nature plants

Article

https://doi.org/10.1038/s41477-023-01473-2

Haplotype-resolved genomes of wild octoploid progenitors illuminate genomic diversifications from wild relatives to cultivated strawberry

Received: 23 October 2022

Accepted: 3 July 2023

Published online: 3 August 2023

Check for updates

Xin Jin[1,2], Haiyuan Du[1,2], Chumeng Zhu[1,2], Hong Wan[3], Fang Liu[1], Jiwei Ruan[4] ✉, Jeffrey P. Mower[5,6] ✉ & Andan Zhu[1] ✉

Strawberry is an emerging model for studying polyploid genome evolution and rapid domestication of fruit crops. Here we report haplotype-resolved genomes of two wild octoploids (*Fragaria chiloensis* and *Fragaria virginiana*), the progenitor species of cultivated strawberry. Substantial variation is identified between species and between haplotypes. We redefine the four subgenomes and track the genetic contributions of diploid species by additional sequencing of the diploid *F. nipponica* genome. We provide multiple lines of evidence that *F. vesca* and *F. iinumae*, rather than other described extant species, are the closest living relatives of these wild and cultivated octoploids. In response to coexistence with quadruplicate gene copies, the octoploid strawberries have experienced subgenome dominance, homoeologous exchanges and coordinated expression of homoeologous genes. However, some homoeologues have substantially altered expression bias after speciation and during domestication. These findings enhance our understanding of the origin, genome evolution and domestication of strawberries.

Polyploidy and hybridization are widespread in plants and have played vital roles in plant speciation and crop domestication[1–3]. Many important crops, such as cotton, wheat, oilseed rape and strawberry, are allopolyploids, formed by fixing heterozygosity or hybrid vigour during domestication. A major challenge for allopolyploids is determining their genomic organization and coordinated expression between homoeologues. This is particularly difficult in high-order allopolyploids such as hexaploid wheat ($2n = 8x = 42$) and octoploid strawberry ($2n = 8x = 56$) due to their genomic complexity and the superimposed effects of recurrent polyploidization and interspecific hybridization. Allopolyploids can maintain relatively high levels of heterozygosity via the enforced pairing of homologous chromosomes and disomic inheritance[4,5], which in turn bring substantial challenges for accurately recovering the different homoalleles and haplotypes during genome assembly[6].

Recent technical innovations, such as trio binning[7], gamete binning[8] and PacBio circular consensus sequencing (CCS)[9], have enabled haplotype phasing on a genome-wide scale. In particular, PacBio CCS technology (also known as HiFi) combines the advantages of long read lengths (averaging 10–25 kb) and high accuracy (>99.5%)[10] to achieve

[1]Germplasm Bank of Wild Species, Yunnan Key Laboratory of Crop Wild Relatives Omics, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Horticultural Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. [4]Flower Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. [5]Center for Plant Science Innovation, University of Nebraska, Lincoln, NE, USA. [6]Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE, USA. ✉e-mail: rjw@yaas.org.cn; jpmower@unl.edu; zhuandan@mail.kib.ac.cn

https://doi.org/10.1038/s41477-023-01473-2

Table 1 | Summary of genome features and quality evaluation of the three sequenced *Fragaria* species

Category	*F. chiloensis* hap1	*F. chiloensis* hap2	*F. virginiana* hap1	*F. virginiana* hap2	*F. nipponica*
Genome estimate					
Ploidy level (x=7)	2n=8x=56		2n=8x=56		2n=2x=14
Genome size by flow cytometry (Mb)	841.6		758.8		259.6
Genome size by *k*-mer analysis (Mb)	834.1		740.7		253.4
Assembly feature					
Assembled genome size (Mb)	839.9	824.2	787.8	769.2	290.9
Number of contigs	1,305	632	995	531	334
N50 (Mb)	11.3	8.9	14.3	14.2	1.9
Base completeness					
Assigned rate (%)	95.3	96.4	95.5	97.2	99.4
Number of haploid chromosomes	28	28	28	28	7
Illumina reads mappability (%)	99.3		99.4		94.9
Continuity					
Number of gaps	129	136	110	83	361
LAI value	13.2	13.7	17.9	17.3	14.5
Base accuracy					
k-mer QV (HiFi reads)	68.1		68.3		24.9
k-mer completeness (HiFi reads) (%)	98.5		98.8		87.9
Haplotype phasing					
Proportion of error phasing (%)	0.37		1.05		–
Annotation feature and completeness					
Repeat density (%)	44.2	44.4	41.9	41.6	43.0
Number of genes	96,759	95,830	94,294	94,143	36,059
Transcript mappability (%)	99.3		99.2		99.2
BUSCO completeness (%)	96.3	96.1	96.3	96.4	91.7

very good assemblies with relatively low depth of sequencing coverage[11]. To date, this technology has mainly been applied to diploids in plants (for example, potato[12] and apple[13]), although it has also shown promise for autotetraploid alfalfa[14] and may be useful for high-order polyploids[11], especially crop wild relatives, which generally have complex heterozygous genomes.

The genus *Fragaria* comprises approximately 25 recognized species and one popular cultivated fruit crop, *Fragaria* × *ananassa*[15–17]. *Fragaria* species show diverse ploidy levels, ranging from diploids to decaploids, making this genus an excellent system to study genomic and morphological consequences of recurrent polyploidization[18]. Hybridization also plays an important role in *Fragaria* speciation, with both diploid hybrids and many allopolyploid species. In fact, the modern cultivated strawberry, *F.* × *ananassa*, originated from interspecific hybridization between two octoploid progenitors, *F. virginiana* and *F. chiloensis*, in the mid-1700s in Europe[16,18,19]. Strawberry cultivars display distinctive morphology and physiology compared with their wild diploid and octoploid relatives and have been intensively selected for yield-related traits, such as fruit size, crown branching and seasonal flowering[20–22]. The two wild progenitors, *F. chiloensis* and *F. virginiana*, thus provide a link to track the genetic contributions from wild species to domesticated germplasms, yet their complete and accurate genome sequences are lacking.

Characterization of the genomic organization of octoploid strawberries is complicated due to their heterozygosity and high ploidy level[23–25]. A chromosome-level genome of a cultivated strawberry (*F.* × *ananassa* cv. 'Camarosa') has been released[26], and there is a contig-level assembly of another cultivar, 'Royal Royce', using HiFi data[10]. On the basis of the 'Camarosa' genome, Edger et al. argued that *F. vesca*, *F. iinumae*, *F. viridis* and *F. nipponica* were the four extant diploid progenitors of the octoploid strawberries[26]. However, technical difficulties in accounting for possible extinct ancestors and distinguishing homoeologous exchanges (HEs)[27], as well as the confounding effects of interspecific hybridization and incomplete lineage sorting on phylogenetic resolution[28], have continued to raise questions about the inference of the diploid ancestry of *Fragaria*. The ancestry from wild diploids to modern strawberry cultivars thus remains elusive.

In this study, we report fully haplotype-resolved genomes of the two wild octoploids, *F. chiloensis* and *F. virginiana*, both of which were sequenced to high levels of continuity, completeness and accuracy. We also sequenced a diploid species, *F. nipponica*, and combined it with other published *Fragaria* diploid genomes to identify the potential wild diploid relatives of the wild octoploids and cultivated strawberries. We further characterized genetic divergence and evolutionary dynamism of homoeologous expression bias (HEB) among the wild and cultivated octoploid strawberries. Altogether, these genomic data provide new resources to enhance our understanding of genome evolution in high-order polyploids and the biology of strawberry domestication.

Results

Haplotype-resolved assemblies for wild *Fragaria* octoploids

We generated haplotype-resolved and reference-level genomes for the two wild octoploid progenitors (*F. chiloensis* and *F. virginiana*; 2n = 8x = 56) of modern cultivated strawberries by using a combination of platforms (Extended Data Figs. 1 and 2 and Supplementary Tables 1 and 2). The *F. chiloensis* genome had an assembled size of 1.66 Gb (Table 1). The assembly was partitioned into two haplotypes (hap1 and hap2),

covering 839.9 Mb and 824.2 Mb (a contig N50 of 11.3 Mb and 8.9 Mb, respectively). The *F. virginiana* genome was 1.56 Gb, composed of two haplotypes of 787.8 Mb and 769.2 Mb (a contig N50 of 14.3 Mb and 14.2 Mb), respectively (Table 1). The haploid assembly of each species was generally consistent with the sizes estimated by flow cytometry but slightly higher than that by *k*-mer analysis (Supplementary Table 1 and Supplementary Fig. 1).

Considering the complex genome architecture in high-order polyploids, we developed a pre-clustering, iterative scaffolding and contig recovery pipeline that was modified from a typical ALLHiC[29] workflow for pseudochromosome construction (Fig. 1a and Supplementary Fig. 2). We used the diploid *F. vesca* genome ($2n = 2x = 14$)[30] as a reference to guide pre-clustering of homologous contigs, followed by iterative contig anchoring and orientation with Hi-C interaction signals (Supplementary Fig. 3). On the basis of this approach, 95.3–97.2% of the assembled contigs were anchored to 28 pseudochromosomes for each haplotype. Hi-C interaction signals exhibited clear boundaries between chromosomes (Fig. 1b and Extended Data Fig. 3a,b), and the density of signals was smoothly distributed along each chromosome, providing evidence for proper contig anchoring (Fig. 1b and Extended Data Fig. 3a,b). The two octoploid genomes exhibited almost complete colinearity with the *F. vesca* genome[30], except for a large rearrangement (~10 Mb) at chromosome 2D and a few rearrangements affecting one end of chromosomes 1A, 1B and 1C, which probably occurred before the divergence of *F. chiloensis* and *F. virginiana* (Extended Data Fig. 4a,b).

We annotated 96,759, 95,830, 94,294 and 94,143 gene models for the *F. chiloensis* and *F. virginiana* haplotypes, respectively (Table 1). We further classified all the annotated genes using functional domains and identified 4,517 and 4,525 (super-)gene families for the two wild octoploids. Expansions and contractions of many gene families were found between the wild progenitors and cultivated strawberry 'Camarosa' (Supplementary Fig. 4a). In particular, the NB-ARC family size has significantly expanded in the strawberry cultivar (724 in 'Camarosa' versus 264 in *F. chiloensis*; odds ratio, 2.46; $P < 0.0001$; 724 in 'Camarosa' versus 264 in *F. virginiana*; odds ratio, 2.40; $P < 0.0001$; Fisher's exact test) (Supplementary Fig. 4b). We further compared the changes in NB-ARC gene family size among diploid species and wild and cultivated octoploid strawberries, and found substantially and significantly reduced family sizes in each subgenome of the wild octoploids ($P < 0.05$; one-way analysis of variance (ANOVA) with Duncan's multiple range test) (Supplementary Fig. 4c), probably due to the evolutionary processes of diploidization. We also identified 311.0–353.8 Mb of repetitive sequences, totalling 41.6–44.4% of each haploid genome (Supplementary Fig. 5 and Supplementary Table 6).

To assist with phylogenetic and genetic composition analyses of the octoploids, we separately assembled a diploid *F. nipponica* genome ($2n = 2x = 14$) with 126× nanopore data, generating a 290.9 Mb genome with a contig N50 length of 1.9 Mb (Table 1). The assembly was syntenic with the *F. vesca* genome (Extended Data Figs. 3c and 4c). A total of 36,136 protein-coding genes and 124.3 Mb of repetitive sequences were predicted in the *F. nipponica* genome (Supplementary Table 6).

For all three species, complete organelle genomes were also assembled. These exhibited typical genome sizes for the mitochondrial (285–296 kb) and chloroplast (156 kb) genomes (Supplementary Fig. 6).

Verification of assembly quality

We performed a series of measures to assess the assembly, phasing and annotation quality of the two octoploid genomes. In terms of continuity, there were only three to five gaps on average per pseudochromosome for each haplotype (Table 1 and Supplementary Fig. 7a). Telomere repeat motifs ('CCCTAAA') were found for 71 and 86 of the 112 chromosome ends in the *F. chiloensis* and *F. virginiana* assemblies, respectively (Supplementary Fig. 8). The LTR Assembly Index (LAI) scores were also used for estimating assembly continuity on the basis of long terminal repeat retrotransposons (LTR-RTs). The LAI scores were 13.2 in *F. chiloensis* and 17.9 in *F. virginiana*, which were considerably higher than that of the 'Camarosa' assembly[26] (LAI = 10.4) (Table 1 and Fig. 1d). Moreover, between 99.3% (of 43.7 Gb) and 99.4% (of 73.5 Gb) of Illumina paired-end reads were successfully mapped to the *F. chiloensis* and *F. virginiana* genomes (Table 1), and more than 98% of *k*-mers from HiFi and Illumina reads were detected in *F. chiloensis* and *F. virginiana* (Table 1 and Supplementary Fig. 9), suggesting a high degree of genome completeness. In terms of accuracy, we analysed the *k*-mer-based consensus quality value (QV)[31], which depends on the coverage and quality of the read set used for measuring a log-scaled probability of error for the consensus base calls. The QVs of the *F. chiloensis* and *F. virginiana* genomes were 68.1 and 68.3 (corresponding to 0.155 and 0.149 errors per 1,000,000 bp) when using *k*-mers ($k = 19$) present in HiFi reads, respectively, indicating that the assemblies were highly accurate (>99.99%) at the single-base level (Table 1 and Supplementary Fig. 9).

When aligning long reads back to the assemblies (comprising both haplotypes) of the two wild octoploids, each showed one clear major peak depth distribution (accounting for 99.88% and 99.53% of the whole assemblies, respectively), indicating that nearly all of the assembled regions in the genomes are individual haplotypes rather than collapsed assemblies (Supplementary Fig. 10). The phasing qualities of two polyploid genomes were further evaluated using nphase[32], a ploidy agnostic phasing pipeline that determines the allele frequency distribution within the haplotig. Allele frequency values around 0.5 indicate a potential erroneous merger of two distinct haplotypes, whereas a significant departure from 0.5 provides evidence of good phasing[32]. We found a small number of allele frequency values around 0.5 for *F. virginiana* (0.37%) and *F. chiloensis* (1.05%), indicating that most (99.63% and 98.95%) regions in both haplotypes were correctly phased (Fig. 1c and Supplementary Table 3).

More than 90% of gene models had Annotation Edit Distance scores of less than 0.5 (Supplementary Fig. 7b), indicating high concordance with empirical evidence for most gene models. Benchmarking Universal Single-Copy Ortholog (BUSCO) evaluations identified 96.1–96.4% of the 1,614 conserved orthologues in the genome annotations (Supplementary Figs. 7c and 11). We also aligned the annotated 'Camarosa' transcripts[26] to the new assemblies, showing that the mapping rates were >99.5% (Supplementary Table 5). Altogether, these results indicate that the two wild octoploids achieved haplotype-resolved and reference-level assemblies.

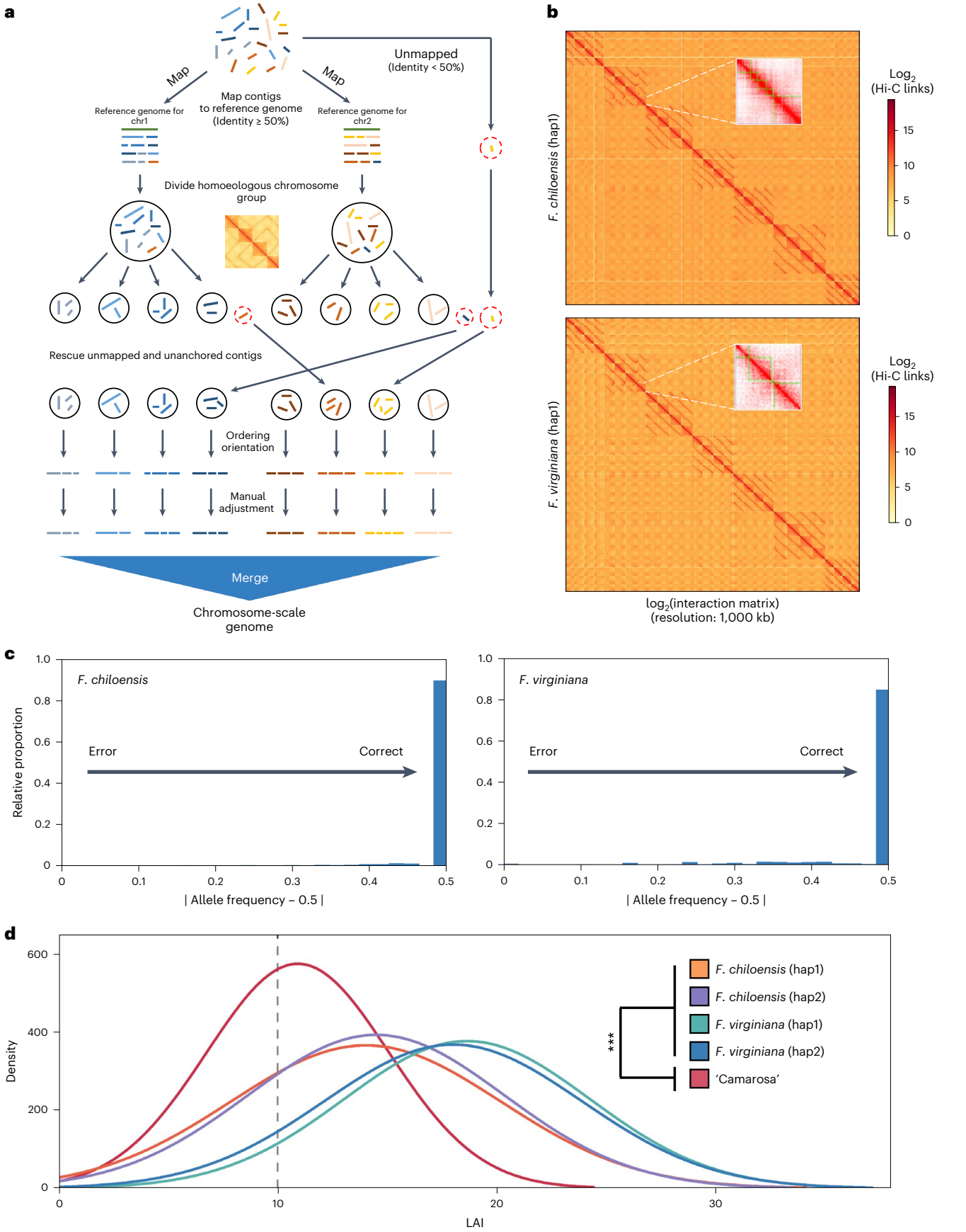
Genetic divergence between species and between haplotypes

Using the haplotype-resolved genome assemblies, we identified genomic variations between the two haplotypes of each species and between the two *Fragaria* species. Alignments of the homologous chromosomes indicate that gene orders are largely syntenic, except for a few

Fig. 1 | Hi-C-based anchoring of the wild octoploid strawberry genomes. **a**, The Hi-C-based anchoring pipeline for the wild octoploid strawberry genomes. The assembled contigs are pre-clustered into seven heterologous groups on the basis of sequence similarity to the chromosomes of the diploid *F. vesca* (shown as two heterologous chromosome (chr) groups). Each heterologous group is further divided into four homoeologous groups on the basis of the density of Hi-C signals. The unanchored and unmapped contigs are assigned to the divided groups by searching for optimal Hi-C signals. The contigs are ordered and oriented in each divided group and adjusted manually. **b**, Heat maps of *F. chiloensis* and *F. virginiana* showing the distribution of Hi-C interaction signals in a 1,000 kb resolution. High densities of interactions are indicated in red. **c**, Allele frequencies of the haplotigs for *F. chiloensis* and *F. virginiana*. Values around 0.5 are a sign that two distinct haplotypes are erroneously merged into one. **d**, Assessment of the completeness of the five assemblies using the LAI. The results are shown as a normal distribution (one-way ANOVA with Duncan's multiple range test; d.f. = 11,809; $^{***}P < 0.001$).

https://doi.org/10.1038/s41477-023-01473-2

small-scale structural variants (>50 bp) (Supplementary Figs. 12 and 13). For example, a ~0.81 Mb inversion occurs on chromosome 3A between *F. chiloensis* and *F. virginiana*, and another 0.8 Mb inversion occurs uniquely on chromosome 2A between two haplotypes of *F. virginiana* (Supplementary Fig. 14 and Supplementary Tables 7 and 9). These two notable inversions were each in a single contig with continuous



Hi-C signals, ruling out the possibility of misassembly (Supplementary Fig. 14). Among these structural variants, inversions were approximately fivefold to tenfold less abundant than translocations or duplications (Supplementary Table 7).

Relative to structural variants, single nucleotide variations (SNVs) were the most abundant type of variant. A total of 5.13 million and 4.63 million SNVs were found between haplotypes of *F. chiloensis* and *F. virginiana*, while 5.04 million and 4.94 million were identified between the assemblies of the two wild octoploids (Supplementary Table 8). The frequencies of SNVs were 4.51–8.07 per kb, providing a basis for separating homoeologous chromosomes and phasing haplotypes. Moreover, the SNV densities between and within the wild octoploid strawberries were generally similar (Supplementary Table 8 and Supplementary Fig. 15a,b). We found a moderate correlation in SNV density between haplotypes and species ($R = 0.54$; $P < 2.2 \times 10^{-16}$; linear regression) (Supplementary Fig. 15c). We further identified 3,090 bins with significantly different SNV densities, among them 1,481 bins biased between haplotypes and 1,609 bins biased between species (Supplementary Fig. 15c,d). These results suggest that although similar numbers of SNVs were identified among haplotypes and among species, 19.28% (154.4 Mb) of the regions showed significant differences in SNV density (Supplementary Fig. 15c,d).

Subgenome assignment for the wild and cultivated octoploids

The assignment of chromosomes from octoploid *Fragaria* species to one of four subgenomes has been contentious. Although most studies consistently identify sets of chromosomes to subgenomes A and B on the basis of their sequence similarity to *F. vesca* and *F. iinumae*, chromosomal assignments have produced conflicting results for subgenomes C and D[26–28,33–35]. To distinguish the subgenomes of the octoploid strawberry genomes, we integrated prior knowledge and used both mapping and *k*-mer-based methods.

Several studies[26,33] have indicated that *F. nipponica* and *F. viridis* were either the closest diploid progenitors or the best surrogate diploid species for identifying the C and D subgenomes[33]. We therefore mapped transcriptomic data from the two recognized diploid progenitors (*F. vesca* and *F. iinumae*) and two potential surrogates (*F. viridis* and our newly sequenced *F. nipponica*) to the *F. chiloensis* and *F. virginiana* genomes. The mapping of the *F. vesca* and *F. iinumae* data could each categorize 7 of the 28 pseudochromosomes into the A or B subgenome, as the mapping ratios (median values, 8.89% and 5.52–5.60%, respectively) were significantly higher than for the remaining chromosomes ($P < 0.05$; one-way ANOVA and Duncan's multiple range test) (Extended Data Fig. 5). In contrast, the mapping of the *F. nipponica* and *F. viridis* data failed to clearly resolve assignments of the remaining 14 chromosomes to the C and D subgenomes, as their mapping rates were similar between the C and D subgenomes and actually higher for the A subgenome than for the other subgenomes (Extended Data Fig. 5 and Supplementary Table 10).

As an alternative approach, we used the *k*-mer-based method to identify subgenome-specific sequence signatures, which has been successfully applied in characterizing the subgenomes of several plant polyploids with unknown or even extinct diploid progenitors[36], including cultivated strawberry[35]. We identified a total of 10,174 differential *k*-mers ($k = 13$ and frequency ≥ 50) in *F. virginiana* (hap1), which confidently divided 28 chromosomes into four distinct subgenomes (bootstrap = 100) (Fig. 2a, Supplementary Fig. 16 and Supplementary Table 11). Fully consistent results were also obtained from the other octoploid assemblies: *F. chiloensis* (hap1 and hap2), *F. virginiana* (hap2) and 'Camarosa' (Supplementary Figs. 16 and 17 and Supplementary Table 11). Principal component analysis (PCA) clearly separated the four subgenomes of all haplotypes, each of which comprised seven heterogeneous chromosomes (Fig. 2b), with PC1 mainly explaining the variance (44.3–47.4%) between the A subgenome and others, and PC2 explaining the variance (22.6–47.4%) between the B subgenome and the C + D subgenomes. Notably, for all these analyses, the C and D subgenomes were closely related yet still distinguishable (Fig. 2b), of which the variance could be explained by PC3 (12.7–14.2%) (Supplementary Fig. 18).

To directly compare subgenomic assignments in this study with those previously proposed for the 'Camarosa' genome[26,33], we identified unique matches for each chromosome of *F. chiloensis* and *F. virginiana* to chromosomes of 'Camarosa' (Supplementary Fig. 19 and Supplementary Table 11). Overall, 22 of the 28 *k*-mer-based chromosomal assignments for *F. chiloensis*, *F. virginiana* and 'Camarosa' were consistent, but six chromosomes (2C, 2D, 5C, 5D, 6C and 6D) were inconsistent (Fig. 2c). HE regions totalling 6.3 Mb (3.5%), 7.8 Mb (4.6%) and 6.2 Mb (3.6%) were identified between these six chromosomes of the C and D subgenomes in *F. chiloensis*, *F. virginiana* and 'Camarosa', respectively, suggesting that the effect of HEs on subgenome assignment would be minor. As a final test, we looked for specific *k*-mers and LTR-RTs within each of the four subgenomes based on the previously proposed assignment[26,33], which identified only a very small number of those sequence signatures (<4 specific *k*-mers and <50 specific LTR-RTs) for the C and D subgenomes (Extended Data Figs. 6 and 7 and Supplementary Fig. 20). Thus, the *k*-mer-based subgenomic assignment approach, which proceeds independently of any inferred diploid ancestor, strongly rejects the previous assignments for the C and D subgenomes, which were based on inferences of *F. viridis* and *F. nipponica* as diploid ancestors (or surrogates).

Genomic evidence of the potential diploid ancestors

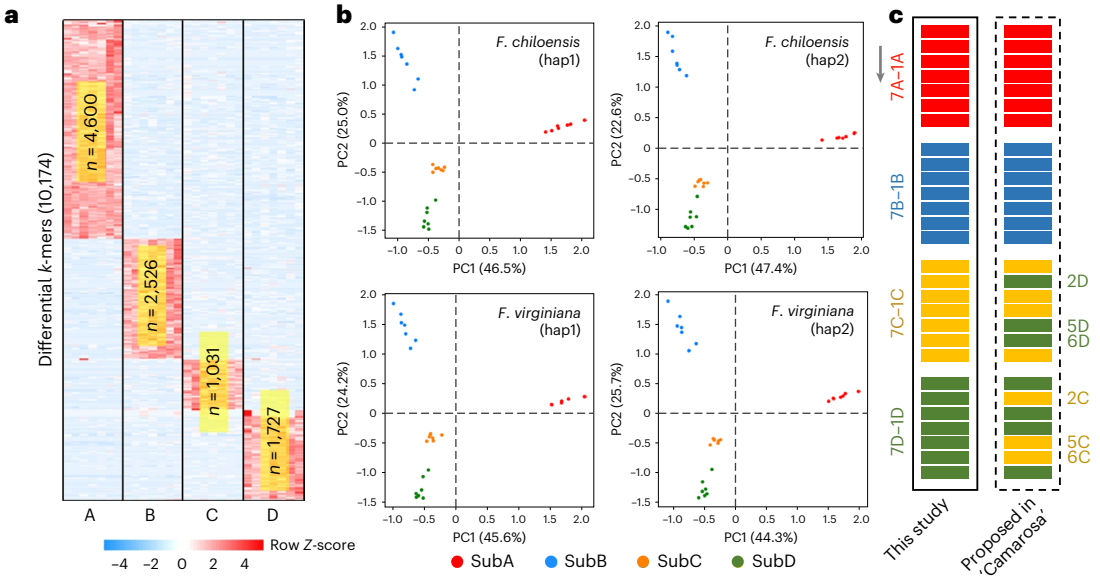
As the inferred potential diploid ancestry of the *Fragaria* octoploid species remains highly controversial[26,28,33], we generated a chromosome-level genome assembly of *F. nipponica*, the last putative diploid progenitor without an available genome, to complement the eight other published diploid genomes from *F. daltoniana*[17], *F. iinumae*[37], *F. mandschurica*[17], *F. nilgerrensis*[38], *F. nubicola*[28], *F. pentaphylla*[17], *F. vesca*[30] and *F. viridis*[28] for analyses. We then used *k*-mer-based, mapping-based and phylogenetic-based methods to perform the inference of putative diploid ancestors or closely related species.

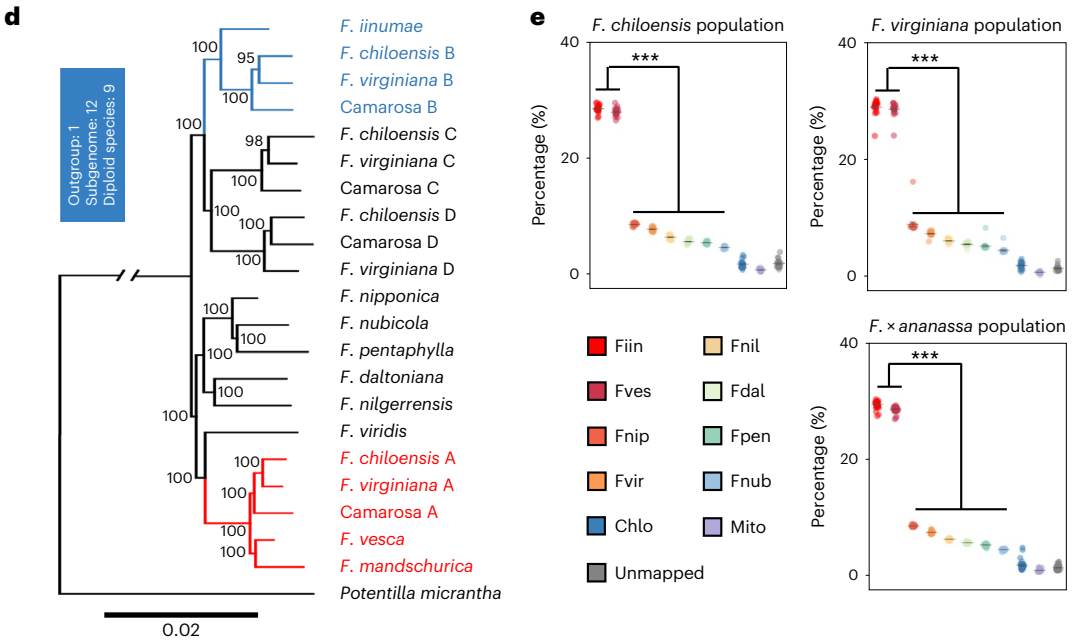
Phylogenetic relationships were inferred from a concatenation of all 941 single-copy orthologues (Fig. 2d) and concatenations of single-copy orthologues in each homologous chromosome (Supplementary Fig. 21). These trees nearly universally support a grouping of subgenome A with *F. vesca* and its close relative *F. mandschurica*,

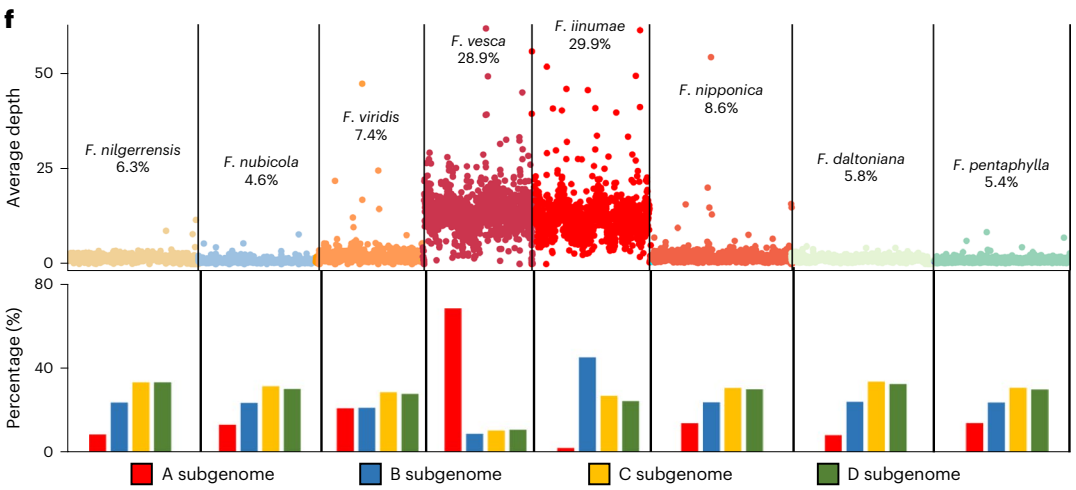
Fig. 2 | Subgenome assignments of the octoploids and tracing the potential diploid ancestors. a, Clustering of differential *k*-mers ($k = 13$ and frequency ≥ 50) among homoeologous chromosome sets that could differentiate *F. virginiana* (hap1) chromosomes into four subgenomes. *n* is the number of specific *k*-mers on each subgenome. **b**, PCA plots of four assemblies based on differential *k*-mers. The values in parentheses indicate the percentage of variance explained. **c**, Comparison of the subgenome assignment in this study and that previously proposed for the strawberry cultivar 'Camarosa' by Hardigan et al.[33] and Edger et al.[26]. **d**, Phylogenetic analysis of the three octoploids (each with four subgenomes) and nine diploid *Fragaria* species. Note: The clades where *F. iinumae and F. vesca* are located are shown in blue and red, respectively. **e**, Genetic contributions of diploid strawberry species to octoploids. The percentages of resequencing data from 22 *F. chiloensis* accessions, 20 *F. virginiana* accessions and 42 *F.* × *ananassa* cultivars that mapped to eight *Fragaria* diploid species (*F. iinumae* (Fiin), *F. vesca* (Fves), *F. nipponica* (Fnip), *F. viridis* (Fvir), *F. nilgerrensis* (Fnil), *F. daltoniana* (Fdal), *F. pentaphylla* (Fpen) and *F. nubicola* (Fnub)) and organelle genomes (Chlo and Mito) are shown. The data are presented as mean values ±1 s.e.m. (one-way ANOVA with Duncan's multiple range test; d.f. = 175, 159 and 335, respectively; ***$P < 0.001$). **f**, The depth of reads of a cultivated strawberry accession (sample_023)[33] mapped to the eight diploid strawberry genomes (top) and the distribution of reads contributed by diploids to each subgenome (bottom).

https://doi.org/10.1038/s41477-023-01473-2

whereas subgenomes B, C and D are clustered with *F. iinumae*. Moreover, a coalescent-based approach of 941 genes produced an identical tree to the concatenated-based analyses (Supplementary Fig. 22). This phylogenetic relationship was also consistent with *k*-mer-based estimation of genetic distance (Extended Data Fig. 8) and the relative abundance of subgenome-specific-*k*-mers in each genome







(Supplementary Figs. 23 and 24). To further verify the above conclusion, we tabulated the closest species of each subgenome in individual gene trees for the 6,223 single-copy orthologues (Extended Data Fig. 9a,b), which again support subgenome A with *F. vesca* and subgenome B with *F. iinumae*, whereas subgenomes C and D do not closely ally with any individual diploid (Extended Data Fig. 9c). Instead, subgenomes C and D predominantly associate with the subgenome B and *F. iinumae* clade (53.5% of trees).

We also assessed the genomic contributions of diploid *Fragaria* species to the octoploid strawberries using a phylogeny-free approach (Fig. 2e,f) by mapping resequencing data from 22 *F. chiloensis* accessions, 20 *F. virginiana* accessions and 42 *F.* × *ananassa* cultivars[33] to eight diploid *Fragaria* genomes, excluding *F. mandschurica* due to its sequence similarity (Supplementary Fig. 19) and close phylogenetic relationship with *F. vesca* (Fig. 2d and Supplementary Figs. 21 and 22). Nearly two thirds of the reads were mapped to either *F. vesca* or *F. iinumae*, similar to the recent findings by Feng et al.[28], whereas the remaining reads were mapped at a lower frequency to the remaining diploids: *F. nipponica* (8.6%), *F. viridis* (7.5%), *F. nilgerrensis* (6.3%), *F. daltoniana* (5.7%), *F. daltoniana* (5.3%) and *F. nubicola* (4.5%) (Fig. 2e,f). Mapping of the high-coverage Illumina data of *F. chiloensis* and *F. virginiana* sequenced in this study produced similar results (Supplementary Fig. 25).

Differential gene retention among octoploid subgenomes

Previous studies have indicated different degrees of gene retention among subgenomes of the 'Camarosa' genome[26]. BUSCO analysis suggests that there is differential gene retention among subgenomes of the wild octoploid haplotypes as well (Supplementary Fig. 11). To more directly compare whether this evolutionary pattern is similar among the *Fragaria* octoploids, we investigated the retention of homoeologous genes in each subgenome. Approximately 46% of the total annotated genes maintained all four copies of the homoeologues, while 7%, 23% and 14% of genes retained one, two and three homoeologues (Supplementary Table 12). Among these homoeologues that lack one or more copies, subgenome A generally showed the fewest gene losses (approximately twofold to threefold lower than the B and C subgenomes and more than fourfold lower than the D subgenome) and maintained the highest number of subgenome-specific homoeologues (Supplementary Table 12). In contrast, subgenome D displayed the highest gene losses (1,704 in *F. chiloensis*, 1,762 in *F. virginiana* and 1,354 in 'Camarosa'). Close examinations indicated that more than 300 genes were co-lost in the D subgenome of the three *Fragaria* octoploids, many of which were related to environmental stimuli and other biological processes such as catalytic and nucleic acid binding activities (Supplementary Fig. 26 and Supplementary Table 13).

Similarity and divergence in HEB

The subgenome assignment efforts provided an excellent basis to investigate the patterns of HEB in high-order polyploids. On the basis of the new subgenome assignments, we performed RNA-seq experiments and analysed HEB in the red-fruit stage among *F. chiloensis*, *F. virginiana* and 'Camarosa' (Extended Data Fig. 1 and Supplementary Fig. 27). A total of 11,189, 11,109 and 5,836 homoeologous gene sets (comprising 44,756, 44,436 and 23,344 genes) from *F. chiloensis*, *F. virginiana* and 'Camarosa', respectively, have retained a 1:1:1:1 syntenic relationship across the four homoeologous subgenomes, which we refer to as quadruplets (Supplementary Table 12). Among the quadruplets, the overall expression levels of homoeologues in the A subgenome were generally higher than in the other three subgenomes in red-fruit developmental stages (Supplementary Fig. 28 and Extended Data Fig. 10). To examine HEB, we filtered out the quadruplets with the lowest expression (Supplementary Fig. 29) and normalized the expression level of the remaining quadruplets so that the sum was 1.0 in each octoploid. The relative expression of each homoeologue defines the quadruplet's position within a quaternary phase diagram (Fig. 3a), and each quadruplet was qualitatively defined into balanced, dominant and suppressed categories (Supplementary Figs. 30a, 31 and 32). Most syntenic quadruplets were assigned to the balanced category, ranging from 52.0% in 'Camarosa' to 56.9% in *F. virginiana*. Consistent with the pairwise comparisons of expression dominance (Fig. 3a,b), the dominant homoeologue was more likely to be from the A subgenome (2.9–3.0%) than from the other subgenomes (1.7–2.0% for B, 1.6–2.3% for C and 1.5–1.7% for D) ($P < 0.001$; Fisher's exact test and one-way ANOVA with Duncan's multiple range test), and the suppressed homoeologue was least likely to be from the A subgenome (5.0–6.1%) compared with the other subgenomes (8.6–9.3% for B, 9.8–11.1% for C and 11.1–12.7% for D) ($P < 0.001$; Fisher's exact test and one-way ANOVA with Duncan's multiple range test), supporting the dominance of the A subgenome in wild and cultivated octoploid *Fragaria* (Fig. 3b and Supplementary Fig. 33).

To determine the evolutionary dynamism of HEB in the *Fragaria* octoploids, we evaluated the stability of HEB categorization for the 3,005 quadruplets (36,060 genes) shared among the three species. We found that 67.6% of quadruplets with balanced HEBs remained in the same category in *F. chiloensis* and *F. virginiana*. Dominant and suppressed quadruplets tended to be more variable, with only 26.0% and 26.6% maintaining their HEB categories in the two wild octoploids (Supplementary Fig. 34). We further used the HEB of quadruplets to quantitatively assess the relative distance of HEB shifts between quadruplet pairs from the two wild progenitors (*F. chiloensis* and *F. virginiana*) (Fig. 3d). About 7% of the quadruplet pairs exhibited a dynamic shift (>0.4 distance), which were enriched for functions related to plant organ morphogenesis and stress response (Figs. 3c,d and Supplementary Fig. 35a). Most (68.0%) of the quadruplet pairs were quite stable (<0.2 distance), indicating a similar HEB status for most quadruplets between the two wild octoploids (Figs. 3c,d).

To assess whether cultivated strawberry has experienced unique HEB shifts relative to the wild octoploids, we examined the HEB status in 'Camarosa' for the stable quadruplets identified in the wild octoploids. We identified the relative balance point (RBP) (Supplementary Fig. 30b) for each quadruplet pair between the two wild species and then calculated the relative distance between the RBP of the wild octoploids and the 'Camarosa' quadruplet. As expected, more than 80% of the stable wild quadruplets showed similar expression patterns in 'Camarosa', with predominantly housekeeping functions such as cellular metabolic process and protein binding (Fig. 3e,f and Supplementary Fig. 35b). About 3% of quadruplets were dynamically shifted in 'Camarosa'. Many

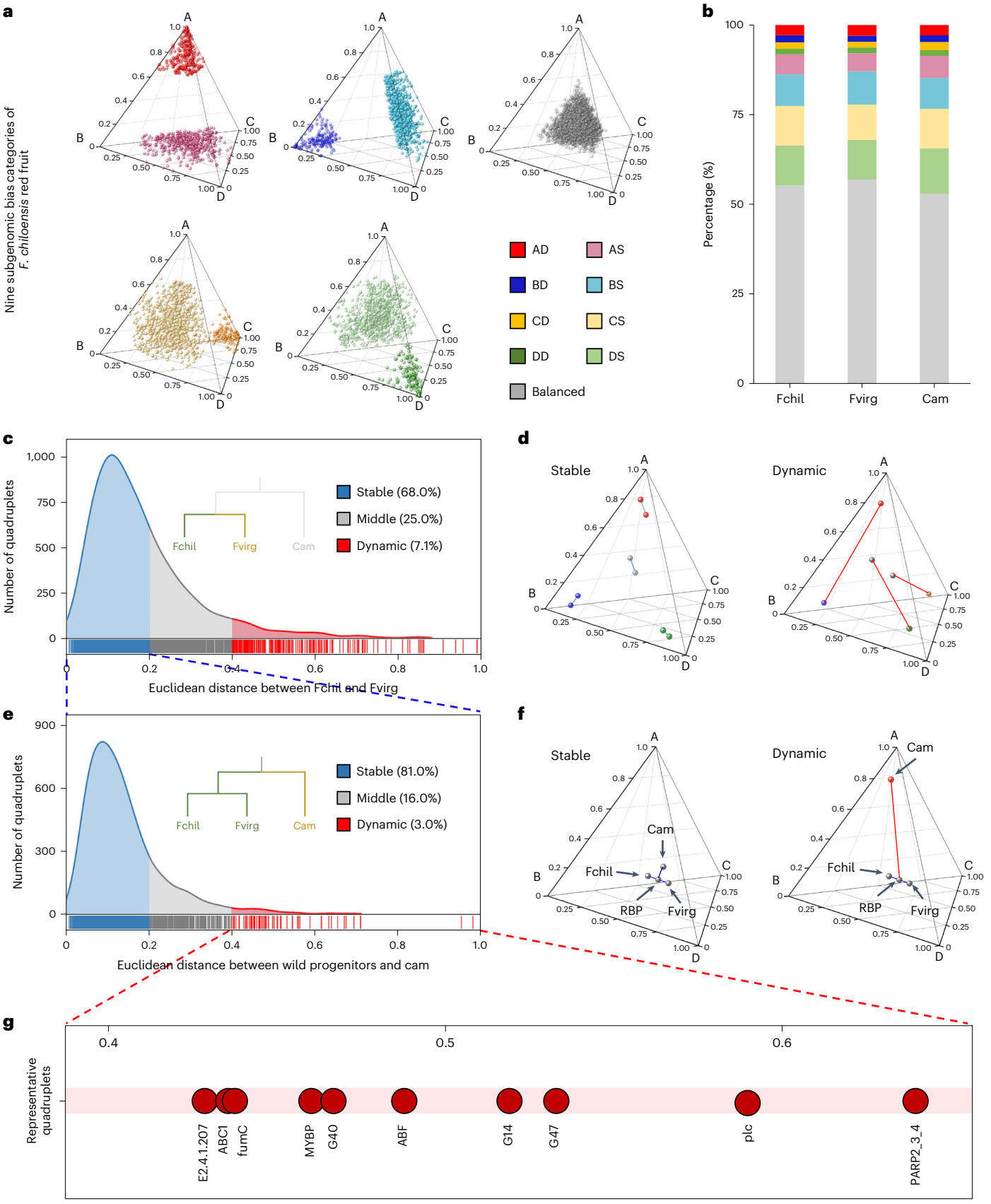
Fig. 3 | Static and dynamic pattern of HEB among quadruplets of the wild and cultivated octoploid strawberries. a, Quaternary phase diagram showing the relative expression levels of 8,882 quadruplets in *F. chiloensis*. Each point has four coordinates, representing the relative expression level of a homologue that could bias to the A, B, C and D subgenomes. Quadruplets in vertices correspond to single-subgenome-dominant categories, whereas quadruplets close to edges and between vertices correspond to suppressed categories. Balanced quadruplets are shown in grey. AD, A dominant; BD, B dominant; CD, C dominant; DD, D dominant; AS, A suppressed; BS, B suppressed; CS, C suppressed; DS, D suppressed. **b**, The proportion of quadruplets in each category of HEB across three octoploids: *F. chiloensis* (Fchil), *F. virginiana* (Fvirg) and 'Camarosa' (Cam). **c**, Distribution of Euclidean distance variation of quadruplets between the two wild species. Quadruplets with distance ≤ 0.2, 0.2 < distance < 0.4 and distance ≥ 0.4 were defined as stable, middle and dynamic, respectively. **d**, Quaternary phase diagram of representative stable and dynamic quadruplet pairs between the two wild octoploids. **e**, Distribution of Euclidean distance variation of quadruplets between the RBP of two wild species and cultivated strawberry 'Camarosa'. **f**, Quaternary phase diagram of representative stable and dynamic quadruplet groups between the wild and cultivated strawberries. **g**, Representatives of dynamic quadruplets between wild and cultivated strawberries.

https://doi.org/10.1038/s41477-023-01473-2

of these were related to stress response, but they also included several transcription factors (such as ABF and MYB) (Fig. 3g, Supplementary Fig. 36 and Supplementary Table 15), which might be relevant to strawberry domestication.

Discussion

The genomes of high-order polyploids are complex, requiring extensive effort to disentangle their genomic structure, genetic composition and evolutionary diversification[39–41]. In this study, we generated



chromosome-level and haplotype-resolved genome assemblies for two wild octoploid strawberries (*F. chiloensis* and *F. virginiana*), which provided ample information to investigate their subgenome-specific structure and genetic content, the extent of genomic and transcriptional coordination and variation between subgenomes and species, and the sources and relative contributions of the diploid and octoploid progenitors of modern cultivated strawberry (*F.* × *ananassa*).

F. chiloensis and *F. virginiana*, the two parental species of modern strawberry, show considerable variation in morphological and physiological characteristics, such as fruit size and shape (Extended Data Fig. 1) and drought tolerance[18,42,43]. However, the two wild *Fragaria* genomes have nearly identical gene family sizes and similar gene content (Supplementary Fig. 4a). They are also largely syntenic, with a few large-scale structural changes among subgenomes or between species (Supplementary Figs. 12 and 13). This structural conservation extends to the cultivated strawberry and more distantly related species from *Fragaria* and related genera[25,28,44] and parallels the pattern of genome evolution in allotetraploid cottons[45]. This conservation has allowed for extensive HE regions between subgenomes, many of which are shared with the cultivated strawberry (Supplementary Figs. 37 and 38).

However, the two wild octoploids have experienced substantial levels of genomic diversification among subgenomes by accumulating substitution mutations (single nucleotide polymorphisms (SNPs)) (Table 1, Extended Data Fig. 1 and Supplementary Table 8), accompanied by the alteration of their genome sizes and transposable element (TE) content as well as shifts of coordinated homoeologue expression (Fig. 3d, Table 1 and Supplementary Table 6). The genomic heterozygosity between the two species probably provides at least two benefits for strawberry breeding when the genetically distinct parents hybridize together. First, hybrid vigour of this fruit crop may be easily achieved[46], resulting from heterogeneous evolutionary rates of individual genes and the alteration of molecular interactions among homoeoalleles. To this end, we analysed the evolution of gene family size between the wild and cultivated strawberries and revealed that several gene families have experienced dramatic expansion or contraction (Supplementary Fig. 4), which might be the product of artificial selection during strawberry breeding. We also provided a four-dimensional analysis of the evolutionary dynamism of HEBs in octoploids (Fig. 3) and revealed several transcription factors that probably shifted in HEBs during domestication (Supplementary Fig. 36). These transcription factors (such as ABF, which could enhance drought tolerance in *Arabidopsis*[47], and MYBP, which plays an important role in abiotic stress resistance[48]) are good candidates for further testing their expression evolution among strawberry cultivars and individuals (Supplementary Table 15). Second, it is possible that the coexistence of heterozygous octoploid genomes may facilitate homologous recombination during strawberry germplasm improvements, which are intensively selected for phenotypic variations for human needs (Supplementary Fig. 39).

F. chiloensis and *F. virginiana* are closely related allo-octoploids thought to have originated from a common octoploid ancestor[15]. Yet the species identity of the diploid progenitors of the octoploid strawberries remains strongly contested. We reexamined the genomic contributions of extant diploid species by additional sequencing of an unpublished genome from *F. nipponica* (Fig. 2e,f and Supplementary Fig. 25). Our results, based on more species for phylogenetic analysis (including the *F. nipponica* genome), more comprehensive read mapping analyses (a phylo-free approach) and genome-wide *k*-mer signatures, provide highly consistent evidence for *F. vesca* and *F. mandschurica* as the closest living relatives of the A subgenome and *F. iinumae* as the closest living relative of not only the B subgenome but also the C and D subgenomes. In contrast, *F. nipponica* and *F. viridis*, which have been occasionally described as the closest relatives of subgenomes C and D, do not have support from phylogenetic analysis of 941 or 6,223 single-copy genes (Fig. 2d and Extended Data Fig. 9). Mapping analysis also supports these results, where *F. vesca* and *F. iinumae* represent the most abundant mapped reads to the octoploids, and *F. vesca* shows the strongest association with A, while *F. iinumae* associates closest not only with the B subgenome but also most strongly of all diploids with the C and D subgenomes. These results are consistent with a recent finding on the diploid progenitors of cultivated strawberry based on transposon-based *k*-mers[35]. On the basis of these results, we suggest that the B, C and D subgenomes arose by polyploidization of *F. iinumae* and an extinct close relative, followed by the addition of subgenome A from subspecies of *F. vesca*.

Methods

Genome sequencing

Detailed information on the three sequenced *Fragaria* species (*F. chiloensis*, *F. virginiana* and *F. nipponica*) is summarized in Supplementary Table 1, and their morphological photos are shown in Extended Data Fig. 1. In brief, seedlings were provided by J. Lei, who collected them from diverse resources in long-term efforts. The seedlings were grown and vegetatively propagated in a greenhouse at Kunming Institute of Botany (Yunnan, China). Total DNA was extracted from young leaves and used for library construction and further genomic sequencing with a combination of platforms.

The two octoploids, *F. chiloensis* and *F. virginiana*, were sequenced using PacBio CCS technology. The PacBio library was constructed for each species and sequenced on a PacBio Sequel II instrument following the manufacturer's instructions at Grandomics Biosciences Co. (Wuhan, China). Approximately 34.6 Gb and 35.0 Gb of PacBio HiFi reads were generated for *F. chiloensis* and *F. virginiana*, respectively (Supplementary Table 2).

Diploid *F. nipponica* was sequenced on a Nanopore GridION X5 sequencer (ONT), generating approximately 31.5 Gb nanopore raw reads (Supplementary Table 2). All libraries were constructed according to the manufacturer's protocols and sequenced at Biomarker Technologies Corporation (Beijing, China).

Genome size estimation

The genome sizes of the three wild *Fragaria* species were estimated via both *k*-mer frequency analysis and flow cytometry. For the *k*-mer analysis, one Illumina sequencing library was constructed for each species and sequenced on the Illumina HiSeq 2500 platform, generating 43.7–73.5 Gb of paired-end reads (Supplementary Table 2). Then, 21-mer counts were collected from the sequenced Illumina reads using jellyfish v.2.2.10 (ref. 49), and genome size was estimated using GenomeScope v.1 (ref. 50). Absolute genome sizes were further measured by flow cytometry. Briefly, suspensions of the test sample and the internal reference sample (*Oryza sativa* L. 'Japonica') were mixed, and a BD FACSCalibur flow cytometer was used to detect the stained cell nuclei in the suspension samples. The ploidy levels of the sequencing materials were further estimated using Smudgeplot[51], taking the coverages of the pre-computed 21-mers as input (Extended Data Fig. 2).

Hi-C sequencing

Hi-C libraries were constructed from crosslinked chromatin of plant cells following a standard protocol (DpnII enzyme). The libraries were sequenced on a MGISEQ-T7 device, generating a total of 137.4 Gb, 105.6 Gb and 48.8 Gb of Hi-C data for *F. chiloensis*, *F. virginiana* and *F. nipponica*, respectively (Supplementary Table 2).

Genome assembly and pseudochromosome construction

Long (median length, >13 kb) and accurate (median read quality, >Q30) HiFi reads were used to de novo assemble the *F. chiloensis* genome using Hifiasm v.0.15.1 (ref. 52) with the parameters $D = 10$, $r = 4$ and $a = 5$ and to assemble the *F. virginiana* genome with the default parameters, both integrating the Hi-C data to achieve haplotype-resolved assemblies[52]. The HiFi with Hi-C integrated assembly strategy implemented

in Hifiasm performs phased string graph construction and long-range phasing information for those without parental information[52].

To construct the pseudochromosomes of the two wild octoploids, the *F. vesca* genome[30] was used as a reference to pre-cluster contigs followed by chromosome-level scaffolding with the Hi-C data (Fig. 1a). As shown in Supplementary Fig. 5, our scaffolding pipeline differed from a typical ALLHiC workflow by the partition of homologous sequences, iterative scaffolding and contig recovery. Briefly, the main process included the following. First, the assembled haplotigs were separated into seven heterologous chromosomal groups by mapping to the reference genome using RaGOO[53] with the parameter $i = 0.5$. Second, the Hi-C reads were mapped to the draft assemblies by using BWA v.0.7.14-r1188 (ref. 54), and the resulting bam files were filtered by using scripts (PreprocessSAMs.pl and filterBAM_forHiC.pl) implemented in ALLHiC[29] to get the Hi-C interaction signals. Each heterologous group was categorized into four distinct subgenome groups according to the Hi-C signals by using ALLHiC_partition[29], with the parameters e = GATC and enzyme_sites = MboI. Third, the unanchored and unmapped haplotigs were reassigned with optimal Hi-C signals by using ALLHiC_rescue. Fourth, the haplotigs in each chromosome were ordered in orientation with the optimized function in ALLHiC. Fifth, the ALLHIC output file (.agp file) was converted into a Juicebox[55] input file (.hic file and .assembly file) by using Juicebox utilities (https://github.com/phasegenomics/juicebox_scripts). The Hi-C interaction signals in each chromosome were manually checked and adjusted with Juicebox[55]. Moreover, to avoid any technique bias that could be introduced by the diploid reference, all the anchored, unanchored and unmapped contigs were treated with caution (Supplementary Fig. 3). These scaffolding processes finally resulted in approximately 95.3–97.2% of the assembled haplotypes being anchored to the 28 pseudochromosomes.

For *F. nipponica*, the raw nanopore reads were self-corrected and assembled using Canu v.1.9 (ref. 56) with the default parameters. The initial assembly (contig-level) was polished using Racon v.1.4.3 (ref. 57) for two rounds with the corrected nanopore reads. Additional genome polishing was conducted using Pilon v.1.22 (ref. 58) for three rounds with deep-sequenced (~109.3×) Illumina reads. The corrected contigs (N50 = 1.9 Mb; error rate, 0.32%) were anchored into the seven pseudochromosomes using LACHESIS[59] with the parameters CLUSTER_MIN_RE_SITES = 34, CLUSTER_MAX_LINK_DENSITY = 2, CLUSTER_NONINFORMATIVE_RATIO = 34, ORDER_MIN_N_RES_IN_TRUN = 34 and ORDER_MIN_N_RES_IN_SHREDS = 34. The Hi-C interaction signals in each chromosome were manually checked and adjusted with Juicebox[55].

Organelle genome assembly

For each species, the complete mitochondrial and chloroplast genomes were also assembled. Briefly, Illumina reads were used to assemble the chloroplast genomes using NOVOPlasty v.2.7.2 (ref. 60). For the mitochondrial assemblies, long reads of *F. chiloensis*, *F. virginiana* and *F. nipponica* were mapped to the *F. orientalis* mitogenome (MW537838) separately, and the mapped reads were used for de novo assembly using HiCanu v.2.2 (ref. 61). The organelle genomes were annotated with Geneious v.7.1.4, followed by manual adjustment.

Contamination screen

To identify putative contaminated contigs in the assemblies, megaBLAST was used to search against the databases of common contaminants (ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz), adaptor sequences (ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/adaptors_for_screening_euks.fa), and bacterial sequences and fungal sequences derived from the nt database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz), following the methods and standards used in the Vertebrate Genomes Project[62].

The assembled organelle genomes were used to identify any misassembled or transferred organelle sequences in the nuclear genomes using megaBLAST with the parameters $e \leq 1 \times 10^{-4}$; sequence identity, ≥90%; and match length, ≥500. Organelle sequence matches embedded in nuclear sequences were kept but masked before the mapping-based analyses (Supplementary Table 4).

Quality assessment of genome assembly and annotation

Continuity. The LAI[63,64] was calculated for each assembly using both whole-genome TE annotations and intact LTR-RTs identified by LTR_FINDER v.1.1 (ref. 65) and LTRharvest v.1.1 (ref. 66) as inputs.

Completeness and base accuracy. Short reads were mapped to each assembly using sppIDer[67] to count the mapping rates. The *k*-mer QVs and *k*-mer completeness were reported by Merqury[31] with 17-, 19- and 21-mers present in the HiFi and Illumina reads. Similar accuracy estimates were obtained using different *k*-mers, but the QVs were slightly lower using *k*-mers from the Illumina data (Supplementary Fig. 9).

Phasing. The long reads were aligned back to their corresponding assemblies (comprising both haplotypes) using minimap2 v.2.17-r941 (ref. 68) to generate a coverage histogram plot with purge_dups v.1.2.3 (ref. 69), which was used to determine whether the removal of false duplicated contigs was needed. To quantitively evaluate the distribution of long-read coverage, we defined a clear peak with strict standards as all data points before reaching the peak point with an increased frequency and with decreased frequency afterwards. The percentage of peaks was then calculated using the formula Main peak% = Sum of the frequencies per depth contained in the main peak/Sum of frequencies per depth. Moreover, to test the validation of this method in detecting collapsed regions, we removed 14 chromosomes (4C–D, 5A–D, 6A–D and 7A–D) (371.48 Mb sequence, which accounts for 24.76% of the whole genome) from the *F. virginiana* hap2 assembly but included the hap1 assembly to generate a reduced genome representation. We evaluated this simulated collapsed assembly using identical processes, and we identified two major peaks. One peak (76.72%) represented the region without collapse, and the other peak (23.08%) represented the collapsed region (Supplementary Fig. 10), suggesting that this method is valid. The accuracy of haplotype phasing was also estimated using nphase[32]. Briefly, the highly accurate short reads (Illumina data) and long reads (HiFi data) were mapped to the genome assemblies to identify heterozygous positions and cluster alleles in the long reads into haplotypes. The distribution of heterozygous allele frequency within each haplotig was then reported and used as a proxy for phasing quality.

Telomere identification. Telomere repeat motifs ('CCCTAAA') were identified in the ends of pseudochromosomes using Tandem Repeats Finder[70].

Annotation estimation. The completeness of gene annotation was determined with BUSCO v.4.1.2, which measures the retention rate of conserved genes in genome annotation, in the protein model based on the embryophyta_odb10 database[71].

Repeat annotation and gene prediction

A de novo TE library was constructed on the basis of the *F. nipponica*, *F. chiloensis* and *F. virginiana* genomes using RepeatModeler v.2.0.1. Repetitive sequences in the three species genomes were then identified using RepeatMasker v.4.1.0 (ref. 72) with the combination of the de-novo-built TE library and Repbase (v.20181026) with the parameter *s*. All three genomes were soft-masked before further analyses.

To generate gene models for the *F. chiloensis* and *F. virginiana* genome assemblies, Iso-seq data were generated from samples of mixed tissues and stages (unrevealed and exposed flower buds, blooming flowers, and newly sprouted and mature leaves), generating 1.8 Gb and 3.3 Gb of CCS long reads (Supplementary Fig. 40a and Supplementary Table 16). These data were clustered and polished using the

isoseq3 pipeline (https://github.com/PacificBiosciences/IsoSeq) to obtain high-quality transcripts. The transcripts were aligned to their corresponding genomes to train gene models with BRAKER v.2.1.5 (ref. 73). Total RNA was also extracted from the mixed samples for constructing an RNA-seq library sequenced on the HiSeq 4000 system following the manufacturer's instructions. Raw RNA-seq data from the mixed samples and mature fruits (detailed samplings are shown in the 'RNA-seq experiment setup' section) were cleaned using fastp v.0.20.1 (ref. 74) and assembled using Trinity v.2.8.5 (ref. 75). The assembled transcripts from RNA-seq and Iso-seq were included as transcript evidence. Moreover, all the annotated protein sequences of *Arabidopsis thaliana*[76], *F. vesca*[77], *Rosa chinensis*[78] and *Malus domestica*[79] were downloaded from the TAIR (https://www.arabidopsis.org/) and GDR (https://www.rosaceae.org/) database, and used for homology-based gene model prediction. The final gene models in each genome were annotated by integrating RNA-seq-based and protein-homology-based evidence and ab initio prediction using MAKER v.2.31.10 (ref. 80). The detailed workflow of genome annotation is shown in Supplementary Fig. 40b. For *F. nipponica*, RNA-seq data from mixed tissue samples were generated, and gene models were annotated with identical processing steps except without Iso-seq data. The Annotation Edit Distance scores were measured for each predicted gene model, which measured the consistency of gene models with evidence alignment[77]. Furthermore, gene families were identified using HMMER v.3.3.2 (ref. 81) with the parameter $E = 1 \times 10^{-5}$ on the basis of the Pfam domains. Fisher's exact test was performed to test the expansion and contraction of the NB-ARC gene family.

Genome-wide synteny analysis and structural variations

Genome-wide synteny analysis was performed using the JCVI v.1.0.10 software package (https://github.com/tanghaibao/jcvi) with the default parameters. To identify structural variations, syntenic blocks were identified using NUCMER v.4.0.0rc1 (ref. 82) and further filtered using the delta-filter program with the parameters $i = 80$ and $l = 16{,}000$. Genetic variation analysis between the assemblies was based on genome-wide alignments. SNVs were identified using the MUMMER software package[82]. Large structural variations (>50 bp) were identified using SYRI[83] with the default parameters.

Genetic classification of sampled *F. chiloensis* and *F. virginiana* individuals

To re-examine the genotypes of the sequencing materials and primarily assess genetic variation among strawberry octoploids (Supplementary Fig. 41), we randomly selected sampling data from ten individuals for each of the three species (*F. chiloensis*, *F. virginiana* and *F.* × *ananassa*) from a previous study[33] and added Illumina data from 'Camarosa'[26], *F. chiloensis* (this study) and *F. virginiana* (this study) for population genetic analysis. The Illumina reads were filtered by fastp v.0.20.1 (ref. 74) and mapped to *F. chiloensis* (hap1) assembly using BWA v.0.7.14-r1188 (ref. 54) with the default parameters. GATK v.4.1.3.0 (ref. 84), VCFtools v.0.1.16 (ref. 85) and PLINK v.1.9 (ref. 86) were used to call and filter SNPs. The high-quality SNPs identified on 33 octoploid individuals were used to analyse PCs with PLINK v.1.9 (ref. 86) and construct a phylogenetic tree on the basis of the maximum likelihood method with FastTree v.2.1.11 (ref. 87).

Subgenome assignment

Two different methods, based on read mapping and differential *k*-mers, were used to assign the homoeologous chromosomes to each of the four subgenomes. For read-mapping-based analyses, RNA-seq data for *F. vesca* (SRR13657681 and SRR13657682) were downloaded from GenBank. These data together with the sequenced *F. nipponica* and *F. iinumae* RNA-seq data (Supplementary Table 16) were mapped to the octoploid *F. chiloensis* and *F. virginiana* genomes using sppIDer[67]. The 28 pseudochromosomes ($x = 7$) were categorized into seven distinct chromosome groups, each comprising four homoeologous chromosomes, based on the mapping rates of RNA-seq data from the different diploid species.

To further differentiate the chromosomes of the octoploids (the *F. chiloensis*, *F. virginiana* and 'Camarosa' assemblies) into four subgenomes, we identified the differential *k*-mers ($k = 13$ and frequency ≥ 50) among homoeologous chromosome sets in each assembly using SubPhaser[88]. Heterogenous chromosomes embedded with those enriched *k*-mers were clustered by a *k*-means algorithm, and the confidence level was estimated by the bootstrap method (Supplementary Table 11). To verify the efficiency of this approach, we repeated the analyses with different lengths of *k*-mer ($k = 13, 15, 17, 19, 21$ and 25) and different frequencies (50, 100, 150 and 200). This reassessment produced fully consistent assignments in 115 of 120 analyses; the five inconsistent assignments occurred with the highest *k*-mer length (19, 21 and 25) and frequency (200) cut-offs for *F. chiloensis* (hap2) and 'Camarosa', which may have resulted from having too few specific *k*-mers (Supplementary Fig. 42). Moreover, to compare the subgenome assignments in this study and those by Hardigan et al.[33] and Edger et al.[26], the genomes of *F. chiloensis* and *F. virginiana* were mapped to the 'Camarosa' genome using the MUMMER software package[82], RaGOO[53] and Chorder[83]. Repetitive *k*-mers were re-analysed to obtain sequence signatures among homoeologous chromosomes.

Phylogenetic analyses

OrthoFinder[89] was used to identify orthogroups using the annotated coding sequences from *F. nipponica*, *F. vesca*[30], *F. iinumae*[37], *F. viridis*[28], *F. mandschurica*[17], *F. nilgerrensis*[38], *F. nubicola*[28], *F. daltoniana*[17] and *F. pentaphylla*[17] and each subgenome of the octoploids *F. chiloensis*, *F. virginiana* and 'Camarosa', with *Potentilla micrantha*[90] as an outgroup. In total, 941 single-copy orthogroups were identified, and genes in each orthogroup were aligned using protein sequences with MUSCLE v.3.8.1551 (ref. 91) and converted into their corresponding codon alignments using PAL2NAL[92]. The alignments were trimmed with Gblocks v.0.91b[93] with the parameters b4 = 5, b5 = h and *t* = c to remove poorly aligned regions. IQ-TREE v.2.0.3 (ref. 94) was applied to construct the maximum likelihood tree for supergenes and individual genes, respectively, with the best-fit substitution model automatically selected. The Coalescent tree was constructed using ASTRAL[95], on the basis of all the individual maximum likelihood trees.

To control the effect of HEs on phylogenetic inference, we first identified the putative HE regions following the approach used in the allotetraploid *Micanthus*[36]. Briefly, we identified subgenome-specific *k*-mers ($k = 13$) using SubPhaser[88] and analysed significantly enriched *k*-mers by 100 kb windows. Specific regions that showed the distribution of *k*-mers differing from the global sequence signatures of that chromosome were considered to be putative HE regions (Supplementary Table 14 and Supplementary Fig. 38). We further verified those candidate HE regions using the method described by Edger et al.[26]. Briefly, we aligned the *F. vesca* genome to the *F. virginiana* (hap1) assembly with NUCMER v.4.0.0rc1 (ref. 82) with the parameters maxmatch, $l = 80$ and $c = 200$. The coverage was calculated and split into 100 kb bins. We then compared the read coverage of HE regions with non-HE regions using a *t*-test (Supplementary Fig. 37). Then, we used OrthoFinder[89] to identify single-copy orthogroups from the four closest diploid species (*F. nipponica*, *F. vesca*[30], *F. iinumae*[37] and *F. viridis*[28]), subgenomes of *F. virginiana*, and *Potentilla micrantha*[90] (as the outgroup) and filter those orthogroups located in the identified HE regions (Supplementary Table 14). A total of 6,223 individual genes were used for phylogenetic analyses as described above.

k-mer-based analysis of genetic distance

After filtering the HE regions, we used the Assembly and Alignment-Free method[96] to calculate the genetic distances between the subgenomes of the octoploids and diploid species with 21 *k*-mers.

Analysis of genetic contributions in *Fragaria* octoploid genomes

Diploid genomes to the octoploids. We downloaded resequencing data from 22 *F. chiloensis*, 20 *F. virginiana* and 42 *F.* × *ananassa* plants[33] from GenBank (PRJNA578384), cleaned them using fastp v.0.20.1 (ref. 74) and mapped them to a composite reference of eight diploid strawberry genomes using sppIDer[67]. Organelle genomes were used to enrich potential organelle short reads to avoid errors (Supplementary Fig. 43). The mapping rates against each of the eight genomes were used as indicators of the genetic contributions of diploid species to the octoploids, as described previously[28]. The Illumina short reads of *F. chiloensis*, *F. virginiana* and 'Camarosa' were mapped to composite data of the eight diploid strawberry genomes using the same process. Among them, the 'Camarosa' data were downloaded from GenBank (SRR8358384 and SRR8358387).

Wild octoploid progenitors of the cultivated strawberry. To further determine the genetic contributions of the two wild octoploid strawberries to each chromosome of the cultivated strawberry, we divided the cultivated strawberry genome into windows of different sizes (50 kb) and identified their best matches in the *F. chiloensis* and *F. virginiana* genomes using minimap2 v.2.17-r941 (ref. 68).

Transcriptome sequencing and gene expression analysis

RNA-seq experiment setup. To get insights into the divergence of gene expression in wild and cultivated octoploid fruits, seedlings of *F. chiloensis*, *F. virginiana* and *F.* × *ananassa* cv. 'Camarosa' were grown in a growth chamber under a 12:12 h light–dark cycle with temperatures of 25 °C (light) and 18 °C (night) before blossoming. Red-staged strawberry fruits were sampled. Total RNA was extracted from the fruits for RNA-seq. Raw RNA-seq data were cleaned using fastp v.0.20.1 (ref. 74) and aligned to their corresponding haplotype 1 assemblies using HISAT2 (ref. 97) with the default parameters. Samtools view with the $f = 2$ parameter was used to filter discordant mapping. Gene expression levels (fragments per kilobase per million mapped reads (FPKM)) were measured using StringTie v.2.1.7 (ref. 98) with the default parameters (with multiple mapping corrections allowed) to minimize the potential effect of sequence similarity among homoeologues. For each gene, the expression levels of all biological replicates were averaged for homoeologue comparisons.

Homoeologous expression. Strawberry syntenic homoeologues were identified in the octoploid strawberry genomes using OrthoFinder v.2.5.1 (ref. 89). Homoeologues that had a 1:1:1:1 correspondence across the four subgenomes were referred to as quadruplets, and only quadruplets with a summed expression value of all four homoeologues >0.5 FPKM were retained for downstream analyses. We further calculated relative expression levels by normalizing the expression levels of each homoeologue within a quadruplet using the following formula: relative expression level equals the FPKM of each homoeologue divided by the summed FPKM values of its corresponding quadruplets. This was modified from the method of studying homoeologue expression patterns in hexaploid wheat[99]. In this way, the relative abundance of homoeologue expression is comparable within quadruplets as well as across strawberry species (Supplementary Fig. 31c).

HEB categories. To minimize the effects of mapping bias to multiple reference genomes and reduce the complexity of multiple statistical tests, we calculated the Euclidean distance from the observed normalized expression of each quadruplet to each of the nine ideal categories to qualitatively define the HEB categories (Supplementary Figs. 30a and 31a). We assigned the HEB category for each quadruplet according to the shortest distance as described by Ramírez-González et al.[99] (shown in Supplementary Fig. 31a).

Divergence of HEB in wild and cultivated strawberries. We first used OrthoFinder v.2.5.1 (ref. 89) to find the correspondence of quadruplets in the *F. chiloensis*, *F. virginiana* and 'Camarosa' genome assemblies. Then, to study patterns of HEB in wild and cultivated strawberries, we calculated the Euclidean distance between quadruplet pairs of *F. chiloensis* and *F. virginiana* and defined those quadruplet pairs as stable (distance ≤ 0.2), middle (0.2 < distance < 0.4) or dynamic (distance ≥ 0.4). We identified each RBP (Supplementary Fig. 30b) to represent each stable quadruplet pair between two wild species and calculated the relative distance between each RBP and 'Camarosa'. The classification of homologous quadruplet pairs was done as described above.

Functional enrichments of HEB-related genes. We analysed the functions HEB-related genes on the basis of Gene Ontology annotation and enrichment, KEGG annotation and gene description using EggNog[100]. These HEB-related genes were queried using related studies to determine the gene function manually.

Statistical analysis

The significance of the differences between groups was determined by Student's *t*-test using SPSS v.19.0 (IBM Corp., Armonk, NY, USA) and Fisher's exact test using the R package stats v.4.0.3. The significance of the differences with more than two groups was determined by one-way ANOVA with Duncan's multiple range test using SPSS v.19.0. Differences were considered to be significant at $P < 0.05$.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the raw genome sequencing data have been submitted to the National Genomics Data Center (https://ngdc.cncb.ac.cn/), and the accession number is CRA005392. All the genome assemblies reported in this paper have been deposited in the Genome Warehouse of the National Genomics Data Center (https://ngdc.cncb.ac.cn/gwh), and the accession numbers are GWHDEDQ00000000 (*F. chiloensis*), GWHDEDR00000000 (*F. virginiana*) and GWHDEDN00000000 (*F. nipponica*). All the genome assembly and annotation files are also available in the Genome Database for Rosaceae (GDR) (https://www.rosaceae.org/Analysis/16216791,16216792,16216793).

Code availability

The scripts used for HEB category analysis for each quadruplet in this paper are available on GitHub (https://github.com/jinxin112233/HEB_categories). Bash commands for studying wild strawberry genomes have been uploaded on GitHub (https://github.com/jinxin112233/WSG).

References

1. Soltis, P. S. & Soltis, D. E. *Polyploidy and Genome Evolution* (Springer, 2012).
2. Chen, J. Z. & Birchler, J. A. *Polyploid and Hybrid Genomics* (Wiley-Blackwell, 2013).
3. Ye, C. Y. et al. The genomes of the allohexaploid *Echinochloa crus-galli* and its progenitors provide insights into polyploidization-driven adaptation. *Mol. Plant* **13**, 1298–1310 (2020).
4. Osborn, T. C. et al. Understanding mechanisms of novel gene expression in polyploids. *Trends Genet.* **19**, 141–147 (2003).
5. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).
6. Michael, T. P. & VanBuren, R. Building near-complete plant genomes. *Curr. Opin. Plant Biol.* **54**, 26–33 (2020).

7. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
8. Campoy, J. A. et al. Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol.* **21**, 306 (2020).
9. Wenger, A. M. et al. Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
10. Hon, T. et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci. Data* **7**, 399 (2020).
11. Mascher, M. et al. Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* **33**, 1888–1906 (2021).
12. Zhou, Q. et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **52**, 1018–1023 (2020).
13. Sun, X. et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* **52**, 1423–1432 (2020).
14. Chen, H. et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* **11**, 2494 (2020).
15. Folta, K. M. & Davis, T. M. Strawberry genes and genomics. *Crit. Rev. Plant Sci.* **25**, 399–415 (2006).
16. Hummer, K. E. & Hancock, J. Strawberry genomics: botanical history, cultivation, traditional breeding, and new technologies. In *Genetics and genomics of Rosaceae* (eds Folta, K. M. & Gardiner, S. E.) 413–435 (Springer, 2009).
17. Qiao, Q. et al. Evolutionary history and pan-genome dynamics of strawberry (*Fragaria* spp.). *Proc. Natl Acad. Sci. USA* **118**, e2105431118 (2021).
18. Liston, A., Cronn, R. & Ashman, T. L. *Fragaria*: a genus with deep historical roots and ripe for evolutionary and ecological insights. *Am. J. Bot.* **101**, 1686–1699 (2014).
19. Njuguna, W., Liston, A., Cronn, R., Ashman, T. L. & Bassil, N. Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. *Mol. Phylogenet. Evol.* **66**, 17–29 (2013).
20. Whitaker, V. M. et al. A roadmap for research in octoploid strawberry. *Hortic. Res.* **7**, 33 (2020).
21. Moyano-Cañete, E. et al. FaGAST2, a strawberry ripening-related gene, acts together with FaGAST1 to determine cell size of the fruit receptacle. *Plant Cell Physiol.* **54**, 218–236 (2013).
22. Gaston, A. et al. The FveFT2 florigen/FveTFL1 antiflorigen balance is critical for the control of seasonal flowering in strawberry while FveFT3 modulates axillary meristem fate and yield. *N. Phytol.* **232**, 372–387 (2021).
23. Hirakawa, H. et al. Dissection of the octoploid strawberry genome by deep sequencing of the genomes of *Fragaria* species. *DNA Res.* **21**, 169–181 (2014).
24. Hirsch, C. N. & Buell, C. R. Tapping the promise of genomics in species with complex, nonmodel genomes. *Annu. Rev. Plant Biol.* **64**, 89–110 (2013).
25. Hardigan, M. A. et al. Genome synteny has been conserved among the octoploid progenitors of cultivated strawberry over millions of years of evolution. *Front. Plant Sci.* **10**, 1789 (2020).
26. Edger, P. P. et al. Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* **51**, 541–547 (2019).
27. Liston, A. et al. Revisiting the origin of octoploid strawberry. *Nat. Genet.* **52**, 2–4 (2020).
28. Feng, C. et al. Tracing the diploid ancestry of the cultivated octoploid strawberry. *Mol. Biol. Evol.* **38**, 478–485 (2021).
29. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
30. Edger, P. P. et al. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *GigaScience* **7**, 1–7 (2018).
31. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
32. Abou, Saada et al. nPhase: an accurate and contiguous phasing method for polyploids. *Genome Biol.* **22**, 126 (2021).
33. Hardigan, M. A. et al. Unraveling the complex hybrid ancestry and domestication history of cultivated strawberry. *Mol. Biol. Evol.* **38**, 2285–2305 (2021).
34. Tennessen, J. A., Govindarajulu, R., Ashman, T. L. & Liston, A. Evolutionary origins and dynamics of octoploid strawberry subgenomes revealed by dense targeted capture linkage maps. *Genome Biol. Evol.* **6**, 3295–3313 (2014).
35. Session, A. M. & Rokhsar, D. S. Transposon signatures of allopolyploid genome evolution. *Nat. Commun.* **14**, 3180 (2023).
36. Mitros, T. et al. Genome biology of the paleotetraploid perennial biomass crop *Miscanthus*. *Nat. Commun.* **11**, 5442 (2020).
37. Edger, P. P. et al. Reply to: Revisiting the origin of octoploid strawberry. *Nat. Genet.* **52**, 5–7 (2020).
38. Zhang, J. et al. The high-quality genome of diploid strawberry (*Fragaria nilgerrensis*) provides new insights into anthocyanin accumulation. *Plant Biotechnol. J.* **18**, 1908–1924 (2020).
39. Wei, N., Tennessen, J. A., Liston, A. & Ashman, T. L. Present-day sympatry belies the evolutionary origin of a high-order polyploid. *N. Phytol.* **216**, 279–290 (2017).
40. Zhang, X., Wu, R., Wang, Y., Yu, J. & Tang, H. Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.* **18**, 66–72 (2019).
41. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B. & Hirsch, C. N. How the pan-genome is changing crop genomics and improvement. *Genome Biol.* **22**, 3 (2021).
42. Hancock, J. F. & Bringhurst, R. S. Evolution of California populations of diploid and octoploid *Fragaria* (Rosaceae): a comparison. *Am. J. Bot.* **68**, 1–5 (1981).
43. Harrison, R. E., Luby, J. J., Furnier, G. R. & Hancock, J. F. Morphological and molecular variation among populations of octoploid *Fragaria virginiana* and *F. chiloensis* (*Rosaceae*) from North America. *Am. J. Bot.* **84**, 612–620 (1997).
44. Qu, M. et al. Karyotypic stability of *Fragaria* (strawberry) species revealed by cross-species chromosome painting. *Chromosome Res.* **29**, 285–300 (2021).
45. Chen, Z. J. et al. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533 (2020).
46. Hancock, J. F. et al. Reconstruction of the strawberry, *Fragaria×ananassa*, using genotypes of *F. virginiana* and *F. chiloensis*. *HortScience* **45**, 1006–1013 (2010).
47. Nakashima, K. & Yamaguchi-Shinozaki, K. ABA signaling in stress-response and seed development. *Plant Cell Rep.* **32**, 959–970 (2013).
48. Li, J. et al. Research advances of MYB transcription factors in plant stress resistance and breeding. *Plant Signal. Behav.* **14**, 1613131 (2019).
49. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
50. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
51. Ranallo-Benavidez, T. R. et al. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

52. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
53. Alonge, M. et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
55. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
56. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
57. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
58. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
59. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
60. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18 (2017).
61. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
62. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
63. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
64. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
65. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
66. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
67. Langdon, Q. K., Peris, D., Kyle, B. & Hittinger, C. T. sppIDer: a species identification tool to investigate hybrid genomes with high-throughput sequencing. *Mol. Biol. Evol.* **35**, 2835–2849 (2018).
68. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
69. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
70. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
71. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
72. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **25**, 4.10.1–4.10.14 (2009).
73. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. *Methods Mol. Biol.* **1962**, 65–95 (2019).
74. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
75. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
76. Cheng, C. Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
77. Li, Y., Pi, M., Gao, Q., Liu, Z. & Kang, C. Updated annotation of the wild strawberry *Fragaria vesca* V4 genome. *Hort. Res.* **6**, 61 (2019).
78. Raymond, O. et al. The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
79. Zhang, L. et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat. Commun.* **10**, 1494 (2019).
80. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **12**, 491 (2011).
81. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
82. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
83. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
84. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing nextgeneration DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
85. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
86. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
87. Price, M. N. et al. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
88. Jia, K. H. et al. SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific *k*-mers. *N. Phytol.* **235**, 801–809 (2022).
89. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
90. Buti, M. et al. The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry). *GigaScience* **7**, 1–14 (2018).
91. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
92. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
93. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
94. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
95. Mirarab, S. et al. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).
96. Fan, H., Ives, A. R., Surget-Groba, Y. & Cannon, C. H. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* **16**, 522 (2015).
97. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

98. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
99. Ramírez-González, R. H. et al. The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).
100. Cantalapiedra, C. P. et al. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

Acknowledgements

This study was financially supported by the National Key Research and Development Program of China (grant no. 2018YFD1000107), CAS Pioneer Hundred Talents, and the open research project of the 'Cross-Cooperative Team' of the Germplasm Bank of Wild Species to A.Z., by the University of Nebraska–Lincoln to J.P.M, and by the National Science Foundation of China (grant no. 31860534) to J.R.

Author contributions

A.Z. and J.R. conceived the project. A.Z., J.P.M. and J.R. designed the research. J.R., H.W. and H.D. collected and cared for the plant materials. H.D., C.Z. and F.L. sampled the plant tissues for genome and transcriptome sequencing. X.J., A.Z. and H.D. performed the computational analyses. X.J., J.P.M. and A.Z. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

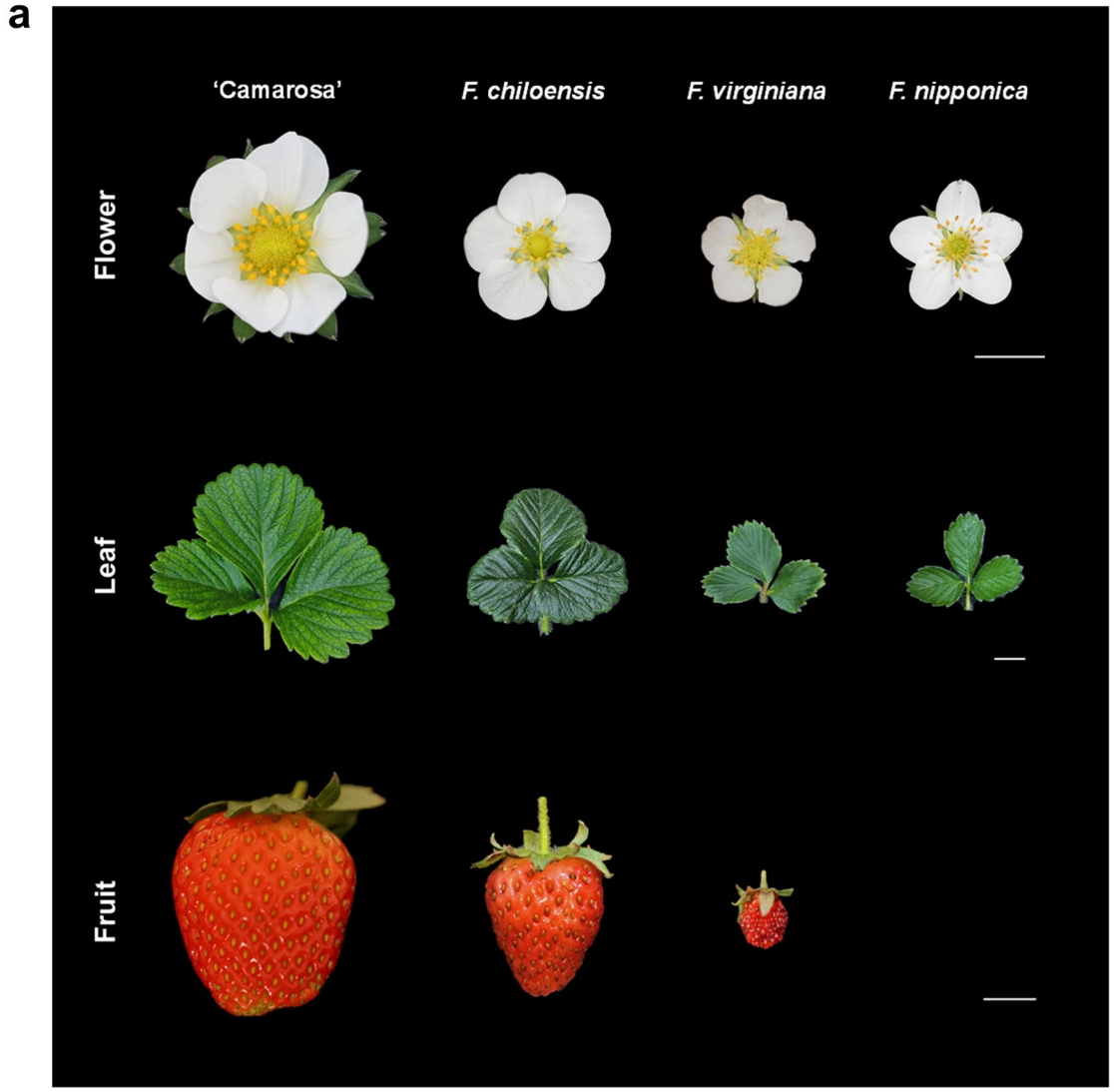
Extended data is available for this paper at https://doi.org/10.1038/s41477-023-01473-2.

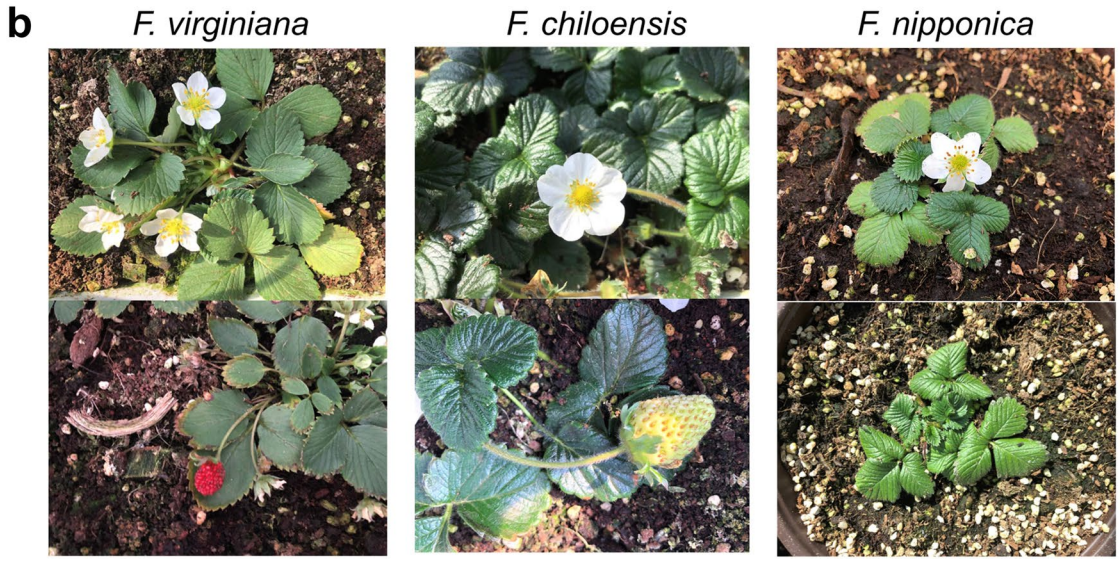
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-023-01473-2.

Correspondence and requests for materials should be addressed to Jiwei Ruan, Jeffrey P. Mower or Andan Zhu.

Peer review information *Nature Plants* thanks Andrew H. Paterson, John Lovell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

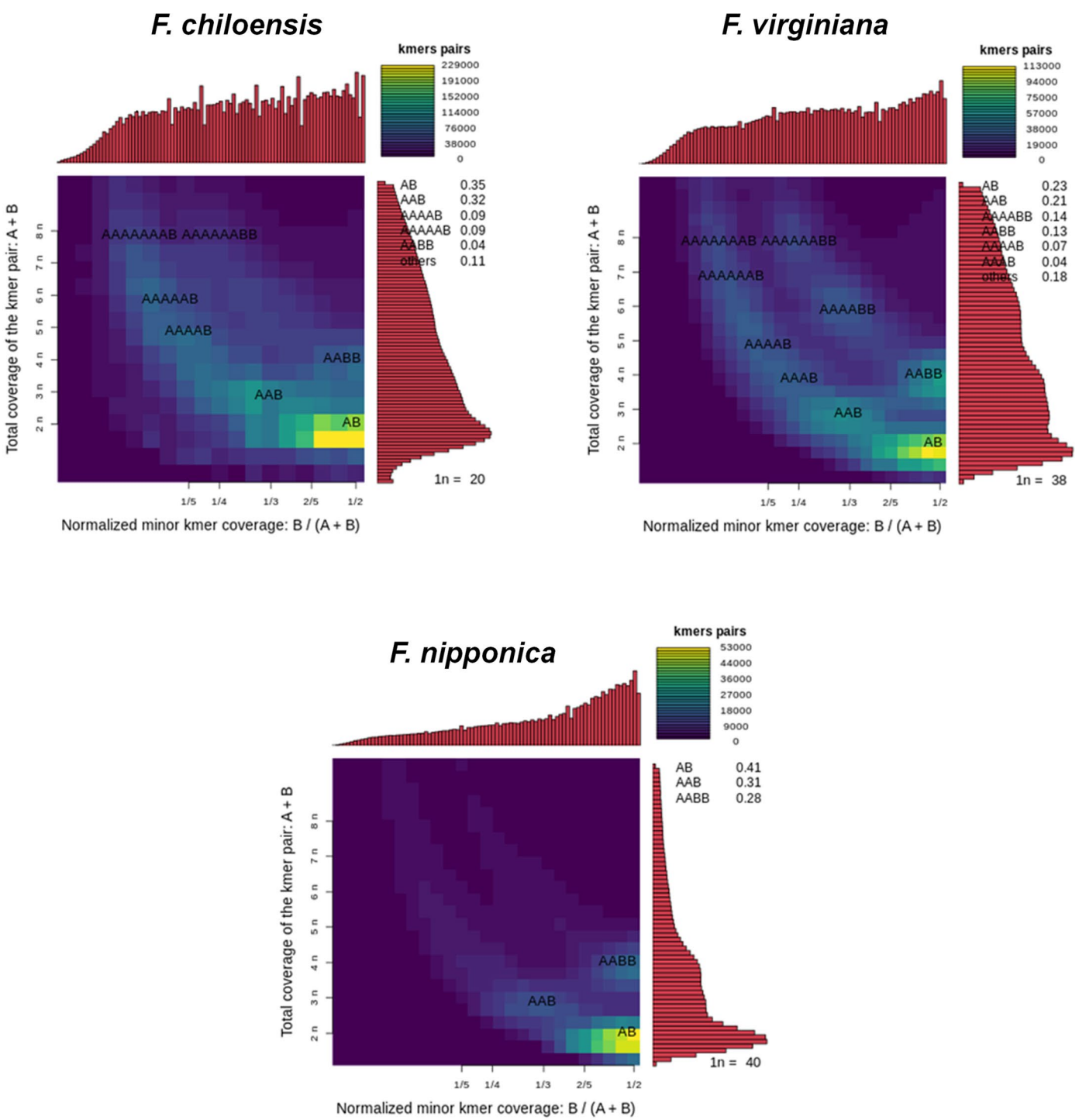
https://doi.org/10.1038/s41477-023-01473-2





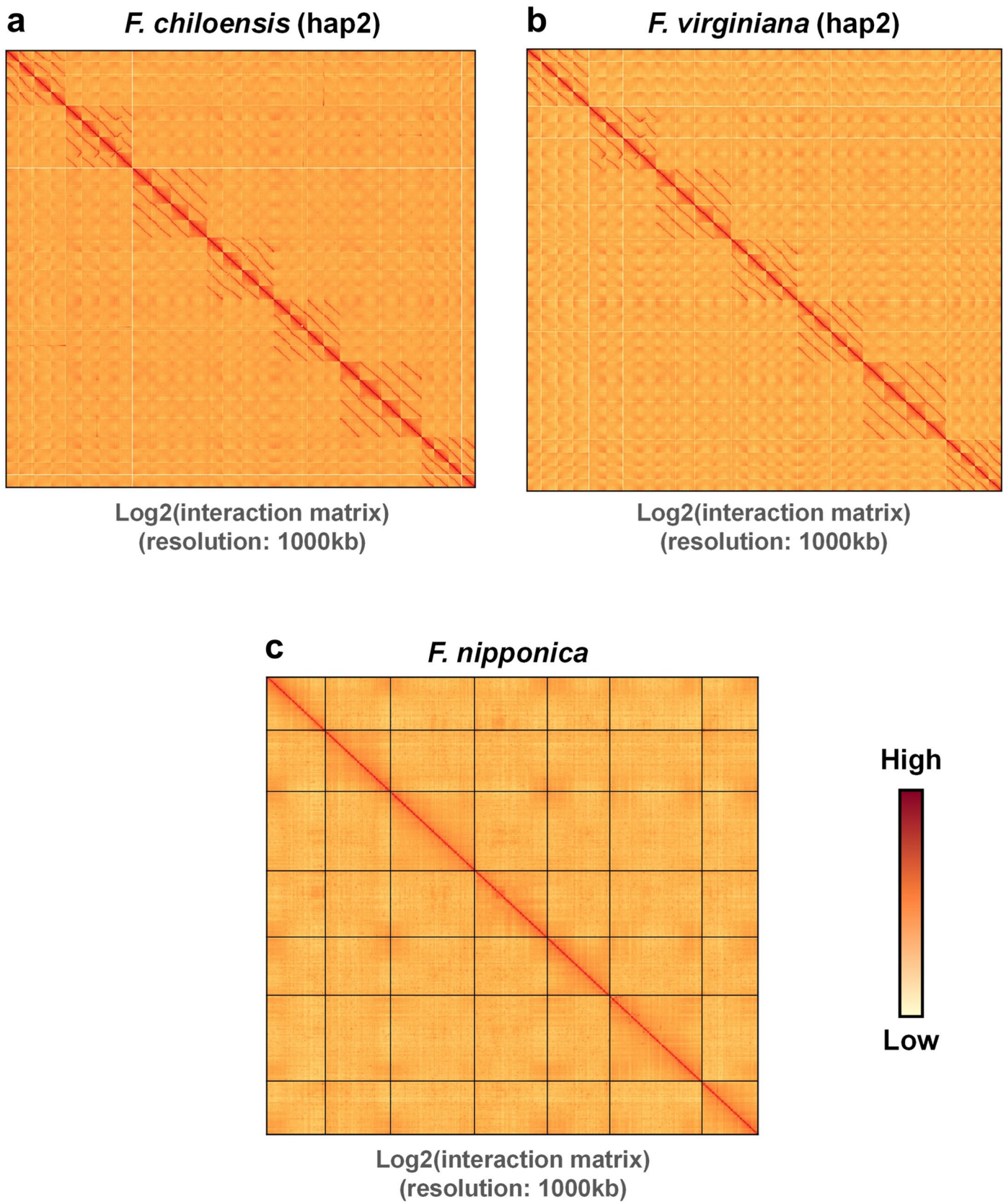
Extended Data Fig. 1 | Morphological identification of the sequenced *Fragaria* species. (**a**) Morphological features of *F. chiloensis*, *F. virginiana*, *F. nipponica*, and the cultivar 'Camarosa'. Scale bar, 1 cm. (**b**) Seedings of *F. chiloensis*, *F. virginiana*, *F. nipponica* in the greenhouse.

https://doi.org/10.1038/s41477-023-01473-2



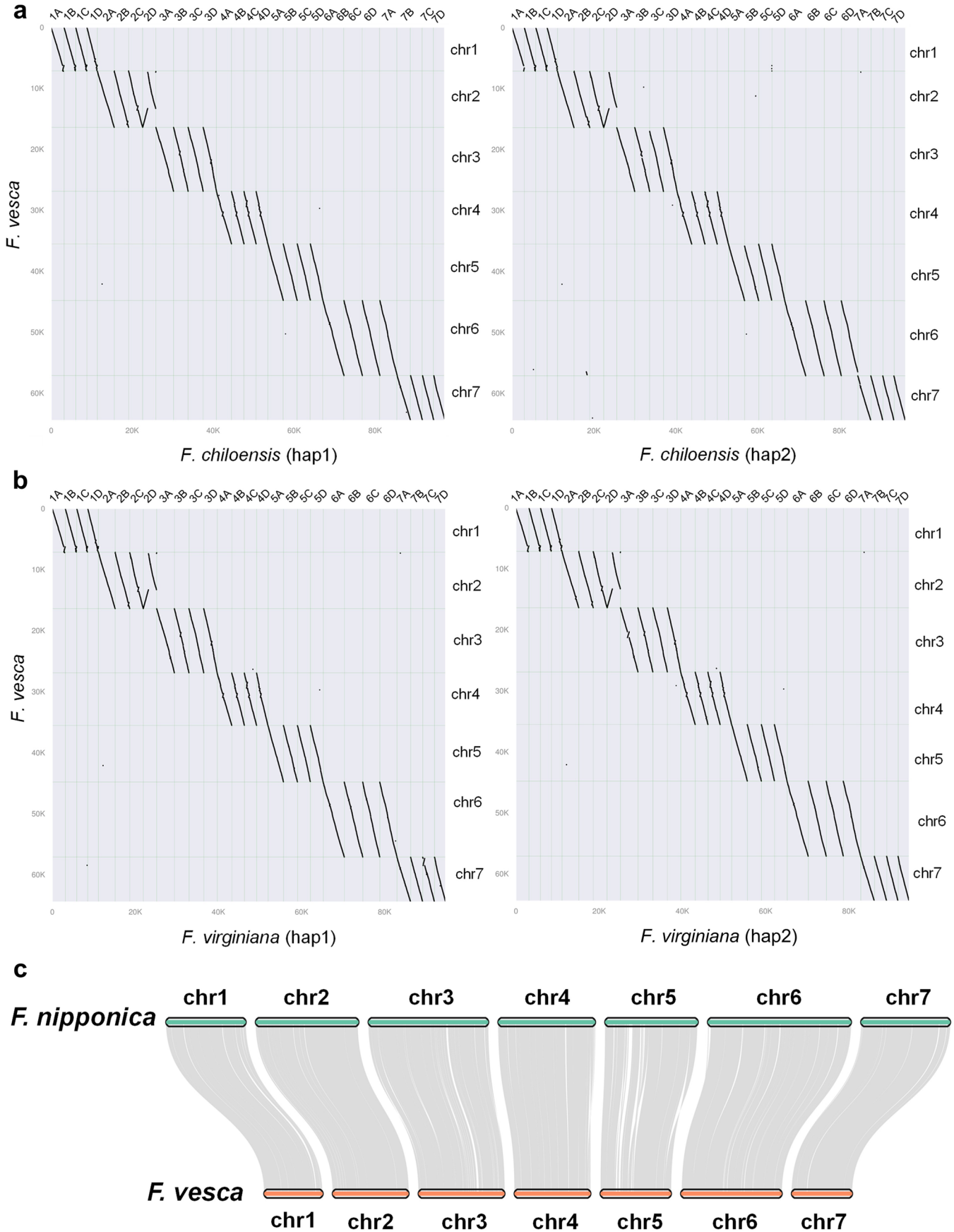
Extended Data Fig. 2 | Ploidy level estimation. Smudge plots showing the ploidy level estimation for the sequenced *F. chiloensis*, *F. virginiana* and *F. nipponica* plants.

https://doi.org/10.1038/s41477-023-01473-2



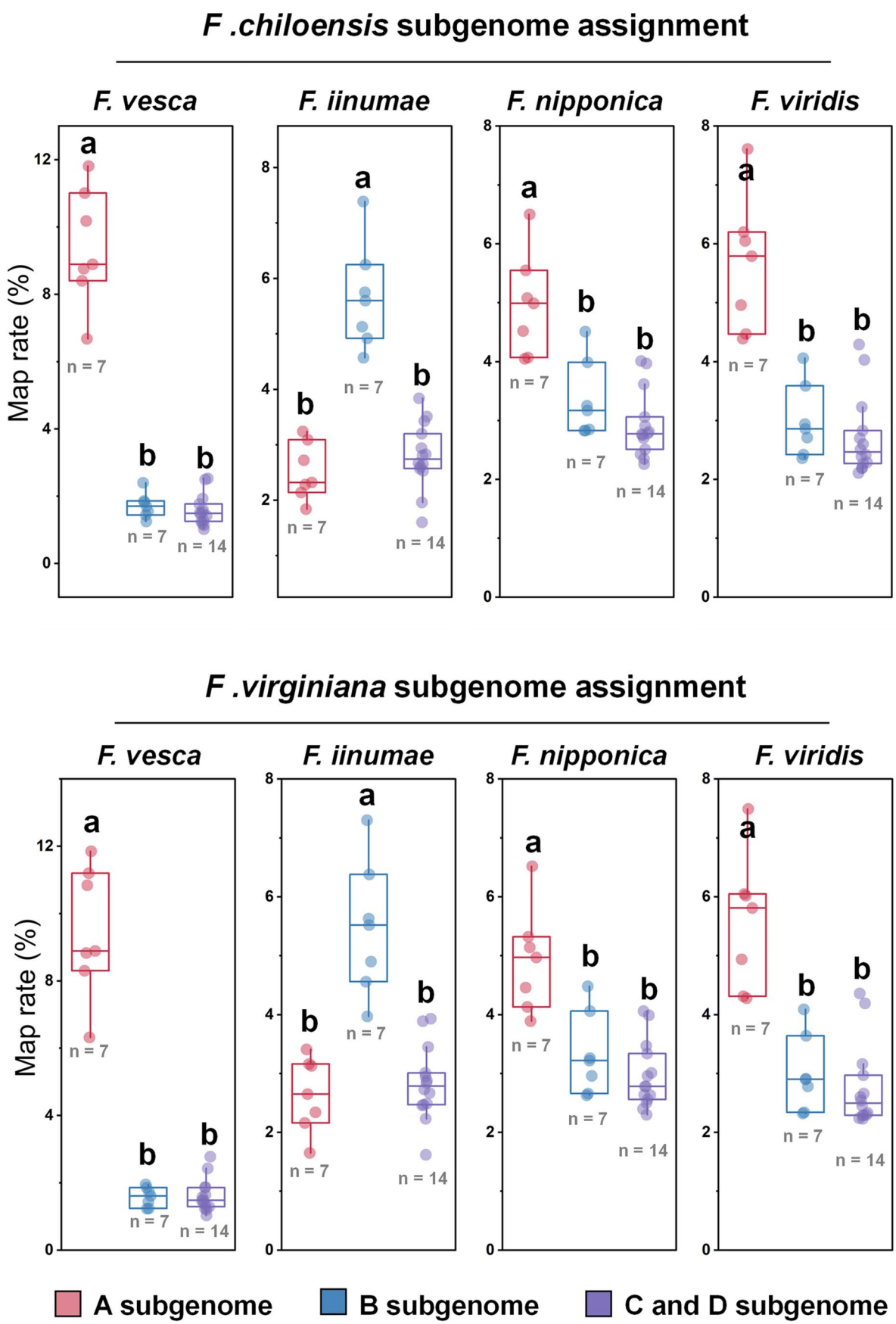
Extended Data Fig. 3 | Hi-C interaction maps of the haplotype2 of the octoploid *F. chiloensis*. (**a**), *F. virginiana* (**b**) and diploid *F. nipponica* (**c**). Note: Low to high densities of interaction signals were scaled with colours from orange to deep red.

https://doi.org/10.1038/s41477-023-01473-2



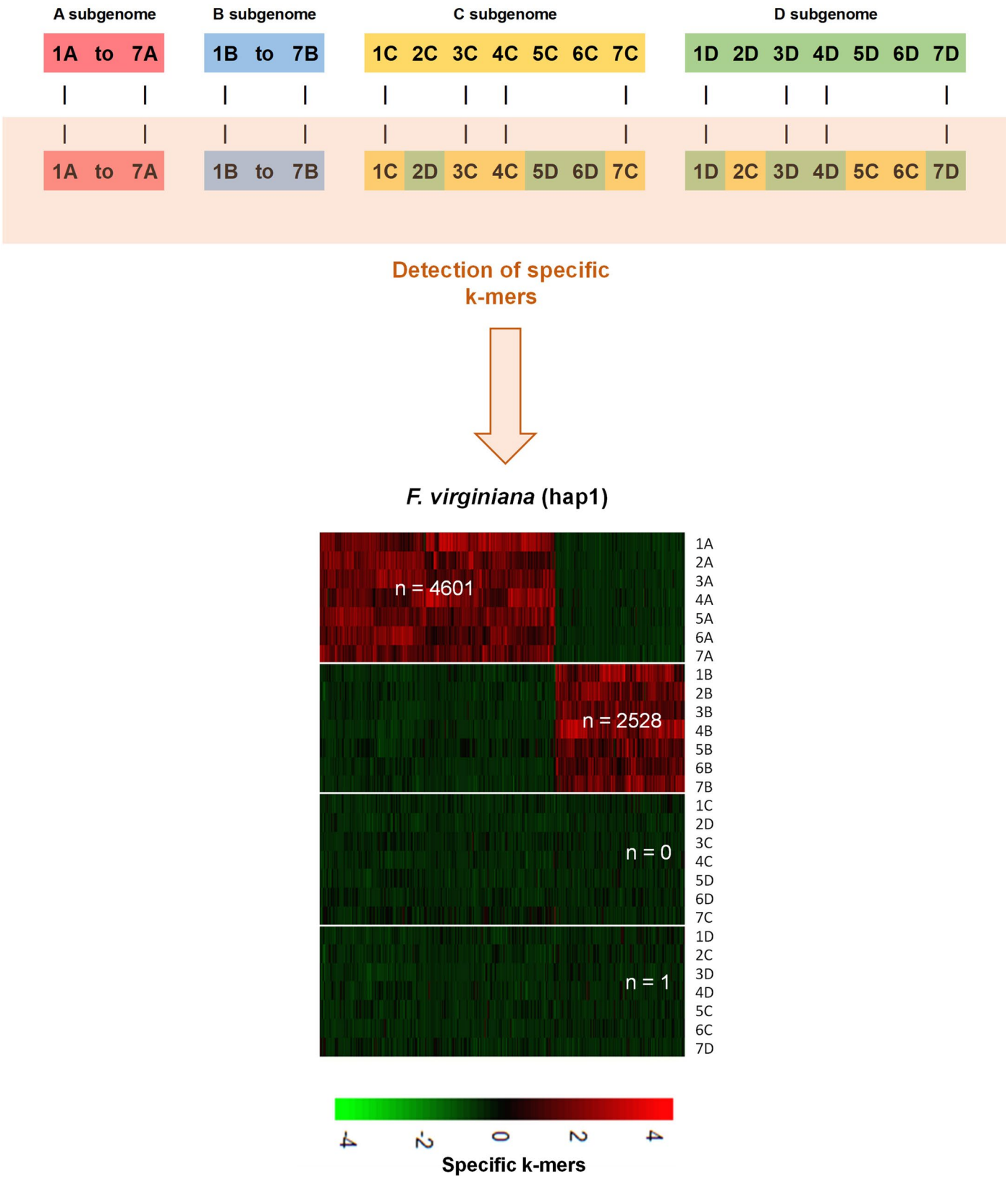
Extended Data Fig. 4 | Graphical alignment of *F. vesca* genome with *F. chiloensis* genome, the *F. virginiana* genome, and the *F. nipponica* genome. Macrosynteny between the *F. vesca* genome and the *F. chiloensis* (**a**) and the *F. virginiana* genome (**b**). Syntenic gene pairs are denoted by black points. (**c**) Macrosynteny between the *F. nipponica* genome and the *F. vesca* genome. Syntenic gene pairs are denoted by gray line.

https://doi.org/10.1038/s41477-023-01473-2



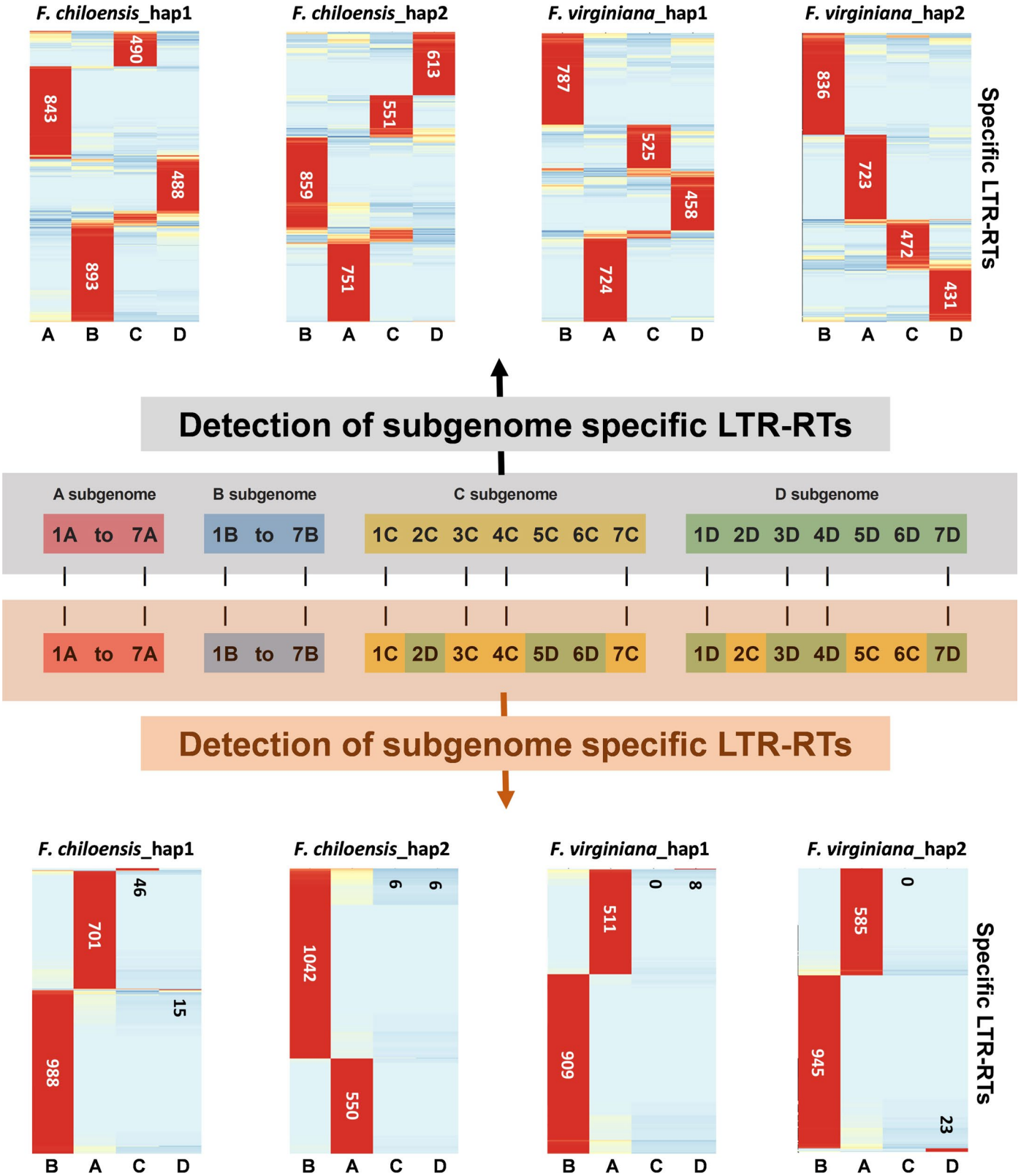
Extended Data Fig. 5 | Mapping-based subgenome assignments of the *F. chiloensis* and *F. virginiana* chromosomes. Note: The top and bottom line of box plot represent 25th and 75th percentiles, the centre line is the median and whiskers are the full data range. Different lowercase letters indicate the significance of differences in mapping rates among subgenomes, using one-way ANOVA with Duncan's multiple range test (df = 27; $P < 0.05$).

https://doi.org/10.1038/s41477-023-01473-2



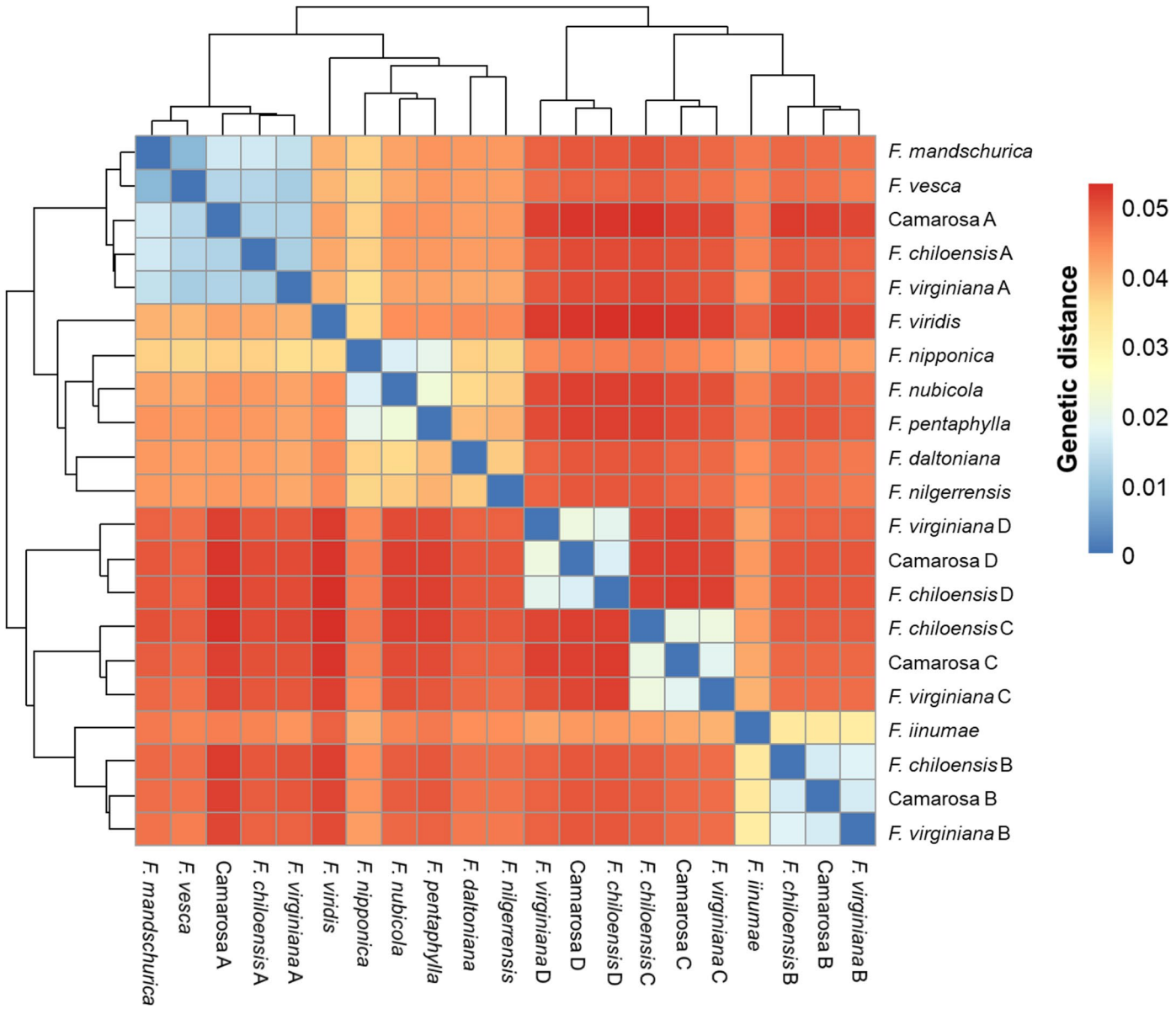
Extended Data Fig. 6 | Identification of specific subgenome k-mers (K = 13 and frequency = 50) *F. virginiana* (hap1) based on the subgenomic assignment originally proposed in the 'Camarosa' genome by Edger et al., (2019) and Hardigan et al., (2021). The difference of chromosome assignments (2 C, 2D, 5 C, 5D, 6 C, 6D) are shown. Note: n = number of specific k-mer on each subgenome.

https://doi.org/10.1038/s41477-023-01473-2



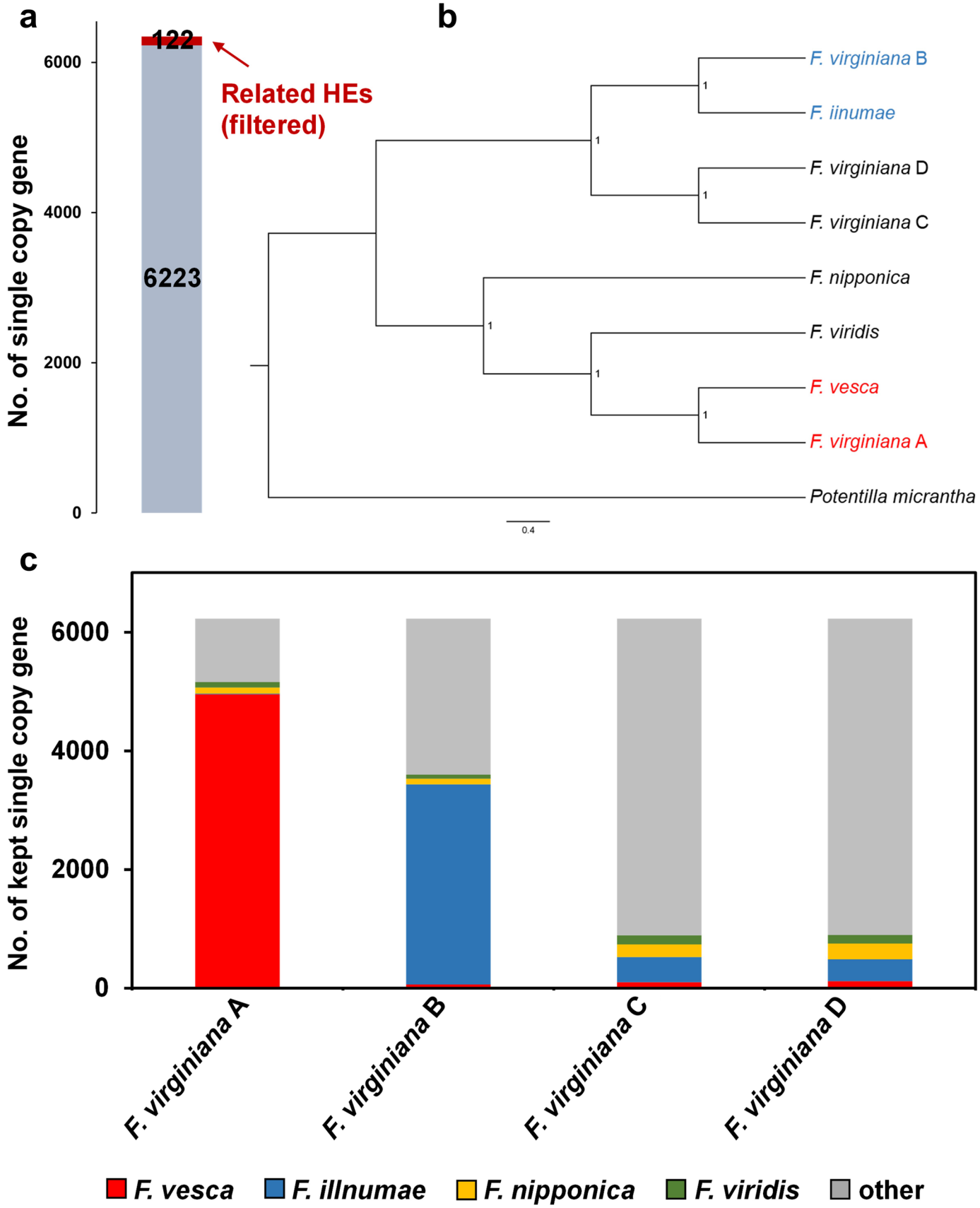
Extended Data Fig. 7 | Identification of *F. chiloensis*(hap1), *F. chiloensis*(hap2), *F. virginiana*(hap1) and *F. virginiana*(hap2) subgenome specific LTR-RTs based on new subgenome assignment and subgenome assignment by Edger et al., (2019) and Hardigan et al., (2021). Note: n = number of specific LTR-RTs on each subgenome.

https://doi.org/10.1038/s41477-023-01473-2



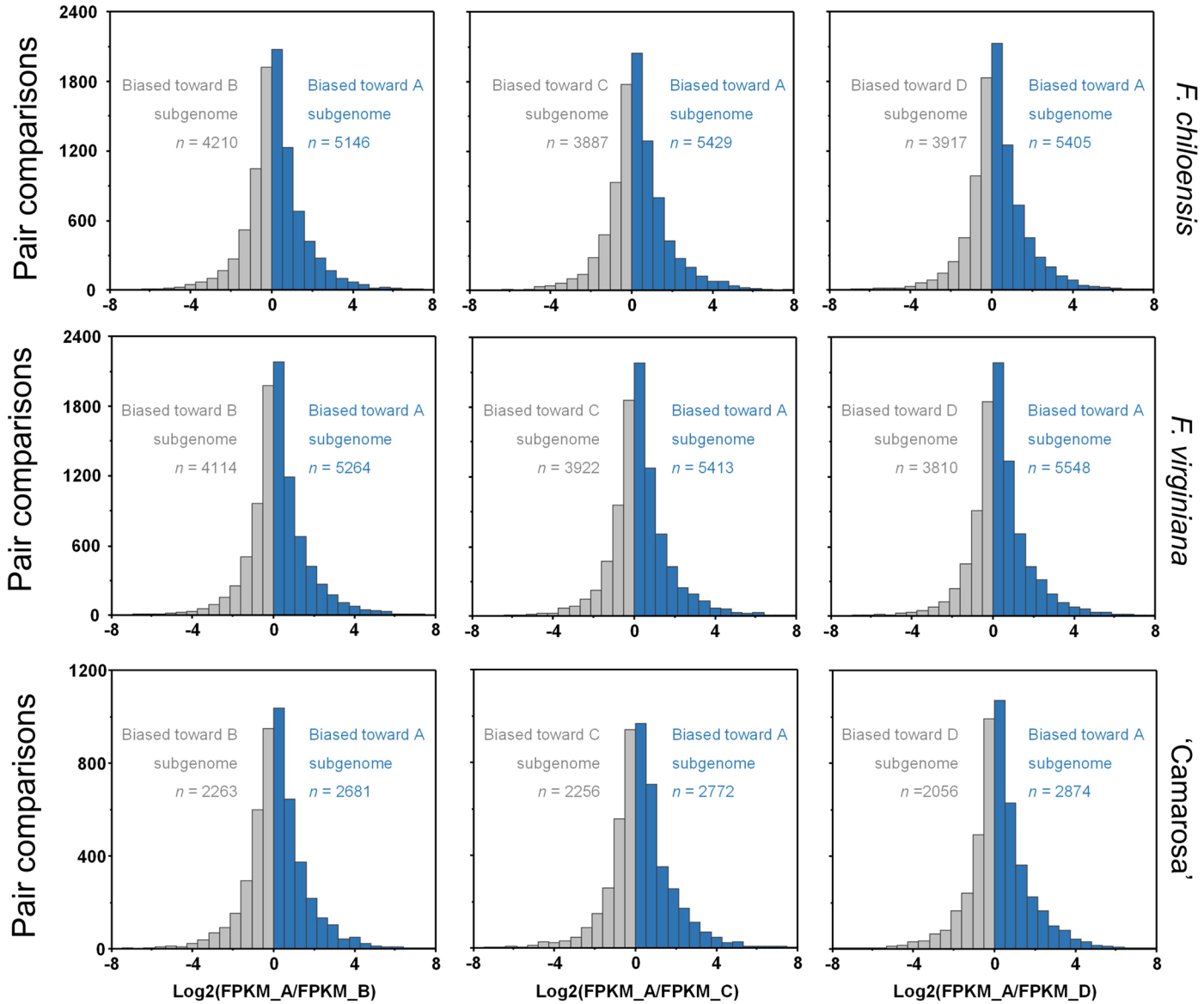
Extended Data Fig. 8 | Genetic distance matrix between diploid species and each subgenome based on 21 k-mer calculation. (homoeologous exchange regions were filtered).

https://doi.org/10.1038/s41477-023-01473-2



Extended Data Fig. 9 | Phylogenomic analysis of the octoploid subgenomes. (**a**) Total of 6345 single copy gene were identified and 122 single copy gene located in homoeologous exchange regions (HEs) were filtered. (**b**) Coalescent-based analysis of 6223 genes from four diploid species and each subgenome of the *F. virginiana genome*. (**c**) Summary of phylogenetic positions of the four octoploid subgenomes. Different colour indicates the number of kept homologous gene clade with diploid species.

https://doi.org/10.1038/s41477-023-01473-2



Extended Data Fig. 10 | Expression dominance. The distribution of HEB between all gene pairs in the red fruits of *F. chiloensis*, *F. virginiana* and 'Camarosa'.

nature portfolio

Corresponding author(s): Andan Zhu & Jeffrey P. Mower & Jiwei Ruan

Last updated by author(s): Feb 17, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection. All sequencing platforms have been described.
Data analysis	All open source softwares or pipelines (versions) used in this study were cited in the Method. These include jellyfish v.2.2.10, GenomeScope v1, Hifiasm v.0.15.1, RaGOO, ALLHiC, Canu v.1.9, Racon v.1.4.3, LACHESIS, Juicebox, NOVOPlasty v.2.7.2, LTR_FINDER v.1.1, LTRharvest v.1.1, minimap2 v2.17-r941, purge_dups v1.2.3, BUSCO v.4.1.2, RepeatModeler v.2.0.1, RepeatMasker v.4.1.0, BRAKER v.2.1.5, fastp v.0.20.1, Trinity v.2.8.5, HMMER v.3.3.2, JCVI v.1.0.10, NUCMER v.4.0.0rc1, SYRI, BWA v.0.7.14-r1188, GATK v.4.1.3.0, VCFtools v.0.1.16, PLINK v.1.9, FastTree v.2.1.11, Chorder, MUSCLE v3.8.1551, PAL2NAL, Gblocks v0.91b, IQ-TREE v2.0.3, ASTRAL, SubPhaser, AFF, minimap2 v2.17-r941, HISAT2, StringTie v.2.1.7, Orthofinder v2.5.1, EggNog, etc. Scripts used for HEB categories analysis for each quadruplet in this paper are available on GitHub (https://github.com/jinxin112233/HEB_categories). Bash commands for studying wild strawberry genomes have been uploaded on GitHub (https://github.com/jinxin112233/WSG).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data in this study have been submitted to the National Genomics Data Center (NGDC) under the accession no. CRA005392. All the genome assemblies reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center (https://ngdc.cncb.ac.cn/gwh), and the accession number is GWHDEDQ00000000 (F. chiloensis), GWHDEDR00000000 (F. virginiana) and GWHDEDN00000000 (F. nipponica).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	One seedling per Fragaria species (sample size=1) was used for genome sequencing and construction of Hi-C libraries. For the collection of fruit samples, ~30 plants per species were used.
Data exclusions	No data were excluded for the analyses.
Replication	The findings for genomic variations were measured using different sequencing data (one deep sequencing Illumina library, one HiFi library etc.) as described in the Method . For subgenome expression dominance analyses, three or two biological replicates per developmental stages were performed and all sampling attempts were sucessful.
Randomization	For gene expression analyses, experiment was in a randomized design, and tissues were randomly sampled. eedlings of F. chiloensis, F. virginiana, and F. x ananassa cv. 'Camarosa' were grown in a growth chamber under a 12:12 h light-dark cycle with a temperature of 25°C (light) and 18°C (night) before blossoming. Tissue samplings were random.
Blinding	The investigators were blinded to group allocation during data collection

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were harvested using MGB dissociation fluid. Then, suspensions of the test sample (Fragaria) and the internal reference sample (Oryza sativa L. 'Japonica') were mixed to measure the fluorescence (propidium iodide) immediately.
Instrument	The BD FACSCalibur flow cytometer
Software	Modifit 3.0
Cell population abundance	>10000 cells
Gating strategy	Forward and side scatter

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.