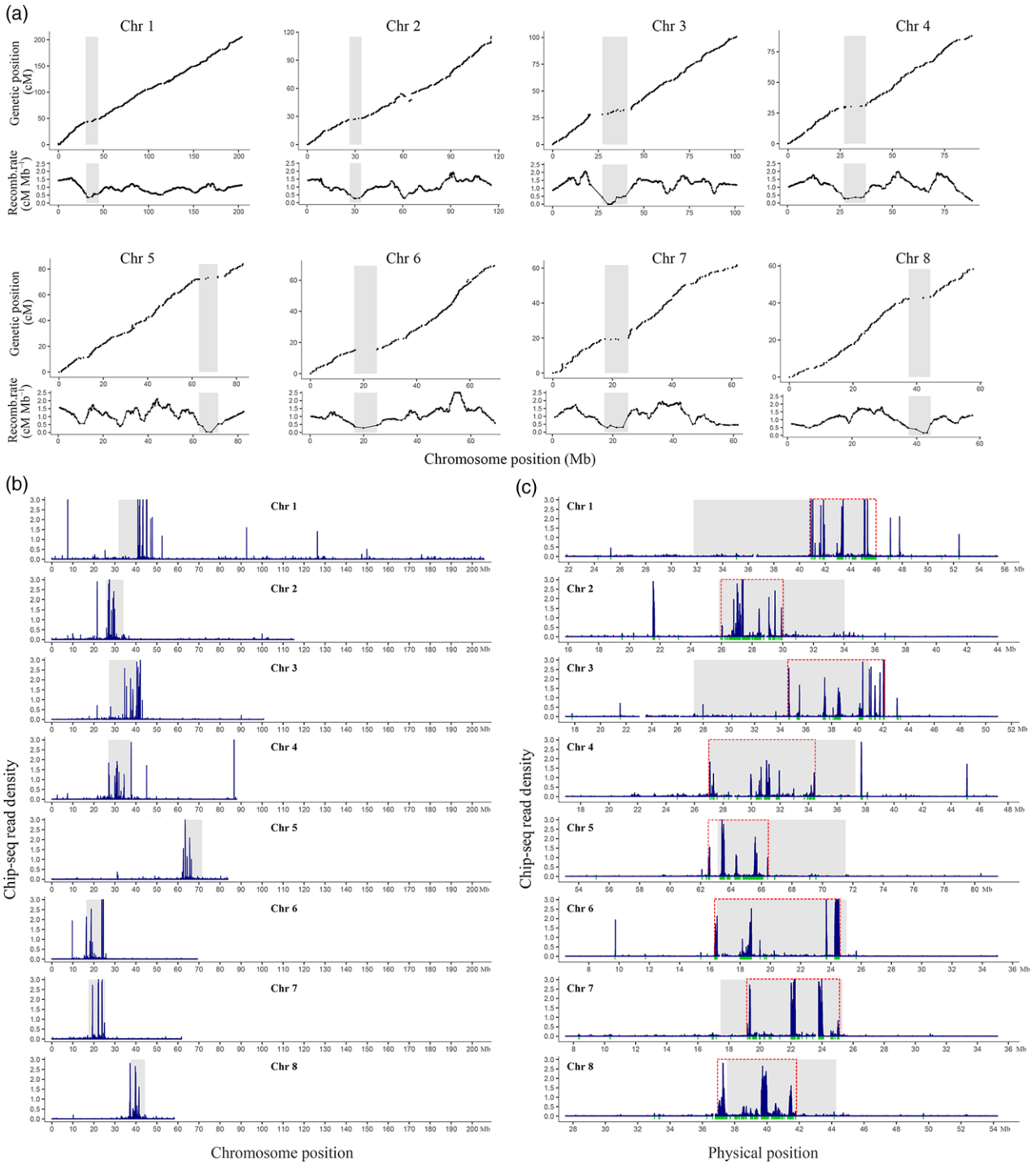

Figure 3. Identification of centromere regions in the pseudo-chromosomes of the 'Chinese Tai-zi'.

(a) Crossover-suppressed regions in eight pseudo-chromosomes of the 'Chinese Tai-zi'. *X*-axis represents the genetic location (upper) of the bin markers and the local recombination rate (lower). *Y*-axis represents the physical position of the bin markers. The predicted crossover-suppressed regions are highlighted with grey rectangles.

(b) Plot of ChIP-Seq reads density along individual pseudo-chromosomes of the 'Chinese Tai-zi'. The window size was 20 kb, spaced every 10 kb. The predicted crossover-suppressed regions were highlighted with grey rectangles.

(c) Details of 10 Mb of flanking sequence from each side of the crossover-suppressed domain. The NnCenH3-binding domains are indicated as green bars. The inferred boundaries of the NnCenH3-binding domain are indicated as vertical red lines. [Colour figure can be viewed at wileyonlinelibrary.com].

genome triplication (WGT γ event) generating a 21-chromosome intermediate for the formation of the modern chromosomes of most eudicots. However, previous studies have reported that *N. nucifera* did not experience the γ event. Instead, a lineage-specific paleotetraploidy event referred to as λ was detected in *N. nucifera* (Ming *et al.*, 2013; Wang *et al.*, 2013). To assess the paleohistory of the *N. nucifera* genome, we compared AEK, grape, coffee, cacao and peach to *Nelumbo nucifera* using the genome alignment parameters and ancestral genome reconstruction methods described by Salse (2016). The results showed a one-to-two chromosome relationship between AEK and *N. nucifera* (Figure 4(a)), confirming that *N. nucifera* experienced a specific whole-genome duplication (WGD) event. The *N. nucifera* genome had large synteny blocks with AEK. *N. nucifera* chromosome 1 showed collinearity with a nearly full set of AEK1 genes and two sets of AEK7 genes, and chromosomes 2, 3, 4, 5 and 8 retained the gene orders of AEK4, 3, 6, 5 and 2, respectively, while syntenic blocks representing partial ancient chromosomes were observed in all *N. nucifera* chromosomes, except chromosome 8 (Figure 4(a)). This result indicated that *N. nucifera* retained the eudicot ancestor genome structure, but with several chromosome rearrangements. Moreover, the gene models from the *N. nucifera* genome were also aligned onto themselves. The one-to-one relationships represent by the seven duplicated blocks (Figure 4(b)) also indicated a WGD event in *N. nucifera*. Based on the paralogy between AEK and the *N. nucifera* genome, we proposed an evolutionary scenario that shaped the modern eight *N. nucifera* chromosomes from the seven ancestral eudicot chromosomes. Generally, after the lineage-specific WGD event, generating a 14 chromosomes intermediate, chromosome rearrangements, including 14 ancestral chromosome fissions, 20 fusions and 1 inversion were required to reach the modern genome structure of eight chromosomes of *N. nucifera* (Figure 4(c)), compared with the evolutionary scenario of major families of Rosids, Asterids, Malvids and Fabids respectively exemplified in the present study by grape, coffee, cacao and peach. From the 21-chromosome intermediate generated by WGT γ, grape, coffee, cacao and peach experienced 1 fission and 3 fusions, 13 fissions and 23 fusions, 2 fissions and 13 fusions, and 4 fissions and 17 fusions to derive the modern karyotypes of 19, 11, 10 and 8 chromosomes, respectively (Figure 4(c)).

To achieve a better understanding of the molecular mechanisms driving the ancient chromosome rearrangement in *N. nucifera*, we compared the distribution of telomeric repeats and centromere reads with the breakpoints of the ancient chromosome fusions. The results showed telomeric repeats located not only at the chromosome ends, as identified in all eight *N. nucifera* chromosomes, but also scattered at internal chromosome sites (Figure 4(b)). Eight of the 16 internal telomeric repeats showed a clear correlation with the ancient chromosome fusion points, as indicated by the red arrows in Figure 4(b). Moreover, telomeric repeats were also observed close to the centromere reads enrichment regions, particularly in chromosome 1, which showed extra centromere reads enrichment signals in addition to the main centromere region. This finding was consistent with previous results showing that *N. nucifera* chromosome 1 displayed two FISH signals of centromere-associated repeat cluster CL6, and two centromere-associated repeat clusters (CL19 and CL21) exhibited FISH signals at both the centromere region and the distal ends of the chromosomes (Zhu *et al.*, 2016). These results indicated that some of the ancient chromosome fusions in *N. nucifera* could be explained by chromosome end fusion (the nested chromosome fusion of AEK7 and AEK3 in chromosome 1) or reciprocal translocation with breakpoints close to the centromere in one ancient chromosome and close to the chromosome ends in another (the fusion of AEK3 and AEK6 in chromosome 3 and the fusion of AEK5 in chromosome 5), while other fusions probably resulted from multiple reciprocal translocations (Schubert and Lysak, 2011).

DISCUSSION

High-resolution linkage maps and segregation distortion regions of *N. nucifera*

Genetic linkage mapping has been a universal tool to order genomic loci along chromosomes of sexually reproducing species (Mascher and Stein, 2014) for more than a century since the construction of the first genetic map (Sturtevant, 1913). Although many factors affect the efficiency of the genetic mapping process, it is the size and type of the population under study that primarily determine the accuracy of genetic mapping (Ferreira *et al.*, 2006). Compared with previously published linkage maps of Nelumbo (Yang *et al.*, 2012b; Zhang *et al.*, 2014; Liu *et al.*, 2016), a much larger F_2 segregating population was constructed in the present study using two low heterozygous lotus strains that were supported by high qualify genome sequences (Wang *et al.*, 2013; Hu *et al.*, 2015), which made us able to successfully construct the highest density linkage map for *N. nucifera* based on 217 577 SNPs and 195 SSRs. This linkage map also showed compatibility or even higher resolution compared with the recently published linkage maps of other species (Qi *et al.*, 2014; Wu *et al.*, 2014a; Kujur *et al.*, 2015). The higher resolution genetic maps and the polymorphic SNP and SSR markers generated in the present study laid the foundation for scaffold anchoring, map-based gene cloning, quantitative trait loci (QTL) analysis of economically important traits, and molecular breeding using marker-assisted selection (MAS) in *N. nucifera*.

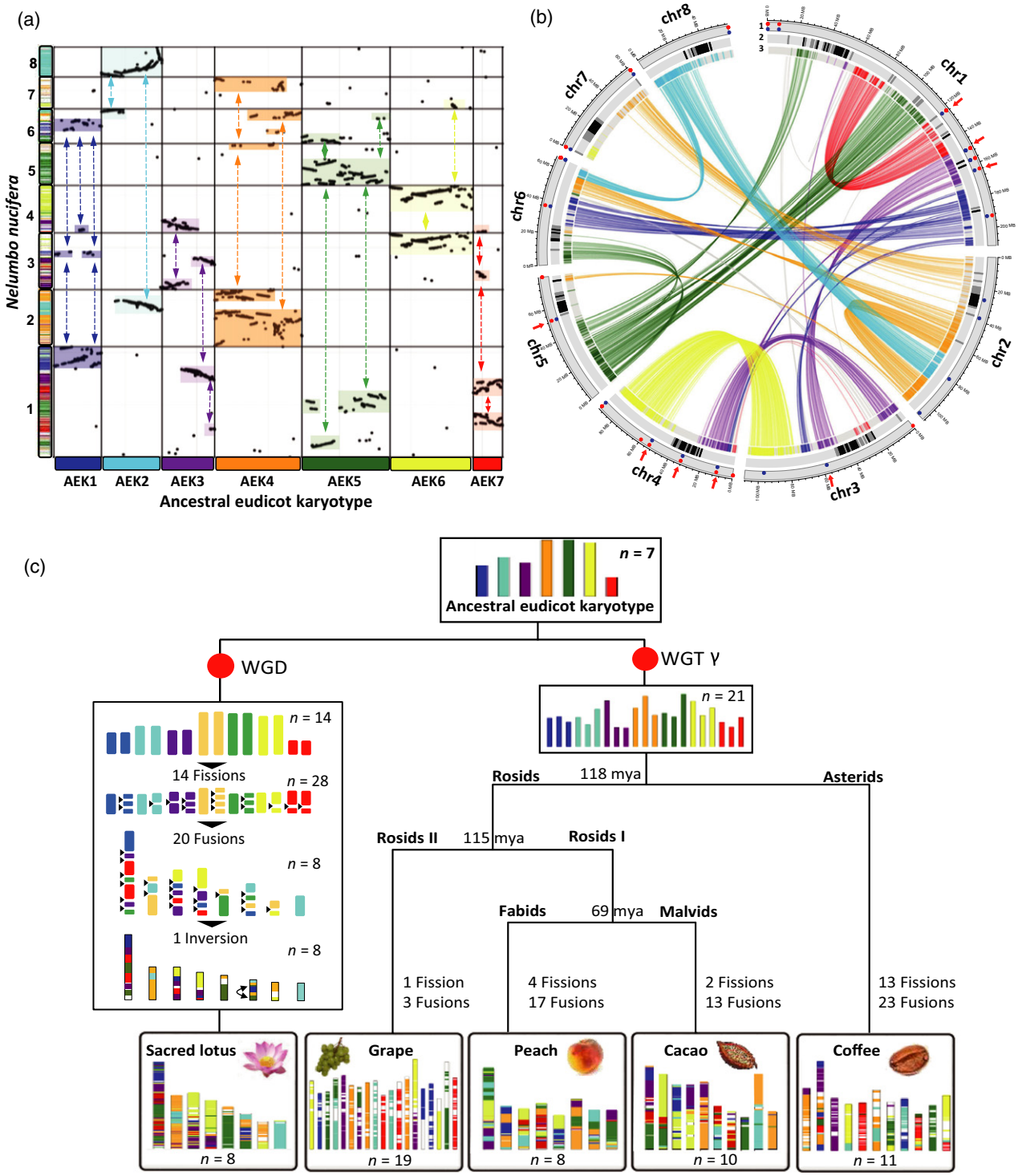

Figure 4. *N. nucifera* genome paleohistory.
(a) Complete dot plot-based deconvolution of the observed synteny (coloured diagonals) and paralogy (vertical arrows) between AEK (*x*-axis with seven chromosomes in colour) and *N. nucifera* (*y*-axis, with the colour code of seven AEKs).
(b) *N. nucifera* genome duplication and synteny. Tracks from the outside in represents: (1) the distribution of telomeric repeats identified in TZ-Chrs (blue) and CA-Chrs (red), (2) centromere reads density in 1 Mb windows, blocks with top 20% reads density are highlighted (grey, 80–90%; black, 90–100%), and (3) *N. nucifera* genes coloured using AEK colour codes.
(c) Evolutionary scenario of the modern *N. nucifera*, grape, peach (representative of the Fabids), cacao (representative of the Malvids) and coffee (representative of the Asterids) genomes from the ancestral eudicot karyotype (AEK) illustrated with seven colours (top). The modern genomes are illustrated at the bottom with different colours reflecting the origin from the seven ancestral chromosomes from AEK. Duplication (WGD) and triplication (WGT) events are shown with red dots on the tree branches, along with the shuffling events (fusions and fissions). [Colour figure can be viewed at wileyonlinelibrary.com].

Segregation distortion has been commonly discovered in variety of taxa, and could be a potentially powerful evolutionary force (Taylor and Ingvarsson, 2003). The proportion of segregation-distorted markers in the present study (9.07%) is smaller than that reported in an interspecific cross F_1 population (38.42%) (Yang *et al.*, 2012b) and similar to that in another intraspecific cross F_2 population (10.88%) of *N. nucifera* (Liu *et al.*, 2016), indicated that segregation distortion commonly occurred in various populations of *N. nucifera* despite the different distortion rate caused by factors such as genetic divergence or population types. Segregation distortion could be caused by technical bias during genotyping process or biological factors such as gain or loss of function of gametophyte genes (Mizuta *et al.*, 2010; Yang *et al.*, 2012a) and chromosome rearrangements between genetically divergent parents (Kianian and Quiros, 1992). The consistence of marker orders across replications of analysis indicated the SDRs in the present study were caused more likely by biological factors rather than genotyping errors. The two largest SDRs located at the long arm of Chr1 and the centromeric regions of Chr2 in the present study were correspondent to the SDRs on LG1 and LG7 reported by Liu *et al.* (2016) according to our linkage map comparison analysis (Figure 1(c)), indicated that these SDRs may be related to several gametophyte genes. Most of the SDRs in the present study were skewed to the female parent 'Chinense Tai-zi', this could be caused by the maternal cytoplasmic environment, which could influence the viability selection of gametes and zygotes (Tang *et al.*, 2013), or the environmental factors (Zamir *et al.*, 1982) since the original tropical growth environment of the male parent 'Thailand Chiang Mai' may weaken its pollen when growing at temperate regions. Further studies including reciprocal crosses may help to understand the causes of segregation distortion in *N. nucifera*.

N. nucifera chromosomes and ancient chromosome rearrangements

Genetic anchoring of WGS is necessary to improve the fragmented WGS assemblies and reconstruct the most likely chromosomal assemblies in many species (Jaillon *et al.*, 2007; Potato Genome Sequencing Consortium *et al.*, 2011; Ren *et al.*, 2012; Bartholome *et al.*, 2015). However, most anchoring is based on information from single linkage map. Scaffold anchoring based on one linkage map may differ from that based on another map, as genetic maps can vary in many aspects, such as recombination frequency, segregation distortion and presence–absence variation (Tang *et al.*, 2015). In the present study, the multiple-maps-based results of both the 'Chinese Tai-zi' and the 'China Antique' draft genomes were better than any single-map-based result in both the scaffold anchoring rate and scaffold orienting rate. With the development of genotyping and mapping strategies, many organisms often have several genetic maps available, making it feasible to anchor the scaffold using a combination of evidence from multiple maps. The accuracy of the anchored *N. nucifera* pseudo-chromosomes could not only be ensured by the reliability of the linkage maps and the genome completeness estimation, but also be verified by the result that the centromere location inferred from CSRs was consistent with that inferred from the distribution of NnCenH3 ChIP-Seq reads and the karyotype of *N. nucifera* (Diao *et al.*, 2005). These results indicated that the *N. nucifera* pseudo-chromosomes represent most of the information in its draft genome assemblies and reflected the actual gene orders along the *N. nucifera* chromosomes. However, the pseudo-chromosomes of *N. nucifera* still have several gaps and unmapped sequences. Additional studies combining these results with new sequencing technologies, such as single molecule sequencing (Eid *et al.*, 2009) and chromosome interaction mapping (Hi-C) (Lieberman-Aiden *et al.*, 2009), are needed to obtain a more complete and accurate *N. nucifera* reference genome.

Angiosperms experience recursive polyploidizations during their evolution, which act as a major driver in their divergence and speciation (Tang *et al.*, 2008; Soltis *et al.*, 2014). Ancestral eudicot karyotype (AEK), the most recent common ancestors of eudicots, comprising seven protochromosomes (Salse, 2016), undergoes a WGT event (γ event), followed by lineage-specific WGDs and chromosome rearrangements to reach the modern karyotype of most core eudicots. Previous studies indicated a lineage-specific paleotetraploidy event rather than the γ event in *N. nucifera* (Ming *et al.*, 2013; Wang *et al.*, 2013), but how did the post-duplication chromosomes reach the modern karyotype of *N. nucifera* was still vague. Based on the anchored *N. nucifera* pseudo-chromosomes in the present study, we were able to infer the possible chromosome rearrangements of *N. nucifera*. In contrast with the low mutation rate of *N. nucifera* when compared with core eudicots as reported in its nuclear genome (Ming *et al.*, 2013; Wang *et al.*, 2013) and organelle genomes (Wu *et al.*, 2014b; Gui *et al.*, 2016), the ancient chromosome rearrangements happened more frequently in *N. nucifera* than that in many core eudicots. Considering that the lotus-specific WGD event occurred between 76 and 54 MYA (Ming *et al.*, 2013), the ancient chromosome rearrangements in *N. nucifera* may be caused by changes in climate during the Cretaceous–Paleogene transition. Previous studies have suggested that the γ polyploidy ancestor was formed from an initial tetraploidization event and a subsequent hybridization of a third subgenome based on the findings that two of the three subgenomes are more fractionated than the third (Lyons *et al.*, 2008; Murat *et al.*, 2015). However, it remains equivocal whether the hybridization of the third subgenome resulted from genome duplication after

fertilization between a 2n and a 1n gamete or a direct fusion between a tetraploid and a diploid (Lyons *et al.*, 2008). The comparative genomics-based evolutionary scenario deduced in the present study suggested that *N. nucifera* has retained the eudicot ancestor genome structure, making *N. nucifera* a pivotal genome representing basal eudicots for comparative and evolutionary genomics studies in angiosperms, which would facilitate the reconstruction of a prehexaploidization ancestor and improve the studies on the fate of the ancestral triplicates.

Chromosome fusions in grass were correlated with the centromeric repeats, indicating non-homologous centromeric–telomeric recombination that led to the nested chromosome fusions in grass (Murat *et al.*, 2010). However, most chromosome fusions in Rosids were telomeric (Murat *et al.*, 2015). In *N. nucifera*, although two fusions resulted from centromeric–telomeric recombination, most fusions correlated with telomeric repeats. The telomeric fusions could lead to dicentric chromosome intermediates (Villasante *et al.*, 2007), with one centromere decayed during evolution. The additional centromeric signals in the non-centromeric region in *N. nucifera* chromosome 1 reported here and by Zhu *et al.* (2016) could be remnants of a decayed centromere. The molecular mechanism of the end-to-end chromosome fusion remains obscure. Schubert and Lysak (2011) suggested that rather than the simple fusion of intact chromosomes, the end-to-end chromosome fusion more likely resulted from symmetric reciprocal translocations between a telo- or acrocentric chromosome and another chromosome with breakpoints close to the centromere of the telo- or acrocentric chromosome and close to the end of the other chromosome. Additional studies on reconstruction of chromosome rearrangements using strategies, including comparative chromosome painting (Scherthan *et al.*, 1994) could be performed to specify the actual chromosome shuffling events that led to the current karyotypes.

EXPERIMENTAL PROCEDURES

Plant materials

The F_2 mapping population of *N. nucifera* consisted of 181 individuals was generated using a cross between two wild strains of sacred lotus, 'Chinese Tai-zi' (Wang *et al.*, 2013) (female parent) and 'Thailand Chiang Mai' (Hu *et al.*, 2015) (male parent). The two parents, F_1 and F_2 individuals were maintained at the Wuhan National Germplasm Repository for Aquatic Vegetables (30°12′N, 111°20′E), Wuhan, Hubei, People's Republic of China. The genomic DNAs were extracted from fresh young leaves using the modified CTAB method as previously described (Pan *et al.*, 2010). DNA concentration and quality were estimated using NanoDrop 2000 (Thermo) spectrophotometry and electrophoresis on 0.8% agarose gels with a lambda DNA standard. The genetic relationship between the two parents, F_1 and F_2 individuals were identified using 12 polymorphic SSR markers to avoid condemnation.

Linkage analysis

The polymorphic SSR markers between the two parents were predicted by comparing the whole-genome resequencing data of 'Thailand Chiang Mai' with the 'Chinese Tai-zi' draft genome. RAD-sequencing of the 181 F_2 individuals were generated with the HiSeq 2000 platform (Illumina). SNPs were genotyped using realSFS (Korneliussen *et al.*, 2014) and converted into bin markers using a sliding window approach developed by Huang *et al.* (2009). The R/ASMap package (Taylor and Butler, 2014), R/onemap package (Margarido *et al.*, 2007) and the MAPCOMP pipeline (Sutherland *et al.*, 2016) were used to construct and validate the linkage maps. Detailed methods for the prediction, validation and genotyping of polymorphic SSR markers, RAD-sequencing and SNP calling, linkage map construction and validation are provided in Methods S1.

BioNano mapping and hybrid scaffolding

The draft assembly of 'Chinese Tai-zi' was inspected for frequency of recognition sites of particular nicking enzymes. The genomic DNA of 'Chinese Tai-zi' was purified and embedded in a thin agarose layer and was labelled and counterstained using the IrysPrep Reagent Kit (BioNano Genomics, San Diego, CA, USA) following manufacturer's instructions. The DNA was nicked using 60U of Nt.*BssSI* (New England BioLabs, Beverly, MA, USA) for 2 h at 37°C and subsequently labelled with a fluorescent-dUTP nucleotide analogue. After labelling, the nicks were ligated with *Taq* ligase (New England BioLabs) in the presence of dNTPs for 18 min at 37°C. The labelled DNA was stained with IrysPrep DNA Stain (BioNano Genomics) and subsequently loaded on Irys chips and run for 10 runs. A series of assessing and filtering measures was retained to facilitate assembly. One molecule or label was removed if it matched any condition of the following conditions: (1) molecule length <100 kb; (2) molecule signal-noise ratio (SNR) <3.2; and (3) molecule intensity >0.8. BioNano *de novo* genome mapping was performed using IrysSolve scripts pipeline (version 5134) and IrysSolve tools *Assembler* and *RefAligner* (version 5122). The hybrid scaffolding was performed using IrysSolve HybridScaffold (version 5134) to generate super-scaffolds. The parameters used in BioNano *de novo* genome mapping and hybrid scaffolding procedures are accessible at https://de.cyverse.org/dl/d/F50BCF40–5E0B-4199-B94A-D36A8EEC63B3/DataS1_Bionano_parameters.zip.

Scaffold anchoring and pseudo-chromosome construction

The four linkage maps ('MapBA', 'MapBO', 'MapAll' and 'MapDD') were used to anchor the supper-scaffolds of the 'Chinese Tai-zi' and the 'China Antique' draft genomes using ALLMAPS (Tang *et al.*, 2015). The maps were merged with weights of MapBA, MapBO, MapAll and MapDD set to 2, 2, 2 and 1, respectively. For each scaffold, consecutive SNPs belonging to a same bin were merged to reduce the computation amounts. The scaffold ordering and orientation were performed using Genetic Algorithm with iterations of 500 and population size of 100. The pseudo-chromosomes were constructed by joining the scaffolds with 10 kb gaps. The completeness of the genomes was assessed using BUSCO version 3.0.1 (Simão *et al.*, 2015) with the plants set of single-copy orthologues.

Genome annotations

The pseudo-chromosomes of 'Chinese Tai-zi' were annotated after masking the repeats. Three approaches were used for gene

prediction: transcript sequences-based prediction, *ab initio* gene prediction and homology-based prediction. Detailed methods for the annotation of the 'Chinese Tai-zi' pseudo-chromosomes are provided in Methods S1. The annotations of pseudo-chromosomes of 'China Antique' were converted from the annotations of the *de novo* assembly of 'China Antique' with the liftOver UCSC tool (Kuhn *et al.*, 2012).

Identification of centromere regions and telomere repeats

The genome-wide recombination rates were estimated from the comparison of the linkage maps and the anchored pseudo-molecules using the R/MareyMap package (Rezvoy *et al.*, 2007) by locally adjusting a polynomial curve to the plot of genetic versus physical distances, the slope of the curve was subsequently obtained using the Locally WEighted Scatterplot Smoothing (lowess or loess) method with parameters of span = 0.1 and degree = 2. The regions that show lowest recombination rates in each chromosome were predicted as the possible centromere regions.

NnCenH3 ChIP-Seq reads were aligned to the anchored pseudo-molecules using SOAPaligner with default parameters and only reads that mapped to a unique position were retained for further analysis. Read density was estimated following Gong *et al.* (2012) to map the NnCenH3 enrichment along each chromosome in an unbiased approach. Generally, 49 bp (same as the ChIP-Seq reads length) reads starting from every base pair of the *N. nucifera* chromosome were generated and mapped to the genome using SOAPaligner with default parameters. The 'uniquely mappable region' was defined as the genomic position of the starting nucleotide of a unique read. Subsequently each chromosome was divided into 20 kb windows (spaced every 10 kb) and the unique read number per base pair mappable region was calculated in each window. Thus, read density equals the number of unique reads in a 20 kb window per the length of mappable region in the same window. The NnChenH3 subdomains were identified using SICER version 1.1 (Zang *et al.*, 2009) with a window size of 1 kb, which required the *P*-value of a NnCenH3 subdomain to be <0.0001, and a gap size of 1 kb was allowed in the defined NnCenH3 subdomains.

To identify telomeric sequence, we used the typical angiosperm telomeric 7-mer motif $(TTTAGGG)_n$ to search for exact matches (forward or reverse strand) in the tandem repeat results. The matched motifs with more than 10 repeat times were flagged as potential telomeric repeats.

Genome evolution

The *N. nucifera* evolutionary history was obtained based on the orthologous and paralogous relationships identified between *N. nucifera* (the pseudo-chromosomes of 'China Antique', eight chromosomes, 27 931 genes), *Vitis vinifera* (grape, 19 chromosomes, 33 514 genes) (Jaillon *et al.*, 2007), *Theobroma cacao* (cacao, 10 chromosomes, 28 798 genes) (Argout *et al.*, 2011), *Coffea canephora* (coffee, 11 chromosomes, 25 574 genes) (Denoeud *et al.*, 2014), *Prunus persica* (peach, 8 chromosomes, 27 852 genes) (Verde *et al.*, 2013) and the AEK, following the method described in Salse (2016). Briefly, the first step involves the alignment of the investigated genomes to define conserved/duplicated gene pairs on the basis of alignment parameters (CIP for Cumulative Identity Percentage and CALP for Cumulative Alignment Length Percentage). The second step involves clustering or chaining groups of conserved and duplicated genes into CARs (for Contiguous Ancestral Regions, illustrated as dot plot diagonals) corresponding to independent sets of blocks sharing paralogous and/or orthologous relationships in modern species investigated. From the reconstructed AEK an evolutionary scenario can subsequently be inferred taking into account the fewest number of genomic rearrangements (including inversions, deletions, fusions, fissions, translocations) which may have operated between AEK and the modern genome of *N. nucifera*.

ACCESSION NUMBERS

The RAD-sequencing data generated for this study were deposited in the Sequence Read Archive (Bioproject PRJNA417779). The genotypes of the SNPs were deposited in NCBI dbSNP (Bioproject PRJNA400121). The BioNano genome map data generated for this study is accessible at NCBI Supplementary Files (accession SUPPF_0000001138). The anchored pseudo-chromosomes of 'Chinese Tai-zi' were deposited at GenBank under the accession no. DLUA00000000, the version described in this paper is version no. DLUA01000000. The anchored pseudo-chromosomes of 'China Antique' were deposited at GenBank under the accession no. DLUB00000000, the version described in this paper is version no. DLUB01000000.

The draft assemblies of 'Chinese Tai-zi' and 'China Antique' were downloaded from GenBank under accessions nos. GCA_000805495.1 and GCA_000365185.2, respectively. The resequencing reads of the 'Thailand Chiang Mai' wild were downloaded from the Sequence Read Archive (SRR2131192). The ChIP-Seq data of NnCenH3 were downloaded from the Sequence Read Archive (SRR2970623).

ACKNOWLEDGEMENTS

The authors thank Jianwei Chen and Bin Wu (BGI-Shenzhen, Shenzhen, China) for help in the analysis of BioNano optical maps.

FUNDING

This research is financially supported by the National Natural Science Foundation of China (31271310).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Statistics of the predicted polymorphic SSRs.

Figure S2. The amount of Rad-Seq data in each F_2 individual.

Figure S3. SNP genotyping and recombination breakpoints.

Figure S4. Distribution of BioNano molecules.

Figure S5. Identification of chimeric scaffolds.

Figure S6. The Schema of filtering the chimeric scaffolds and determining the breakpoints.

Figure S7. Scaffold anchoring and pseudo-chromosomes of the 'Chinese Tai-zi' genome.

Figure S8. Scaffold anchoring and pseudo-chromosomes of the 'China Antique' genome.

Figure S9. Synteny between the pseudo-chromosomes and the megascaffolds of the 'China Antique' genome.

Figure S10. Syntenic path dot plot of the 'Chinese Tai-zi' pseudo-chromosomes versus the 'China Antique' pseudo-chromosomes.

Figure S11. Plot of ChIP-Seq sequence reads density along individual pseudo-chromosomes of the 'China Antique'.

Table S1. Details of the predicted polymorphic SSRs.

Table S2. Statistics of the Rad-Seq reads for each F_2 individual.

Table S3. Summary of the SNPs, recombination breakpoints and bin markers.

Table S4. Details of the recombination breakpoints.

Table S5. Details of the bin markers.

Table S6. The genotypes of bin markers and SSRs in the 178 F_2 individuals.

Table S7. Detail marker genetic distances of MapBA, MapBO and MapAll.

Table S8. Summary of the clusters of differently ordered bin markers between MapBA and MapBO.

Table S9. Summary of the clusters of differently ordered bin markers between MapBA and MapAll.

Table S10. Summary of the clusters of differently ordered SSRs between their genetic location in MapAll and their physical locations in the ordered scaffolds according to MapBA.

Table S11. The correspondence between MapBA and MapDD.

Table S12. Segregation distortion markers in N. nucifera F2 population.

Table S13. Statistics of the BioNano raw data.

Table S14. Statistics of the BioNano clean data.

Table S15. Summary of scaffold anchoring and orientating of the 'Chinese Tai-zi' and 'China Antique' assemblies.

Table S16. Summary of the annotations of the 'Chinese Tai-zi' pseudo-chromosomes.

Table S17. Summary of the annotations in the draft assembly and the pseudo-chromosomes of 'China Antique'.

Table S18. BUSCO assessment of the completeness of the draft assemblies and pseudo-chromosomes of 'Chinese Tai-zi' and 'China Antique'.

Table S19. The correlations of repeats with the enrichment of NnCenH3 ChIP-Seq reads.

Methods S1. Supporting Methods.

Appendix S1. Supporting Notes.

REFERENCES

Albert, V.A., Barbazuk, W.B., Der, J.P. *et al.* (2013) The Amborella genome and the evolution of flowering plants. *Science*, **342**, 1241089.

Anantharaman, T.S., Mishra, B. and Schwartz, D.C. (1999) Genomics via optical mapping III: contiging genomic DNA and variations. In *The Seventh International Conference on Intelligent Systems for Molecular Biology: AAAI Press*, pp. 18–27.

Argout, X., Salse, J., Aury, J.M. *et al.* (2011) The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108.

Bartholome, J., Mandrou, E., Mabiala, A. *et al.* (2015) High-resolution genetic maps of Eucalyptus improve *Eucalyptus grandis* genome assembly. *New Phytol.* **206**, 1283–1296.

Bremer, B., Bremer, K., Chase, M. *et al.* (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* **161**, 105–121.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510.

Denoeud, F., Carretero-Paulet, L., Dereeper, A. *et al.* (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, **345**, 1181–1184.

Denton, J.F., Lugo-Martinez, J., Tucker, A.E. *et al.* (2014) Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comp. Biol.* **10**, e1003998.

Diao, Y., Liu, J., Yang, G. *et al.* (2005) Karyotype analysis of *Nelumbo nucifera* and *Nelumbo lutea* by chromosome banding and fluorescence in situ hybridization. *Korean J. Genetic*, **27**, 187–194.

Dong, Y., Xie, M., Jiang, Y. *et al.* (2013) Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* **31**, 135–141.

Eid, J., Fehr, A., Gray, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Ferreira, A., Silva, M.F.D. and Cruz, C.D. (2006) Estimating the effects of population size and type on the accuracy of genetic maps. *Genet. Mol. Biol.* **29**, 187–192.

Fierst, J.L. (2015) Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front. Genet.* **6**, 220.

Gong, Z., Wu, Y., Koblizkova, A. *et al.* (2012) Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell*, **24**, 3559–3574.

Green, P. (1997) Against a whole-genome shotgun. *Genome Res.* **7**, 410–417.

Gui, S., Wu, Z., Zhang, H. *et al.* (2016) The mitochondrial genome map of *Nelumbo nucifera* reveals ancient evolutionary features. *Sci. Rep.* **6**, 30158.

Henikoff, S., Ahmad, K. and Malik, H.S. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098–1102.

Hu, J., Gui, S., Zhu, Z. *et al.* (2015) Genome-wide identification of SSR and SNP markers based on whole-genome re-sequencing of a Thailand wild sacred lotus (*Nelumbo nucifera*). *PLoS ONE*, **10**, e0143765.

Huang, X., Feng, Q., Qian, Q. *et al.* (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19**, 1068–1076.

Jaillon, O., Aury, J.M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463.

Kaur, P., Bayer, P.E., Milec, Z. *et al.* (2017) An advanced reference genome of *Trifolium subterraneum* L. reveals genes related to agronomic performance. *Plant Biotechnol. J.* **15**, 1034–1046.

Kawakami, T., Smeds, L., Backstrom, N. *et al.* (2014) A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Mol. Ecol.* **23**, 4035–4058.

Kianian, S. and Quiros, C. (1992) Generation of a *Brassica oleracea* composite RFLP map: linkage arrangements among various populations and evolutionary implications. *Theor. Appl. Genet.* **84**, 544–554.

Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinform.* **15**, 356.

Kuhn, R.M., Haussler, D. and Kent, W.J. (2012) The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161.

Kujur, A., Upadhyaya, H.D., Shree, T. *et al.* (2015) Ultra-high density intraspecific genetic linkage maps accelerate identification of functionally relevant molecular tags governing important agronomic traits in chickpea. *Sci. Rep.* **5**, 9468.

Lam, E.T., Hastie, A., Lin, C. *et al.* (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776.

Lewin, H.A., Larkin, D.M., Pontius, J. and O'Brien, S.J. (2009) Every genome sequence needs a good map. *Genome Res.* **19**, 1925–1928.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Liu, Z., Zhu, H., Liu, Y. *et al.* (2016) Construction of a high-density, high-quality genetic map of cultivated lotus (*Nelumbo nucifera*) using next-generation sequencing. *BMC Genom.* **17**, 466.

Lyons, E., Pedersen, B., Kane, J. and Freeling, M. (2008) The value of non-model genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop. Plant Biol.* **1**, 181–190.

Margarido, G., Souza, A. and Garcia, A. (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas*, **144**, 78–79.

Mascher, M. and Stein, N. (2014) Genetic anchoring of whole-genome shotgun assemblies. *Front. Genet.* **5**, 208.

Ming, R., VanBuren, R., Liu, Y. *et al.* (2013) Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* **14**, R41.

Mizuta, Y., Harushima, Y. and Kurata, N. (2010) Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. *Proc. Natl Acad. Sci. USA*, **107**, 20417–20422.

Murat, F., Xu, J.H., Tannier, E. *et al.* (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**, 1545–1557.

Murat, F., Zhang, R., Guizard, S. *et al.* (2015) Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. *Genome Biol. Evol.* **7**, 735–749.

Nossa, C.W., Havlak, P., Yue, J.-X. *et al.* (2014) Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. *GigaScience*, **3**, 9.

Olsen, J.L., Rouze, P., Verhelst, B. *et al.* (2016) The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature*, **530**, 331–335.

Pan, L., Xia, Q., Quan, Z. *et al.* (2010) Development of novel EST-SSRs from sacred lotus (*Nelumbo nucifera* Gaertn) and their utilization for the genetic diversity analysis of *N. nucifera*. *J. Hered.* **101**, 71–82.

Potato Genome Sequencing Consortium, Xu, X., Pan, S. *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

Qi, Z., Huang, L., Zhu, R. *et al.* (2014) A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PLoS ONE*, **9**, e104871.

Ren, Y., Zhao, H., Kou, Q. *et al.* (2012) A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PLoS ONE*, **7**, e29453.

Rezvoy, C., Charif, D., Guéguen, L. and Marais, G.A. (2007) MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*, **23**, 2188–2189.

Salse, J. (2016) Ancestors of modern plant crops. *Curr. Opin. Plant Biol.* **30**, 134–142.

Scherthan, H., Cremer, T., Arnason, U. *et al.* (1994) Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nat. Genet.* **6**, 342–347.

Schubert, I. and Lysak, M.A. (2011) Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* **27**, 207–216.

Sharma, B.R., Gautam, L.N., Adhikari, D. and Karki, R. (2017) A comprehensive review on chemical profiling of *Nelumbo Nucifera*: potential for drug development. *Phytother. Res.* **31**, 3–26.

Shen-Miller, J. (2002) Sacred lotus, the long-living fruits of China Antique. *Seed Sci. Res.* **12**, 131–143.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologues. *Bioinformatics*, **31**, 3210–3212.

Soltis, P.S., Liu, X., Marchant, D.B., Visger, C.J. and Soltis, D.E. (2014) Polyploidy and novelty: Gottlieb's legacy. *Phil. Trans. R. Soc. B*, **369**, 20130351.

Stankova, H., Hastie, A.R., Chan, S. *et al.* (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* **14**, 1523–1531.

Sturtevant, A.H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J. Exp. Zool. Part A*, **14**, 43–59.

Sutherland, B.J., Gosselin, T., Normandeau, E. *et al.* (2016) Salmonid chromosome evolution as revealed by a novel method for comparing RAD-seq linkage maps. *Genome Biol. Evol.* **8**, 3600–3617.

Talbert, P.B. and Henikoff, S. (2010) Centromeres convert but don't cross. *PLoS Biol.* **8**, e1000326.

Tang, H., Bowers, J.E., Wang, X. *et al.* (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

Tang, Z., Wang, X., Zhang, M. *et al.* (2013) The maternal cytoplasmic environment may be involved in the viability selection of gametes and zygotes. *Heredity*, **110**, 331–337.

Tang, H., Zhang, X., Miao, C. *et al.* (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3.

Taylor, J. and Butler, D. (2014) ASMap: An (A)ccurate and (S)peedy linkage map construction package for inbred populations that uses the extremely efficient MSTmap algorithm. R package version 0.3.

Taylor, D.R. and Ingvarsson, P.K. (2003) Common features of segregation distortion in plants and animals. *Genetica*, **117**, 27–35.

Venter, J.C., Smith, H.O. and Hood, L. (1996) A new strategy for genome sequencing. *Nature*, **381**, 364.

Verde, I., Abbott, A.G., Scalabrin, S. *et al.* (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494.

Villasante, A., Abad, J.P. and Méndez-Lago, M. (2007) Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome. *Proc. Natl Acad. Sci. USA*, **104**, 10542–10547.

Wang, Y., Fan, G., Liu, Y. *et al.* (2013) The sacred lotus genome provides insights into the evolution of flowering plants. *Plant J.* **76**, 557–567.

Wu, R., Ma, C.-X., Painter, I. and Zeng, Z.-B. (2002) Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. *Theor. Popul. Biol.* **61**, 349–363.

Wu, Y., Bhat, P.R., Close, T.J. and Lonardi, S. (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* **4**, e1000212.

Wu, K., Liu, H., Yang, M. *et al.* (2014a) High-density genetic map construction and QTLs analysis of grain yield-related traits in Sesame (*Sesamum indicum* L.) based on RAD-Seq technology. *BMC Plant Biol.* **14**, 274.

Wu, Z., Gui, S., Quan, Z. *et al.* (2014b) A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biol.* **14**, 289.

Yang, J., Zhao, X., Cheng, K. *et al.* (2012a) A killer-protector system regulates both hybrid sterility and segregation distortion in rice. *Science*, **337**, 1336–1340.

Yang, M., Han, Y., VanBuren, R. *et al.* (2012b) Genetic linkage maps for Asian and American lotus constructed using novel SSR markers derived from the genome of sequenced cultivar. *BMC Genom.* **13**, 653.

Zamir, D., Tanksley, S.D. and Jones, R.A. (1982) Haploid selection for low temperature tolerance of tomato pollen. *Genetics*, **101**, 129–137.

Zang, C., Schones, D.E., Zeng, C. *et al.* (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.

Zhang, Q., Li, L., VanBuren, R. *et al.* (2014) Optimization of linkage mapping strategy and construction of a high-density American lotus linkage map. *BMC Genom.* **15**, 372.

Zhu, Z., Gui, S., Jin, J. *et al.* (2016) The NnCenH3 protein and centromeric DNA sequence profiles of *Nelumbo nucifera* Gaertn. (sacred lotus) reveal the DNA structures and dynamics of centromeres in basal eudicots. *Plant J.* **87**, 568–582.