Origin and Evolution of the Octoploid Strawberry Genome

Patrick P. Edger

Thomas J. Poorten

... Ron D. Smith
William & Mary, rdsmith@wm.edu

Scott J. Teresi
William & Mary

See next page for additional authors

Follow this and additional works at: https://scholarworks.wm.edu/aspubs

Part of the Plant Sciences Commons

Recommended Citation
Edger, Patrick P.; Poorten, Thomas J.; ...; Smith, Ron D.; Teresi, Scott J.; Puzey, Joshua R.; and et al., Origin and Evolution of the Octoploid Strawberry Genome (2019). Nature Genetics, 51, 541-547. https://doi.org/10.1038/s41588-019-0356-4

This Article is brought to you for free and open access by the Arts and Sciences at W&M ScholarWorks. It has been accepted for inclusion in Arts & Sciences Articles by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
Origin and evolution of the octoploid strawberry genome

Patrick P. Edger1,2*, Thomas J. Poorten3, Robert VanBuren1,4, Michael A. Hardigan3, Marivi Colle1, Michael R. McKain5, Ronald D. Smith6, Scott J. Teresi6, Andrew D. L. Nelson7, Ching Man Wai1, Elizabeth I. Alger1, Kevin A. Bird1,2, Alan E. Yocca1, Nathan Pumplin3, Shujun Ou1,2*, Thomas J. Poorten3, Robert VanBuren1,4, Michael A. Hardigan3, Marivi Colle1,7, Ching Man Wai1,5, Ronald D. Smith6.

Cultivated strawberry emerged from the hybridization of two wild octoploid species, both descendants from the merger of four diploid progenitor species into a single nucleus more than 1 million years ago. Here we report a near-complete chromosome-scale assembly for cultivated octoploid strawberry (Fragaria × ananassa) and uncovered the origin and evolutionary processes that shaped this complex allopolyploid. We identified the extant relatives of each diploid progenitor species and provide support for the North American origin of octoploid strawberry. We examined the dynamics among the four subgenomes in octoploid strawberry and uncovered the presence of a single dominant subgenome with significantly greater gene content, gene expression abundance, and biased exchanges between homoeologous chromosomes, as compared with the other subgenomes. Pathway analysis showed that certain metabolomic and disease-resistance traits are largely controlled by the dominant subgenome. These findings and the reference genome should serve as a powerful platform for future evolutionary studies and enable molecular breeding in strawberry.

The cultivated garden strawberry (Fragaria × ananassa), an allo-octoploid (2n = 8x = 56), has a unique natural and domestication history, originating as an interspecific hybrid between wild octoploid progenitor species approximately 300 years before present. The genomes of the progenitor species, Fragaria virginiana and Fragaria chiloensis, are the products of polyploid evolution: they were formed by the fusion of and interactions among genomes from four diploid progenitor species (that is, subgenomes) approximately 1 million years before present. Whereas two of the diploid progenitor species have been identified, the other two diploid progenitor species have remained unknown. Moreover, the history of events leading to the formation of the octoploid lineage and the evolutionary dynamics among the four subgenomes that restabilized cellular processes after ‘genomic shock’ in allopolyploids remain poorly understood. Here, we present what is, to our knowledge, the first chromosome-scale assembly of an octoploid strawberry genome, the identities of the extant diploid progenitor species of each subgenome, and novel insights into the collective evolutionary processes involved in establishing a dominant subgenome in this highly polyploid species.

The Rosaceae are a large eudicot family including a rich diversity of crops with major economic importance worldwide, such as nuts (for example, almonds), ornamentals (for example, roses), pome fruits (for example, apples), stone fruits (for example, peaches), and berries (for example, strawberries). Strawberries are prized by consumers, largely because of their complex array of flavors and aromas. The genus Fragaria was named by the botanist Carl Linnaeus, on the basis of the Latin word ‘fragrans,’ meaning ‘sweet scented,’ describing its striking, highly aromatic fruit. A total of 22 wild species of Fragaria have been described, ranging from diploid (2n = 2x = 14) to decaploid (2n = 10x = 70). The genus Fragaria is highly interfertile between and within ploidy levels, thus leading to the natural formation of higher-polyploid species.

Polyploid events, also known as whole-genome duplications, have been an important recurrent process throughout the evolutionary history of eukaryotes and have probably contributed to novel and varied phenotypes. Polyploids are grouped into two main categories: autoploids and allopolyploids, involving either a single or multiple diploid progenitor species, respectively. Many crop species are allopolyploids, thus contributing to the emergence of important agronomic traits such as spinnable fibers in cotton, diversified morphotypes in Brassica, and varied aroma and flavor profiles in strawberry. Allopolyploids face the challenge of organizing distinct parental subgenomes—each with a unique genetic and epigenetic makeup shaped by independent evolutionary histories—residing within a single nucleus. Previous studies have proposed, as part of the ‘subgenome dominance’ hypothesis, that the establishment of a single dominant subgenome may resolve...
various (epi)genetic conflicts in allopolyploids\textsuperscript{21–23}. However, understanding of the underlying mechanisms and ultimate consequences of subgenome dominance remains largely incomplete\textsuperscript{24}.

Subgenome-level analyses in most allopolyploid systems are greatly hindered by the inability to confidently assign parental gene copies (that is, homoeologs) to each subgenome, owing to both large-scale chromosomal changes and homoeologous exchanges that shuffle and replace homoeologs among parental chromosomes\textsuperscript{26–29}. Octoploid strawberry still has a complete set of homoeologous chromosomes from all four parental subgenomes, thus greatly simplifying homoeolog assignment. Furthermore, gene sequences from extant relatives of the diploid progenitor species, which probably still exist for octoploid strawberry\textsuperscript{9}, can be used to accurately assign homoeologs to each parental subgenome\textsuperscript{30}. However, a high-quality reference genome for the octoploid is needed to fully exploit strawberry as a model system for studying allopolyploidy as well as to provide a platform for identifying biologically and agriculturally important genes and applying genomic-enabled breeding approaches\textsuperscript{36}. The assembly of the octoploid strawberry genome, with an estimated genome size of 813.4 Mb, has been particularly challenging because of its high heterozygosity and ploidy level\textsuperscript{31}. For example, the most recently published version of the octoploid strawberry genome is highly fragmented, with more than 625,000 scaffolds, and largely incomplete, with less than 660 Mb assembled after removal of the numerous gaps\textsuperscript{31}. Thus, that version of the genome, owing to its overall highly fragmented nature, has not been a useful resource for genome-wide analyses including the discovery of molecular markers for breeding.

Results

Assembly and annotation of the octoploid strawberry genome.

Our goal was to obtain a high-quality reference genome for the Fragaria \texttimes ananassa cultivar Camarosa, one of the most historically important and widely grown strawberry cultivars worldwide. We sequenced the genome through a combination of short- and long-read approaches, including Illumina, 10X Genomics, and PacBio, totaling 615-fold coverage of the genome (Supplementary Table 1). Illumina (455-fold coverage) and 10X Genomics (117-fold coverage) data were assembled and scaffolded with the software package DenovoMAGIC3 (NRGene ( Supplementary Table 2), which has recently been used to assemble the allotetraploid wheat (Triticum turgidum) genome\textsuperscript{32}. We further scaffolded the genome to chromosome scale by using Hi-C data (401-fold coverage) in combination with the HiRise pipeline (Dovetail) (Supplementary Figs. 1–3), then performed gap-filling with 43-fold coverage error-corrected PacBio reads with PBJelly\textsuperscript{33} (Supplementary Table 3). The total length of the final assembly is 805,488,706 bp, distributed across 28 chromosome-level pseudomolecules (Fig. 1) and representing ~99% of the estimated genome size, on the basis of flow cytometry measurements. A genetic map for Fragaria \texttimes ananassa\textsuperscript{34} was used to correct any misassemblies, and comparisons to Fragaria vesca were used to identify homoeologous chromosomes.

We annotated 108,087 protein-coding genes along with 30,703 genes encoding long noncoding RNAs (lncRNAs), which were subdivided into 15,621 long intergenic noncoding RNAs, 9,265 antisense overlapping transcripts (AOT-lncRNAs), and 5,817 sense overlapping transcripts (SOT-lncRNAs) (Supplementary Table 4). Gene annotation and genome-assembly quality were evaluated with the Benchmarking Universal Single-Copy Orthologs v2 (BUSCO)\textsuperscript{35} method (Supplementary Table 5). Most (99.17%) of the 1,440 core genes in the embryophyta dataset were identified in the annotation, thus supporting a high-quality genome assembly. The repetitive components of the nuclear genome were annotated with a custom-repeat-library approach\textsuperscript{36}, including DNA transposons, long-terminal-repeat retrotransposons (LTR-RTs; for example, Copia and Gypsy), and non-LTR retrotransposons (Supplementary Table 6 and Supplementary Fig. 4). Transposable element (TE)-related sequences make up ~36% of the total genome assembly, and LTR-RTs are the most abundant TEs (~28%). The plastid and mitochondrial genomes were also assembled, annotated, and verified for completeness (Supplementary Fig. 5).

Origin of octoploid strawberry.

Using the Fragaria \texttimes ananassa reference-genome assembly, we sought to identify the extant diploid relatives of each subgenome donor\textsuperscript{37}. Previous phylogenetic studies aimed at identifying these progenitor species, often analyzing a limited number or different sets of molecular markers, have obtained inconsistent results\textsuperscript{15,16,38}. However, F. vesca has long been suspected to be a progenitor, on the basis of meiotic chromosome pairing\textsuperscript{39}; subsequent molecular phylogenetic analyses supported it being one of the diploid progenitors along with Fragaria inumae and two additional unknown species\textsuperscript{1}. We sequenced and de novo assembled 31 transcriptomes of every described diploid Fragaria species, which we used to identify progenitor species on the basis of the phylogenetic analysis of 19,302 nuclear genes in the genome (Fig. 2, Supplementary Figs. 6–8 and Supplementary Table 7). To our knowledge, this is the most comprehensive molecular phylogenetic analysis of the genus Fragaria to date, including the greatest number of molecular markers and sampling of diploid species, aimed at identifying the extant relatives of the progenitor species of octoploid strawberry (Supplementary Fig. 9 and Supplementary Table 8).

Our phylogenetic analyses provided strong genome-wide support for the two diploid progenitor species that had been previously hypothesized and identified the two previously unknown diploid progenitors. This discovery, together with the geographic distributions, natural history, and genomic footprints of the diploid species, provided a model for the chronological formation of intermediate polyploids that culminated in the formation of the octoploid (Fig. 2). Our phylogenetic analyses revealed F. inumae and Fragaria nipponica as two of the four extant diploid progenitor species, both of which are endemic to Japan and in geographic proximity to all five described tetraploid species in China. The third species identified in our analyses, Fragaria viridis, is geographically distributed in Europe and Asia, and partially overlaps with the sole hexaploid species, Fragaria moschata. Therefore, we hypothesized that these tetraploid and hexaploid species may be evolutionary intermediates between the diploids and the wild octoploid species. This possibility is supported by a previous phylogenetic analysis identifying F. viridis as a possible parental contributor to both F. moschata and the octoploid event\textsuperscript{41}. Finally, we identified F. vesca subsp. brachetii, which is endemic to the western part of North America, spanning Mexico to British Columbia, as the fourth parental contributor. Our species sampling also included two other F. vesca subspecies: F. vesca subsp. vesca, which is distributed from Europe to the Russian Far East, and F. vesca subsp. californica, which is endemic to the coast of California.

Octoploid strawberry species are geographically restricted to the New World and are largely distributed across North America, with the exception of isolated F. chiloensis populations in Chile and the Hawaiian Islands\textsuperscript{42}. Therefore, our phylogenetic analyses combined with the geographic distributions of extant species not only support a North American origin for the octoploid strawberry but also suggest that F. vesca subsp. brachetii was probably the last diploid progenitor species to contribute to the formation of the ancestral octoploid strawberry. This possibility is further supported by a previous study revealing F. vesca subsp. brachetii as the likely maternal donor of the octoploid event, on the basis of the phylogenetic history of the plastid genome\textsuperscript{1}. This finding is consistent with our analysis of the plastid genome of ’Camarosa’ (Supplementary Fig. 10). Thus, these data suggest that the hexaploid ancestor probably crossed into North America from Asia and hybridized with native populations of F. vesca subsp. brachetii, an event dated at ~1.1 million years before
present). Our phylogenetic analysis also identified related diploid species possibly arising from ancient hybridization and introgression events with putative progenitor species or issues related to incomplete lineage sorting and/or missing data (Supplementary Fig. 6). Future studies will be able to more thoroughly investigate these possibilities after reference quality genomes are assembled for these other diploid progenitor species.

Subgenome dominance in allopolyploids. After most ancient allopolyploid events, one of the subgenomes, commonly referred to as the ‘dominant’ subgenome, emerges with significantly greater gene content and more highly expressed homoeologs (that is, post-polyploidy duplicate genes) than those of the other ‘submissive’ subgenome(s)4,51. Biased fractionation, which results in greater gene content of the dominant subgenome45, was first described in the model plant Arabidopsis thaliana21 and later described in Zea mays (maize)43, Brassica rapa (Chinese cabbage)44, and Triticum aestivum (bread wheat)45. The dominant subgenome has also been shown to be under stronger selective constraints46–48 and to be heritable through successive allopolyploid events49, and, as predicted50, it is not observed in ancient autopolyploids50–52. Moreover, subgenome expression dominance has recently been shown to occur instantly after interspecific hybridization and to increase over successive generations in monkeyflower23. However, some allopolyploids, including Capsella bursa-pastoris53 and Cucurbita species54, do not exhibit subgenome dominance.

The emergence of a dominant subgenome may resolve various genetic and epigenetic conflicts that arise from the genomic merger of divergent diploid progenitor species4,55, including mismatches between transcriptional regulators and their target genes41. The mechanistic basis of subgenome dominance, at least in part, appears to be related to subgenome differences in the content and regulation of TEs22,56. Gene expression levels are negatively correlated with the density of nearby TEs56 (Supplementary Fig. 11). Thus, the merger of subgenomes with different TE densities results in higher gene expression for the dominant homoeolog with fewer
Because strawberry production is threatened by several agricul-
turally important diseases, we analyzed, in greater depth, the major
adaptive traits57,58. For example, major disease-resistance genes in
the other subgenomes (Supplementary Table 9). The
overall TE densities near genes were also lowest for
F. vesca
other homoeologous chromosomes (Supplementary Table 9). The
subgenomes compared with the other two subgenomes (9.8
×
subgenome (Supplementary Table 9). The F. vesca
subgenome, compared with the other subgenomes, also contains a greater
number of tandem gene arrays as well as larger average tandem-
gene-array sizes on six of seven homoeologous chromosomes. These
findings suggest that the dominant F. vesca subgenome, compared
with the other three subgenomes, has been under stronger selective
constraints to retain genes, including tandemly duplicated genes
known to be biased toward gene families that encode important
adaptive traits57,58. For example, major disease-resistance genes in
plants, including nucleotide-binding-site leucine-rich-repeate genes
(NBS-LRRs), which are usually clustered in tandem arrays 59, are
biased toward the dominant F. vesca subgenome (χ² test, P < 0.0001; Supplementary Fig. 12).

Because strawberry production is threatened by several agricultur-
ally important diseases, we analyzed, in greater depth, the major
family of plant resistance (R) genes60,61. Collectively, 423 NBS-
LRR genes were identified, including 195 encoding an N-terminal
coiled-coil (CC), 79 encoding toll interleukin 1 receptor (TIR), and
24 encoding resistance to powdery mildew 8 (RPW8) domains
( Supplementary Fig. 12). Recent work has demonstrated that many
R proteins recognize pathogen effectors through integrated decoy
domains62, and the F. vesca genome encodes 20 such protein mod-
els62. Fragaria × ananassa has a greatly expanded set of 105 diverse
domains that are fused to the R-protein structures and have the
potential to function as integrated decoys62 (Supplementary Fig. 13
and Supplementary Dataset 1). Only a few resistance genes have
been phenotypically identified in Fragaria × ananassa, but none have been functionally characterized62,63. The annotated genome
thus provides a framework for accelerating R-gene discovery,
connecting phenotype to genotype, and pyramiding R genes by devel-
oping targeted, homoeolog-specific molecular markers.

Although chromosomes contributed by the F. vesca progenitor
retained the most genes overall, certain regions on chromosomes
from the other progenitor species retained higher numbers of ancestral
genes (Fig. 1b and Supplementary Fig. 14). Further analy-
sis revealed that these regions are the products of homoeologous
exchanges (HEs) or gene-conversion events64,65,66 (Supplementary
Figs. 15 and 16). Notably, most HEs in octoploid strawberry involved
replacements of the submissive homoeologs by correspond-
ing regions of the dominant F. vesca subgenome (Supplementary
Table 10). For example, our phylogenetic and comparative genomic
analyses showed that HEs are 7.3× biased toward the F. vesca sub-
genome compared with F. iinumae, but they are not unidirectional as
previously reported61. HEs were even more biased toward the F. vesca
subgenomes compared with the other two subgenomes (9.8× for
F. viridis and 10.4× for F. nipponica). These analyses validate find-
ings from a previous study in wild octoploid strawberry7 and show
that portions of the F. iinumae subgenome have been replaced with
the F. vesca subgenome (Fig. 1b). Here, we identified HEs ranging
in size from single genes to megabase-sized regions on chromo-
somes (Supplementary Table 10), findings similar to the patterns
observed in other alloployploids including Brassica napus (rape-
seed)7,28, Gossypium hirsutum (cotton)67,68, and bread wheat7,28. The
observed bias of HEs genome wide may be due to selection favoring
the maintenance of proper network stoichiometry7 and altered dos-
age of certain gene products25 during the establishment of the domi-
nant subgenome. Interestingly, 32.6% of NBS-LRR genes encoded on
the three submissive subgenomes are derived from HE with the
F. vesca subgenome. This result suggests that although the F. vesca
subgenome may also dominate disease resistance in strawberry, the
maintained diversity of resistance mechanisms contributed by the
other three diploid progenitors may also have been under selection.
Finally, we examined gene expression in diverse organs to test whether the dominant *F. vesca* subgenome is more highly expressed than the submissive genomes (Fig. 3), as predicted by the subgenome-dominance hypothesis. The density of TEs near genes was found to be negatively correlated with gene expression across all subgenomes (Supplementary Fig. 11a). Because HES reshuffled and replaced homoeologs across each of the four parental chromosomes, only homoeolog pairs that had support for subgenome assignment were evaluated for subgenome expression dominance (that is, homoeolog expression bias). Our analyses revealed that the dominant *F. vesca* subgenome, which had the lowest overall TE densities near genes of all subgenomes (Supplementary Fig. 11b; Kolmogorov–Smirnov test, $P<10^{-3}$), encodes more significantly dominantly expressed homoeologs than the other three submersive subgenomes combined (Fig. 3c). This finding supports the hypothesis that subgenome expression dominance is influenced by overall TE-density differences between subgenomes. At the individual homoeolog level, many dominantly expressed homoeologs were also contributed by one of the three submersive subgenomes. This observation was expected, given the variation in TE densities near homoeologs in each of the diploid progenitor genomes.

Most HES in octoploid strawberry resulted in the dominant *F. vesca* subgenome replacing the corresponding homoeologous regions of one of the submersive subgenomes. Thus, the observed homoeolog expression bias toward the *F. vesca* subgenome in Fig. 3 is an underestimate of transcriptome-wide expression dominance (68.7% of all transcripts). This bias has resulted in certain biological pathways being largely controlled by a single dominant subgenome. Our analyses revealed that certain metabolic pathways, including those that give rise to strawberry flavor, color, and aroma, are largely controlled by the dominant subgenome. For example, *F. vesca* homoeologs in octoploid strawberry are responsible for 88.8% of the biosynthesis of anthocyanins, the metabolites responsible for the red pigments in ripening strawberry fruit; 69.2% of the biosynthesis of geranyl acetate, a terpene associated with fruit aroma; and 95.3% of the biosynthesis of fructose associated with sweetness (Supplementary Dataset 2). Similar results have been found in allotetraploid *Brassica juncea*, in which many dominant homoeologs have been found to be related to glucosinolate biosynthesis and to show signs of positive selection.

**Discussion**

We present what is, to our knowledge, the first chromosome-scale genome assembly for an octoploid strawberry—the highest-level polyploid genome of this quality assembled to date. Analysis of this genome allowed us to identify each of the diploid progenitor species, reconstruct the evolutionary history of the octoploid event, and investigate the evolution of a dominant subgenome. Our data support the hypothesis that subgenome dominance in an allopolyploid is established by TE-density differences near homoeologous genes in each of the diploid progenitor genomes. Furthermore, our results show that the *F. vesca* subgenome has increased in dominance over time by having retained significantly more ancestral genes and a greater number of tandemly duplicated genes than the other three subgenomes, and replaced large portions of the submersive subgenomes via homoeologous exchanges. These trends, combined with subgenome expression dominance, have resulted in many traits being largely controlled by a single dominant subgenome in octoploid strawberry. This finding is consistent with results from a recent report indicating that the dominant subgenome in maize contributes more to phenotypic variation than the submersive subgenome. This reference genome should serve as a powerful platform for breeders to develop homoeolog-specific markers to track and leverage allelic diversity at target loci. Thus, we anticipate that this new reference genome, combined with insights into subgenome dominance, will greatly accelerate molecular breeding efforts in the cultivated garden strawberry.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at [https://doi.org/10.1038/s41588-019-0356-4](https://doi.org/10.1038/s41588-019-0356-4).

**References**

1. Duchesne, A.-N. Histoire Naturelle des Fraisiers Contenant les Vues d’Economie Réunies à la Botanique, et Suivie de Remarques Particulières sur Plusieurs Points qui ont Rapport à l’Histoire Naturelle Générale, par M. Duchesne Fils. (Didot le Jeune, Paris, 1766).

2. 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).

3. Tennesen, J. A., Gowindarajulu, 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).

4. McClintock, B. The significance of the responses of the genome to challenge. *Science* **226**, 792–801 (1985).

5. Folta, K. M. & Gardiner, S. E. Genetics and Genomics of Rosaceae (Springer, New York, 2009).

6. Staudt, G. Taxonomic studies in the genus *Fragaria* typification of *Fragaria* species known at the time of Linnaeus. *Can. J. Bot.* **40**, 869–886 (1962).

7. 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).

8. Bringhurst, R. S. & Khan, D. A. Natural pentaploid *Fragaria chiloensis*—*F. vesca* hybrids in coastal california and their significance in polyploid *Fragaria* evolution. *Am. J. Bot.* **50**, 658–661 (1963).

9. Milne, R. J. & Abbott, R. J. Reproductive isolation among two interfertile *Rhododendron* species: low frequency of post-F1 hybrid genotypes in alpine hybrid zones. *Mol. Ecol.* **17**, 1108–1121 (2008).

10. Soltis, P. S. Ancient and recent polyploidy in angiosperms. *New Phytol.* **166**, 5–8 (2005).

11. Freeling, M. & Thomas, B. C. Gene-balanced duplications, like tetraptodops, provide predictable drive to increase morphological complexity. *Genome Res.* **16**, 805–814 (2006).

12. Doyle, J. J. et al. Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461 (2008).

13. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).

14. Stubbins, G. L. Jr. Types of polyploids; their classification and significance. *Adv. Genet.* **1**, 403–429 (1947).

15. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).

16. Leitch, A. R. & Leitch, I. J. Genomic plasticity and the diversity of polyploid species: low frequency of polyploid genotypes in alpine hybrid zones. *Mol. Ecol.* **17**, 1108–1121 (2008).

17. Ulrich, D. & Olbricht, K. Diversity of volatile patterns in sixteen *Fragaria vesca* L. accessions in comparison to cultivars of *Fragariaxananassa*. *J. Appl. Bot. Food Qual.* **86**, 37–46 (2013).

**URLs**

- Sequence Read Archive, [https://www.ncbi.nlm.nih.gov/sra/](https://www.ncbi.nlm.nih.gov/sra);
- Dryad, [https://doi.org/10.5061/dryad.b2c58pc](https://doi.org/10.5061/dryad.b2c58pc);
- PhyDS, [https://github.com/mrmckain/PhyDS/](https://github.com/mrmckain/PhyDS);
- GDR, [https://www.rosaceae.org](https://www.rosaceae.org);
- CoGe, [https://genomevolution.org/r/tx72/](https://genomevolution.org/r/tx72);
- RefTranS, [https://github.com/mrmckain/RefTranS](https://github.com/mrmckain/RefTranS);
- annoBTD, [https://github.com/mrmckain/annoBTD](https://github.com/mrmckain/annoBTD);
- Mitofy, [http://dogma.ccbb.utexas.edu/mitofy/](http://dogma.ccbb.utexas.edu/mitofy);
- dotPlotly, [https://github.com/tpoorten/dotPlotly](https://github.com/tpoorten/dotPlotly);
- NCBI Conserved Domain Database, [www.ncbi.nlm.nih.gov/Structure/bwrbpsb/bwrbpsb.cgi](http://www.ncbi.nlm.nih.gov/Structure/bwrbpsb/bwrbpsb.cgi);
- Pafm database, [www.ebi.ac.uk/Tools/pfa/pafmscan](http://www.ebi.ac.uk/Tools/pfa/pafmscan);
- FastQC, [https://www.bioinformatics.babraham.ac.uk/projects/fastqc](https://www.bioinformatics.babraham.ac.uk/projects/fastqc);
- R, [https://www.r-project.org](https://www.r-project.org);
- Repeat-Masker, [http://www.repeatmasker.org](http://www.repeatmasker.org);
- RepeatModeler, [http://www.repeatmasker.org/RepeatModeler](http://www.repeatmasker.org/RepeatModeler);
- Google Maps, [https://www.google.com/maps](https://www.google.com/maps).

**Received:** 27 April 2018; Accepted: 15 January 2019;
**Published online:** 25 February 2019
20. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl Acad. Sci. USA 108, 4069–4074 (2011).

21. Thomas, B. C., Pedersen, B. & Freeling, M. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16, 934–946 (2006).

22. Freeling, M. et al. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr. Opin. Plant Biol. 15, 131–139 (2012).

23. Edger, P. P. et al. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell 29, 2130–2167 (2017).

24. Bottani, S., Zabet, N. R., Wendel, J. F. & Vittia, R. A. Gene expression dominance in allopolyploids: hypotheses and models. Trends Plant Sci. 23, 393–402 (2018).

25. Bird, K. A., VanBuren, R., Puzey, J. R. & Edger, P. P. The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytol. 200, 87–93 (2013).

26. Lysak, M. A., Cheung, K., Kitchche, M. & Burel, P. Ancestral chromosomal blocks are triplicated in Brassicaceae species with varying chromosome number and genome size. Plant Physiol. 145, 402–410 (2007).

27. Xiong, Z., Gaeta, R. T. & Pires, J. C. Homeologous shuffling and chromosome compensation maintain genome balance in resynthesized Brassica napus. Proc. Natl Acad. Sci. USA 108, 7908–7913 (2011).

28. Chalhoub, B. et al. Early allopolyploid evolution in the post-Neolithic Brassica oleracea B. Phylogenetic relationships among Capsella, B. Plant Cell 26, 288–299 (2014).

29. Edger, P. P., McKain, M. R., Bird, K. A. & VanBuren, R. Subgenome assignment in allopolyploids: challenges and future directions. Curr. Opin. Plant Biol. 42, 76–80 (2018).

30. Collard, B. C. Y. & Mackill, D. J. Marker-assisted selection: an approach for improved plant breeding in the twenty-first century. Phil. Trans. R. Soc. Lond. B 363, 557–572 (2008).

31. 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).

32. Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357, 93–97 (2017).

33. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences. Plant Biol. 51, 289–304 (2019).

34. Edger, P. P. et al. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria virginiana). Nature 545–547 (2014).

35. Collard, B. C. Y. & Mackill, D. J. Marker-assisted selection: an approach for improved plant breeding in the twenty-first century. Phil. Trans. R. Soc. Lond. B 363, 557–572 (2008).

36. Campbell, M. S. et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164, 513–524 (2014).

37. 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).

38. Potter, D., Luby, J. J. & Harrison, R. E. Phylogenetic relationships among species of Fragaria (Rosaceae) inferred from non-coding nuclear and chloroplast DNA sequences. Syst. Bot. 25, 337–348 (2000).

39. Yang, Y. & Davis, T. M. A new perspective on polyploid Fragaria (strawberry) genome composition based on large-scale, multi-locus phylogenetic analysis. Genome Biol. Evol. 9, 3433–3448 (2017).

40. Fedorova, N. J. Crossability and phylogenetic relations in the main European species of Fragaria. Compil. Natl Acad. Sci. USSR. 52, 545–547 (1946).

41. Lundberg, M. Systematics and Polyploid Evolution in Potentilleae (Rosaceae). PhD thesis, Stockholm University (2011).

42. Johnson, A. L., Govindaraju, R. & Ashman, T. L. Bioclimatic evaluation of geographical range of Fragaria (Rosaceae): consequences of variation in breeding system, ploidy and species age. Bot. J. Linn. Soc. 176, 99–114 (2014).

43. Sankoff, D., Zheng, C., & Wang, B. A model for biased fractionation after whole genome duplication. BMC Genomics 13 (Suppl. 1), S8 (2012).

44. Wang, X. et al. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43, 1035–1039 (2011).

45. Eckardt, N. A. Genome dominance and interaction at the gene expression level in allohexaploid wheat. Plant Cell 26, 1834 (2014).

46. Parkin, I. A. P. et al. Transcriptome and methylation profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. Genome. Biol. 15, R77 (2014).

47. Cheng, F. et al. Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. New Phytol. 211, 288–299 (2016).

48. Renny-Byfield, S., Rodgers-Melnick, E. & Ross-Ibarra, J. Gene fractionation and analysis of Preferred homeolog gene expression in strawberry. Nat. Genet. 48, 1225–1232 (2016).

49. Woodhouse, M. R. et al. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc. Natl Acad. Sci. USA 111, 5283–5288 (2014).

50. Garmauz, O. et al. Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31, 448–454 (2014).

51. Zhao, M., Zhang, B., Lisch, D. & Ma, J. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. Plant Cell 29, 2974–2994 (2017).

52. Wendel, J. F., Lisch, D., Hu, G. & Mason, A. S. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr. Opin. Genet. Dev. 49, 1–7 (2018).

53. Douglas, G. M. et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. Proc. Natl Acad. Sci. USA 112, 2806–2811 (2015).

54. Sun, H. et al. Karyotype stability and unbiased fractionation in the paleo-allopolyploid Crucibaria genomes. Mol. Plant 10, 1293–1306 (2017).

55. Osborn, T. C. et al. Understanding mechanisms of novel gene expression in polyploids. Trends Genet. 19, 141–147 (2003).

56. Hollister, I. D. & Garden, S. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19, 1419–1428 (2009).

57. Rizzon, C., Ponger, L. & Gaut, B. S. Strikingly similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. Plant Syst. Evol. 292, 705–715 (2016).

58. Hanada, K., Zou, C., Shiu, S.-H., Shin. Zou, C., Shiu, S.-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 148, 993–1003 (2008).

59. Qian, L.-H. et al. Distinct patterns of gene gain and loss in the diploid-to-triploid evolution of Solanum species. New Phytol. 209, 541–547 (2015).

60. Meyers, B. C., Kozik, A., Gregio, A., Kuang, H. & Michelmore, R. W. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell 15, 809–834 (2003).

61. Zhang, T. et al. Sequencing of allotetraploid cotton (Gossypium hirsutum L.). Nat. Biotechnol. 26, 343–346 (2008).

62. Sun, H. et al. Karyotype stability and unbiased fractionation in the paleo-allopolyploid Cucurbita genomes. Mol. Plant 10, 1293–1306 (2017).

63. Fedorova, N. J. Crossability and phylogenetic relations in the main European species of Fragaria. Compil. Natl Acad. Sci. USSR. 52, 545–547 (1946).

64. Garsmeur, O. et al. Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31, 448–454 (2014).

65. Garmauz, O. et al. Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31, 448–454 (2014).
Acknowledgements
This work was supported by Michigan State University AgBioResearch to P.P.E., USDA-NIFA HATCH 1009804 to P.P.E., NSF-DEB 1737898 to P.P.E., S.J.T as a participant in the Plant Genomics at MSU REU program funded by NSF-DBI 1757043, USDA-NIFA SCRI 2017-51181-26833 to S.J.K., the California Strawberry Commission to S.J.K., and the University of California to S.J.K.

Author contributions
P.P.E. and S.J.K. conceived and designed the project. M.C., G.S.C., and C.B.A. collected the samples and extracted DNA and RNA. P.P.E., S.J.K., M.C., C.B.A., K.B., and T.S. coordinated the Illumina and PacBio sequencing. P.P.E., T.J.P., R.V., M.A.H., C.M.W., G.B.-Z., A.B., K.B., T.S., L.S., E.I.A., and J.P.M. assembled the nuclear and organellar genomes. M.C., K.L.C., A.D.L.N., S.O., N.J., E.I.A. and J.P.M. annotated the nuclear and organellar genomes. M.R.K. and P.P.E. performed phylogenetic analyses. P.P.E., T.J.P., R.V., M.A.H., M.F. and E.L. ran comparative genomic analyses. A.E.Y., K.A.B., N.P., P.P.E., and S.J.K. identified disease-resistance genes. P.P.E., R.D.S., S.J.T., J.R.P., and A.D.L.N. performed gene expression analyses. M.C., M.R.M., M.A.H., S.J.K., and P.P.E. were in charge of data submission. P.P.E. wrote the manuscript draft, and all coauthors reviewed and revised the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0356-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.P.E. or S.J.K.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
Methods

Plant material. The cultivar 'Camarosa' was selected because of its importance to the industry; historically, it has been one of the most widely grown short-day varieties worldwide, and it remains an important genotype in breeding programs. The haploidy genome size (>133 Mbp) was estimated through flow cytometry with four technical replicates at the Flow Cytometry Core at Benaroya Research Institute at Virginia Mason (Supplementary Dataset 3).

Genomic sequencing. High-molecular-weight genomic DNA was isolated from young leaf tissue, after a 2 h dark treatment, through a modified nucleic-acid preparation method, and the quality was verified through pulsed-field gel electrophoresis. A total of five PacBio 20-kb libraries were generated with a SMRTbell Template Prep Kit (PacBio) and were sequenced with 67 SMRT cells on the PacBio RSII platform at the UC Davis DNA Sequencing Facility. A total of 67 Gb (~82× of PacBio sequence data was generated with an N50 read length of 17,699 bp (Supplementary Table 3). To further increase read length and eliminate shorter reads, we constructed a separate mate pair (MP) library with jumps of 2–5 kb, 5–7 kb, and 7–10 kb, with an Illumina Nextera Mate-Pair Sample Preparation Kit. The 800-bp library was sequenced on an Illumina HiSeq2500 system with paired-end, 150-bp reads, and the MP libraries were sequenced on an Illumina HiSeq4000 system with paired-end, 150-bp reads. A total of ~570 Gb (~45× fold coverage) of additional Illumina sequencing data was generated (Supplementary Table 1). Illumina library construction and sequencing were conducted at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign.

Genome assembly. The genome was assembled with the DeNovo MAGIC software platform (NRGene), a DeBruijn-graph-based assembler designed for highly polyploid, heterozygous and/or repetitive genomes27,28. The Chromium 10x data were used to phase haplotypes and support scaffold validation and further elongation of the phased scaffolds. Dovetail HiC libraries were prepared as described previously29 and sequenced on an Illumina HiSeq2500 system with paired-end, 150-bp reads to ~401 kb in length. After trimming, we constructed a separate mate pair (MP) library with jumps of 2–5 kb, 5–7 kb, and 7–10 kb, with an Illumina Nextera Mate-Pair Sample Preparation Kit. The 800-bp library was sequenced on an Illumina HiSeq2500 system with paired-end, 150-bp reads, and the MP libraries were sequenced on an Illumina HiSeq4000 system with paired-end, 150-bp reads. A total of ~570 Gb (~45× fold coverage) of additional Illumina sequencing data was generated (Supplementary Table 1). Illumina library construction and sequencing were conducted at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign.

Tissue collection, RNA library preparation, and sequencing. Plant tissue samples (flower before anthesis, flower at anthesis, leaf collected during the day and at night, leaves treated with methyl jasmonate [30 min, 4 h, and 24 h after treatment], runner, and salt-treated and untreated roots) were collected from Fragaria × ananassa cultivar 'Camarosa' grown in a growth chamber and immediately flash frozen in liquid nitrogen. Leaf tissues were also collected from wild diploid species grown at F. vesca. For phylogenetic analyses (Supplementary Table 1), total RNA was isolated with a KingFisher Pure RNA Plant Kit (Thermo Fisher) and quantified with a Qubit 3 fluorometer (Thermo Fisher). RNA libraries were prepared with the KAPA mRNA HyperPrep Kit protocol (KAPA Biosystems). All samples were submitted to the Michigan State University Research Technology Support Facility Genomics core and sequenced with paired-end, 150-bp reads on an Illumina HiSeq 4000 system.

Transcriptome assembly and translation. Reads were cleaned with Trimmomatic v0.32 (ref. 30) with adaptor trimming for TruSeq3 paired-end reads with a 1-bp mismatch, a palindrome clip threshold of 30, and a simple clip threshold of 10. Reads were then filtered on the basis of an average phred score calculated from a sliding window of 10 bp with a minimum threshold of 20 (Supplementary Dataset 4). The quality of trimmed reads was assessed afterward with FastQC. Genomically guided and de novo transcriptome assemblies were generated with Trinity v2.2.0 (ref. 31) for the genome annotation/expression and phylogenetic analyses, respectively. For genome annotation and expression analyses, reads were aligned to the Fragaria × ananassa cultivar 'Camarosa' genome with STAR v 2.5.3a with default options, except for --alignIntronMax, which was set to 10000. For genome annotation, the coordinate-sorted output files from STAR were used for the genome-guided transcriptome assembly, and name-sorted SAM files were used for gene expression analysis (HTSeq in section 3). For the diploid species libraries used in the phylogenetic analyses, because transcriptome libraries were generated with a stranded method, the ‘SS_lib_type’ parameter with ‘RF’ option was used in the assembly. In addition, reads were normalized to maximum read coverage of 100 with ‘normalize_max_read_cov’ in Trinity. The normalization option, which decreases the quantity of input reads for highly expressed genes, was used to improve assembly efficiency32. For homoeolog expression bias (HEB) analyses (described in the section below), counts of uniquely mapping reads were generated with HTSeq v 0.6.1 (ref. 33) with default options of htsLib-count, except for feature type, which was set to ‘gene’ for all RNA-seq datasets of ‘Camarosa’. The fragments per kilobase per million reads mapped (FPKM) values were derived with the standard formula for FPKM = (read count/‘per million’ scaling factor)/gene length in kilobases. For phylogenetic analysis, according to McKain et al.34, reads were aligned to the assembled transcripts with bowtie v 1.1.0 (ref. 35). Transcript abundance was estimated with RSEM v 1.2.29 (ref. 36) through the align, and estimate abundance.pl script packaged with Trinity. Transcripts were filtered by FPKM, an output from the aforementioned Perl script, with a minimum threshold of 1.0% of fragments per isoform mapped, as implemented in the filter fasta__by_rsem_values.pl script. Filtered transcripts were BLASTed against the Fragaria v2.40 genome with BLASTX with an E-value threshold of 1 × 10^-10. The RefTrans package (see URLs) was used to translate assembled transcripts by filtering BLAST hits to identify the best hit with at least 75% bidirectional overlap between the transcript and F. vesca coding sequences. Best hits were used to guide translations with GeneWise (Wisev 2.2.0).10. The longest translations were used in downstream analyses.

Gene annotation. The genome was annotated with the MAKER-P annotation pipeline37. Protein sequences (Araport11 and UniprotKB plant database), expressed sequence tags (NCBI), and ten RNA-seq datasets (described below) and additional RNA-seq data for Fragaria × ananassa downloaded from NCBI-SRA (BioProject PRJNA949190; red ripening fruit) were used as evidence during annotation. The RNA-seq datasets were assembled into transcripts through the StringTie genome-guided approach10. A custom repeat library ('Repeat annotation' section below) and MAKER repeat library were used for genome masking. Ab initio gene prediction was performed with the gene predictors SNAP8 and Augustus, which were previously iteratively trained for F. vesca. During annotation, gene models with annotation edit distance <1.0 were included in the MAKER gene set and scanned for the presence of protein domains. The predicted gene models were further filtered to remove those with TE-related domains. Briefly, the protein-coding genes were searched (BLASTp, e = 1 × 10^-10) against a transposable database from a previous study38, and if more than 50% of gene length aligned to the transposases, the gene was removed from the gene set. However, if 60% or more of the amino acid matches were due to only three individual amino acids, the alignment was considered to be caused by low complexity and was excluded. In addition, to assess whether core plant genes were annotated, the gene set was searched against the BUSCO v 2 (ref. 39) plant dataset (embryophyta_odbl9). InRNAs, including long intergenic noncoding RNAs, antisense overlapping transcripts, and sense overlapping transcripts, were identified with the Evolinc InRNA-discovery pipeline (v 1.5.1).10. Transcripts with fewer than three reads per base pair were discarded. Putative InRNAs with similarity (BLASTn e value <1 × 10^-10) to known TEIs or rRNAs (catalog v 13.0) of housekeeping RNAs were removed.

Repeat annotation. The Fragaria × ananassa genome search for LTR-RTs with LTRharvest with parameters ‘-m mlen110 -mmaxlen7000 -minsd 4 -maxsd 6 -motif TGCA -motifsm 1 -similar 85 -visc 10 -seed 20 -seqids yes’ and LTR_finder with parameters ‘-D 15000 -d 1000 -L 7000 -I 1000 -p 20 -M 0.9’. The identified LTR-RT candidates were filtered with LTR_retriever with default options. Candidate MITEs were manually checked for TSD and TIR, which were used for superfamly classification. Those with ambiguous TSD and TIR were classified as unknowns. The Fragaria × ananassa genome was then masked with both MITE and LTR libraries through RepeatMasker (see URLs) and other repetitive elements were identified with RepeatModeler (see URLs). The repeats were then grouped into two categories: sequences of known identity and sequences of unknown identity. The latter were then searched against the transposase database, and if they had a match, they were included in the TE library. The library was further filtered with ProtExcluder and an in-house Perl script to exclude gene fragments. The final TE library was used to annotate the Fragaria × ananassa genome with RepeatMasker with parameters ‘--nq 0’ and ‘--nr 40’. Annotation results were summarized with the ‘samcover.pl’ script from the LTR-retriever package11.

Organellar genome annotation. The chloroplast genome was annotated with Verdant, a web-based software suite specifically designed for plant chloroplast
with 90% or more gaps and transcripts with unaligned lengths less than 30% of amino acids. Codon alignments were filtered by removal of alignment columns under default parameters to create a codon alignment from MAFFT-aligned present. Coding sequences and amino acid translations were separated into Orthogroups were filtered so that a minimum of five unique accessions were present. The best-scoring references were used to annotate the RNA. Finally, the boundaries of each feature were identified on the basis of the sequence and positional information for the orthologous features from the five reference chloroplast genomes (Supplementary Fig. 5). The mitochondrial genome was annotated with the webserver for Mitofy (see URLS), a program designed to annotate the genes and RNAs in the mitochondrial genomes of seed plants10. Mitofy uses NCBI-BLAST against two genome sequences so the basis of Blast and its illustrated domain organization is displayed in Supplementary Fig. 13. Eight Fxa proteins with predicted Sec7/ADP-ribosylation-factor-and G-nucleotide-exchange-factor domains were aligned by ClustalW and FastME 2.0 (ref. 124), and their illustrated domain organization is displayed in Supplementary Fig. 13. The full protein sequences of the 423 Fxa-NB-ARC-domain-containing proteins were aligned with MUSCLE v 3.8.31 (ref. 125) under default settings. This alignment was transformed with Tri.matches v 1.4.rev22 build 2015-05-21 (ref. 125) under default settings. An unrooted maximum-likelihood tree was constructed with RAxML v 8.2.11 (ref. 126) with the PROTGRAMMA substitution model. The tree was visualized with the APE package v 4.1 (ref. 127) in R v 3.3.3 (ref. 128) (see URLS).

Statistical analysis. The comparison of homoelog-expression abundance between the dominant subgroup and the three submissive subgenomes was carried out with a likelihood-ratio test and compared with Benjamini–Hochberg correction for multiple testing with a 1% significance level. The Kolmogorov–Smirnov test was used to determine which subgroup had the lowest-overall TE densities near genes. The p-test, with three degrees of freedom, was used to analyze the subgenome bias of disease-resistance genes. Bootstrapping, with 500 replicates under the GTR + gamma evolutionary model, was used to assess node support in trees generated by phylogenetic analyses.

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

Data availability

The genome assembly, annotation files, alignments, and phylogeny trees are available on Dryad (see URLS). Custom software for running PhyDS phylogenetic analyses is available on GitHub (see URLS). The genome assembly and annotation files are also available on the Genome Database for Rosaceae (GDR; see URLS) and the CyVerse CoGe platform (see URLS). 'Camarosa' clones are available from most strawberry nurseries. The raw sequence data are available in the Sequence Read Archive under NCBI BioProject PRJNA508389 (see URLS).

References

75. Zhang, H.-B., Zhao, X., Ding, X., Paterson, A. H. & Wing, R. A. Preparation of megabase-size DNA from plant nuclei. Plant J. 7, 175–184 (1995).
76. VanBuren, R. et al. Single-molecule sequencing of the desiccation-tolerant grass Oryzopsis hymenoides. Nature 527, 508–511 (2015).
77. Luo, M.-C. et al. Genome sequence of the progenitor of the wheat genome Aegilops tauschii. Nature 551, 498–502 (2017).
78. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009).
79. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26, 342–350 (2016).
80. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9, e112963 (2014).
81. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
82. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
83. Lyons, E., Pedersen, B., Kane, J. & Freeing, M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the Rosids. *Trop. Plant Biol.* 1, 181–190 (2008).

84. Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Wuite, D. & Davey, R. P. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front. Genet.* 4, 288 (2013).

85. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652 (2011).

86. Dobin, A. & Gingeras, T. R. Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics*. 51, 11.14.1–11.14.19 (2015).

87. 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).

88. Anders, S., Pyl, P. T. & Huber, W. HTSeq: a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

89. McKain, M. R. et al. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biol. Evol.* 8, 1150–1164 (2016).

90. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome. Biol.* 10, R25 (2009).

91. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323 (2011).

92. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Bioinformatics* 17, 25–26 (2001).

93. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295 (2015).

94. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196 (2008).

95. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* 5, 59 (2004).

96. Stanke, M. & Waiczak, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 (Suppl. 2), i215–i225 (2003).

97. Nelson, A. D. L. et al. Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs. *Front. Genet.* 8, 5 (2017).

98. Kalvari, I. et al. Pfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* 46, D335–D342 (2018).

99. Ellinghaus, D., Kurtz, S. & Willhoet, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9, 18 (2008).

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

101. 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).

102. Han, Y. & Wessler, S. R. MIT-E-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 38, e199 (2010).

103. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*. 25, 4.9.1–4.9.10 (2009).

104. Bao, L. & Liu, Z. in *Bioinformatics in Aquaculture* (ed. Liu, Z.) 1, 86–97 (Wiley, Hoboken, NJ, USA, 2017).

105. McKain, M. R., Hartsook, R. H., Wohl, M. M. & Kellogg, E. A. Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes. *Bioinformatics* 33, 130–132 (2017).

106. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421 (2009).

107. Alverson, A. J. et al. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol. Biol. Evol.* 27, 1436–1448 (2010).

108. Lowe, T. M. & Eddy, S. R. TRAScan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964 (1997).

109. Tang, H. et al. Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 12, 102 (2011).

110. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome. Biol.* 5, R12 (2004).

111. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210 (2012).

112. Velasco, R. et al. The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat. Genet.* 42, 833–839 (2010).

113. Palmenberg, A. C. & Bruns, T. *Pytozome*: a comparative platform for plant genomes. *Nucleic Acids Res.* 40, D1178–D1186 (2012).

114. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome. Biol.* 16, 157 (2015).

115. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60 (2015).

116. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780 (2013).

117. 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).

118. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* 57, 289–300 (1995).

119. de Oliveira Dal’Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumley, S. M. & Nielsen, L. K. AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol.* 152, 579–589 (2010).

120. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368 (2017).

121. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* 7, e1002195 (2011).

122. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44 (D1), D279–D285 (2016).

123. Marchler-Bauer, A. et al. CDD: NCBI’s conserved domain database. *Nucleic Acids Res.* 40, D1202–D1210 (2012).

124. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Nat. Methods* 5, D122–D126 (2008).

125. Lefort, V., Desper, R. & Gascuel, O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32, 2798–2800 (2015).

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

127. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973 (2009).

128. Stamatakis, A. RAxML version 8: a fast tree computation method for large phylogeny reconstruction under the maximum likelihood criterion. *Bioinformatics* 26, 3794–3797 (2010).

129. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290 (2004).

130. Ihaka, R. & Gentleman, R. R. A language for data analysis and graphics. *J. Comput. Graph. Stat.* 5, 299–314 (1996).
Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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 |
| ☑ | An indication of 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 statistics 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 |
| ☑ | For all reported estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated |
| ☐ | Clearly defined error bars |
| ☑ | State explicitly what error bars represent (e.g. SD, SE, CI) |

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

All commercial DNA and RNA sequencing platforms used in this study are fully described.

Data analysis

All commercial and custom software used in this study for data analysis is fully described including specifying versions used. All custom software developed for this study has already been deposited on Github with weblinks (e.g. Phylogenetic iDentification of Subgenomes; https://github.com/mrmckain/PhyDS).

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The genome assembly, annotations, and other supporting data will be made available on Dryad (http://datadryad.org/resource/doi:10.5061/dryad.km0s7k0).
Genome assembly and annotation will also be made publicly available on the Genome Database for Rosaceae (https://www.rosaceae.org/) and the CyVerse CoGe platform (https://genomevolution.org/coge/). The raw sequence data will be deposited in the Short Read Archive under NCBI BioProject PRJNA508389. No data restrictions.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size
A single polyploid genome was sequenced; comparing the expression of all syntenic homoeologous genes across all four subgenomes.

Data exclusions
No data was excluded from any analysis, unless described in the manuscript.

Replication
Flow cytometry measurements were replicated four times. Bootstrapping for phylogenetic analyses were replicated 500 times. Gene expression was compared across all four subgenomes with three distinct tissues to serve as independent replicates.

Randomization
Randomizations were not needed for this study, which involved analyzing subgenomes residing within a nucleus of a single genotype. Plants were grown in a sterile growth chamber.

Blinding
A blinded-experiment is not possible for genome analyses.

Reporting for specific materials, systems and methods

Materials & experimental systems

| Unique biological materials |
|-----------------------------|
| ☒ Yes |

| Antibodies |
|——|

| Eukaryotic cell lines |
|——|

| Palaeontology |
|——|

| Animals and other organisms |
|——|

| Human research participants |
|——|

Methods

| Involved in the study |
|-----------------------|
|——|

| ChiP-seq |
|——|

| Flow cytometry |
|——|

| MRI-based neuroimaging |
|——|

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

Clones from the sequenced genotype 'Camarosa' are widely available to the community from nurseries around the world. Material is no longer patent protected.

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.
Flow cytometry analyses were conducted by Dr. Arumuganathan in the Flow Cytometry and Imaging Core Laboratory at Virginia Mason Research Center in Seattle, Washington. The procedure used to analyze nuclear DNA content in plant cells was modified from Arumuganathan and Earle (1991). Briefly, the procedure consists of preparing suspensions of intact nuclei by chopping 50 mg plant plant tissues in MgSO4 buffer mixed with DNA standards and stained with propidium iodide (PI) in a solution containing DNAase-free-RNAse. Fluorescence intensities of the stained nuclei are measured by a flow cytometer. Values for nuclear DNA content are estimated by comparing fluorescence intensities of the nuclei of the test population with those of an appropriate internal DNA standard that is included with the tissue being tested. We use nuclei from Chicken Red blood cells (2.5 pg/2C), Glycine max (2.45 pg/2C), Oryza sativa cv Nipponbare (0.96 pg/2C), Arabidopsis thaliana (0.36 pg/2C) or Zea mays B73 (5.77 pg/2C) as the internal standard. For each measurement, the propidium iodide fluorescence area signals (FL2-A) from 1000 nuclei are collected and analyzed by CellQuest software (Becton-Dickinson, San Jose, CA) on a Macintosh computer. The mean position of the G0/G1 nuclei peak of the sample and the internal standard are determined by CellQuest software. The mean nuclear DNA content of each plant sample, measured in picograms, are based on 1000 scanned nuclei.

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