nature genetics

MATTERS ARISING

https://doi.org/10.1038/s41588-019-0544-2

OPEN

Reply to: Revisiting the origin of octoploid strawberry

Patrick P. Edger[1,2,8*], Michael R. McKain[3,8], Alan E. Yocca[1,4], Steven J. Knapp[5], Qin Qiao[6*] and Ticao Zhang[7*]

REPLYING TO A. Liston et al. *Nature Genetics* https://doi.org/10.1038/s41588-019-0543-3 (2019)

The origin of octoploid strawberry has been the focus of several phylogenetic studies over the past decade (for example, refs. [1–3]). Our previous study, using the octoploid genome and transcriptomes of every extant diploid *Fragaria* species, provided support for four species (*Fragaria vesca*, *Fragaria iinumae*, *Fragaria viridis* and *Fragaria nipponica*) as the closest extant relatives of the diploids that contributed to the origin of octoploid strawberry[4]. In a response paper[5], Liston et al. stated "that only two extand diploids were progenitors" with one subgenome being contributed by *F. vesca* and three by *F. iinumae–like* ancestors. Our reanalysis of the transcriptome data and comparative genomic analyses of a chromosome-scale *F. iinumae* genome support our previous model for the origin of octoploid strawberry[4].

Liston et al.[5] raised a concern regarding one of the steps in the phylogenetic analysis of the subgenome tree-searching algorithm (PhyDS) tool we developed to identify extant relatives of diploid progenitors of allopolyploids. Specifically, they argue that we may have incorrectly identified *F. viridis* and *F. nipponica* as extant relatives because in-paralogs were excluded from our previous phylogenetic analysis[4]. Our reanalysis of the data using PhyDS, now including

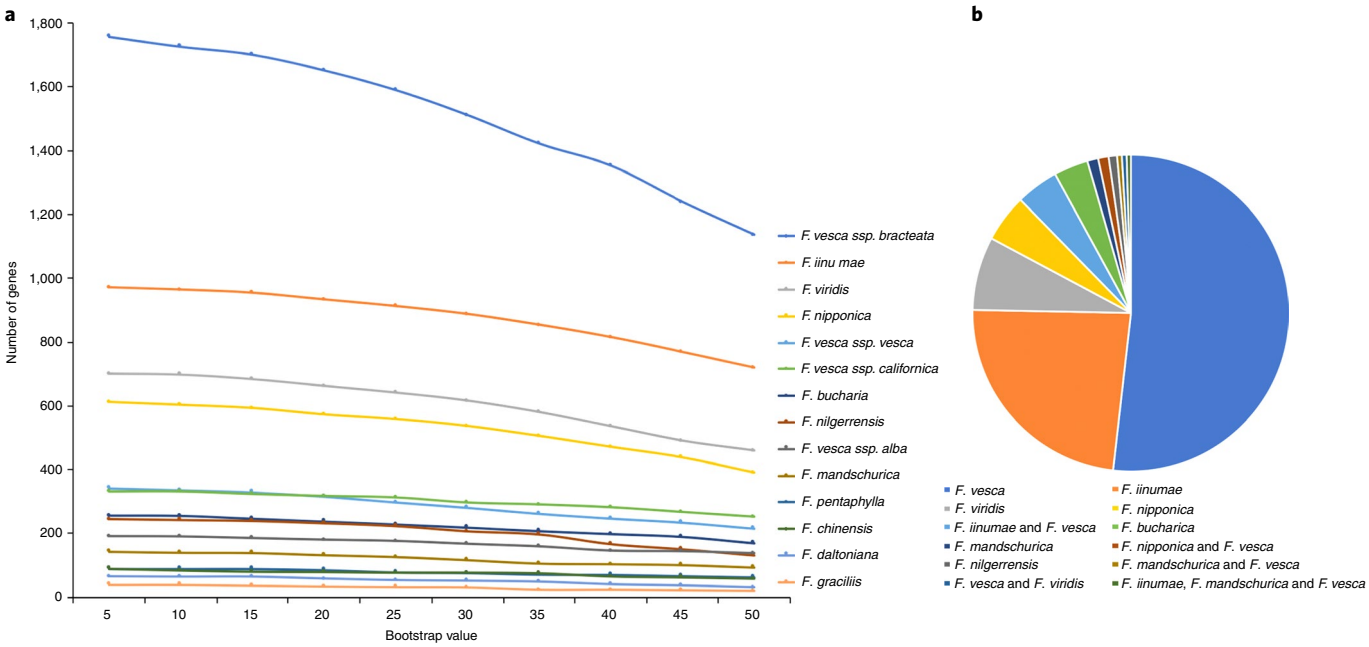


Fig. 1 | Phylogenetic analyses. a, Number of genes from species identified as being sister to a homoeolog from the octoploid genome, by using PhyDS with bootstrap support value (BSV) cutoffs. Based on previous results[4]. **b**, Reanalysis of the data, including in-paralogs and BSV_{50} cutoff, identified the same progenitor species. The prevalence and biased patterns of homoeologous exchanges between subgenomes resulted in the dominant *F. vesca* subgenome replacing a greater number of corresponding regions in each of the recessive subgenomes[4]. Thus, a greater number of genes from the dominant *F. vesca* subgenome were identified, with the *F. iinumae*–like subgenome being second.

[1]Department of Horticulture, Michigan State University, East Lansing, MI, USA. [2]Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI, USA. [3]Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, USA. [4]Department of Plant Biology, Michigan State University, East Lansing, MI, USA. [5]Department of Plant Sciences, University of California–Davis, Davis, CA, USA. [6]School of Agriculture, Yunnan University, Kunming, China. [7]College of Chinese Material Medica, Yunnan University of Chinese Medicine, Kunming, China. [8]These authors contributed equally: Patrick P. Edger, Michael R. McKain. *e-mail: edgerpat@msu.edu; qiaoqin@ynu.edu.cn; zhangticao@mail.kib.ac.cn

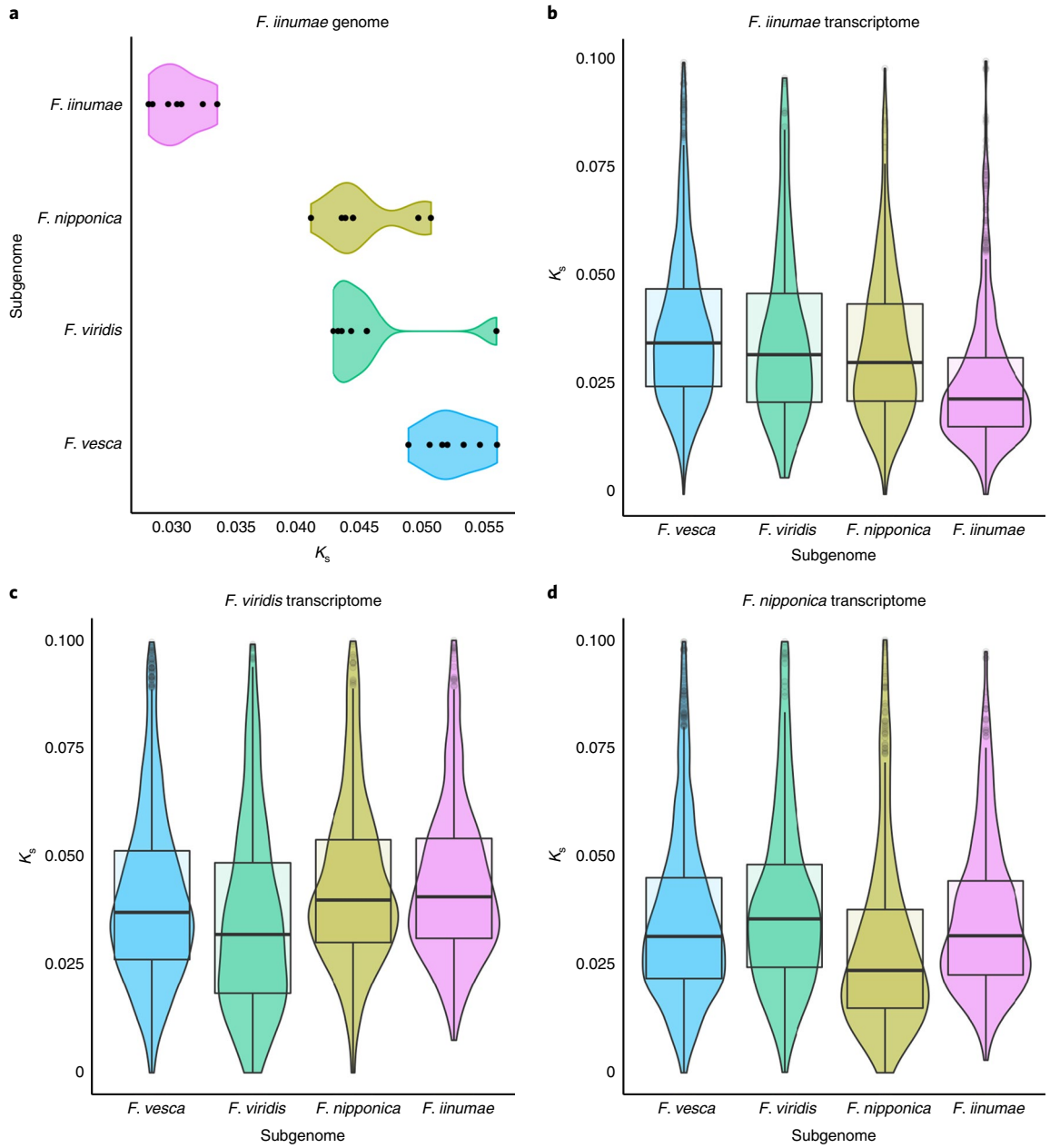


Fig. 2 | Divergence of K_s rates among subgenomes. a, Synonymous substitution divergence for all syntenic genes between the *F. iinumae* and *Fragaria* × *ananassa* genomes[4]. The median K_s divergence values for the seven chromosomes previously assigned to each progenitor species are plotted. The *F. iinumae* and *F. vesca* subgenomes exhibit the lowest and highest K_s divergence, respectively. **b**–**d**, K_s analysis of *F. iinumae* (**b**), *F. viridis* (**c**) and *F. nipponica* (**d**) transcriptomes against the phylogenetically supported homoeolog in the octoploid genome. The K_s distributions of *F. viridis* and *F. nipponica* transcriptomes are both unique and distinct from that of *F. iinumae*.

in-paralogs, yielded results consistent with those presented in our previous study (Fig. 1; Supplementary Information and Supplementary Dataset 1). Furthermore, their alternative model for the origin of octoploid strawberry (1× *F. vesca*–like and 3× *F. iinumae*–like subgenomes) is not supported by comparative genomic analyses of a new chromosome-scale *F. iinumae* genome (Fig. 2).

Phylogenetic analysis of the subgenome tree-searching algorithm searched a set of gene trees to identify sequences most closely related to a set of user-provided paralogs (or homoeologs in polyploids). Homoeologs are orthologous genes that were brought back into the same nucleus by allopolyploidization[6]. For our analyses, we used syntenic (that is, positionally conserved) homoeologs that were present on all subgenomes in octoploid strawberry. Gene trees were estimated using RAxML[7] based on orthologs identified using established orthogrouping approaches[8] applied to de novo assembled transcriptomes for each diploid *Fragaria* species[4]. PhyDS performs a relatively simple and straightforward analysis of gene trees. First, it identifies the user-provided paralog present in a gene tree and then moves to the direct ancestral node of the paralog. Second, PhyDS then returns to the user the direct descendants (that is, sequence identities including the paralog) of that ancestral node with its bootstrap support value (Fig. 1).

We have two major concerns regarding the methods used in refs. [2,5]. First, phylogenetic analyses aimed at estimation of species relationships are reliant first on correct identification of orthologs[9]. These authors used a sequence similarity-based approach to identify

putative orthologs that has relatively high error rates[10]. Furthermore, pangenome studies have shown that up to one-half of gene content exhibits presence–absence variation at the species level in plants[11]. In other words, many genes are individual- or population-specific. Thus, many of the putative ortholog predictions in their studies may be inaccurate. Second, Liston et al.[5] performed analyses of 100-kb windows across each of the seven base chromosomes. This could be problematic because chromosomal regions from one parental species can be replaced with chromosomal regions from the other parental species during meiosis in polyploids (referred to as homoeologous exchanges[12]). Homoeologous exchanges can range in size from large megabase-sized regions to single genes (see a recent review on its impact on subgenome assignment in ref. [13]). We identifed extensive homoeologous exchanges throughout the octoploid strawberry genome[4]. Thus, the 100-kb windows Liston et al. used consist of genes with different evolutionary histories reflecting each of the different progenitor species. This could result in inaccurate estimates of species relationships.

Here we present a chromosome-scale genome of *F. iinumae* with a scaffold minimum scaffold length needed to cover 50% of the genome of 33.98 Mb and 23,665 protein-coding genes (see Supplementary Information). This genome was used to calculate the synonymous substitution (K_s) divergence between *F. iinumae* to each of the four subgenomes (Fig. 2a). This revealed that only one of the subgenomes of octoploid strawberry is *F. iinumae*–like, which does not support the model presented by Liston et al.[5] that the origin of octoploid strawberry involved three *F. iinumae*–like and one *F. vesca*–like progenitor species. Instead, these results are consistent with our phylogenetic estimates supporting more than two diploid progenitors (Fig. 2b–d). The *F. viridis* (Fig. 2c) and *F. nipponica* (Fig. 2d) subgenomes are not *F. iinumae*–like.

Our new phylogenetic analyses support four distinct progenitor species, which is consistent with our previous results[4] and that of other groups[3]. The conflicting results obtained by Liston et al.[5] are probably due to differences in methodology. As pointed out above, establishing gene orthology is crucial for molecular phylogenetics. Our pipeline started by identifying high-confidence syntenic 1:1 homoeologs present on each of the subgenomes. This step alone filtered out 82.1% of genes from the octoploid strawberry genome[4]. The number of genes analyzed in our study was further reduced due to absence across transcriptome data, stringent orthogroup filtering and bootstrap value filtering. In short, more data are not always better if one introduces 'phylogenetic noise'. It is unclear to us how Liston et al.[5] obtained high unique mapping rates (~89% alignment) across the *F. vesca* genome, which consists of ~31% transposable elements and hundreds of duplicate genes. Furthermore, many genes are species-specific based on previous pangenome studies.

As pointed out by Liston et al.[5], incomplete lineage sorting can impact phylogenetic inferences. However, that is far more likely to impact within-species than between-species estimates. This is exactly what was observed in our study. Other *F. vesca* subspecies were identified as contributors but were present at notably lower levels than *F. viridis* and *F. nipponica* (Fig. 1a). These patterns provide further support for *F. viridis* and *F. nipponica* as extant relatives of the progenitors that contributed to the origin of the intermediate hexaploid ancestor. Lastly, we did state that *F. moschata* may be an extant relative of the intermediate hexaploid ancestor. Given the high frequency of polyploid formation in *Fragaria*[14] and birth–death dynamics of polyploids[15], we agree it is possible that the hexaploid ancestor may be extinct. This remains to be properly evaluated using robust phylogenetic approaches and datasets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0544-2.

Received: 22 May 2019; Accepted: 4 November 2019;
Published online: 16 December 2019

References

1. Rousseau-Gueutin, M. et al. Tracking the evolutionary history of polyploidy in *Fragaria* L. (strawberry): new insights from phylogenetic analyses of low-copy nuclear genes. *Mol. Phylogenet. Evol.* **51**, 515–530 (2009).
2. 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).
3. 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).
4. Edger, P. P. et al. Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* **51**, 541–547 (2019).
5. Liston, A. et al. Revisiting the origin of the octoploid strawberry. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0543-3 (2019).
6. Glover, N. M., Redestig, H. & Dessimoz, C. Homoeologs: what are they and how do we infer them? *Trends Plant Sci.* **21**, 609–621 (2016).
7. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
8. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
9. Duarte, J. M. et al. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10**, 61 (2010).
10. Nichio, B. T. L., Marchaukoski, J. N. & Raittz, R. T. New tools in orthology analysis: a brief review of promising perspectives. *Front. Genet.* **8**, 165 (2017).
11. Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
12. Xiong, Z., Gaeta, R. T. & Pires, J. C. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl Acad. Sci. USA* **108**, 7908–7913 (2011).
13. 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). /4.
14. Hummer, K. The discovery and naming of the cascade strawberry (*Fragaria cascadensis*). *Kalmiopsis* **21**, 26–31 (2015).
15. Mayrose, I. et al. Recently formed polyploid plants diversify at lower rates. *Science* **333**, 1257 (2011).

Reporting Summary

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

Data availability

The phylogenetic trees and alignments are available on Dryad (https://doi.org/10.5061/dryad.b2c58pc). The genome assembly and annotation files are available on the Genome Database for Rosaceae (https://www.rosaceae.org/) and NCBI GenBank under BioProjects PRJNA544784 and PRJNA508389. The raw sequence data are available in the Sequence Read Archive under the same NCBI BioProject numbers, PRJNA544784 and PRJNA508389.

Code availability

Custom software for running PhyDS phylogenetic analyses is available on GitHub (https://github.com/mrmckain/PhyDS/).

Acknowledgements

We thank J. Lei and L. Xue for sample preparation of *F. iinumae*. This work was supported by Michigan State University AgBioResearch (to P.P.E.), USDA-NIFA HATCH (no. 1009804 to P.P.E.), USDA-NIFA (no. SCRI 2014-51181-22378) and NSF-DEB (no. 1737898) to P.P.E., USDA-NIFA (no. SCRI 2017-51181-26833 to S.J.K.), the California Strawberry Commission (to S.J.K.), the University of California (to S.J.K.) and the National Natural Science Foundation of China (nos. 31770408 to T.Z. and 31760082 to Q.Q.).

Author contributions

P.P.E., M.R.M., A.E.Y., S.J.K., Q.Q. and T.Z. perfomed research and/or analyzed data. P.P.E. and M.R.M. drafted the manuscript. P.P.E., M.R.M., A.E.Y., S.J.K., Q.Q. and T.Z. reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

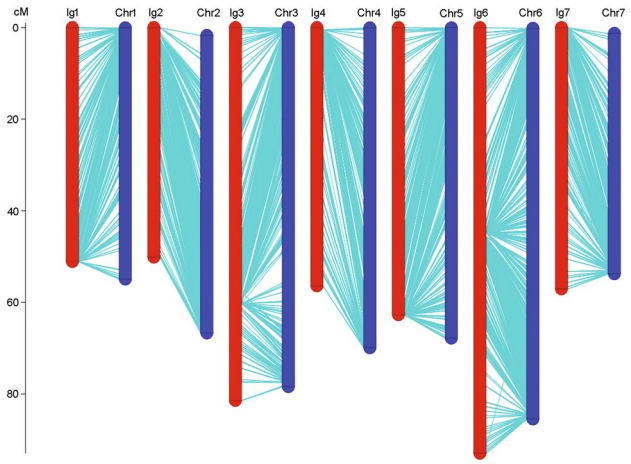
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-019-0544-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0544-2.

Correspondence and requests for materials should be addressed to P.P.E., Q.Q. or T.Z.

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



Extended Data Fig. 1 | Anchoring genome sequence to the genetic map. Previously a high-density linkage map of *F. iinumae* was constructed by 4173 markers, with 3280 from the Array and 893 from genotyping by sequencing[7]. Here we anchored the contigs to this genetic map to obtain a chromosome-scale genome of *F. iinumae*.

nature research

Corresponding author(s): Patrick P. Edger, Qin Qiao, Ticao Zhang,

Last updated by author(s): October 29th, 2019

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.

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	All commericial DNA and RNA sequencing platforms used in this study are fully described.
Data analysis	All 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. PhyDS; (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. 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

Raw sequencing data is available under accession code BioProjects PRJNA544784 and PRJNA508389 in the Sequence Read Archive on NCBI (https://www.ncbi.nlm.nih.gov/sra). Genome assemblies and annotations are on NCBI GenBank (https://www.ncbi.nlm.nih.gov/genome) under the same BioProjects and also available on the Genome Database for Rosaceae (https://www.rosaceae.org/).

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	Phylogenetic analyses were conducted of hundreds of orthologous genes. The genome of a single individual of Fragaria iinumae was sequenced.
Data exclusions	No data was excluded from any analysis, unless described in the manuscript.
Replication	Boostrap analyses were performed as described in the manuscript.
Randomization	Randomizations, included bootstrap analyses, to assess the level of support for each clade in every gene tree.
Blinding	A blinded-experiment is not possible for phylogenetic or genomic analyses.

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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