SUPPLEMENTARY MATERIAL

Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary

Kevin Vanneste[1,2], Guy Baele[3], Steven Maere[1,2,*],

and Yves Van de Peer[1,2,4,*]

[1] Department of Plant Systems Biology, VIB, Ghent, Belgium

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[3] Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium

[4] Department of Genetics, Genomics Research Institute, University of Pretoria, Pretoria, South Africa

***Corresponding authors**

Yves Van de Peer
VIB / Ghent University
Technologiepark 927
Gent (9052), Belgium
Tel: +32 (0)9 331 3807
Fax: +32 (0)9 331 3809
E-mail: yves.vandepeer@psb.vib-ugent.be

Steven Maere
VIB / Ghent University
Technologiepark 927
Gent (9052), Belgium
Tel: +32 (0)9 331 3805
Fax: +32 (0)9 331 3809
E-mail: steven.maere@psb.vib-ugent.be

Overview

Supplementary Table S1 ….. 3

Supplementary Figure S1 ….. 4

Supplementary Figure S2 ….. 8

Supplementary Figure S3 ….. 15

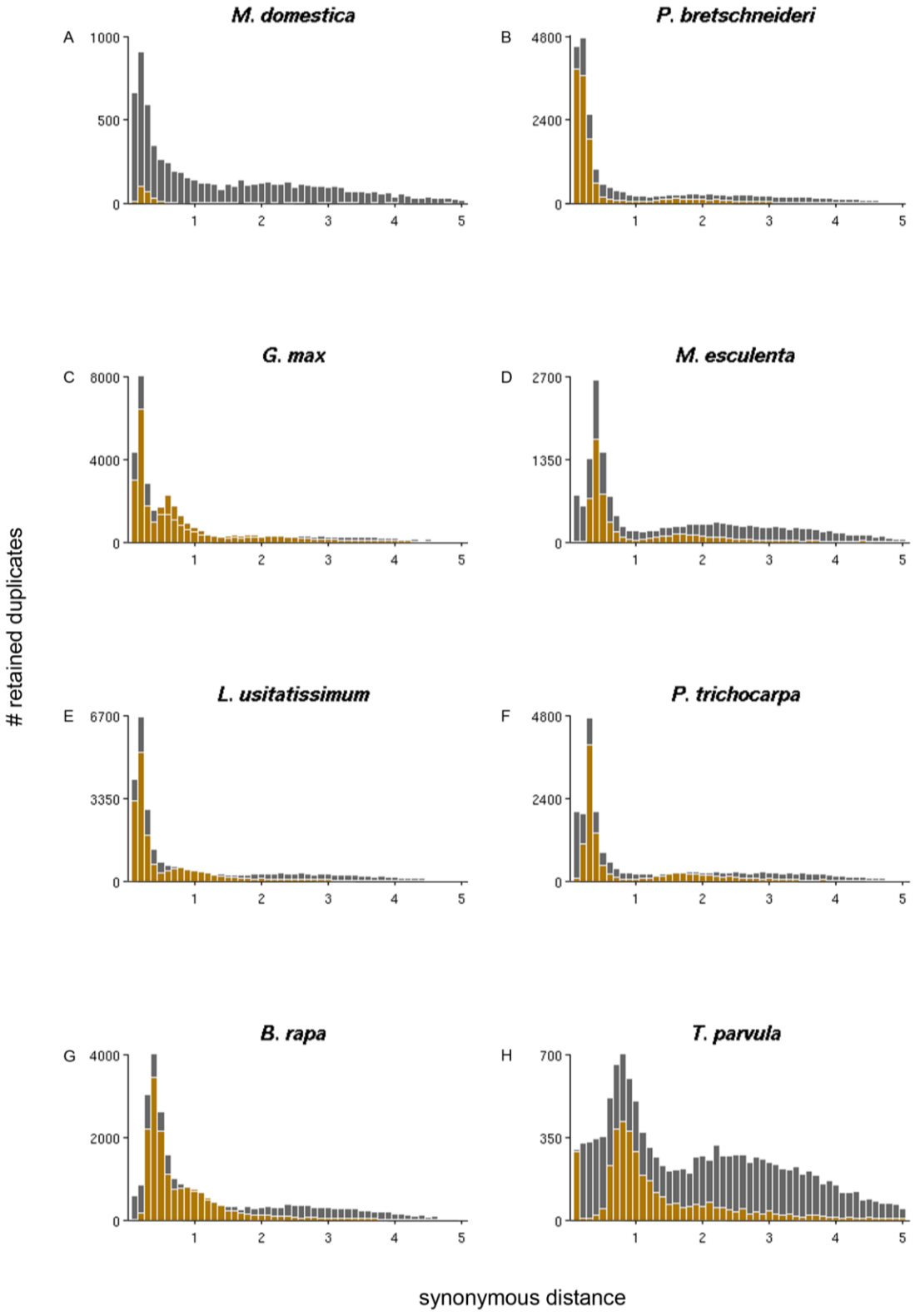
Supplementary Figure S4 ….. 17

Supplementary Table S1

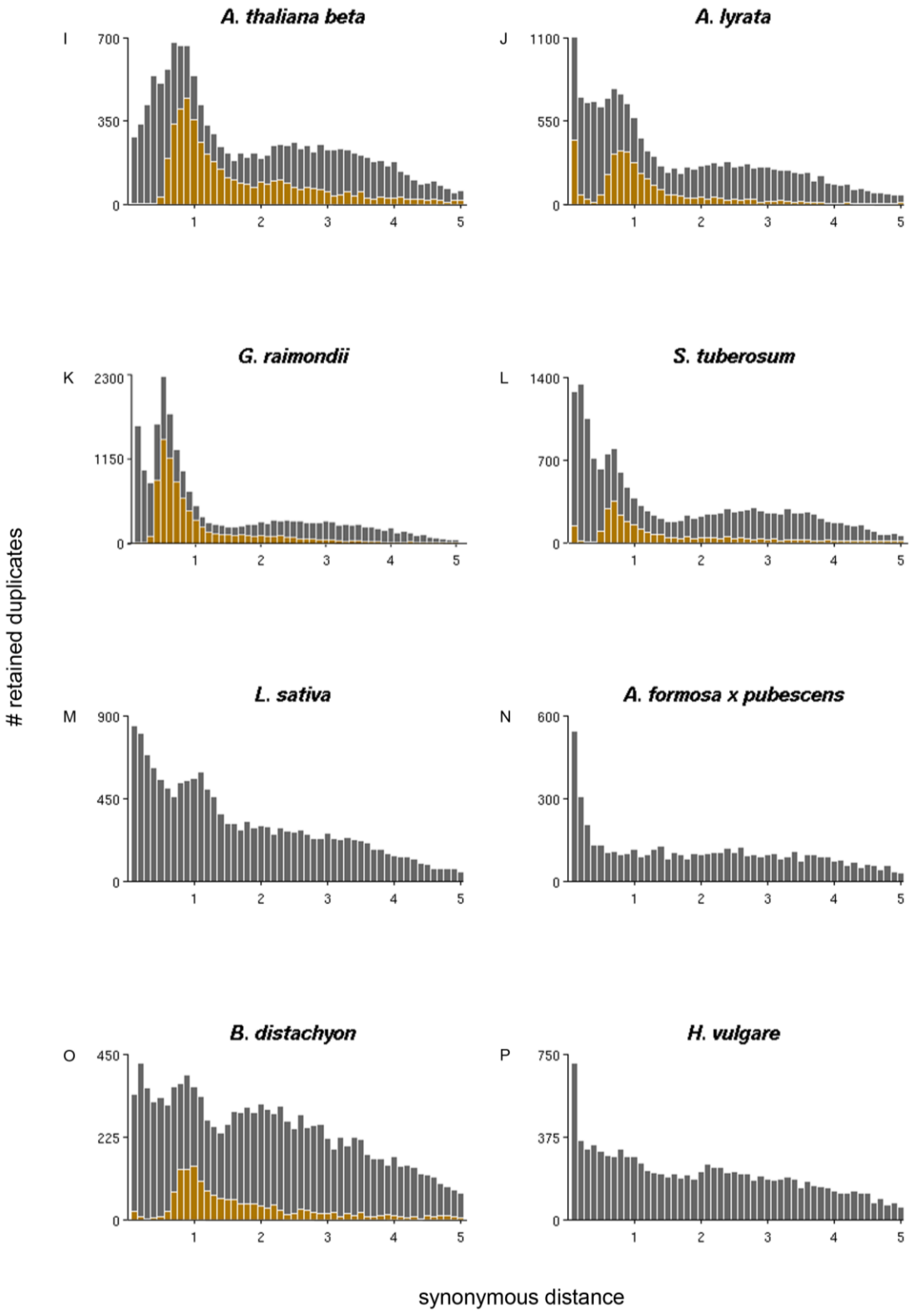
Supplementary Table S1 - Overview of all employed species and their sequence sources.

Species	Release	Source
Aquilegia formosa x pubescens	PLANTGDB (v187a)	http://www.plantgdb.org
Arabidopsis lyrata	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Arabidopsis thaliana	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Brachypodium distachyon	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Brassica rapa	Phytozome (v8)	http://www.phytozome.net
Cajanus cajan	IIPG (v5)	http://www.icrisat.org/gt-bt/iipg/Genome_Manuscript.html
Carica papaya	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Cicer arietinum	LIS (v1)	http://cicar.comparative-legumes.org
Citrullus lanatus	BGI (v1)	http://www.icugi.org/cgi-bin/ICuGI/index.cgi
Cucumis melo	MELONOMICS (v3.5)	https://melonomics.net
Cucumis sativus	BGI (v2)	http://www.icugi.org/cgi-bin/ICuGI/index.cgi
Fragaria vesca	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Glycine max	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Gossypium raimondii	BGI (v1)	http://cgp.genomics.org.cn
Hordeum vulgare	IBSC (v1)	http://www.public.iastate.edu/~imagefpc/IBSC Webpage
Jatropha curcas	JGD (v4.5)	http://www.kazusa.or.jp/jatropha
Lactuca sativa	PLANTGDB (v187a)	http://www.plantgdb.org
Linum usitatissimum	Phytozome (v8)	http://www.phytozome.net
Lotus japonicus	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Malus domestica	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Manihot esculenta	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Medicago truncatula	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Musa acuminata	Genoscope (v1)	http://banana-genome.cirad.fr
Nuphar advena	AAGP (v3)	http://ancangio.uga.edu/content/nuphar-advena
Oryza sativa	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Phoenix dactylifera	Weill Cornell Medical College (v3)	http://qatar-weill.cornell.edu/research/datepalmGenome
Phyllostachys heterocycla	ICBR (v1.0)	http://202.127.18.221/bamboo/index.php
Physcomitrella patens	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Populus trichocarpa	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Prunus mume	BGI (v1)	http://prunusmumegenome.bjfu.edu.cn
Prunus persica	Phytozome (v8)	http://www.phytozome.net
Pyrus bretschneideri	BGI (v1)	http://peargenome.njau.edu.cn
Ricinus communis	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Setaria italica	Phytozome (v8)	http://www.phytozome.net
Solanum lycopersicum	ITAG (v2.3)	http://solgenomics.net/organism/Solanum_lycopersicum/genome
Solanum tuberosum	ITAG (v1)	http://solgenomics.net/organism/Solanum_tuberosum/genome
Sorghum bicolor	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Thellungiella parvula	Thellungiella Consortium (v2)	http://thellungiella.org
Theobroma cacao	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Vitis vinifera	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza
Zea mays	PLAZA (v2.5)	http://bioinformatics.psb.ugent.be/plaza

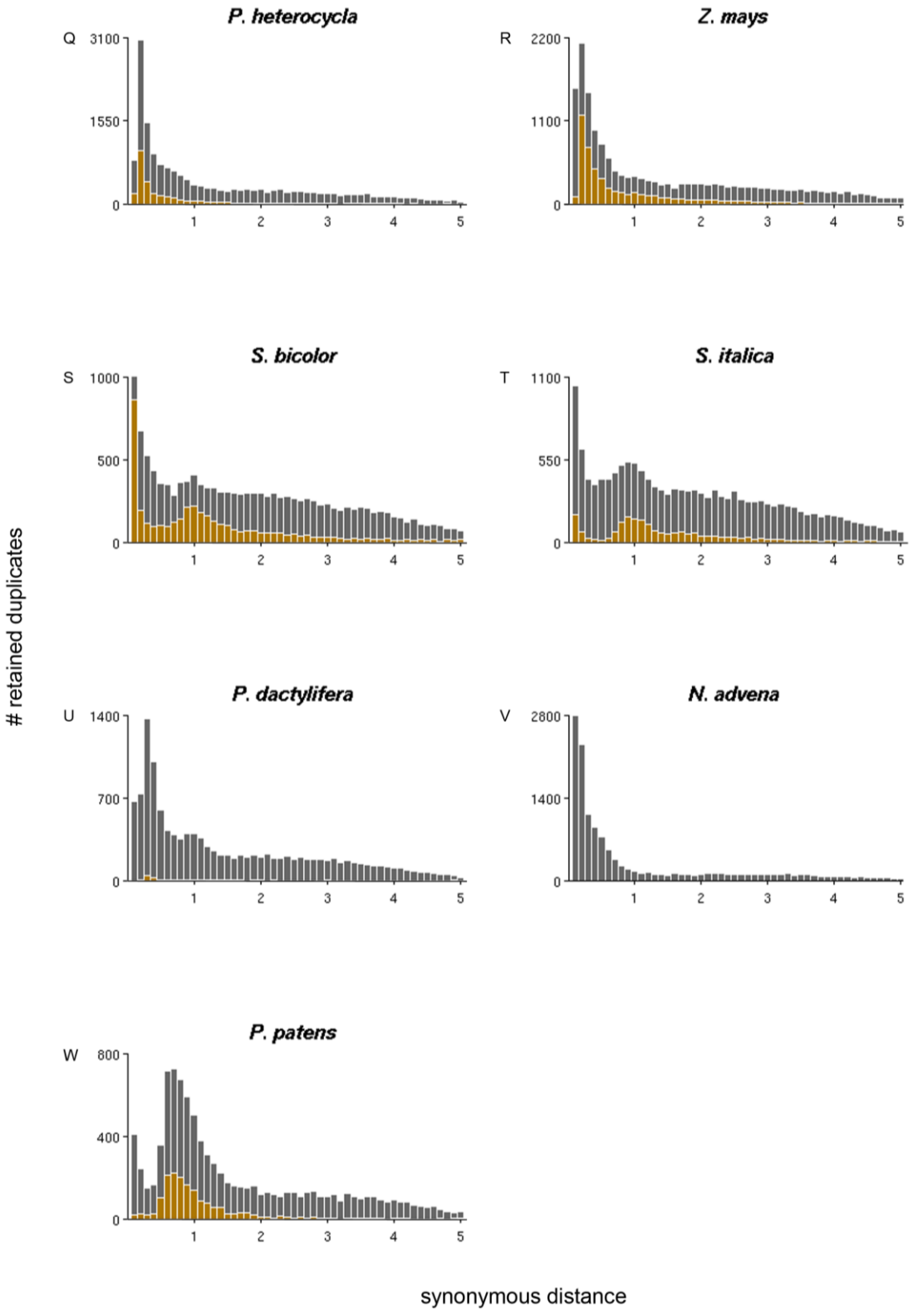
Supplementary Figure S1



Supplementary Figure S1 (continued)

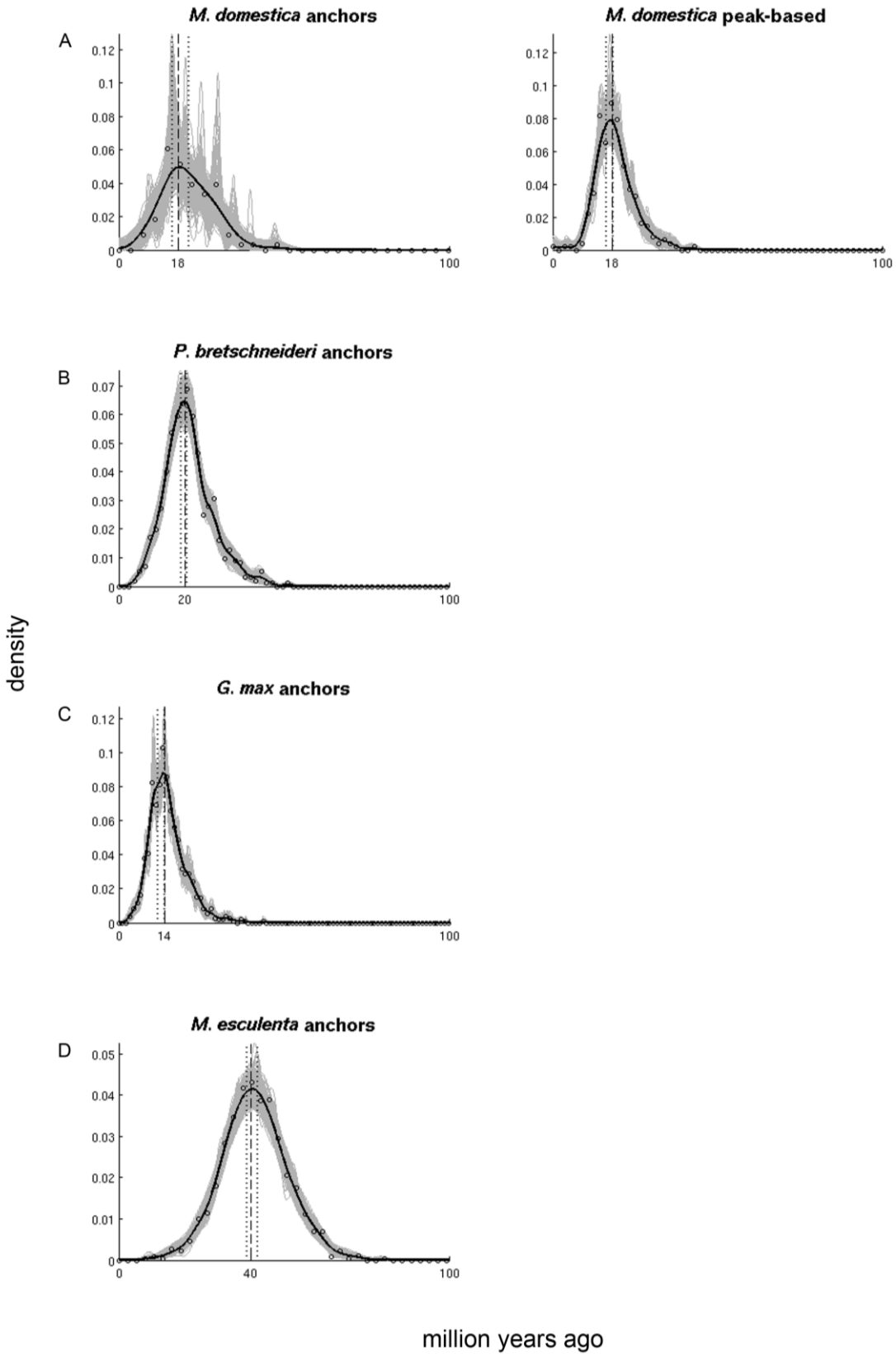


Supplementary Figure S1 (continued)

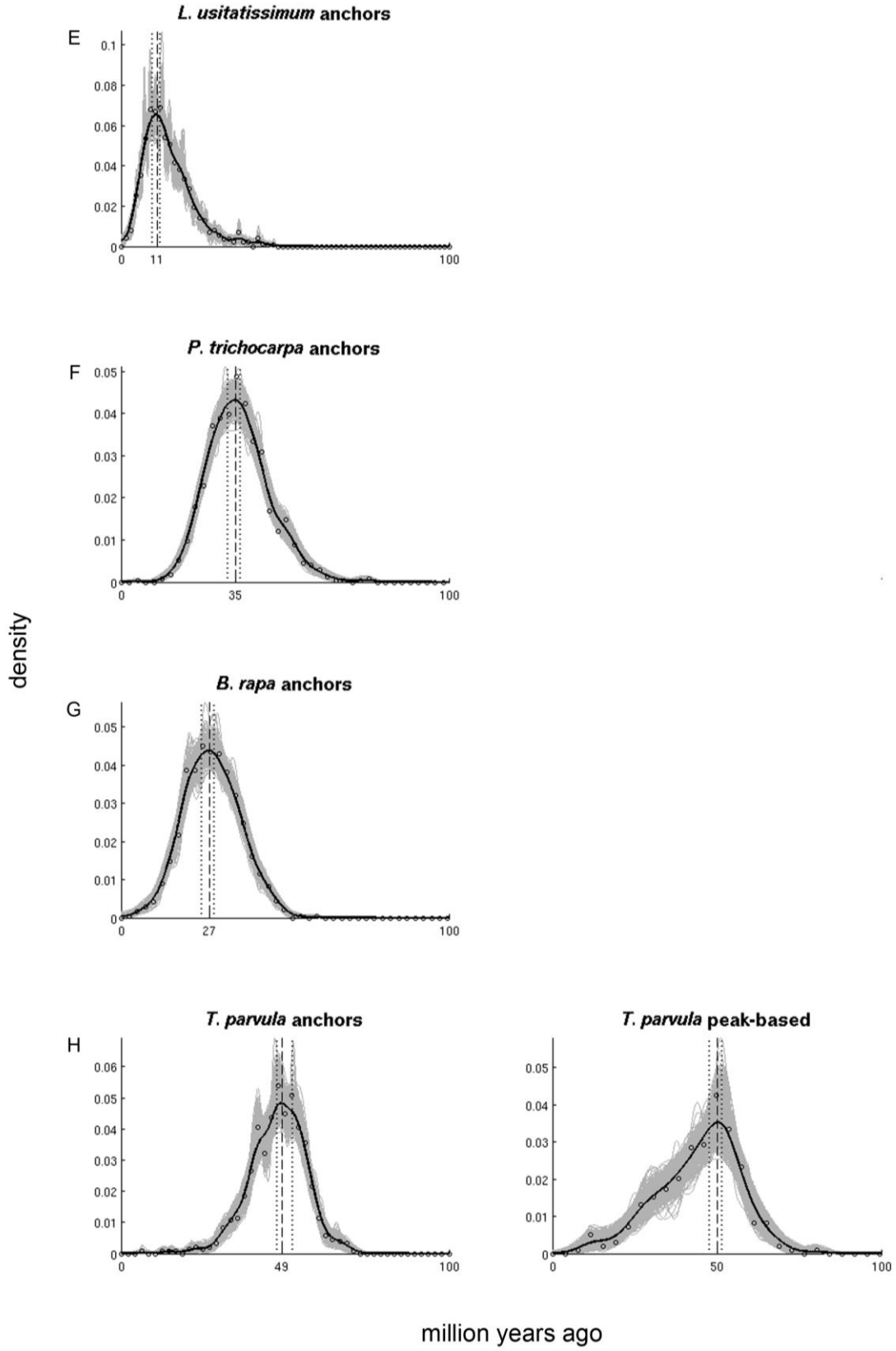


Supplementary Figure S1 - K_S age distributions for (**A**) *M. domestica*, (**B**) *P. bretschneideri*, (**C**) *G. max*, (**D**) *M. esculenta*, (**E**) *L. usitatissimum*, (**F**) *P. trichocarpa*, (**G**) *B. rapa*, (**H**) *T. parvula*, (**I**) *A. thaliana beta*, (**J**) *A. lyrata*, (**K**) *G. raimondii*, (**L**) *S. tuberosum*, (**M**) *L. sativa*, (**N**) *A. formosa x pubescens*, (**O**) *B. distachyon*, (**P**) *H. vulgare*, (**Q**) *P. heterocycla*, (**R**) *Z. mays*, (**S**) *S. bicolor*, (**T**) *S. italica*, (**U**) *P. dactylifera*, (**V**) *N. advena*, and (**W**) *P. patens*. The grey and beige bars represent the distribution of the paranome and duplicated anchors identified with i-ADHoRe, respectively. Anchors and peak-based duplicates used as homeologs for absolute dating were extracted between the WGD peak boundaries (see Table 1).

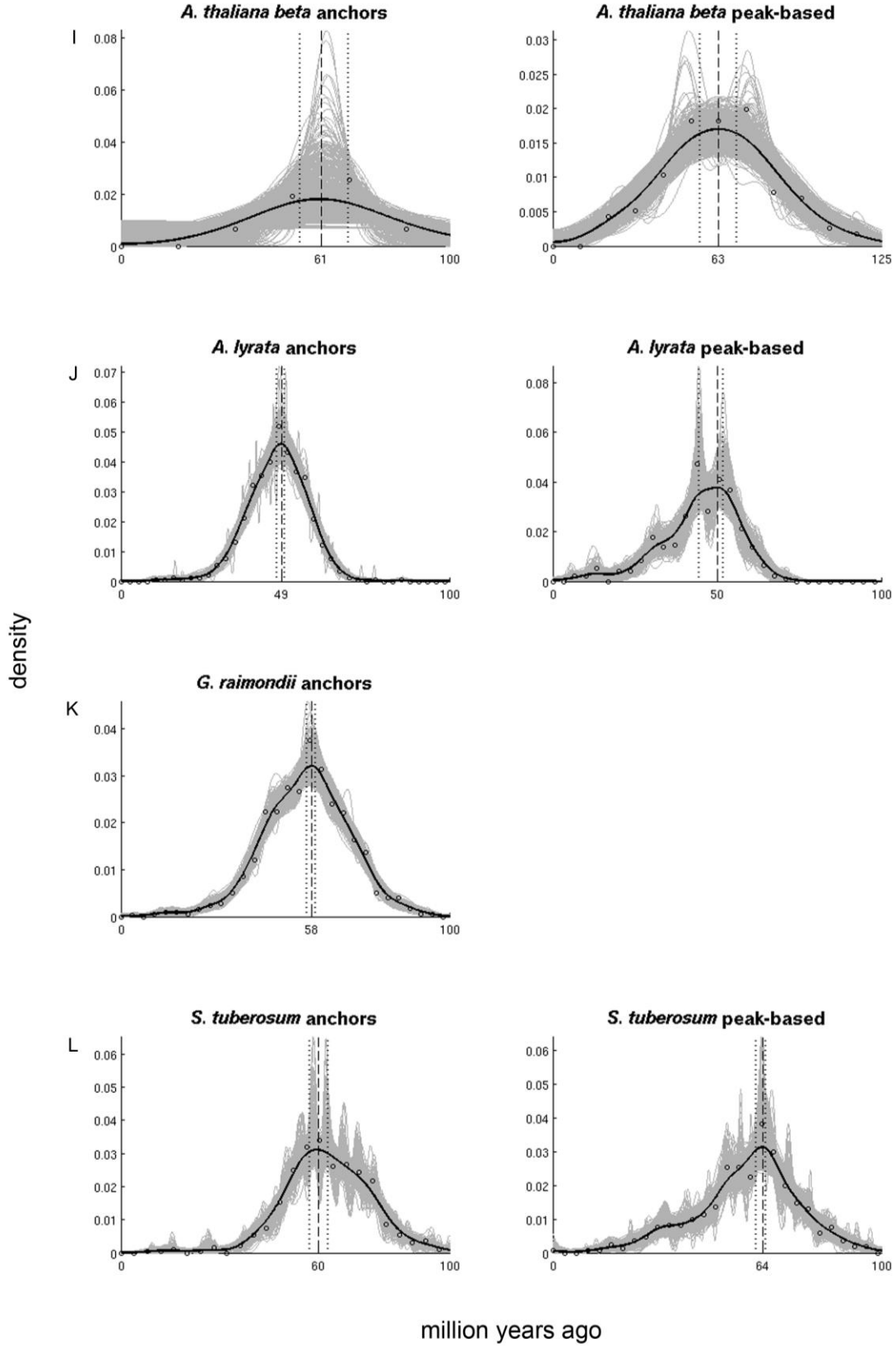
Supplementary Figure S2



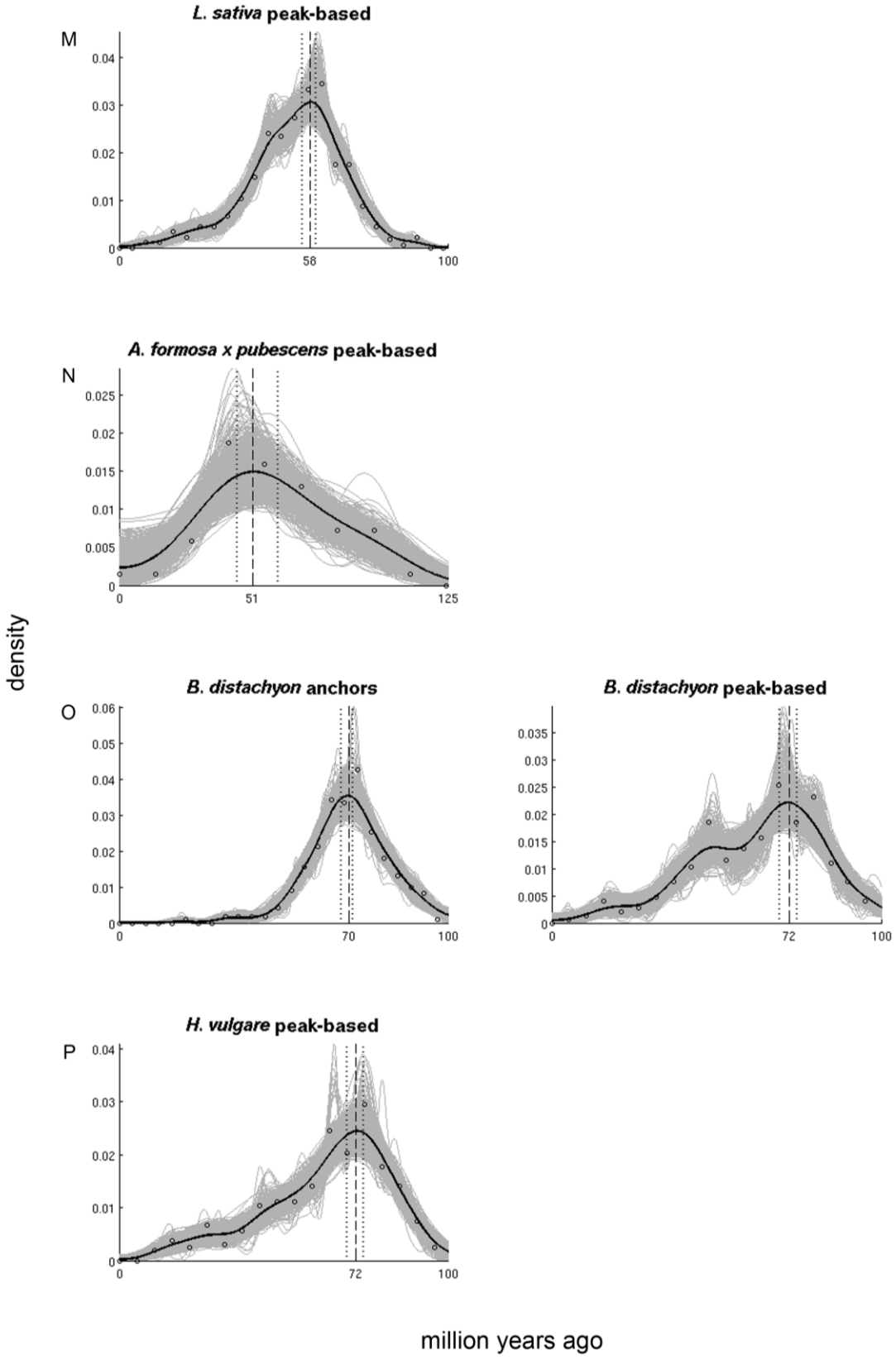
Supplementary Figure S2 (continued)



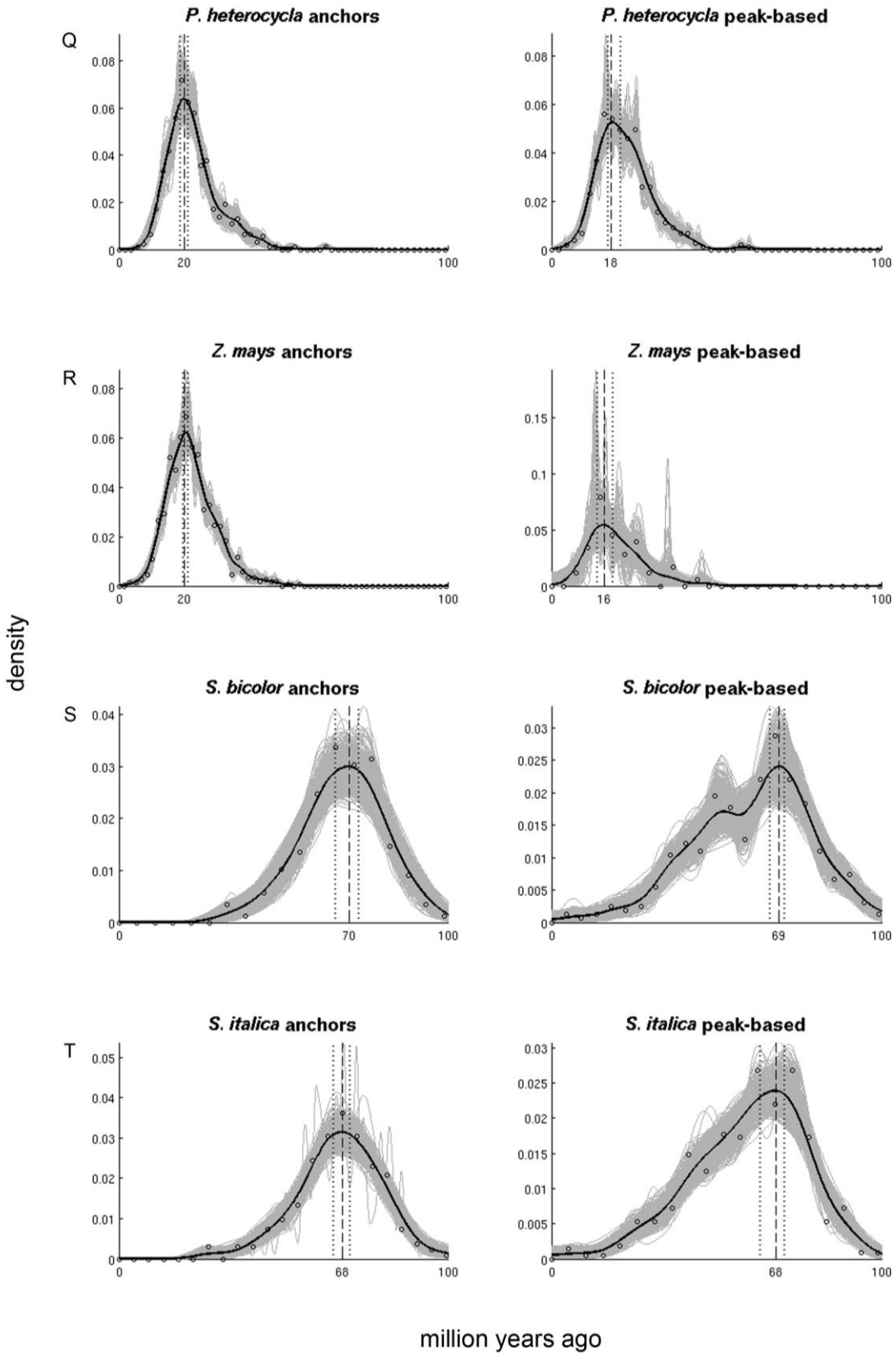
Supplementary Figure S2 (continued)



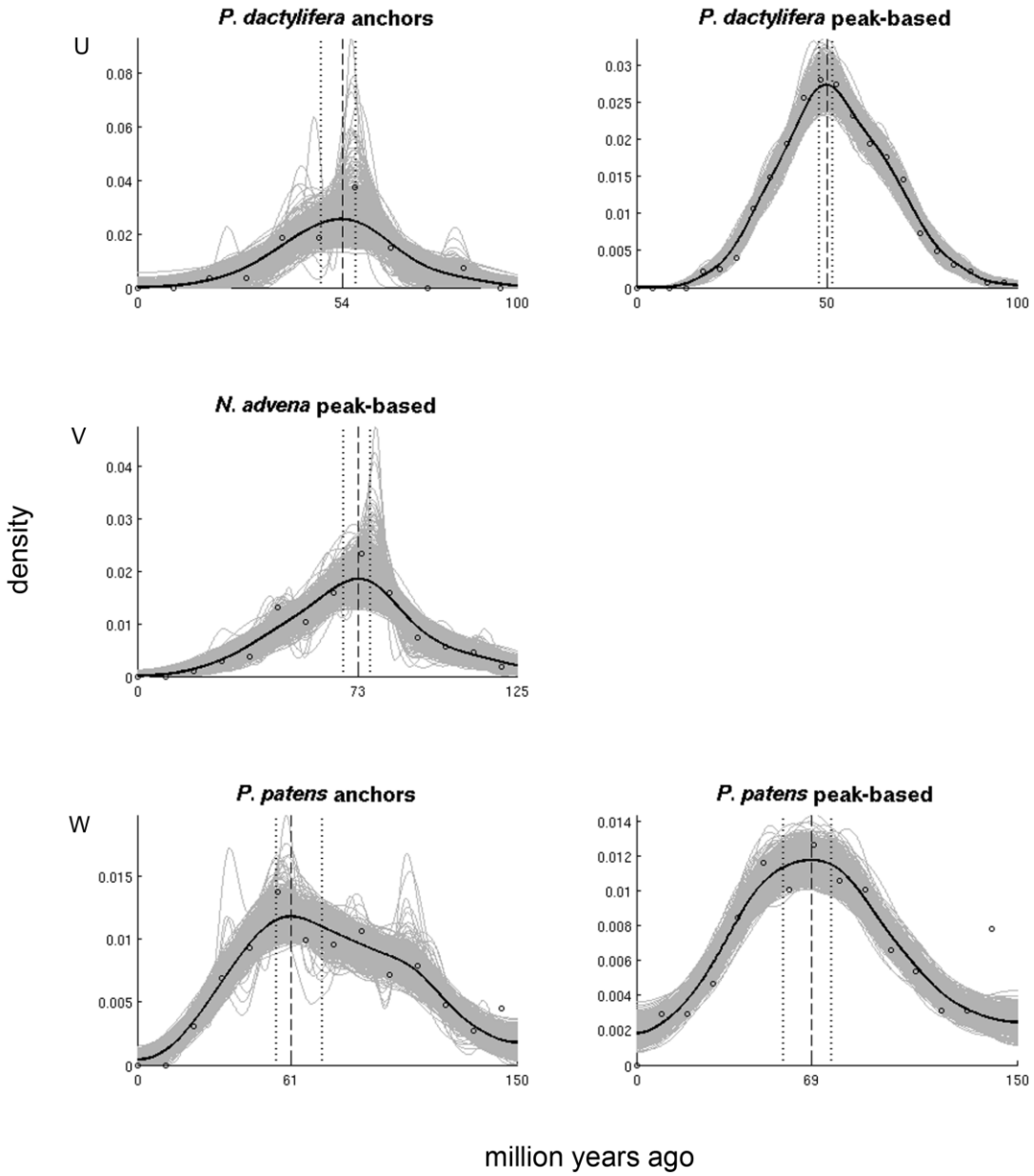
Supplementary Figure S2 (continued)



Supplementary Figure S2 (continued)

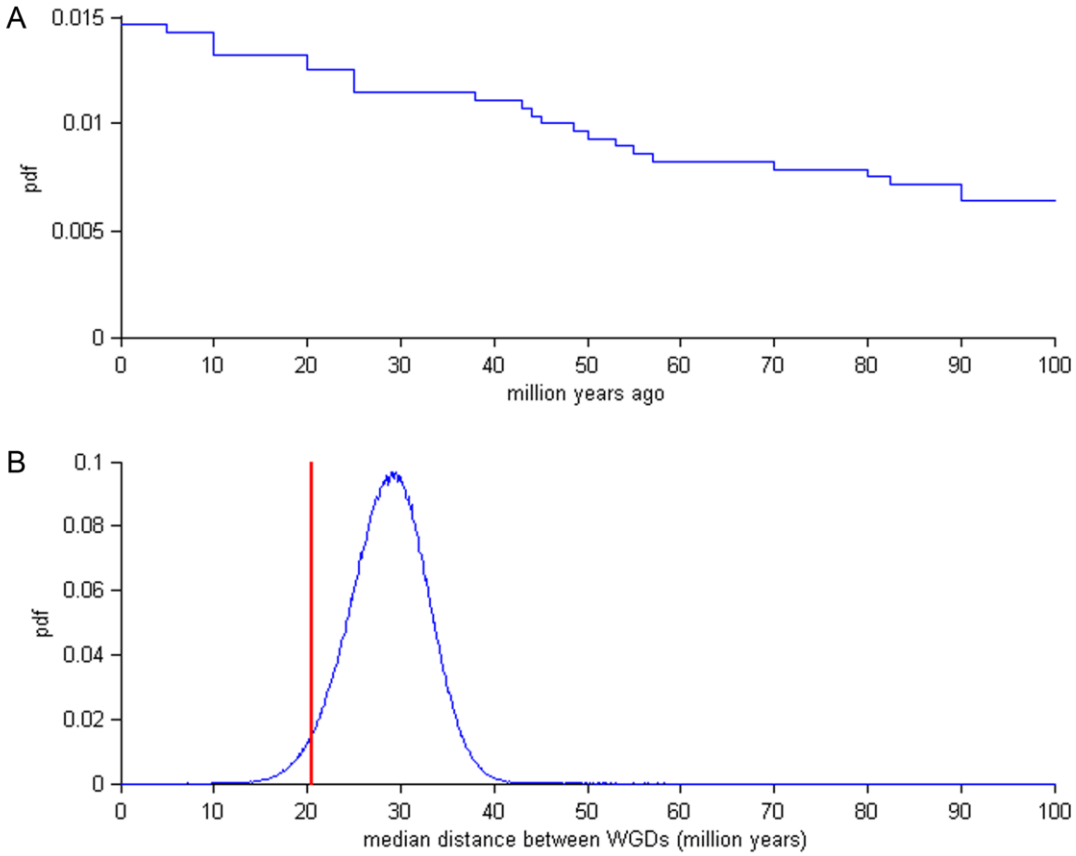


Supplementary Figure S2 (continued)



Supplementary Figure S2 - Absolute age distributions of dated anchors and peak-based duplicates, where applicable (see Table 1), for (**A**) *M. domestica*, (**B**) *P. bretschneideri*, (**C**) *G. max*, (**D**) *M. esculenta*, (**E**) *L. usitatissimum*, (**F**) *P. trichocarpa*, (**G**) *B. rapa*, (**H**) *T. parvula*, (**I**) *A. thaliana beta*, (**J**) *A. lyrata*, (**K**) *G. raimondii*, (**L**) *S. tuberosum*, (**M**) *L. sativa*, (**N**) *A. formosa x pubescens*, (**O**) *B. distachyon*, (**P**) *H. vulgare*, (**Q**) *P. heterocycla*, (**R**) *Z. mays*, (**S**) *S. bicolor*, (**T**) *S. italica*, (**U**) *P. dactylifera*, (**V**) *N. advena*, and (**W**) *P. patens*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey solid lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See Table 1 for sample sizes and exact confidence interval boundaries.

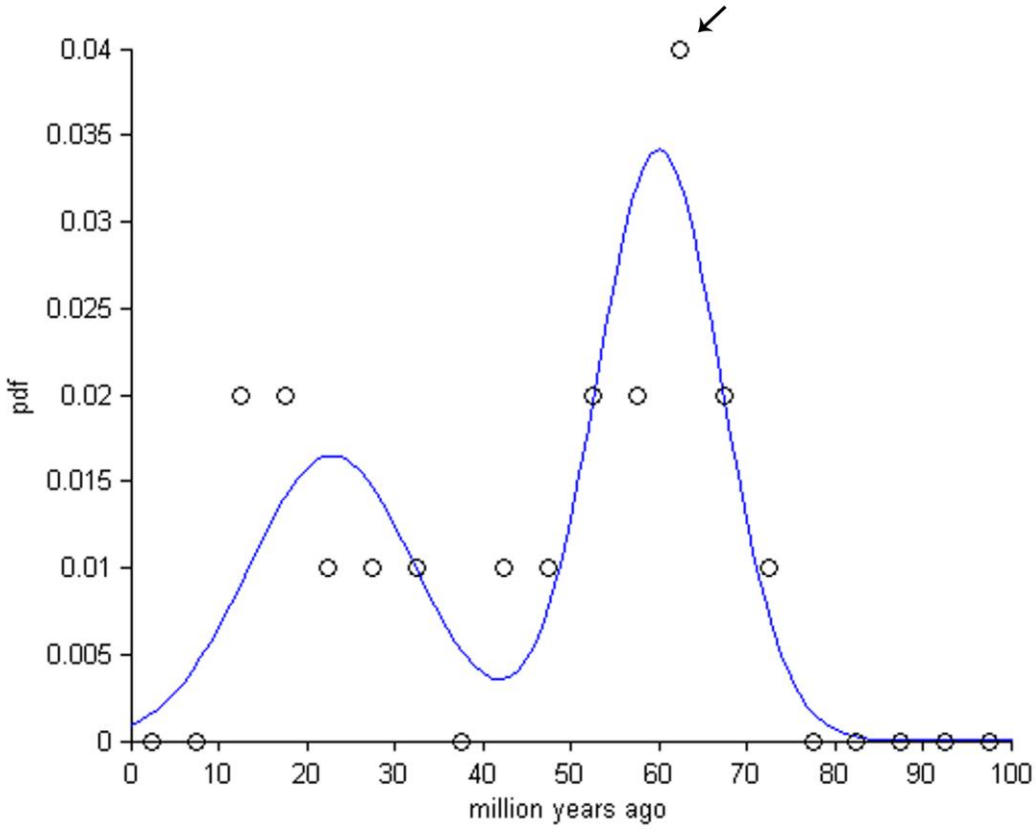
Supplementary Figure S3



Supplementary Figure S3 - (**A**) Probability density function (pdf) for the null model of random WGD occurrence over time. An interval between 0 and 100 mya is considered. Each discontinuity in the pdf corresponds to a speciation event in Figure 3, and the probability of WGD occurrence at a certain point in time is proportional to the total number of species present at that time. (**B**) Assessment of the statistical significance of WGD clustering in time. The true median distance between WGD age estimates presented in Table 1 is indicated by the vertical red line (true median WGD distance = 20.42 million years). Note that shared WGDs were only counted once by taking the average of anchor-based WGD age estimates, or peak-based WGD age estimates if the former were not available, in their descendant species. The distribution of one million random samples is indicated in blue. Each sample is represented by a median WGD distance that was calculated based on pulling WGD ages randomly from the null model in **A** (average random median WGD distance = 28.65 million

years). The true median WGD distance was significantly lower than expected under the null model (*P*-value = 0.0301), indicating that plant paleopolyploidizations cluster statistically significantly in time. Exclusion of the *M. acuminata* WGD, because this most likely represents two WGDs in close succession (see Results and discussion), does not change these results although exclusion of the latter does decrease statistical significance (*P*-value = 0.0430).

Supplementary Figure S4



Supplementary Figure S4 - Probability density function (pdf) of WGD age estimates. The blue curve represents the fit of a mixture of Gaussians that was used to find where WGDs cluster in time (see supplementary Figure S3; Material and methods). A mixture of two components was selected according to the AIC criterion (AIC = 174.90 compared to AIC = 180.33 and 177.96 for a mixture with one and three components, respectively). The total probability of WGD occurrence between 0 and 100 mya is equal to one (i.e., the sum of everything under the blue curve, its integral, sums to one). Note that shared WGDs were only counted once by taking the average of anchor-based WGD age estimates, or peak-based WGD age estimates if the former were not available, in their descendant species. The mixture contains one relatively thin and high component with a peak located at 60.05 mya, corresponding to the clustering of WGDs with the K-Pg boundary, and a broader and lower component with a peak located at 22.91 mya. The raw data is also presented on the figure by open circles. Every circle indicates the relative frequency of WGDs falling within an age bin of 5 million years (i.e., the first circle is located at 2.5 mya and represents the relative frequency of all WGDs

falling between 0 and 5 mya etc.). Note that the particular bin size of 5 million years was arbitrarily chosen to allow a visual comparison of the raw data with the estimated fit of the Gaussian mixture, and does not influence the Gaussian mixture model fitting (i.e., the bin size does not have any influence on the shape of the mixture and its peak at 60.05 mya). The mixture demonstrates an overall good fit to the raw data, especially considering the relatively small sample size of only 20 independent WGDs. The open circle indicated with an arrow represents the relative frequency of WGDs falling between an interval of 60 and 65 mya. Exclusion of the *M. acuminata* WGD, because this most likely represents two WGDs in close succession (see Results and discussion), does not change these results (first and second peak located at 22.47 and 59.21 mya, respectively).