



## ORIGINAL RESEARCH

# Chromosome-level Genomes Reveal the Genetic Basis of Descending Dysploidy and Sex Determination in *Morus* Plants



Zhongqiang Xia<sup>1,#</sup>, Xuelei Dai<sup>2,#</sup>, Wei Fan<sup>1,#</sup>, Changying Liu<sup>3</sup>, Meirong Zhang<sup>1</sup>,  
 Peipei Bian<sup>2</sup>, Yuping Zhou<sup>1</sup>, Liang Li<sup>1</sup>, Baozhong Zhu<sup>1</sup>, Shuman Liu<sup>1</sup>,  
 Zhengang Li<sup>4</sup>, Xiling Wang<sup>5</sup>, Maode Yu<sup>5</sup>, Zhonghuai Xiang<sup>1</sup>, Yu Jiang<sup>2,\*</sup>,  
 Aichun Zhao<sup>1,\*</sup>

<sup>1</sup> State Key Laboratory of Silkworm Genome Biology, Institute of Sericulture and Systems Biology, Southwest University, Chongqing 400716, China

<sup>2</sup> Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China

<sup>3</sup> Key Laboratory of Coarse Cereal Processing, Ministry of Agriculture and Rural Affairs, Chengdu University, Chengdu 610106, China

<sup>4</sup> The Sericultural and Apicultural Research Institute, Yunnan Academy of Agricultural Sciences, Mengzi 661100, China

<sup>5</sup> College of Sericulture, Textile and Biomass Sciences, Southwest University, Chongqing 400716, China

Received 31 October 2021; revised 2 July 2022; accepted 23 August 2022

Available online 30 August 2022

Handled by Peng Cui

## KEYWORDS

Mulberry;  
 Karyotype evolution;  
 Dioecy;  
 Sex determination;  
 Population genomics

**Abstract** Multiple plant lineages have independently evolved sex chromosomes and variable karyotypes to maintain their sessile lifestyles through constant biological innovation. *Morus notabilis*, a dioecious mulberry species, has the fewest chromosomes among *Morus* spp., but the genetic basis of **sex determination** and **karyotype evolution** in this species has not been identified. In this study, three high-quality genome assemblies were generated for *Morus* spp. [including dioecious *M. notabilis* (male and female) and *Morus yunnanensis* (female)] with genome sizes of 301–329 Mb and were grouped into six pseudochromosomes. Using a combination of genomic approaches, we found that the putative ancestral karyotype of *Morus* species was close to 14 protochromosomes, and that several chromosome fusion events resulted in descending dysploidy ( $2n = 2x = 12$ ). We also characterized a ~ 6.2-Mb sex-determining region on chromosome 3. Four potential male-specific genes, a

\* Corresponding authors.

E-mail: [zhaoaichun@swu.edu.cn](mailto:zhaoaichun@swu.edu.cn) (Zhao A), [yu.jiang@nwafu.edu.cn](mailto:yu.jiang@nwafu.edu.cn) (Jiang Y).

# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

<https://doi.org/10.1016/j.gpb.2022.08.005>

1672-0229 © 2022 The Authors. Published by Elsevier B.V. and Science Press on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

partially duplicated DNA helicase gene (named *MSDH*) and three *Ty3\_Gypsy* long terminal repeat retrotransposons (named *MSTG1/2/3*), were identified in the Y-linked area and considered to be strong candidate genes for sex determination or differentiation. Population genomic analysis showed that Guangdong accessions in China were genetically similar to Japanese accessions of mulberry. In addition, genomic areas containing selective sweeps that distinguish domesticated mulberry from wild populations in terms of flowering and disease resistance were identified. Our study provides an important genetic resource for sex identification research and molecular breeding in mulberry.

## Introduction

Mulberry (*Morus* spp.) is a major member of the family Moraceae, which also includes various other important plant species, such as banyan tree and paper mulberry. As one of the earliest domesticated plants, mulberry is considered an “oriental sacred wood”, because it is not only a good food source for rearing silkworms but has also been utilized in a number of other ways, including as a fruit, for landscaping, in medicine, in ecological protection, and as forage for animal production [1–3]. The draft genome sequence assembly of male *Morus notabilis* was published in 2013 based on short sequencing reads [4]. In 2020, genome sequencing research revealed two diploid karyotypes in mulberry species [5]. Congeneric species commonly display varying chromosome counts due to two opposing trends: 1) an increased chromosome copy number as a result of polyploidy [whole-genome duplication (WGD)] and 2) a reduced basic chromosome number via structural rearrangements (descending dysploidy) [6,7]. These processes have long been considered important in speciation due to the shock of chromosome number discrepancies in reproductive isolation [8]. Multi-karyotype evolutionary models for mulberry are limited due to the existence of only a few comparative analyses of multiple chromosome-scale genomes. Without considering genomic synteny and large-scale rearrangement events, misleading conclusions on pathway evolution have likely been obtained from evolutionary and comparative analyses of genes for important agronomic traits. Similar to *M. notabilis*, *Morus yunnanensis* is also commonly found in Southwest China due to its unique altitude and humidity requirements [9]. The phylogenetic relationship between wild *M. notabilis* and *M. yunnanensis* is unknown. In addition, our previous study revealed the population structure of cultivated mulberry but lacked sampling of gene pools within wild mulberry and in the Guangdong region, leaving gaps in our knowledge and raising important questions about the evolutionary history of domesticated perennial mulberry [5].

Mulberry species are either dioecious or hermaphroditic, thus providing abundant resources for research on plant sex determination [10,11]. The sex determination system of mulberry is similar to that in humans (XY type) [12]. Specific DNA markers associated with the sex determination of male flowers in mulberry have been identified using restriction site-associated DNA sequencing [13]. However, the genotypes of these identified markers were not correlated with sex determination among cultivars, and genomic evidence remains unavailable. Previously, clonal propagation was widely used in mulberry propagation, and consequently, sex determination has been little studied in this genus. Moreover, the key genetic basis of sex chromosomes in mulberry has yet to be discovered.

With the development of the fruit mulberry industry, cultivation of mulberry cultivars with stable sex characteristics has become increasingly important to ensure high fruit production. Knowledge of sex-specific gene expression is a prerequisite for developing an effective breeding program for mulberry (especially fruit mulberry). Dissecting the genetic mechanisms underlying dioecy (*i.e.*, separate male and female trees) is crucial for understanding the evolution of this widespread reproductive strategy. *M. notabilis* always shows dioecious, which is not affected by the environment, making it an ideal system for studying the genetic basis of sex determination and sex chromosome evolution.

In this study, we performed *de novo* assembly of two chromosome-level genomes of female *M. yunnanensis* and female *M. notabilis* to improve the previous draft sequence of male *M. notabilis*, and revealed the evolutionary genomic basis of descending dysploidy and sex determination. Compared with our previously reported *Morus alba* genome, the genomes of the two wild species were more diverse in terms of chromosome number and environmental adaptation level, providing new insights into karyotype evolution in *Morus*. Through a multidata combination analysis, we not only identified the putative sex-determining region (SDR) and candidate loci responsible for sex determination but also provided a clear framework for broader studies of sex determination in mulberry species in the future. Population genomic analyses further revealed the phylogenetic relationships within the gene pool of different *Morus* species and the genetic architecture of domestication traits of cultivated mulberry. These sequencing results will provide a valuable resource for future mulberry research and breeding programs.

## Results

### Assembly and annotation of three high-quality *Morus* genomes

To assemble the genomes of male and female individuals of dioecious *M. notabilis* and female *M. yunnanensis*, we used a combination of sequencing technologies including Pacific Biosciences (PacBio) long reads, Illumina paired-end reads, and high-throughput chromosome conformation capture (Hi-C) sequencing reads. Based on *k*-mer counting, the estimated genome sizes of *M. notabilis* and *M. yunnanensis* were approximately 280 Mb and 296 Mb, respectively (Figure S1A–C). The genome assemblies of female *M. notabilis* and female *M. yunnanensis* were highly contiguous, with 96.4% and 95.1% of genome contigs anchored to chromosomes by Hi-C scaffolding, respectively (with genome assembly sizes of 301 Mb and 313 Mb and contig N50 values of 2.7 Mb and 6.5 Mb, respectively). Moreover, Hi-C scaffolding revealed that

91.8% of the original male *M. notabilis* sequences were anchored on six pseudochromosomes with a contig N50 of 6.8 Mb, indicating considerably improved assembly contiguity compared with the original version (Figure S1D–F; Table S1). The coverage of the genome assembly obtained here was evaluated using high-coverage Illumina sequencing data and transcriptome reads mapped against the assembled genome (Table S2). The scores of the long terminal repeat-retrotransposon (LTR-RT) assembly index [14] for male *M. notabilis*, female *M. notabilis*, and female *M. yunnanensis* were 19.98, 20.48, and 21.25, respectively, and the Benchmarking Universal Single-Copy Orthologs (BUSCO) values were 93.5%, 93.6%, and 94.1%, respectively. These results imply that the three genomes are of high quality (Table 1).

For genome annotation, repetitive sequences in the genome were initially annotated by combining *de novo* and homology-based predictions (Table S3). Protein-coding genes were further annotated by combining *ab initio*, homology, and transcriptome analysis methods (Table S3). In total, 25,333, 25,391, and 24,851 protein-coding genes with 95.6%, 96.1%, and 96.0% BUSCO completeness were predicted in male *M. notabilis*, female *M. notabilis*, and female *M. yunnanensis*, respectively, with 98.2%, 98.2%, and 98.8%, respectively, of these genes being functionally annotated in public databases (Table 1).

### Repetitive elements drive genome expansion of mulberry species

Synteny analysis of two *Morus* genomes with grape (*Vitis vinifera*) did not reveal WGD after the triplication event shared by eudicots, similar to the pattern previously observed in the *M. alba* genome [5] (Figure S2). Furthermore, the collinearities

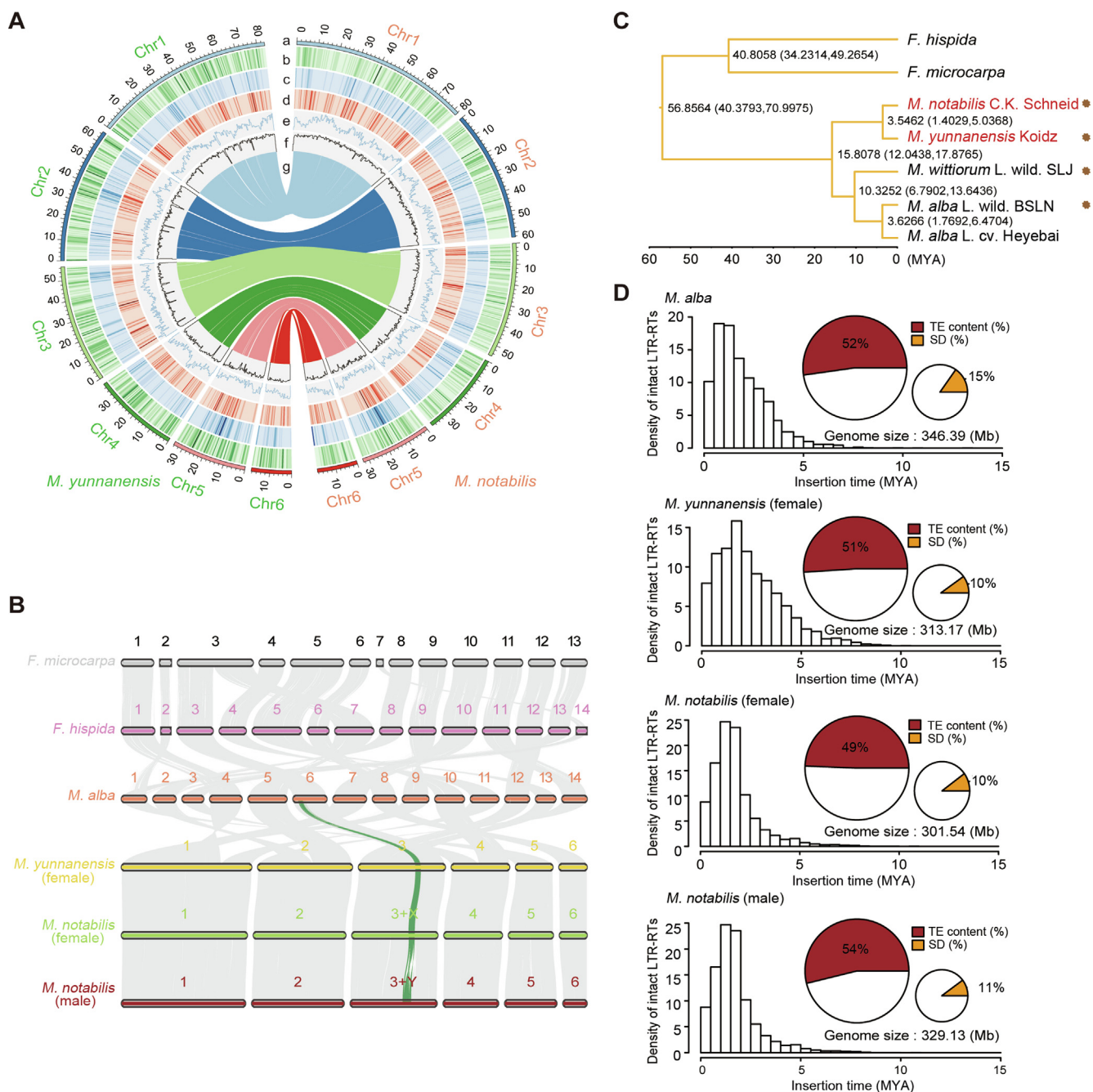
of the genomes of *M. notabilis*, *M. yunnanensis*, *Ficus microcarpa*, and *Ficus hispida* revealed a high frequency of chromosomal rearrangement events, confirming that *M. notabilis* and *M. yunnanensis* were diploid ( $2n = 2x = 12$ ) (Figure 1A and B). Phylogenetic analysis of eight angiosperms using single-copy orthologues identified by OrthoFinder [15] revealed that *M. notabilis* and *M. yunnanensis* diverged only approximately 3 million years ago (MYA) (Figure S3), indicating their close phylogenetic relationship. Combined with similar histological morphologies (Figure S4), these results suggest that *M. notabilis* and *M. yunnanensis* belonged to the same subgenus. Moreover, using a combination of several wild mulberry accessions and *Ficus* as the outgroup, we constructed an evolutionary tree with single nucleotide polymorphisms (SNPs) of fourfold degenerate (neutral) evolving sites based on the maximum-likelihood method in IQ-TREE [16]. The evolutionary distances and divergence time between *M. notabilis* and domesticated *M. alba* were far greater than those observed between other wild mulberry trees (Figure 1C), suggesting that these two species may belong to parallel evolutionary clades.

Whole-genome alignment analysis showed that 30.2 Mb (9% of the assembled genome) covering 870 genes in *M. notabilis* and 51.4 Mb (15%) covering 948 genes in *M. alba* were segmental duplications (SDs) (Figure 1D; Table S3), indicating that SD was a major contributor to genome-size expansion. Genes overlapping within the female *M. notabilis* SD regions were significantly enriched in fatty acid degradation biological processes (“ath00071”, hypergeometric test, adjusted  $P$  value  $< 8.10 \times 10^{-5}$ ) (Figure S5). We identified 54% of the male *M. notabilis* genome, 49% of the female *M. notabilis* genome, and 51% of the female *M. yunnanensis* genome as transposable elements (TEs) (Figure 1D; Table S3).

**Table 1 Comparison of the genome assemblies and annotations of *M. notabilis*, *M. yunnanensis*, *M. alba*, and *M. notabilis***

	<i>M. notabilis</i> (female)	<i>M. notabilis</i> (male)	<i>M. yunnanensis</i> (female)	<i>M. alba</i> [5]	<i>M. notabilis</i> [4]
<b>Assembly</b>					
Genome size (bp)	301,544,460	329,129,568	313,175,542	346,393,484	320,378,613
No. of contigs	539	157	234	398	46,842
N50 of contigs (bp)	2,710,835	6,854,161	6,538,880	2,710,056	40,438
GC content (%)	0.3511	0.3598	0.3558	0.3429	0.3486
BUSCO completeness	93.6%	93.5%	94.1%	94.3%	92.2%
Length of chromosome-scale scaffolds (bp)	290,812,449	302,180,740	298,102,870	326,128,411	NA
Anchor rate	96.4%	91.8%	95.1%	94.1%	NA
LAI	20.48	19.98	21.25	18.58	8.09
<b>Annotation</b>					
Number of predicted protein-coding genes	25,391	25,333	24,851	22,767	29,338
Average gene length (bp)	3140.73	3150.88	3121.06	3209	NA
Average CDS length (bp)	1164.39	1160.71	1172.80	1148	NA
Average exon number	4.86	4.83	5.00	5.09	NA
BUSCO completeness	96.1%	95.6%	96.0%	NA	NA
Length of repeat sequences (bp)	164,410,597	192,806,284	166,442,299	180,113,984	127,983,832
No. of genes annotated to Swiss-Prot	21,959	24,142	23,820	17,381	17,826
No. of genes annotated to InterPro	24,275	24,121	23,804	17,551	17,681
No. of genes annotated to NR	14,318	24,139	23,818	22,109	22,458
No. of genes annotated to TrEMBL	24,293	24,154	23,827	22,109	NA
No. of genes annotated to KEGG	24,258	24,109	23,752	13,124	12,688
Annotation rate	98.2%	98.2%	98.8%	NA	NA

*Note:* The assemblies of dieocious *M. notabilis* (male and female) and *M. yunnanensis* (female) were compared with two previously reported genome assemblies of *M. alba* and *M. notabilis* (contig-level). NA indicates that data were not included in the original articles. BUSCO, Benchmarking Universal Single-Copy Orthologs; LTR, long terminal repeat; LAI, LTR Assembly Index; CDS, coding sequence; KEGG, Kyoto Encyclopedia of Genes and Genomes.



**Figure 1** Genome evolution in the genus *Morus*

**A.** Genome landscape of *Morus notabilis* and *Morus yunnanensis*. The outer circle represents chromosome karyotypes of *M. notabilis* and *M. yunnanensis*. The tracks indicate parameters described below: a, chromosome length (Mb); b, SD; c, density of all Gypsy LTR-RTs; d, density of all Copia LTR-RTs; e, gene density; f, GC content; g, synteny between the two genomes. **B.** Chromosome synteny between *F. microcarpa*, *F. hispida*, *M. alba*, *M. yunnanensis*, and *M. notabilis*, with chromosome numbers shown above. The green line indicates the synteny blocks in the Y-linked regions of female and male *Morus*. **C.** Phylogenetic relationships among different subgenera of mulberry. A maximum-likelihood tree was constructed by extracting neutrally evolving sites from resequencing data from *M. notabilis* C.K. Schneid, *M. yunnanensis* Koidz, *M. wittiorum* L. wild. SLJ, *M. alba* L. wild. BSLN, and *M. alba* L. cv. Heyebai (with *F. microcarpa* and *F. hispida* as outgroups). The divergence time among different groups of species is labeled on the nodes. Clade support values near nodes represent the estimates of divergence time with a 95% credibility interval. The asterisk represents the wild species. **D.** LTR burst patterns and fractions of TEs and SD in *M. alba*, female *M. yunnanensis*, and female and male *M. notabilis*. *M. notabilis*, *Morus notabilis*; *M. yunnanensis*, *Morus yunnanensis*; *F. microcarpa*, *Ficus microcarpa*; *F. hispida*, *Ficus hispida*; *M. alba*, *Morus alba*; Chr, chromosome; LTR-RT, long terminal repeat-retrotransposon; TE, transposable element; SD, segmental duplication; MYA, million years ago.

Genome-wide proliferation of intact LTR-RTs from *M. yunnanensis* and female and male *M. notabilis* occurred approximately 1.5 MYA. Moreover, these LTR-RTs were dated to more recent than *M. alba* formation (Figure 1D; Table S3). Therefore, we concluded the occurrence of a high level of divergence in SDs and TEs between *M. notabilis* and *M. alba*.

We also identified an average of 13,179,790 SNPs between the *M. notabilis* and *M. yunnanensis* genomes in one-to-one aligned regions (Table S4). The total length of small insertions and deletions (50–500 bp) in those one-to-one aligned regions was 644,198 bp, which accounted for approximately 0.21% of the *M. notabilis* genome (Table S4). Notably, fewer SNPs, small insertions, and insertion–deletion mutations (indels) were observed between the *M. notabilis* and *M. yunnanensis* genomes than between the *M. notabilis* and *M. alba* genomes, suggesting a strong relationship between the two wild-grown plants, a finding consistent with the results from the phylogenetic tree described above. Two sequences of 31.91 Mb and 125.95 Mb were affected by structural variants (SVs) in the two comparisons. In particular, most (71%) of the SV sequences in the comparison between *M. notabilis* and *M. alba* showed repeat expansion and contraction.

### Chromosomal fusion in mulberry genomes is associated with adaptive evolution

Chromosomal evolution is associated with genome size, gene family evolution, and speciation. Genome structural changes led to the present-day karyotypes of *M. notabilis* ( $2n = 12$ ) and *M. alba* ( $2n = 28$ ). Using the available genomes of *M. notabilis*, *M. alba*, *Ficus microcarpa*, *Ficus hispida*, and *Cannabis sativa* (as the outgroup), we reconstructed the ancestral karyotype of the Moraceae ( $n = 21$ ) (Figure 2A), which corresponded to the ancestral eudicot karyotype (AEK) [17]. Compared with the putative ancestral chromosomes, 22 and 24 large syntenic blocks were identified in *M. notabilis* and *M. alba*, respectively (Table S5), which enabled us to deduce the arrangements of ancestral chromosome segments in mulberry. Karyotyping of *M. alba* revealed that at least ten major chromosomal fusions (CFUs) and one chromosomal fission of 21 chromosomes of the paleohexaploid ancestor may have been involved. Approximately half of the chromosomes of *M. alba* were found to have descended from a single ancient chromosome, a finding similar to the results reported for paper mulberry [18]. Moreover, pseudochromosomes of *M. notabilis* were constructed from the ancestor karyotype via at least 18 CFU events and one chromosomal fission event, resulting in a substantial decrease in the haploid chromosome number from 21 to 6. Furthermore, *MnChr1*, *MnChr2*, *MnChr3*, and *MnChr4* were each derived from at least two *M. alba* chromosomes via complex translocations. The high-quality genome assemblies revealed that the karyotype of *M. alba* is more similar than that of *M. notabilis* to the ancestral karyotype, whereas the karyotype of *M. notabilis* was derived from more CFUs. We performed repeat element annotation analysis to investigate the genome features of evolutionary fusion regions (EFRs) and found that the EFRs in the *M. notabilis* genomes were mainly enriched with LTR elements (50.29%) (Table S6).

We identified 26 genes located in six CFU regions (Figure 2B). The Kyoto Encyclopedia of Genes and Genomes

(KEGG) enrichment analysis confirmed that the genes in the rearranged regions of *M. notabilis* and *M. alba* chromosomes were mainly enriched in functions related to photosynthesis (“ath00195”, hypergeometric test, adjusted  $P$  value  $< 6.35 \times 10^{-10}$ ) and metabolic pathways (“ath01100”, hypergeometric test, adjusted  $P$  value  $< 1.55 \times 10^{-4}$ ) (Figure S6A). The Gene Ontology (GO) annotation results showed that the rearranged regions were mainly involved in adenosine diphosphate-binding sites (“GO: 0043531”, hypergeometric test, adjusted  $P$  value  $< 4.34 \times 10^{-9}$ ) and positive regulation of gene expression (“GO: 0010628”, hypergeometric test, adjusted  $P$  value  $< 2.13 \times 10^{-5}$ ) (Figure S6B), which prompted us to investigate the function of chromosomal shuffling loci in *Morus* species. We then examined the transcriptome expression levels of these genes in *M. notabilis*, *M. yunnanensis*, and the cultivars of *M. alba* (Figure 2C; Table S7). Most of these fusion genes were differentially expressed in the two karyotypes of mulberry. These findings provide novel insights into chromosome evolution and link chromosomal rearrangements to the evolution of functional genes.

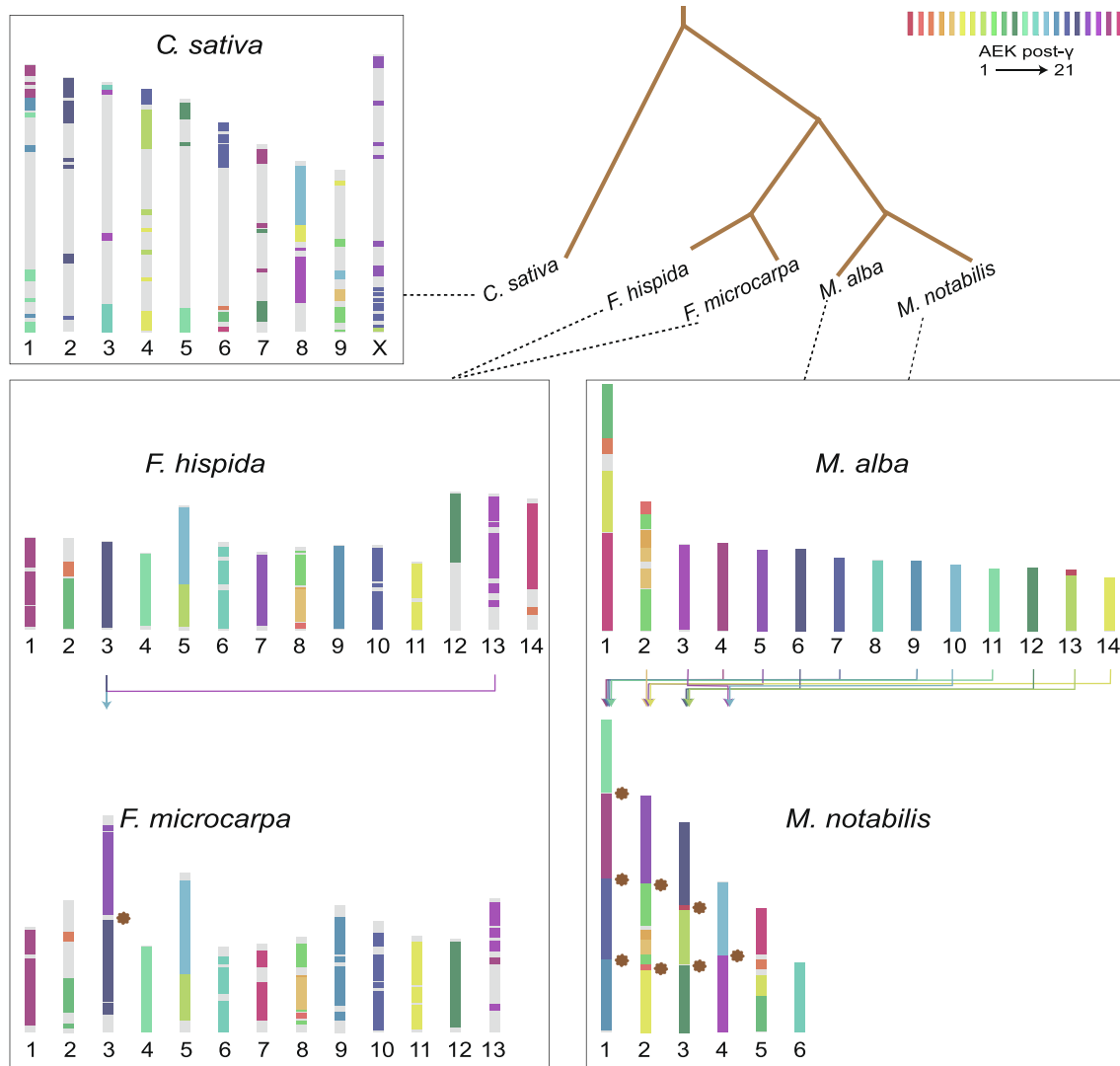
### Location of SDR and identification of candidate sex-determining genes

Short Illumina reads of four male and four female plants of *M. notabilis* were subsequently catalogued into 40-bp  $k$ -mers in different categories using the genome-wide classification method to identify the putative SDR and sex locus regions [19,20]. Sex-specific  $k$ -mers (detected in all samples of one sex, but not in the other sex) were obtained, including 333,348 male-specific  $k$ -mers (MSKs) and 1664 female-specific  $k$ -mers (FSKs). The higher MSK count in males was consistent with the results obtained for persimmon and ginkgo [19,20], suggesting unique genomic regions in male individuals. This result suggested that the sex determination system of *M. notabilis* is an XY system.

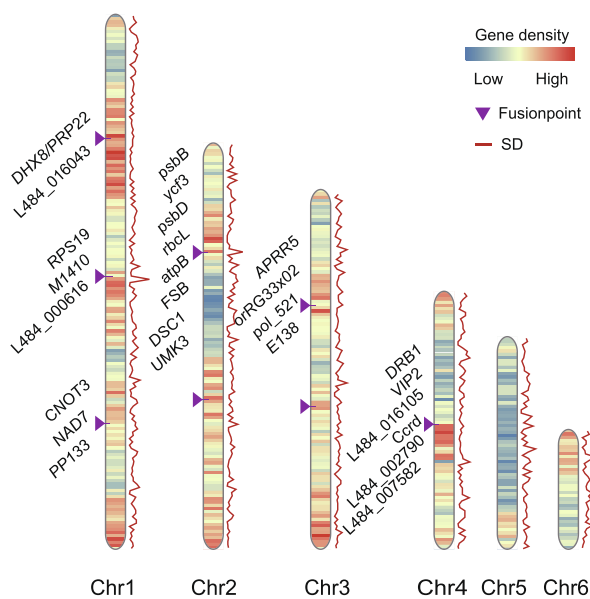
Based on the position of male-specific reads in the genome assembly, we identified a candidate SDR (Chr3: 38,911,287–45,186,478) that contained male-specific reads with 100-kb windows, whereas female-specific reads were found to be uniformly distributed throughout the genome (Figure 3A, tracks a and b; Figure S7). We then detected higher densities of SNPs and indels in the SDR in male *M. notabilis* individuals than in female individuals but not in the rest of the genome, indicating early divergence between the Y and X chromosomes (Figure 3A, tracks c–e). In addition, we further analyzed the genome-wide methylation levels of male and female flowers, and found that CG and CHG methylation levels were higher in the candidate SDR of males (Figure 3A, tracks f–h).

We delineated the SDR-X and SDR-Y haplotypes in this region (Figure 3B). The collinearity results indicated that some fragments of the X and Y haplotypes were not aligned, which motivated us to identify any inversions present on the Y chromosome in *M. notabilis* that masked recombination, consistent with the results obtained from banyan tree [21]. Based on this result, these sex-specific regions contained additional candidate genes for sex determination. From the analysis, 404 genes were predicted in the SDR-Y haplotype, and differentiation of Y and X haplotypes in this region provided strong evidence for the presence of a fully linked region on Chr3. A total of 306

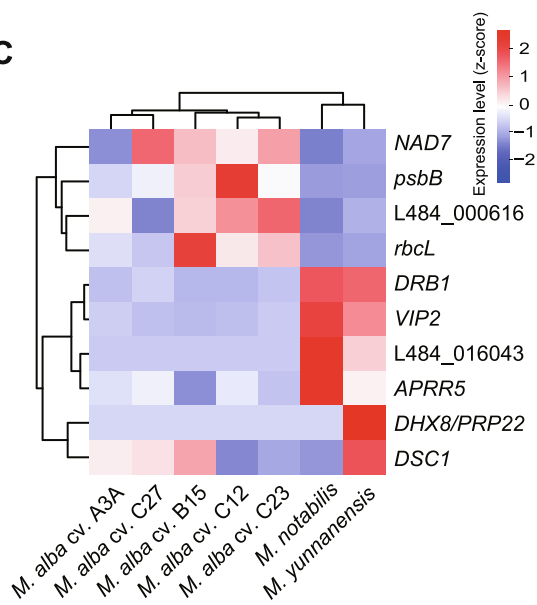
**A**



**B**



**C**



conserved gene pairs (75.76%) from synteny blocks within the SDR and X counterpart were detected using MCscan (Figure S8; Table S8). We further evaluated two genomic datasets for sex determination as follows: 1) based on resequencing data of different sexes, we determined the regions present only in male *Morus* genomes, and 2) the expression of sex bias-related genes in male flowers was measured. Notably, all analyses revealed the existence of a 5-kb region (Chr3: 44,129,165–44,134,608; Figure 3C and D, Figures S9 and S10).

Manual annotation and curation revealed four genes in this region: three male-specific *Ty3\_Gypsy* retrotransposon (RT) gene models (designated *MSTG1*, *MSTG2*, and *MSTG3*) and one male-specific DNA helicase gene (designated *MSDH*) resulting from partial duplication of the *RECQL1* gene (located on Chr2 and found in both male and female genomes) (Figure 3E). The transcriptomic analysis showed that the four male-specific genes were expressed at high levels in male flowers, suggesting their importance in male development and maintenance in *M. notabilis* (Figure 3F; Table S9).

In particular, the coding region of the *MSDH* gene showed 95.25% nucleotide identity with that of canonical *RECQL1*, with considerable divergence at the N-terminus. The *RECQL1* gene contained 10 exons, while *MSDH* had only seven exons (Figure 3E). Moreover, the *MSDH* gene in *M. notabilis* appeared to have undergone a splicing event involving its second exon. *MSDH* was predicted to have helicase\_ATP\_BIND and helicase\_C domains and structural similarities to *RECQL1* (Figure S11), which are critical for anti-crossover signaling during meiotic recombination [22].

We used polymerase chain reaction (PCR) primers designed to amplify four separate gene fragments and detected target fragments of each sex in a collection of dioecious mulberry species, including *M. notabilis* and species from other subgenera. For the *MSDH* gene, we designed a specific gene primer fragment (named *MSDH*-sp1; physical location shown in Figure 3G) for male-specific amplification. Four gene fragments were successfully amplified in all male individuals, whereas no amplified products were obtained from female individuals (Figure 3G, Figures S12 and S13). Based on these results, we revealed a possible XX/XY-determining system and putative candidate sex-determining genes in *Morus*.

### Population structure analysis improves the landscape of genetic affinity in mulberry accessions

Wild relatives are expected to be important sources of genetic diversity in mulberry, and understanding the phylogenetic relationship between wild and cultivated mulberry accessions might greatly facilitate mulberry breeding. In the present study, we resequenced 32 representative wild and landrace specimens from various regions, including Cambodia, Sri Lanka, China, and other countries. Using these data and pub-

licly available genomic sequences for 123 cultivars/landraces, the genetic divergence among multiple wild and domesticated species was studied (Table S10). All sequence reads were mapped to the *M. alba* genome with an average coverage depth of  $\sim 21.6\times$  (Table S11), and a total of 29,185,577 SNPs were identified and used in subsequent population-based genomic analyses (Figure 4A; Table S12).

We performed ADMIXTURE analysis, neighbor-joining (NJ) tree analysis, and principal component analysis (PCA) using genomic SNPs to investigate the phylogenetic relationships between wild relatives and cultivated mulberry. The NJ tree showed clustering of mulberry accessions into six separate genetic groups with the *Ficus* genus (including *F. hispida* and *Ficus macrocarpa*) at the root (Figure 4B). The first group (CHS) included the wild-growing plants *M. notabilis* and *M. yunnanensis* from Southwest China. The second group (WD) included 11 wild plants collected in China and other countries. The third group (JP) mainly included landraces and cultivars from Japan and other countries. The fourth group (GD) consisted of landraces and cultivated species mainly from Guangdong Province. The fifth group (HU) was mainly derived from Taihu Basin in the southern Yangtze delta plain in China, whereas the sixth group (NH) consisted of cultivated mulberry accessions from other places, mainly distributed throughout northern China. The phylogeny of the wild species revealed their close relationships with the cultivated relatives in the JP and GD groups. PCA was performed to confirm these phylogenetic relationships (Figure 4C and D). When higher *k* values were used (Table S13), the NH group was further divided into subgroups, including those from Northeast China, Southwest China, Northwest China, North-central China, and Northeast China (Figure 4E, Figure S14).

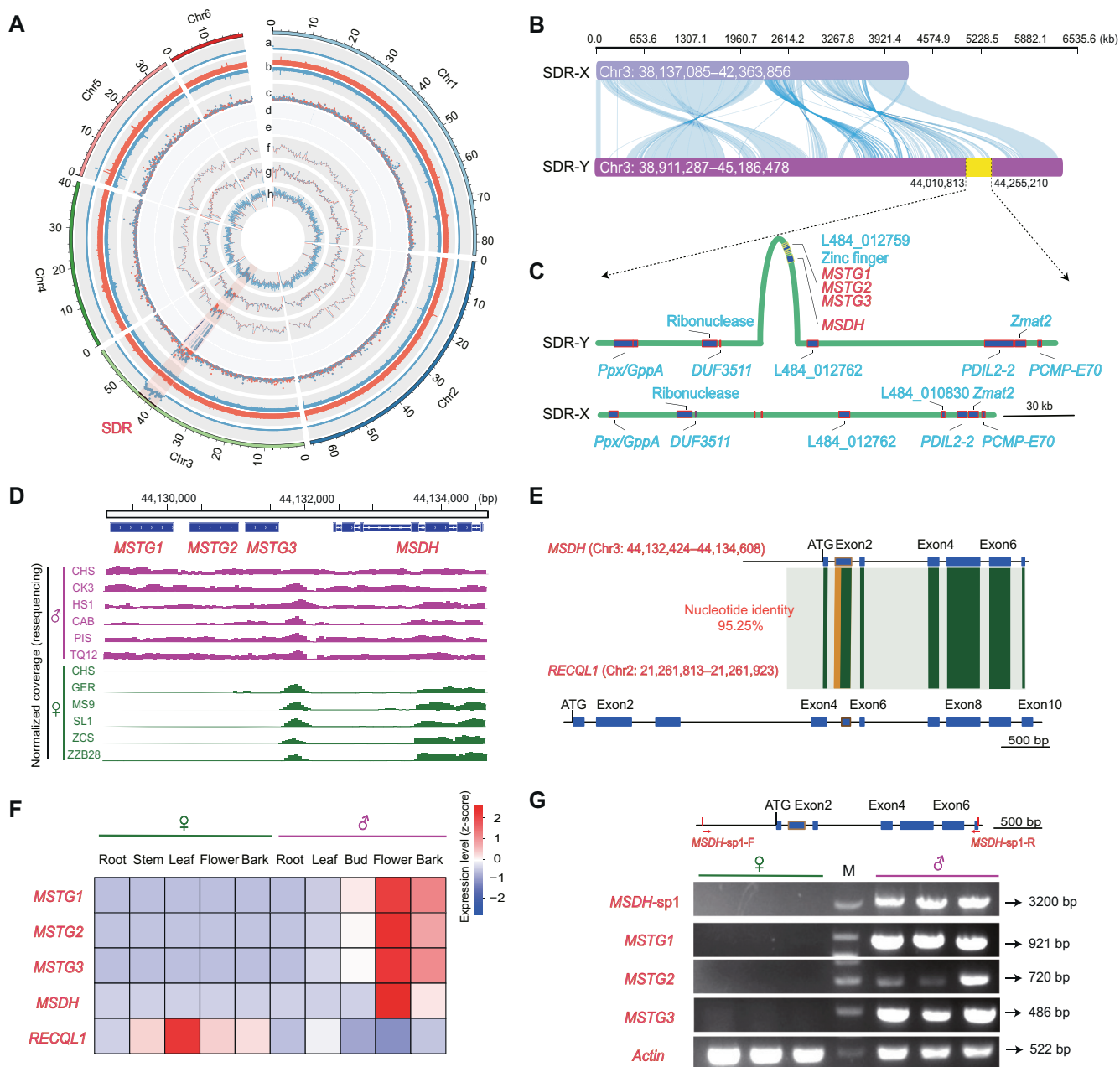
The results presented in Figure 4F show that nucleotide diversity ( $\pi$ ) was highest in WD, followed by JP and GD (Table S14). The highest  $\pi$  ( $5.51 \times 10^{-3}$ ) among the four domesticated groups was observed in JP, which also showed the most singletons and the highest linkage disequilibrium (LD) decay rate (Figure 4G). We further explored the phylogeny and migration history of cultivated and wild mulberry using Treemix [23]. When using CHS as an outgroup, a bifurcation pattern similar to the initial phylogenetic result was observed when up to two migration events were included, and gene flow to HU cultivated mulberry was observed in the JP group (Figure S15).

### Screening for selective sweeps related to domestication

Present-day mulberry cultivars exhibit diversity in many agronomic characteristics, such as flowering time, disease resistance, and leaf development. Among them, traits such as resistance strength and delayed flowering have been recognized as important agronomic traits in cultivated mulberry. We

## Figure 2 Reconstruction of ancestral chromosomes of Moraceae with *C. sativa* as an outgroup

**A.** Probable distribution of ancestral chromosome segments in the genomes of banyan trees and *Morus* plants according to the AEK model proposed by Murat and his colleagues [17]. Blocks are “painted” with colors corresponding to ancestral chromosomes (AEK1–AEK21). Brown asterisks in the *M. notabilis* chromosome diagram indicate sites of chromosomal rearrangement. **B.** Genome-wide landscape of chromosome fusion features in *M. notabilis*. The purple triangle represents the rearrangement site. The heatmap and the red line represent the distribution of gene density and SDs on the chromosomes, respectively. **C.** Transcriptome expression levels of genes in the fusion regions of two karyotypes of *Morus* plants. *C. sativa*, *Cannabis sativa*; AEK, ancestral eudicot karyotype.



**Figure 3** Reconstructed haplotypes in the SDR and identification of candidate sex-determining genes in *M. notabilis*

**A.** Identification of the sex chromosome and SDR among six chromosomes of the *M. notabilis* genome. The tracks indicate the parameters described below: a, Manhattan plot of the mapping depths of MSKs in the male *M. notabilis* genome (100-kb window); b, mapping coverage of female (red) and male (blue) *M. notabilis* Illumina reads; c, SNP and indel densities in female (red) and male (blue) *M. notabilis*; d, heatmap showing the density of candidate male-specific SNPs; e, heatmap showing the density of candidate male-specific indels with the same color coding as track d; f–h, whole-genome methylation levels in CG (f), CHG (g), and CHH (h) contexts are shown.

**B.** SDR-X and SDR-Y haplotypes were reconstructed from genome sequences in our assembly. The plot was created using RectChr (<https://github.com/BGI-shenzhen/RectChr>).

**C.** Gene distribution diagram of candidate regions.

**D.** Validation of four candidate sex-specific genes in two sex phenotypes (six females and six males) using IGV. Purple represents males, and green represents females. The top panel presents the structure of genes.

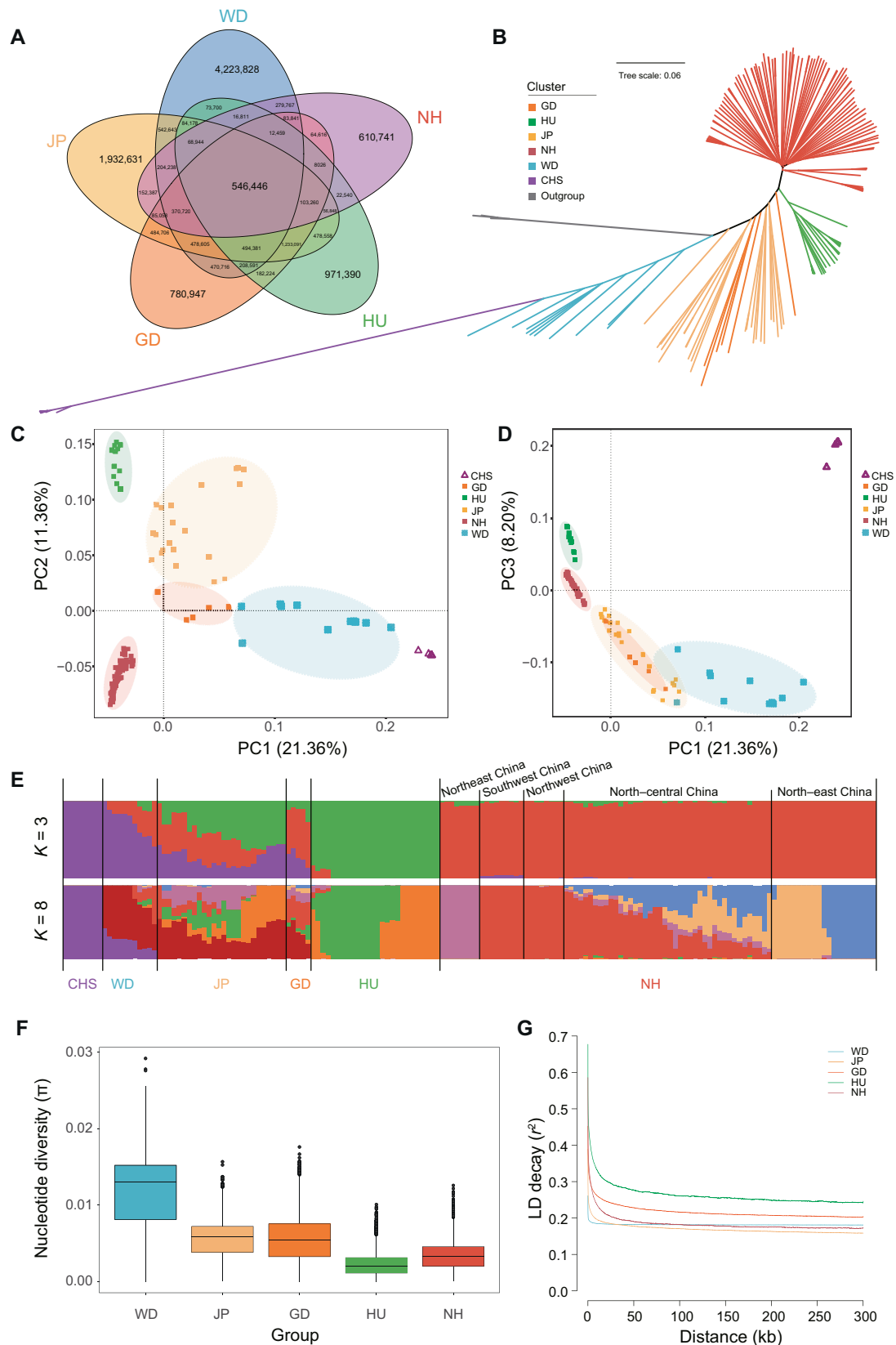
**E.** Schematic depicting the events leading to the formation of the *M. notabilis* *MSDH* gene, which is likely to result from alternative splicing in the *RECQL1* gene. The shaded block indicates the duplicated segments described in the text.

**F.** Transcriptome profile of four sex-specific candidate genes. All genes were expressed at high levels in male flower tissues.

**G.** Agarose gel electrophoresis profile of four candidate genes in female and male *M. notabilis*. The location of the *MSDH*-sp1 primer is shown at the top. *Actin* was used as the control. The primer sequences are listed in Table S16. SDR, sex-determining region; MSK, male-specific *k*-mer; SNP, single nucleotide polymorphism; IGV, Integrative Genomics Viewer; *MSTG1/2/3*, male-specific *Ty3*/*Gypsy* retrotransposon 1/2/3; *MSDH*, male-specific DNA helicase gene; *MSDH*-sp1, *MSDH*-specific 1; M, molecular marker.

compared domestic (JP, GD, HU, and NH) and wild (WD) mulberry populations based on the fixation index ( $F_{ST}$ ) and nucleotide diversity ( $\pi$ ) in 50-kb sliding windows of the gen-

ome (Figure 5A). We defined the windows with outlier signals (top 1%) for both statistics ( $F_{ST} > 0.564$ ,  $\pi$  ln-ratio WD/JP-GD-HU-NH  $> 2.041$ ) as harboring putative selective



sweeps. Merging outlier windows yielded 103 unique regions that contained 411 positively selected genes. We also performed functional enrichment analysis by identifying GO terms for these overlapping genes (Table S15). The GO analysis revealed three significantly enriched biological processes (hypergeometric test, adjusted  $P$  value  $< 0.01$ ) that were associated with disease resistance (adjusted  $P$  value =  $3.10 \times 10^{-5}$  to  $2.42 \times 10^{-9}$ ), flowering (adjusted  $P$  value =  $7.52 \times 10^{-6}$  to  $4.67 \times 10^{-7}$ ), and plant hormones (adjusted  $P$  value =  $4.41 \times 10^{-5}$  to  $2.02 \times 10^{-6}$ ) (Figure 5B).

Changes in flowering time are a major goal of mulberry breeding programs for the production of cultivars optimally adapted to local environments. Therefore, we identified three overlapping genes on Chr4 that were associated with flowering time in cultivated mulberry (*DCNL4*, *NUP96*, and *DESL-like*) using the two selection methods mentioned above ( $F_{ST} = 0.74$  and  $\pi \ln$ -ratio WD/domestic = 2.133). This analysis indicated significant selection on these genes in response to the domestication of mulberry. The positive selection signals around this region were further confirmed by significantly lower values for Tajima's  $D$  and long haplotype patterns in WD (Figure 5C). Of these three genes, *NUP96* functions as a negative regulator of long day-induced flowering [24], whereas *DCNL4* is proposed to be involved in pollen development and embryogenesis [25]. Current knowledge of these genes suggests that mulberry domestication involved selection on flowering time.

Mulberry production is severely threatened by diseases, which highlights the importance of breeding programs with a focus on disease resistance for crop improvement. Our selective sweep analyses revealed genomic regions and candidate genes associated with disease resistance, and some selected genes were involved in increasing plant defence responses (e.g., *NHL10*, *LPXD2*, *CRK10/25*, and *JMJ25*) (Figure 5D) [26–31]. Signals for positive selection in cultivars compared with WD were also detected for the gene *G2OX8*, which encodes the gibberellin 2-beta-hydroxylase enzyme, resulting in changes in hypocotyl length or plant height via the gibberellin signaling pathway (Figure 5E) [32,33]. These results should be valuable for future mulberry research and breeding plans and should facilitate mulberry improvement.

## Discussion

Genome rearrangement, a common genetic process associated with speciation, has long been postulated to be a key

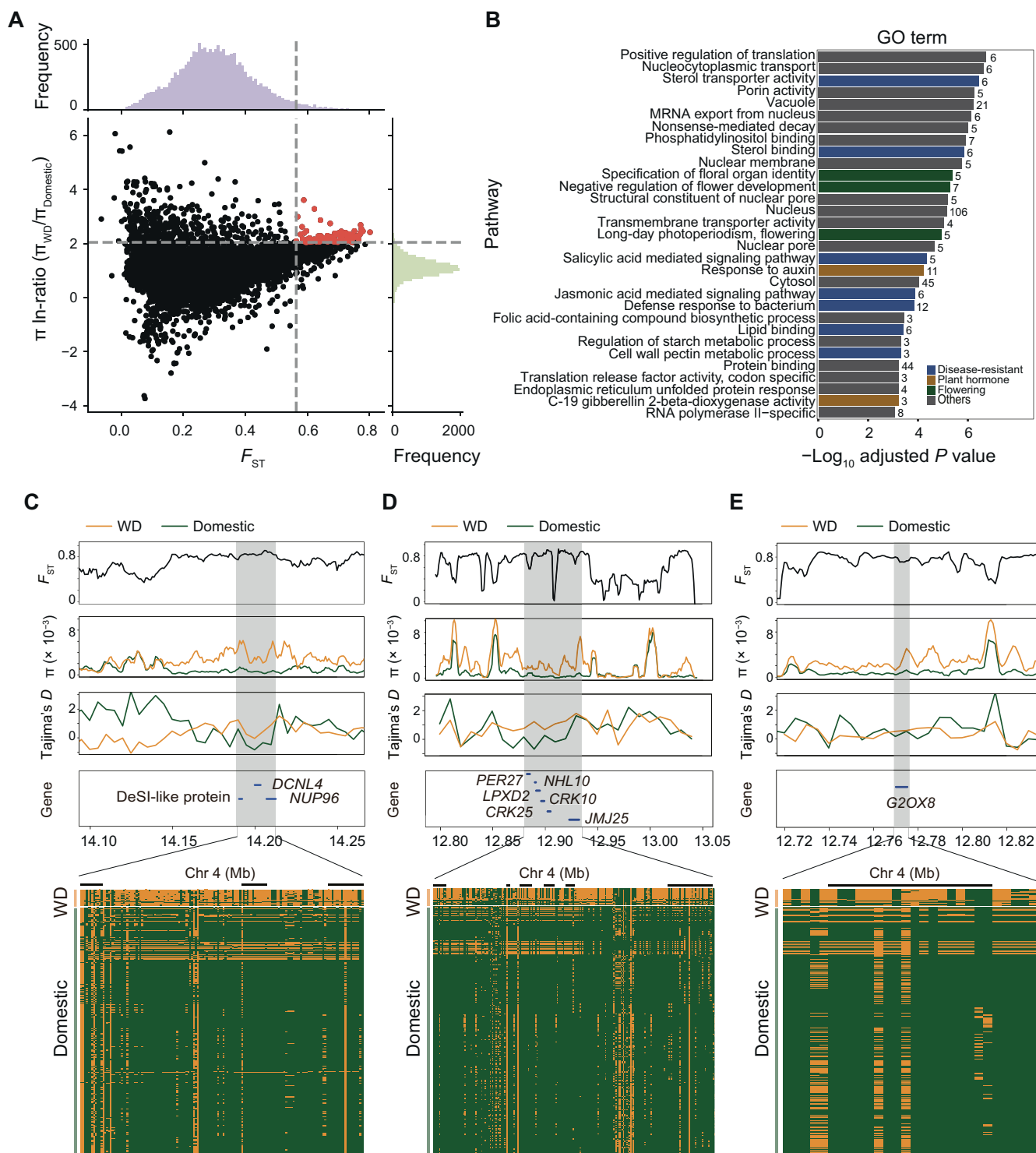
phenomenon in the evolution of higher eukaryotes such as land plants. Recently, with the surge in plant genome sequencing projects and advances in bioinformatics tools, more examples of selection-induced plant genome evolution driven by genome rearrangement, such as in rye (*Secale cereale*) [34], *Brassica* crops [35], and *Miscanthus floridulus* [36], have been reported. Our findings on *M. notabilis* and findings from a previous study [21] have suggested that the formation and diversification of karyotypes in plants are largely attributable to gene gain and reuse via SVs.

Three high-quality mulberry genomes and their distinct, descending dysploidy status enabled us to infer the origins of and steps leading to the current mulberry karyotype. The ancestral karyotype of mulberry was particularly similar to that of modern *M. alba* ( $2n = 28$ ) and was most similar to the inferred ancestral chromosomal arrangement, a finding that is consistent with the results reported for paper mulberry [18]. Due to adaptation in response to unique environmental changes, karyotype formation in *M. notabilis* involved more CFU events than in other plants and caused the observed descending dysploidy (from  $x = 14$  to  $x = 6$ ), which indicated parallel evolution and selection throughout the genus. Researchers previously hypothesized that this type of chromosomal fusion–fission event may be a common phenomenon in *M. notabilis* during growth and development [37]. Some of the genes present in regions that have undergone evolutionary rearrangement may be associated with the adaptation and evolution of the species [38], a hypothesis consistent with our findings. Alternative theories for descending dysploidy generation, such as “telomere sequence directivity”, may also be applicable but will require analysis with further sequencing data and genome analysis. The two unsolved studies on chromosomal behavior by Barbara McClintock promoted further experiments on decreasing chromosome numbers, suggesting that centromere breaks, inactivation, and fusions, in combination with telomeres, may be common mechanisms of karyotype evolution in plants and may extend to all eukaryotes [39]. However, the mechanisms that coregulate plant chromosomal rearrangements, such as descending dysploidy or WGD, at the telomere and centromere levels remain largely unclear and require more research.

Genetically determined dioecy occurs through mutations in two linked genes, one causing male infertility and the other causing female infertility, according to a theoretical concept of the formation of sex chromosomes [40,41]. Dual-gene models have

**Figure 4** Genetic diversity of mulberry accessions

**A.** Venn diagrams showing the number of SNPs of WD, JP, GD, HU, and NH. Arbitrary colors were used to better visualize the different groups, and the overlapping areas represent SNPs shared between groups. **B.** NJ tree from the genome sequences used in this study. **C.** PCA of wild and domestic accessions: PC1–PC2. **D.** PCA of wild and domestic accessions: PC1–PC3. **E.** Model-based clustering of wild and domestic mulberry trees using ADMIXTURE with  $K = 3$  and  $K = 8$ . The colors are the same as those used in panels (A), (C), and (D). **F.** Genome-wide distribution of the nucleotide diversity of each group in 50-kb windows with 50-kb steps. The horizontal line inside the box indicates the median of this distribution; the box limits indicate the first and third quartiles, respectively; and the points show outliers. **G.** Genome-wide average LD decay estimated for each group. CHS represents the first group including the wild-growing plants *M. notabilis* and *M. yunnanensis* from Southwest China. WD represents the second group including 11 wild plants collected in China and other countries. JP represents the third group mainly including landraces and cultivars from Japan and other countries. GD represents the fourth group consisting of landraces and cultivated species mainly from Guangdong Province. HU represents the fifth group derived from Taihu Basin in the southern Yangtze delta plain in China. NH represents the sixth group consisting of cultivated mulberry accessions from other places, mainly distributed throughout northern China. PCA, principal component analysis; PC, principal component; NJ, neighbor-joining; LD, linkage disequilibrium.



**Figure 5** Selection signatures identified from comparisons between wild and domestic mulberry plants

**A.** Distribution of the pairwise fixation index ( $F_{ST}$ ) and  $\pi$  ln-ratio between WD and domestic mulberry trees. The dashed vertical and horizontal lines indicate the significance thresholds ( $F_{ST} > 0.564$ ,  $\pi$  ln-ratio WD/JP-GD-HU-NH  $> 2.041$ ). **B.** GO terms identified as significantly overrepresented (hypergeometric test, adjusted  $P$  value  $< 0.01$ ). **C.–E.**  $F_{ST}$ , nucleotide diversity, and Tajima's  $D$  plots of the three candidate genomic regions. The heatmap at the bottom shows SNPs with MAF  $> 0.05$  that were used to infer haplotype patterns. The major allele at each SNP position in WD is colored green, and the minor allele is colored yellow. GO, Gene Ontology; MAF, minor allele frequency.

been discovered in campion (*Silene* spp.), papaya, asparagus, and kiwifruit in recent empirical studies [42–45]. In persimmon, however, a single gene appears to be sufficient for the expression of male features and inhibition of female development [19]. The pattern of sequence differences between males and females observed in the present study suggested the existence of a large nonrecombining region containing genes involved in sex determination and that male sex in mulberry is associated with heterogamy. We hypothesize that the hemizyosity of this Y-specific region is caused by mutations in sex-determining genes, which might also explain why regions containing sex-determining genes do not recombine between the Y and X haploids. In plants, most sex-biased genes are not carried on sex chromosomes, and their expression levels are most likely regulated by at least one upstream sex-determining gene [46,47].

As TE insertion is rarely fixed in all individuals within a species, the presence of TE sequences in the same location is unusual [48]. Insertion of a complete sequence over a long evolutionary period is necessary for males to be retained in a specific area. Three *MSTG* transposon sequences are absent from the genomes of *Arabidopsis thaliana* and *Oryza sativa*. However, they exist stably in the mulberry genus (Figure S12), suggesting that their specific functions are conserved. TE insertions have been shown to silence the expression of neighboring genes, such as those that control sexual forms in monoecious individuals [49]. This finding is surprising and implies that TE insertions have specialized functions in plants, such as improved male function. Although the specific activities of these genes are unknown, *Ty3\_Gypsy* in budding yeast and humans selects binding sites for essential meiosis transcription factors, linking their transcriptional activity to the meiosis process [50,51]. *Ty3\_Gypsy* LTR-RTs contribute to the development of flowers on male *Populus deltoides* plants by producing long noncoding RNAs (lncRNAs) [52]. Based on a *de novo* repeat library constructed from *M. notabilis* genome sequences (see the Materials and methods section), *MSTGs* are annotated as transposable proteins in the LTR/Gypsy transposon. Consistent with their functions as transposable proteins, transcripts were detected to have the ability to encode proteins using Long non-coding RNA-sequencing (lncRNA-seq) and general RNA-sequencing (RNA-seq) (Figure S16), confirming that these transcripts do not produce lncRNAs. These results suggested that these transposon genes may have new regulatory patterns at the level of sex differentiation.

In plants, DNA helicases, such as *RECQ4A* and *RECQ4B* [22], and their interacting partners play a role in meiotic recombination [53]. In *Drosophila*, one *male-specific lethal (msl)* gene encodes a protein with sequence similarity to members of a superfamily of RNA and DNA helicases [54]. *MSDH* and *RECQL1* have very similar sequence structures, with alternative splicing occurring in exon 2 of *MSDH* (Figure 3E). The phylogenetic analysis revealed that the *MSDH* gene is a partial duplicate of the *RECQL1* gene in male *M. notabilis* (Figure S17). The transcriptomic analysis showed that canonical *RECQL1* was expressed in both leaves and female flowers, with a higher expression level in female flowers, whereas *MSDH* expression was specific to male flowers (Figure 3F, Figure S18). Expression data from strand-specific lncRNA-seq and small RNA-seq revealed that *MSDH* is transcribed into long transcripts that do not generate small interfering RNAs (siRNAs; Figure S16). We also used bisulfite sequencing to analyze methylation levels in the *RECQL1* gene region

and discovered no significant sex-biased differences (Figure S19), implying that *MSDH* and *RECQL1* function independently. *MSDH* is therefore either a new gene that evolved *de novo* or a gene that transposed to new locations, followed by partial loss of the duplicated sequences. However, *MSDH* genes may have a very complex regulatory network, which is currently not well characterized, and this network will be the focus of future research.

Understanding population structure and phylogenetic relationships is very important for the management and utilization of gene pools of germplasm resources. Despite its widespread usage in basic mulberry research, the CHS group is not valuable in mulberry breeding, probably due to reproductive obstacles. Our evolutionary study answered various questions about the taxonomic status of these distantly related species and highlighted a potential future path for germplasm collection and exploitation. Based on the samples surveyed, Japan and the Guangdong region of China were the most promising sources of germplasm resources because the populations in these regions had the greatest nucleotide diversity and shared the same genetic composition, extending our previous understanding of their classification. East Asia has been identified as a significant ancient hotspot for the domestication of crops, including rice, sorghum, millet, soybeans, foxnut, apricot, and peach [55,56]. The domestication and diversification of mulberry, similar to those of other woody plant species, involve several complex steps, leading to geographic radiation and deliberate breeding of varieties. The selection of traits to maximize yield and quality necessitates the collection of accessions and additional evidence in the future to test our proposed evolutionary scenario [57].

## Materials and methods

### Plant materials

Young leaves of two wild mulberry species, *M. notabilis* (male and female) and *M. yunnanensis* (female), were collected for whole-genome and *de novo* assembly. *M. notabilis* grows in Yingjing County, Sichuan Province, China (29°80'N 102°85'E, altitude 1100–1400 m), and *M. yunnanensis* grows in Pingbian County, Yunnan Province, China (22°68'N 103°67'E, altitude 1900–2200 m). Young leaves were collected from wild mulberry species in various regions in China and other countries, grafted, and stored at the Mulberry Breeding Center, Southwest University, for genome resequencing.

### Genome size estimation

Short Illumina reads were obtained to estimate the size of the two genomes using a publicly available Perl script ([https://github.com/josephryan/estimate\\_genome\\_size.pl](https://github.com/josephryan/estimate_genome_size.pl)) to determine the distribution of *k*-mer values with Jellyfish [58]. Genome size was calculated by dividing the total number of *k*-mers by the *k*-mer distribution peak. For visualization, the online web software GenomeScope [59] was used.

### Genome sequencing and assembly

#### Illumina short-read sequencing

Cetyltrimethylammonium bromide (CTAB) was used to extract genomic DNA [60]. The Illumina HiSeq platform was

used to sequence a library with a 350-bp insert size, yielding 150-bp paired-end reads (Table S1). The raw reads were subsequently trimmed and filtered to acquire clean reads.

#### *PacBio library construction and sequencing*

A portion of the DNA samples were transferred to Annoroad (Ningbo, China) for the construction of circular consensus sequence (CCS) libraries (male *M. notabilis*) and 20 K *de novo* libraries (female *M. notabilis* and female *M. yunnanensis*) using PacBio methodology and sequenced using the PacBio Sequel platform.

#### *Hi-C library construction and sequencing*

The library for Hi-C sequencing was created from young leaves crosslinked with the *Mbo*I restriction enzyme, as described previously [5]. Then, the Hi-C libraries were amplified using 12–14 cycles of PCR and sequenced on the Illumina HiSeq platform. An Illumina HiSeq instrument combined with 2150-bp reads was used to infer the sequencing interaction pattern.

#### *Genome assembly and pseudomolecule construction*

The genomes mentioned here were assembled as follows. 1) For male *M. notabilis*, we used hifiasm [61] with the default parameters to construct contigs from PacBio High Fidelity (HiFi) CCS clean reads. PacBio SMRT subreads were corrected and assembled into contigs for the two females using Canu [62] with the parameters “corOutCoverage = 1000, minReadLength = 1000, and correctedErrorRate = 0.085”. 2) Sequencing errors were repaired using Arrow (Pacific Biosciences) with the default parameters (only the females) and the Illumina paired-end reads obtained with Pilon [63] to increase the accuracy of the reference assembly. 3) The improved contigs were further rebuilt into two subassemblies (ref and alt) with HaploMerger2 [64]. 4) Based on the reference subassembly, clean Hi-C reads were analyzed using Juicer v1.6.2 [65], and 3D-DNA [66] was then used to scaffold the contigs into pseudomolecules.

#### **Validation of the genome assembly**

The genome completeness from the contig to chromosome-level assemblies was assessed using BUSCO v3.0217 [67]. The completed assembly was compared to the Plantae BUSCO “Embryophyta odb9” database, which contains 1440 protein sequences and orthogroup annotations for key clades, using the default parameters. This result was then compared with that obtained for *M. alba* genomes (Table S2).

Furthermore, HISAT2 [68] was used to align the RNA-seq data to the two wild mulberry genomes, and the results showed 98.07% and 98.15% single-base mapping accuracy for *M. notabilis* and *M. yunnanensis*, respectively. We used BWA v0.7.8 [69] to map Illumina reads from short-insert-size libraries back to genome assemblies. Our findings indicated that 97.90% of the reads from *M. notabilis* and 99.50% of the reads from *M. yunnanensis* were mapped to the assemblies, implying that the assemblies were highly complete (Table S2).

The Hi-C heatmap revealed a well-organized interaction contact pattern along the diagonals within each pseudochromosome (Figure 1D and F). The LTR-RT assembly index

[14], a metric used to evaluate the completeness of a genome assembly based on the quality of the assembly of repeat sequences, was also used for all the aforementioned genomes in the LTR\_retriever pipeline [70].

#### **RNA-seq and transcriptome assembly**

RNA-seq data were generated from six tissues (root, bark, stem, male flower, female flower, and leaf), and total RNA was extracted using RNAiso Plus (Catalog No. 9108, Takara, Dalian, China) according to the manufacturer’s protocols.

The Illumina HiSeq XTen platform was used to create and sequence 15 paired-end libraries containing sequences with a 150-bp read length, and Trimmomatic v0.36 [71] software was used to trim the adapter sequences of the RNA-seq reads. Genome-guided transcriptome assembly was performed with HISAT2 and StringTie v1.3.475 [72]. HISAT2-build was used to construct the genome index, and HISAT2 was used to map the clean transcriptome reads to the *M. notabilis* genome. The findings were integrated with StringTie in merge mode after the transcripts for each sample were assembled. HISAT2 and the StringTie pipeline were used to calculate reads per kilobase per million (RPKM) values. The R package “edgeR” [73] was used to investigate differentially expressed genes.

#### **Repeat annotation and gene prediction**

A combination of homology searching and *ab initio* prediction was used to identify the repetitive sequences in the mulberry genome. We searched against Repbase with RepeatMasker [74] and RepeatProteinMask for homology-based prediction. We employed Tandem Repeats Finder [75], LTR FINDER [76], and RepeatScout [77] with default parameters for *ab initio* predictions.

The predictions of protein-coding genes were performed using previously reported methods with minor revisions [5]. Briefly, using Augustus [78], GlimmerHMM [79], and SNAP [80], we performed *ab initio* coding region prediction in the repeat-masked genome. PASA-H-set gene models trained Augustus, SNAP, and GlimmerHMM, which were then used to predict three masked mulberry genomes. The annotated proteins from *M. notabilis*, *M. alba*, *C. sativa*, *Fragaria vesca*, *Malus domestica*, *Prunus persica*, *A. thaliana*, and *O. sativa* were used to obtain protein evidence. Next, we used Trinity to assemble the transcriptome based on RNA-seq data and then PASA [81] to align the assembled sequence to the genome for gene predictions. Additionally, TransDecoder (<https://github.com/TransDecoder>) was also used to identify putative coding regions in transcript sequences. EVidenceModeler [82] was then used to combine the aforementioned findings to forecast the complete set of nonredundant genes. All predicted proteins were annotated using InterProScan v5.35–74.0 [83] and by running a BLASTP [84] search against the KEGG [85], Swiss-Prot, and TrEMBL [86] databases with an E-value threshold of 1E–5.

#### **Identification of LTR-RTs**

A comparative analysis of LTR-RTs was performed using the genome sequences of *M. notabilis* (male and female),

*M. yunnanensis* (female), and *M. alba*. LTR-FINDER [76] (parameters: -w 2 -d 0 -l 100) was used to detect LTR-RTs.

### Estimation of the insertion time of LTR-RTs

All LTR sequences with complete 5'-LTR and 3'-LTR were used. MUSCLE [87] (with default parameters) was used to align the 5'-LTR flanking and 3'-LTR flanking sequences, and the distance between the alignment sequences was calculated using distMat (<https://www.bioinformatics.nl/cgi-bin/emboss/distmat>, with the parameter -nucmethod 2).  $T = K/2r$  (divergence between LTRs/substitution per site per year) was used to calculate the insertion time. The mutation rate (per base per year) used was  $1.8 \times 10^{-8}$ .

### Chromosome evolution

We used the genomes of *M. notabilis*, *M. alba*, *F. hispida*, *F. microcarpa*, and *C. sativa* (with *C. sativa* as the outgroup) to reconstruct the ancestral chromosome karyotype of the Moraceae family based on a previously published technique with minor revisions [88]. Briefly, using *M. alba* as the reference genome, we performed pairwise alignments with other species as targets using LAST v1.1 with the default parameters. Subsequently, axtChain, chainMergeSort, chainPreNet, and ChainNet were used to generate “chain” and “net” files as inputs for DESCHRAMBLER (<https://github.com/jkimlab/DESCRAMBLER>). We then identified 455 conserved segments using DESCHRAMBLER at a 1200-kb resolution and reconstructed 21 predicted ancestral chromosomes with a total length of ~ 305 Mb (Table S5).

### Comparative genomic analysis and detection of EFRs in mulberry genomes

The MCscan toolkit (<https://github.com/tanghaibao/jcvi/wiki/MCscan>) was used to identify homologous gene pairs between the genomes of *M. notabilis*, *M. yunnanensis*, and *M. alba*. We estimated large-scale homologous synteny blocks (HSBs) in the pairwise whole-genome alignment utilizing the chromosomal sequences of the *M. notabilis* and *M. alba* genomes to detect probable EFRs. Raw local synteny blocks between the two genomes were detected using the MCscan toolkit. An EFR was defined as the interval between two large-scale HSBs demarcated by the end-sequence coordinates of large-scale HSBs on each side. The relative gene density, SD content, and repetitive content within the EFRs of each chromosome were compared to the complete chromosome using an in-house script.

### Analyses of gene families and phylogenetic evolution

Orthologous gene families in *M. notabilis*, *M. yunnanensis*, *M. alba*, and five other species (*V. vinifera*, *P. persica*, *C. sativa*, *F. hispida*, and *F. microcarpa*) were identified based on annotated genes using OrthoFinder v2.2.7 [15]. The expansion and contraction of the *M. notabilis* and *M. yunnanensis* gene families were examined using CAFÉ [89]. Single-copy orthologous genes were subsequently extracted, aligned with MUSCLE v8.2.10 [87], and analyzed phylogenetically using RAxML

v8.2.10 [90] with the GTRGAMMA model. The species divergence time was estimated using MCMCTree in PAML v4.8 [91], and calibration times were determined using the TimeTree database (<https://www.timetree.org/>).

Fourfold degenerate sites (4DTv) from the whole-genome SNP collection were extracted and concatenated into a “super-genome” format for each species to construct a *Morus* genus tree. The seven aligned 4DTv supergenes were used to construct a phylogenetic tree using the IQ-TREE [16] program.

### Detection of SNPs, small indels, SVs, and SDs

SNPs and small indels (length  $\leq$  500 bp) were compared between the two wild genomes assembled in this work and the *M. alba* genome using MUMmer v3.2394 [92]. First, we used the number from MUMmer with the parameter “-mum -g 1000 -c 90 -l 40” to generate the alignment. The files were then filtered using the delta-filter comparison with the query “-r -q” to create a one-to-one map. Show-snps (a module of MUMmer) with the parameter “-ClrTH” was used to call SNPs and small indels from the one-to-one alignment module. Furthermore, the web-based SV analysis tool Assemblytics [93] was used to analyze large SVs. In addition, we identified SDs in the mulberry genomes with reference to the *Ficus* genomes (<https://github.com/tangerzhang/popCNV>).

### Identification of SDRs in *M. notabilis*

A previously described technique based on *k*-mers [19,20] was used to identify sex-determining genes in the *M. notabilis* genome, with a slight modification. Using Jellyfish [58], the reads from male and female *M. notabilis* (four individual replicates per sex) were cataloged into 40-bp *k*-mers. The total cumulative counts of *k*-mers starting with “AG” in females and males were greater than ten, and the reads were defined as valid *k*-mers. In addition, *k*-mers with a count of zero in the female group were defined as MSKs, and those with a count of zero in the male group were defined as FSKs. We complemented this approach by mapping the sex-specific reads to the assembled male *M. notabilis* genome using BWA-MEM with the default parameters, and we prioritized genomic regions (10-kb windows) with a high depth of sex-specific reads. After deleting duplicates, SAMtools v1.9 [94] was used to calculate the mapping depth of the reference genome (by parsing specific 50-kb windows). Additionally, deduplicated BAM data were normalized using the deepTools bamCoverage tool [95].

### Quantifying the digital expression and analyzing the potential function of candidate sex genes

Illumina sequencing experiments (Illumina NovaSeq 6000, Illumina, California) were performed at the levels of lncRNA, small RNA, and DNA methylation to quantify the digital expression and explore the gene regulation patterns of candidate sex genes. Libraries were constructed using a NEBNext Ultra RNA Library Prep Kit (Catalog No. E7530L, NEB) for mRNA, NEBNext Ultra RNA Library Prep Kit (NEB, Ipswich) for lncRNA, Small RNA Sample Preparation Kit (Catalog No. RS-200-0048, Illumina) for small RNA, and TruSeq Methyl Capture EPIC Library Prep Kit (Catalog No.

FC-151-1002, Illumina) for DNA methylation, respectively, according to the manufacturers' instructions.

In the analysis of RNA-seq, lncRNA-seq, and small RNA (sRNA)-seq data, rRNA/tRNA contaminants were removed by mapping the reads to rRNA/tRNA sequences from public databases. Transcripts of lncRNAs were assembled using Trinity V2.6.6 with the parameter settings for strand-specific reads (“--SS\_lib\_type RF”). The obtained sequences were used to search the National Center for Biotechnology Information (NCBI) RefSeq non-redundant proteins (NR) database. Small RNA reads and bisulfite sequencing reads were mapped to the reference genome using Bowtie2 v2.3.4.1 and Bismark v0.16.3 [96], respectively. The alignment results were visualized using Integrative Genomics Viewer (IGV) v2.4.14 [97].

### Read alignment and variant calling

The filtered reads from all individuals were aligned to the *M. alba* genome using BWA-MEM. The read pairs were filtered using Picard tools v2.18 (<https://broadinstitute.github.io/picard/>). SNPs were identified in all the samples using the HaplotypeCaller module of GATK v3.8.68 [98] in genomic variation call format (GVCF) mode. Briefly, HaplotypeCaller was used to identify the GVCF for each sample. Subsequently, all GVCF files were merged to create a raw population genotype file with the SNPs. SNPs were preliminarily filtered using the GATK VariantFiltration function with the parameters “QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 3.0 and mean variant sequencing depth (including all the individuals) < 1/3× and > 3×”. In addition, SNPs were filtered according to the following criteria: 1) minor allele frequency (MAF) ≥ 5%, 2) maximum missing rate < 0.1, and 3) restriction to two alleles.

### Population structure and phylogenetic analyses

The NJ tree, PCA, and ADMIXTURE methods were used to explore the genetic relationships among wild and domesticated mulberry populations. An NJ tree was constructed for the whole-genome SNP set using MEGA v6.0 [99] based on a pairwise genetic distance matrix, which was calculated using PLINK v1.9 [100] with the option “--distance-matrix”. PCA was performed using smartpca in EIGENSOFT v6.1 [101]. The significance of eigenvectors was assessed using the Tracy–Widom test. Population genetic structure was inferred using ADMIXTURE v1.3.0 [102] considering  $k = 2$  to  $k = 10$  (Figure S11), and the analysis was repeated 20 times for each  $k$  value. For the PCA and ADMIXTURE analysis, we used a nonredundant SNP data set obtained after removing rare alleles with the option “--indep-pairwise 50 5 0.4” in PLINK and further excluding SNPs with intrachromosomal LD  $r^2 < 0.4$  to remove the bias caused by LD.

The value of  $\pi$  was calculated using VCFtools v0.1.15 [103] based on the high-confidence filtered SNPs. The  $\pi$  value for each SNP was calculated, and the nucleotide diversity level was measured using “--window-pi 50000 --window-pi-step 20000” for each subpopulation. LD decay was calculated for

all pairs of SNPs within 500 kb using PopLDdecay v3.27 [104] with the default parameters.

Treemix (v1.13) was used to infer models of population split and migration between groups. Treemix was run using the allele frequencies calculated from the LD-pruned SNP set with the parameters “--bootstrap 5000 -global” and “migration event -m (range 0–4)”.

### Identification of selective sweeps

We performed the analysis described below to detect selective sweeps during mulberry domestication. 1) The high divergence in genetic diversity ( $\pi$  ln-ratio) and high fixation index  $F_{ST}$  were analyzed by parsing specific 50-kb windows between the domesticated group (JP, GD, HU, and NH) and WD. 2) Putative selective sweeps were defined as windows with outlier signals (top 1%) overlapping for the two statistics ( $F_{ST} > 0.564$ ,  $\pi$  ln-ratio WD/JP-GD-HU-NH > 2.041). 3) Tajima's  $D$  and comparison haplotypes were applied to confirm the top signals.

### Functional enrichment analyses

We characterized the most relevant functions of the protein-coding genes with chromosomal break regions and selective sweeps by searching for overrepresented KEGG pathways and GO terms. Target protein sequences were used to conduct functional enrichment tests of the target genes using KOBAS 3.0 (<https://kobas.cbi.pku.edu.cn/kobas3/annotate/>). The  $P$  value was calculated using a hypergeometric distribution, and  $P < 0.05$  was considered significantly enriched.

### RNA extraction and qRT-PCR analysis

Total RNA was extracted as described above, and cDNA was generated with a PrimeScript RT Reagent Kit with gDNA Eraser (Catalog No. RR047B, Takara, Japan). We performed qRT-PCR with TB Green Premix Ex Taq II (Catalog No. RR820Q, Takara, Japan). Relative expression was calculated using the  $2^{-\Delta\Delta C_t}$  method [105]. The primers used for the gene expression analysis are listed in Table S16. The values are presented as the mean  $\pm$  SD from three biological replicates (\*\*,  $P < 0.01$ ; \*\*\*,  $P < 0.001$ ; and \*\*\*\*,  $P < 0.0001$ ), as determined using a one-way analysis of variance (ANOVA).

### PCR validation of candidate sex-determining genes

Degenerate primers were designed for four candidate genes to verify their male specificity (Table S16). The PCR mix was composed of 10  $\mu$ l of 2× Ex Taq MasterMix (Catalog No. CW0682, CWBIO, China), 1  $\mu$ l of each primer, and 1  $\mu$ l of genomic DNA at a concentration of  $\sim 100$  ng/ $\mu$ l, and ddH<sub>2</sub>O was added to obtain a total reaction volume of 20  $\mu$ l. The thermocycling conditions were as follows: an initial cycle of 5 min at 94 °C, followed by 35 cycles of 30 s at 94 °C, 30 s at 57 °C, 30 s at 72 °C, and 2 min of extension at 72 °C. The PCR products were loaded on a 1.0% agarose gel and run at 150 V for 15 min. The samples were imaged under ultraviolet light.

## Data availability

Illumina re-sequencing short reads generated in this study have been deposited in the Genome Sequence Archive [106] at the National Genomics Data Center (NGDC), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) / China National Center for Bioinformation (CNCB) (GSA: CRA006420), and are publicly accessible at <https://ngdc.cncb.ac.cn/gsa>. The genome assemblies and gene annotations have been deposited in the Genome Warehouse [107] at the NGDC, BIG, CAS / CNCB (GWH: GWHBISO00000000 for male *M. notabilis*; GWHBISQ00000000 for female *M. notabilis*; GWHBISP00000000 for female *M. yunnanensis*), and are publicly accessible at <https://ngdc.cncb.ac.cn/gwh>.

## CRedit author statement

**Zhongqiang Xia:** Investigation, Data curation, Validation, Visualization, Writing - original draft, Writing - review & editing. **Xuelei Dai:** Investigation. **Wei Fan:** Investigation, Data curation, Writing - review & editing. **Changying Liu:** Investigation, Writing - review & editing. **Meirong Zhang:** Validation. **Peipei Bian:** Data curation. **Yuping Zhou:** Validation. **Liang Li:** Validation, Visualization. **Baozhong Zhu:** Resources. **Shuman Liu:** Resources. **Zhengang Li:** Resources. **Xiling Wang:** Investigation, Resources. **Maode Yu:** Project administration. **Zhonghuai Xiang:** Project administration. **Yu Jiang:** Conceptualization, Writing - original draft, Writing - review & editing, Supervision. **Aichun Zhao:** Conceptualization, Resources, Writing - original draft, Writing - review & editing, Supervision. All authors have read and approved the final manuscript.

## Competing interests

The authors have declared no competing interests.

## Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2019YFD1000604), the China Agriculture Research System (Grant No. CARS-18-ZJ0201), the Forestry Promotion by Science and Technology Program of Chongqing, China (Grant No. Yulinkeyan2020-2), and the Sichuan Science and Technology Program, China (Grant No. 22NSFSC3680).

## Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gpb.2022.08.005>.

## ORCID

ORCID 0000-0003-1759-4143 (Zhongqiang Xia)  
ORCID 0000-0002-7681-7338 (Xuelei Dai)  
ORCID 0000-0003-3501-8860 (Wei Fan)  
ORCID 0000-0002-5573-236X (Changying Liu)

ORCID 0000-0002-2767-684X (Meirong Zhang)  
ORCID 0000-0002-3948-4052 (Peipei Bian)  
ORCID 0000-0003-1214-0126 (Yuping Zhou)  
ORCID 0000-0003-1658-8956 (Liang Li)  
ORCID 0000-0002-1552-2536 (Baozhong Zhu)  
ORCID 0000-0003-4225-5944 (Shuman Liu)  
ORCID 0000-0002-1603-7735 (Zhengang Li)  
ORCID 0000-0002-2805-8146 (Xiling Wang)  
ORCID 0000-0002-2366-5772 (Maode Yu)  
ORCID 0000-0003-3370-2984 (Zhonghuai Xiang)  
ORCID 0000-0003-4821-3585 (Yu Jiang)  
ORCID 0000-0001-6140-0102 (Aichun Zhao)

## References

- [1] Bao T, Xu Y, Gowd V, Zhao J, Xie J, Liang W, et al. Systematic study on phytochemicals and antioxidant activity of some new and common mulberry cultivars in China. *J Funct Foods* 2016;25:537–47.
- [2] Lee HW, Sin DH, Lee WJ. Morphological and chemical characteristics of mulberry (*Morus*) fruit with varieties. *J Sericult Entomol Sci* 1998;40:1–7.
- [3] Kim SY, Gao JJ, Lee WC, Ryu KS, Lee KR, Kim YC. Antioxidative flavonoids from the leaves of *Morus alba*. *Arch Pharmacol Res* 1999;22:81–5.
- [4] He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, et al. Draft genome sequence of the mulberry tree *Morus notabilis*. *Nat Commun* 2013;4:1–9.
- [5] Jiao F, Luo R, Dai X, Liu H, Yu G, Han S, et al. Chromosome-level reference genome and population genomic analysis provide insights into the evolution and improvement of domesticated mulberry (*Morus alba*). *Mol Plant* 2020;13:1001–12.
- [6] Weiss-Schneeweiss H, Schneeweiss GM. Karyotype diversity and evolutionary trends in angiosperms. In: Greilhuber J, Dolezel J, Wendel J, editors. *Plant Genome Diversity Volume 2*. Vienna: Springer; 2013, p. 209–30.
- [7] Fonsêca A, Ferraz ME, Pedrosa-Harand A. Speeding up chromosome evolution in *Phaseolus*: multiple rearrangements associated with a one-step descending dysploidy. *Chromosoma* 2016;125:413–21.
- [8] Rieseberg LH, Willis JH. Plant speciation. *Science* 2007;317:910–4.
- [9] Liu K, Chen R, Liu L, Ke H, Lv Y. Characteristics, distribution and systematic location of *Morus yunnanensis* Koidz. *Sichuan Canye* 2015;43:43–4.
- [10] Tikader A, Vijayan K, Raghunath M, Chakroborti S, Roy B, Pavankumar T. Studies on sexual variation in mulberry (*Morus* spp.). *Euphytica* 1995;84:115–20.
- [11] Leite VG, Mansano VF, Teixeira SP. Floral development of Moraceae species with emphasis on the perianth and androecium. *Flora* 2018;240:116–32.
- [12] Sinoto Y. Chromosome studies in some dioecious plants, with special reference to the allosomes. *Cytologia (Tokyo)* 1929;1:109–91.
- [13] Atsumi R, Nishihara R, Tarora K, Urasaki N, Matsumura H. Identification of dominant genetic markers relevant to male sex determination in mulberry (*Morus alba* L.). *Euphytica* 2019;215:1–13.
- [14] Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res* 2018;46:e126.
- [15] Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16:1–14.

- [16] Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–74.
- [17] Murat F, Armero A, Pont C, Klopp C, Salse J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet* 2017;49:490–6.
- [18] Peng X, Liu H, Chen P, Tang F, Hu Y, Wang F, et al. A chromosome-scale genome assembly of paper mulberry (*Broussonetia papyrifera*) provides new insights into its forage and papermaking usage. *Mol Plant* 2019;12:661–77.
- [19] Akagi T, Henry IM, Tao R, Comai L. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science* 2014;346:646–50.
- [20] Liao Q, Du R, Gou J, Guo L, Shen H, Liu H, et al. The genomic architecture of the sex-determining region and sex-related metabolic variation in *Ginkgo biloba*. *Plant J* 2020;104:1399–409.
- [21] Zhang X, Wang G, Zhang S, Chen S, Wang Y, Wen P, et al. Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. *Cell* 2020;183:875–889.e17.
- [22] Serra H, Lambing C, Griffin CH, Topp SD, Nageswaran DC, Underwood CJ, et al. Massive crossover elevation via combination of *HEI10* and *recq4a recq4b* during *Arabidopsis* meiosis. *Proc Natl Acad Sci U S A* 2018;115:2437–42.
- [23] Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 2012;8:e1002967.
- [24] Cheng Z, Zhang X, Huang P, Huang G, Zhu J, Chen F, et al. Nup96 and HOS1 are mutually stabilized and gate CONSTANS protein level, conferring long-day photoperiodic flowering regulation in *Arabidopsis*. *Plant Cell* 2020;32:374–91.
- [25] Hosp J, Ribarits A, Retzer K, Jin Y, Tashpulatov A, Resch T, et al. A tobacco homolog of DCN1 is involved in pollen development and embryogenesis. *Plant Cell Rep* 2014;33:1187–202.
- [26] Varet A, Parker J, Tornero P, Nass N, Nürnberg T, Dangl JL, et al. *NHL25* and *NHL3*, two *NDR1/HIN1*-like genes in *Arabidopsis thaliana* with potential role(s) in plant defense. *Mol Plant-Microbe Interact* 2002;15:608–16.
- [27] Chen Q, Tian Z, Jiang R, Zheng X, Xie C, Liu J. *StPOTHR1*, a *NDR1/HIN1*-like gene in *Solanum tuberosum*, enhances resistance against *Phytophthora infestans*. *Biochem Biophys Res Commun* 2018;496:1155–61.
- [28] Newman MA, Sundelin T, Nielsen JT, Erbs G. MAMP (microbe-associated molecular pattern) triggered immunity in plants. *Front Plant Sci* 2013;4:139.
- [29] Mou S, Meng Q, Gao F, Zhang T, He W, Guan D, et al. A cysteine-rich receptor-like protein kinase CaCKR5 modulates immune response against *Ralstonia solanacearum* infection in pepper. *BMC Plant Biol* 2021;21:1–15.
- [30] Gu J, Sun J, Liu N, Sun X, Liu C, Wu L, et al. A novel cysteine-rich receptor-like kinase gene, *TaCRK2*, contributes to leaf rust resistance in wheat. *Mol Plant Pathol* 2020;21:732–46.
- [31] Noh SW, Seo RR, Park HJ, Jung HW. Two *Arabidopsis* homologs of human lysine-specific demethylase function in epigenetic regulation of plant defense responses. *Front Plant Sci* 2021;12:688003.
- [32] Lopez-Juez E, Kobayashi M, Sakurai A, Kamiya Y, Kendrick RE. Phytochrome, gibberellins, and hypocotyl growth — a study using the cucumber (*Cucumis sativus* L.) *long hypocotyl* mutant. *Plant Physiol* 1995;107:131–40.
- [33] Sun TP. The molecular mechanism and evolution of the GA–GID1–DELLA signaling module in plants. *Curr Biol* 2011;21:R338–45.
- [34] Li G, Wang L, Yang J, He H, Jin H, Li X, et al. A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. *Nat Genet* 2021;53:574–84.
- [35] He Z, Ji R, Havlickova L, Wang L, Li Y, Lee HT, et al. Genome structural evolution in *Brassica* crops. *Nat Plants* 2021;7:757–65.
- [36] Zhang G, Ge C, Xu P, Wang S, Cheng S, Han Y, et al. The reference genome of *Miscanthus floridulus* illuminates the evolution of Saccharinae. *Nat Plants* 2021;7:608–18.
- [37] Xuan Y, Li C, Wu Y, Ma B, Liu R, Xiang Z, et al. FISH-based mitotic and meiotic diakinesis karyotypes of *Morus notabilis* reveal a chromosomal fusion-fission cycle between mitotic and meiotic phases. *Sci Rep* 2017;7:1–9.
- [38] Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res* 2009;19:770–7.
- [39] Birchler JA, Han F. Barbara McClintock’s unsolved chromosomal mysteries: parallels to common rearrangements and karyotype evolution. *Plant Cell* 2018;30:771–9.
- [40] Westergaard M. The mechanism of sex determination in dioecious flowering plants. *Adv Genet* 1958;9:217–81.
- [41] Charlesworth B, Charlesworth D. A model for the evolution of dioecy and gynodioecy. *Am Nat* 1978;112:975–97.
- [42] Wang J, Na JK, Yu Q, Gschwend AR, Han J, Zeng F, et al. Sequencing papaya X and Y<sup>h</sup> chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc Natl Acad Sci U S A* 2012;109:13710–5.
- [43] Kazama Y, Ishii K, Aonuma W, Ikeda T, Kawamoto H, Koizumi A, et al. A new physical mapping approach refines the sex-determining gene positions on the *Silene latifolia* Y-chromosome. *Sci Rep* 2016;6:1–10.
- [44] Harkess A, Zhou J, Xu C, Bowers JE, Van der Hulst R, Ayyampalayam S, et al. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat Commun* 2017;8:1–10.
- [45] Akagi T, Pilkington SM, Varkonyi-Gasic E, Henry IM, Sugano SS, Sonoda M, et al. Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nat Plants* 2019;5:801–9.
- [46] Vatén A, Soyars CL, Tarr PT, Nimchuk ZL, Bergmann DC. Modulation of asymmetric division diversity through cytokinin and SPEECHLESS regulatory interactions in the *Arabidopsis* stomatal lineage. *Dev Cell* 2018;47:53–66.e5.
- [47] Jones-Rhoades MW, Bartel DP, Bartel B. MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol* 2006;57:19–53.
- [48] Furman BL, Metzger DC, Darolti I, Wright AE, Sandkam BA, Almeida P, et al. Sex chromosome evolution: so many exceptions to the rules. *Genome Biol Evol* 2020;12:750–63.
- [49] Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 2009;461:1135–8.
- [50] Malki S, van der Heijden GW, O’Donnell KA, Martin SL, Bortvin A. A role for retrotransposon LINE-1 in fetal oocyte attrition in mice. *Dev Cell* 2014;29:521–33.
- [51] Bourc’his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* 2004;431:96–9.
- [52] Xue L, Wu H, Chen Y, Li X, Hou J, Lu J, et al. Evidences for a role of two Y-specific genes in sex determination in *Populus deltoides*. *Nat Commun* 2020;11:1–12.
- [53] Knoll A, Puchta H. The role of DNA helicases and their interaction partners in genome stability and meiotic recombination in plants. *J Exp Bot* 2011;62:1565–79.
- [54] Gorman M, Kuroda MI, Baker BS. Regulation of the sex-specific binding of the maleless dosage compensation protein to the male X chromosome in *Drosophila*. *Cell* 1993;72:39–49.
- [55] Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. *Cell* 2006;127:1309–21.
- [56] Gepts P. The contribution of genetic and genomic approaches to plant domestication studies. *Curr Opin Plant Biol* 2014;18:51–9.
- [57] Meyer RS, Purugganan MD. Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet* 2013;14:840–52.

- [58] Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of  $k$ -mers. *Bioinformatics* 2011;27:764–70.
- [59] Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 2020;11:1–10.
- [60] Porebski S, Bailey LG, Baum BR. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol Biol Rep* 1997;15:8–15.
- [61] Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;18:170–5.
- [62] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive  $k$ -mer weighting and repeat separation. *Genome Res* 2017;27:722–36.
- [63] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9:e112963.
- [64] Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* 2017;33:2577–9.
- [65] Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* 2016;3:95–8.
- [66] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 2017;356:92–5.
- [67] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–2.
- [68] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;12:357–60.
- [69] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–60.
- [70] Ou S, Jiang N. LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* 2018;176:1410–22.
- [71] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–20.
- [72] Perteau M, Perteau GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;33:290–5.
- [73] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–40.
- [74] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 2004;5:4.10.1–14.
- [75] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;27:573–80.
- [76] Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 2007;35:W265–8.
- [77] Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. *Bioinformatics* 2005;21:i351–8.
- [78] Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 2003;19:ii215–25.
- [79] Majoros WH, Perteau M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 2004;20:2878–9.
- [80] Korf I. Gene finding in novel genomes. *BMC Bioinformatics* 2004;5:1–9.
- [81] Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 2003;31:5654–66.
- [82] Haas BJ, Salzberg SL, Zhu W, Perteau M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 2008;9:1–22.
- [83] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. *Nucleic Acids Res* 2005;33:W116–20.
- [84] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–10.
- [85] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
- [86] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;31:365–70.
- [87] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–7.
- [88] Liu C, Gao J, Cui X, Li Z, Chen L, Yuan Y, et al. A towering genome: experimentally validated adaptations to high blood pressure and extreme stature in the giraffe. *Sci Adv* 2021;7:eabe9459.
- [89] Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* 2013;30:1987–97.
- [90] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–3.
- [91] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;24:1586–91.
- [92] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol* 2004;5:1–9.
- [93] Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 2016;32:3021–3.
- [94] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–9.
- [95] Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 2014;42:W187–91.
- [96] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011;27:1571–2.
- [97] Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the integrative genomics viewer. *Cancer Res* 2017;77:e31–4.
- [98] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
- [99] Tamura K, Stecher G, Peterson D, Filipiński A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013;30:2725–9.
- [100] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–75.
- [101] Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;2:e190.

- [102] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003;164:1567–87.
- [103] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156–8.
- [104] Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 2019;35:1786–8.
- [105] Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the  $2^{-\Delta\Delta CT}$  method. *Methods* 2001;25:402–8.
- [106] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics* 2021;19:578–83.
- [107] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. *Genomics Proteomics Bioinformatics* 2021;19:584–9.