This is a post-print of an article published in Molecular Plant. The final authenticated version is available online at:
https://doi.org/10.1016/j.molp.2018.12.019

Finding evidence for whole genome duplications: a reappraisal

Arthur Zwaenepoel[1,2], Zhen Li[1,2], Rolf Lohaus[1,2] & Yves Van de Peer[1,2,3,*]

[1] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium
[2] Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium
[3] Center for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa

[*]Correspondence to Yves Van de Peer (yves.vandepeer@psb.ugent.be)

Dear Editor,

Recently, Ren *et al.* reported an extensive analysis of the incidence of whole genome duplications (WGDs) throughout the evolutionary history of extant angiosperms (Ren et al., 2018). Examining a wealth of genomic data (36 complete genomes and 69 transcriptomic data sets) using commonly applied methods, they detected and located 55 WGDs throughout the angiosperm phylogeny. Furthermore, they provided estimates of the dates for these important events and discuss correlations with global climatic change during the Cenozoic as well as species diversification shifts. However, we are concerned that methodological flaws and misinterpretations render some of their analyses unreliable. We would like to address these issues in this letter, as we feel these are of general importance for the field of plant evolutionary genomics.

Ren *et al.* used two common approaches to detect WGDs in genomic and transcriptomic data sets, namely gene tree – species tree reconciliation and age distributions of paralogs based on the estimated number of synonymous substitutions per synonymous site (K_S), where the number of gene duplicates is plotted against the age of the duplication event, measured by K_S, and where peaks in the distribution are thought to reflect signatures of potential WGD events (Lynch and Conery, 2000; Vanneste et al., 2013). Gene tree – species tree reconciliation methods have been widely used for the detection (and phylogenetic placement) of WGDs (Jiao et al., 2011; Li et al., 2015; McKain et al., 2016; Yang et al., 2018) and the main reason for their adoption was that (1) within-species co-linearity information — often considered the strongest evidence for WGD — requires high-quality genome assemblies, whereas (2) K_S distributions cannot be used to reliably detect very ancient WGDs and are not directly comparable between lineages due to differences in molecular evolutionary rates. The general idea of gene tree – species tree reconciliation methods is that a high concentration of gene duplication nodes reconciled onto a particular branch of the species tree can be considered as support for a potential WGD on that branch. We would like to stress here that these methods are not without serious caveats, related to, for example, the prevalence of small-scale duplication (SSD, see below), the assumption of a known gene and species tree and inherent biases in the reconciliation algorithms used (e.g. Hahn, 2007). The reliability of commonly employed reconciliation approaches for WGD inference remains to be explored. Therefore, analyses using these approaches should be interpreted with appropriate caution.

A major confounding factor in both reconciliation and K_S based WGD inference is the presence of duplicates originating from SSD events. The failure to account for this is the major issue underlying most of the analyses in the study by Ren *et al.* It has been observed that duplication ages (measured by K_S) from SSDs (only) are approximately exponentially distributed (somewhat 'L'-shaped), which is the expected outcome of a quasi-equilibrium birth-death process, with most extant small-scale duplicates being of fairly recent origin and few paralogs being retained from very old duplication events (Lynch and Conery, 2000; Blanc & Wolfe, 2004; Lynch, 2007). We stress that, under common molecular evolutionary assumptions, this exponential decay is expected to be present in any K_S distribution (i.e., independent of whether the genome did or did not have any WGD(s)); although in some rare cases, large numbers of retained duplicates from a very recent WGD event may overshadow this 'L'-shape signature in a visualization of the distribution (e.g., as in *Glycine max*, Supplementary Figure 1A, where structural genomic data confirmed this signature to be a WGD). In the absence of the structural information provided by genomes, it is impossible to discriminate whether individual duplicates are the result of an SSD or WGD event, especially for those of recent origin. For this reason, researchers who adopted reconciliation-based methods for WGD inference in the absence of structural information have generally avoided inferring recent lineage- or species-specific WGDs because of the abundancy of SSDs on such terminal branches (Jiao et al., 2011; Li et al., 2015; Yang et al., 2018). However, it seems that Ren *et al.* fail to recognize these issues and use gene tree – species tree reconciliation to infer 27 recent lineage-specific WGDs in their transcriptomic data sets, including 15 newly-identified ones. Inspection of the corresponding K_S distributions for these species should have alerted them that most of these events cannot be distinguished from recent rounds of SSD, i.e., from the 'L'-shaped signature caused by SSDs. However, it seems that the authors failed to interpret these correctly.

We do endorse a strategy to corroborate reconciliation-based findings with analysis of K_S distributions, as these can be more useful for identifying relatively recent, lineage-specific WGDs if their signature peaks in a K_S age distribution are clearly distinct from the background exponential decay of SSDs. In such cases these distinct peaks provide strong support for WGDs. However, almost all K_S distributions for the newly-identified WGDs (as well as some of the 'Calibrated WGDs', see below) in Ren *et al.* do not seem to display such a clear distinctive peak. It appears that the authors base their inference of WGD peaks on kernel density estimates (KDEs) of K_S values from paralogs falling on specific branches (panel B in Figure 1 and Supplemental Figures 4–65 of Ren *et al.*). Most of the distributions show a prominent KDE peak in the low-K_S region with a mode close to zero ($K_S \approx 0.1$). Kernel density estimation is a powerful non-parametric method to visualize the empirical distribution of a data set. However, Ren *et al.* seem not to correct for boundary effects that are well known to plague KDEs and tend to result in underestimation at the boundaries (here at $K_S = 0$). This effect is even aggravated when very low K_S values are filtered out, as is often done (to remove allelic variants for instance). We are convinced that accounting for these effects would lead them to conclude that most of these so-called peaks are KDE artifacts and actually reflect the exponentially-distributed SSDs (see Figure 1A & B for an example). This should have been obvious if the authors would have shown the associated histograms. Therefore, we strongly suspect that there is no conclusive evidence for novel WGDs on most tip branches, and the peak K_S values that are reported actually reflect the K_S values of the recent SSDs that map on this branch.

Additionally, Ren *et al.* followed a common strategy by fitting Gaussian mixture models (GMMs), and show significant Normal components as additional support for WGDs (we note however that Ren *et al.* did not use these GMMs for inference purposes). However, mixture models are likely to 'overcluster' the data (Naik et al., 2007),

especially when applied to K_S distributions where the number of data points is large. Also, there is a widespread misconception that peaks originating from WGDs are expected to show a Normal distribution. A simple model of synonymous substitution after WGD already indicates otherwise; if we consider synonymous substitution as a Poisson process, and synonymous substitution rates for different duplicate pairs sampled from some Gamma distribution, the expected distribution of number of synonymous substitutions will follow a Negative binomial distribution (which is well approximated in the continuous case by the Gamma or log-normal distribution). This indicates that the K_S distribution induced by a WGD will have a positive skew, and this effect will be stronger the more recent the WGD. Normal GMMs are not able to account for this effect, and neither can they cope with the background exponential decay from SSDs. As a result, one or even multiple spurious components will often be fitted to the heavy right tail of either a recent WGD or SSD peak (Figure 1 B), further strengthening misinterpretations such as found in the study of Ren *et al.*

Acknowledging the presence of SSDs and the described issues with mixture modeling and boundary effects in KDEs would have cautioned Ren *et al.* from identifying many of the recent WGDs they report, which would thoroughly alter their study as a whole. Indeed, the most important criterion on which Ren *et al.* based their final inference of 'newly identified' WGDs is dictated by their survivorship model, which is itself based on the median K_S values they inferred for previously reported putative WGDs, referred to as 'Calibrated WGDs' by the authors (Figure 2 & 3 (purple dots) and Supplemental Table 3 in Ren *et al.*). However, most of the peak (median) K_S values they estimated for these 'Calibrated WGDs' are inaccurate because of the reasons described above and either reflect SSD events or are at least strongly affected by these (Supplementary Table 1). Moreover, their estimates often differ considerably from previously reported values. We believe neglecting the effects of SSDs has caused Ren *et al.* to effectively fit an exponential decay distribution to a set of spurious 'peak' K_S values that themselves stem from or are affected by the exponential decaying distributions of SSDs. Consequently, many (though not necessarily all) of both the 'newly identified' and 'Calibrated' recent WGDs based on transcriptomic data are likely to be false, and reflect no more than the continuous birth and death of SSDs. We do not exclude the possibility that some of the proposed WGDs do in fact correspond to real events, however, we stress that this cannot be concluded from the types of analyses performed by Ren *et al.* (Supplementary Figure 1).

This also explains the inflation of apparent WGDs identified from transcriptomes that are dated at around 10 million years ago (Figure 6 in Ren *et al.*), mostly reflecting date estimates for spurious peaks stemming from SSDs. However, even if the chosen peak (median) K_S values would represent *bona fide* WGD peaks (as we might expect for most of the reported and often older WGDs based on genome data), we think the absolute dating approach used by Ren *et al.* is much too simplistic to allow any confident interpretation of the resulting estimates. The authors assume a strict global molecular clock as well as species divergence times known without uncertainty and report point estimates without confidence intervals, practices that have been well criticized for over a decade already (Graur and Martin, 2004). We also find it striking that the authors did not compare and discuss their newly-estimated dates with previously reported estimates which often differ by tens of millions of years (Cheng et al., 2018; Van de Peer et al., 2017; Vanneste et al., 2014). To conclude, we are convinced that the careless adoption of reconciliation methods, the misinterpretation of K_S distributions and model-fitting issues render the WGD inferences of Ren *et al.* highly unreliable. Consequently, we fear that any speculation by Ren *et al.* concerning links of polyploid establishment with climatic change and rates of species diversification is unjustified.

In summary, we fear that while Ren *et al.* took considerable effort in generating and analyzing huge amounts of data, their study suffers from flawed methodology and misinterpretations. We hope that by highlighting some of these issues here, our commentary will benefit future efforts in plant palaeopolyploidy research. With the wealth of genomic data that is now becoming available, many analyses can be performed in a high-throughput fashion. Nevertheless, awareness of the limitations of different approaches and careful interpretation of results remains critical. While we could not discuss all issues at length here (but see Supplementary Information), we hope to have clearly expressed our concerns and hope that the results presented in Ren *et al.* will be interpreted with appropriate caution.

Sincerely,

Arthur Zwaenepoel, Zhen Li, Rolf Lohaus and Yves van de Peer

References

1. Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., and Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. Nature Plants *4*, 258–268.
2. Blanc, G., and Wolfe, K.H. (2004). Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. The Plant Cell *16*, 1667–1678.
3. Graur, D., and Martin, W. (2004). Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet. *20*, 80–86.
4. Hahn, M.W. (2007). Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol. *8*, R141.
5. Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.
6. Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., and Barker, M.S. (2015). Early genome duplications in conifers and other seed plants. Sci. Adv. *1*.
7. Lynch, M. (2007). The origins of genome architecture (Sunderland, Mass.: Sinauer Associates).
8. Lynch, M., and Conery, J.S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. Science *290*, 1151–1155.
9. McKain, M.R., Tang, H., McNeal, J.R., Ayyampalayam, S., Davis, J.I., dePamphilis, C.W., Givnish, T.J., Pires, J.C., Stevenson, D.W., and Leebens-Mack, J.H. (2016). A Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales. Genome Biol. Evol. *8*, 1150–1164.
10. Naik, P.A., Shi, P., and Tsai, C.-L. (2007). Extending the Akaike Information Criterion to Mixture Regression Models. Journal of the American Statistical Association *102*, 244–254.
11. Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., Ma, H., and Qi, J. (2018). Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. Mol. Plant *11*, 414–428.
12. Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. Nat. Rev. Genet. *18*, 411–424.
13. Vanneste, K., Van de Peer, Y., and Maere, S. (2013). Inference of Genome Duplications from Age Distributions Revisited. Mol. Biol. Evol. *30*, 177–190.
14. Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous − Paleogene boundary Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous – Pale. Genome Res. *32*, 1334–1347.
15. Yang, Y., Moore, M.J., Brockington, S.F., Mikenas, J., Olivieri, J., Walker, J.F., and Smith, S.A. (2018). Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. New Phytol. *217*, 855–870

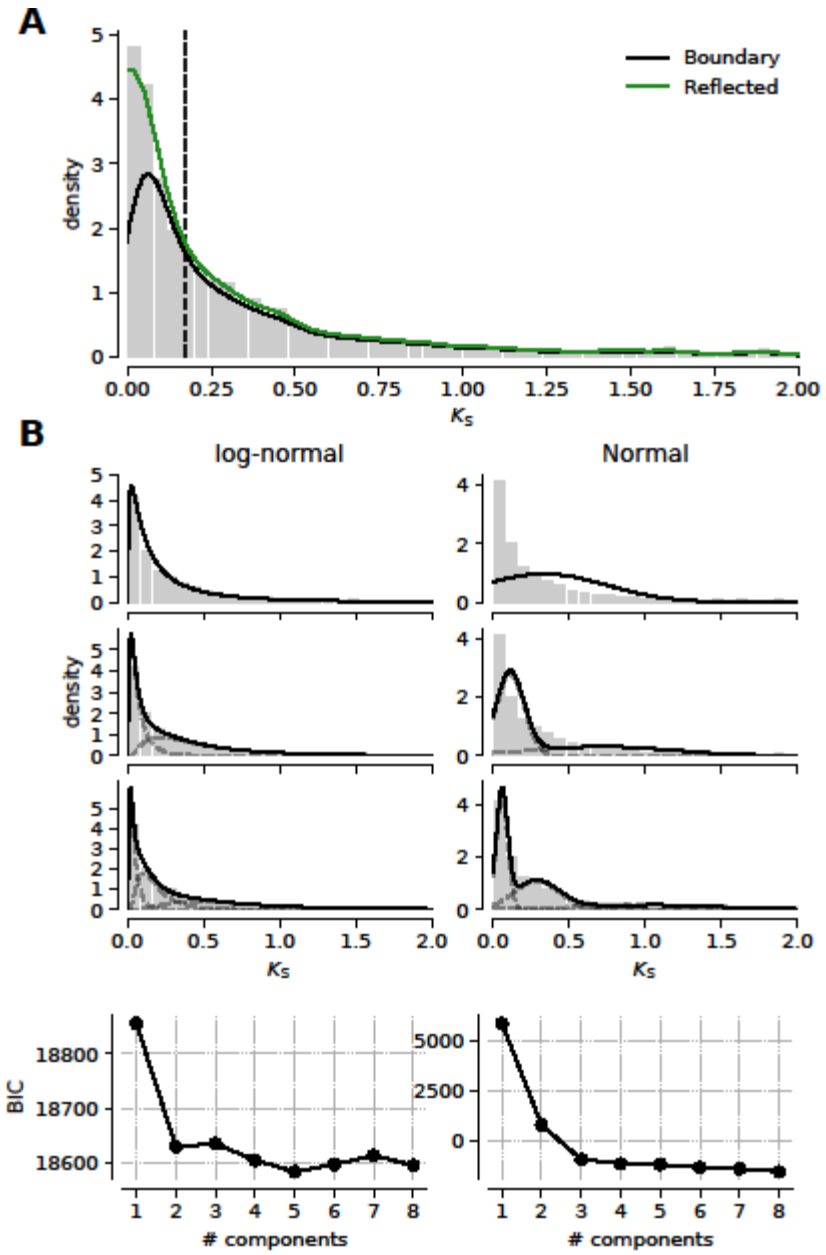


Figure 1: Modeling of K_S distributions and the effect of small scale duplications on WGD inference. (A) A K_S distribution with two KDEs for *Selaginella moellendorffii* is shown, a species for which there is consensus that it probably did not undergo WGDs. When the boundary effect is not accounted for (black line), a suggestive peak appears. This peak clearly disappears when adopting reflection around K_S = 0 (which is arguably the most straightforward technique to account for the boundary effect). The dashed line indicates the median K_S value. (B) Log-normal (left) and normal (right) Gaussian mixture models (GMMs) for the *S. moellendorffii* K_S distribution using 1 to 8 components (only models with up to three components are shown in the histogram plots). The BIC values keep decreasing for more components in the case of the Normal mixture, which is indicative for overclustering. For the log-normal mixture similar overclustering is observed, but here a five-component model gave an optimal fit. Note how the log-normal model with one component already gives a very appealing visual fit and how the ΔBIC values between successive components are much smaller than for the normal GMMs. We note that, based on uncorrected KDEs and Normal GMMs, one could be inclined to infer a spurious recent WGD in *S. moellendorffii*.

Supplementary information - Finding evidence for whole genome duplications: a reappraisal

Arthur Zwaenepoel, Zhen Li, Rolf Lohaus & Yves Van de Peer

Additional methodological remarks

Here we report some additional remarks concerning the study of Ren *et al.* (2018) concerning issues which hamper clear interpretation of their paper. We do not collect conceptual errors here but rather several of the methodological ambiguities in Ren *et al.* that are relevant for the argument in our letter.

- The authors do not describe in their Methods how they fitted the distributions in panel B of Figure 1 and Supplementary Figures 4–65. We assume these are kernel density estimates (KDEs), since we are not aware of any other methods than KDEs and histograms to plot the empirical distribution of a data set. Note that variants of histograms (e.g., smoothed or average-shifted histograms) are all specific cases of KDEs, and hence also sensitive to the discussed boundary effects. We also assume the authors chose some automatic rule for bandwidth selection in their KDEs, however this is not specified. Although KDEs are very common, the authors should still describe their usage and their methodology for fitting them (if these were not KDEs they should definitely specify the method they used). More importantly, we note that relevant detail in a distribution could be lost in KDEs and thus (also) showing the underlying histogram is highly preferred.

- The authors do not describe the details of the Gaussian mixture modeling (GMM). For example, how many components were tested? Which criterion was used for assessing the optimal number of components? Although most of their argument is not based on these GMMs, we feel these should have definitely been mentioned.

- The method used to construct K_S distributions is poorly described. It is unclear which (if any) procedure was used to correct for redundant K_S estimates (Barker et al., 2008; Lynch and Conery, 2003; Maere et al., 2005; Vanneste et al., 2013). This renders (some of) their K_S distributions hard to interpret. The authors report median K_S values for putative WGDs, but they do not describe anywhere from which specific set of K_S values the median was exactly calculated. We assume this would be the median K_S among all gene duplicates (GD) mapping on the branch/node of interest, provided they met the minimal length requirement and did not belong to species-specific families (column 'Gene Duplication' in Supplemental Table 3 of Ren *et al.*). However, for some WGDs it seems impossible that the reported median is in fact the median of the K_S values of these duplicates. As an example, we may consider *Eucalyptus grandis*, for which the authors report a median K_S of 1.27, whereas, judging from Supplemental Figure 17 in Ren *et al.*, this value should be somewhere around 0.25. Furthermore, the authors do not provide any rationale for the usage of the median K_S value. Why not the arithmetic or geometric mean or the mode? If peaks from WGDs are approximately log-normal, as we suggested, the geometric mean would give a more natural estimate of the central location (Morrison, 2008). Intuitively, the mode may be used to represent the peak value. While the median may be more robust to outliers, we do not see a clear meaning for this value in the way it seems to be used.

- The authors do not clearly describe which gene families were used in the reconciliation pipeline. It is however implied that all families that were not species-specific were used (using only genes meeting the minimum

length requirement). Additionally they do not describe how the resulting trees were rooted (which can be non-trivial for large gene families even when the species tree is known).

- The authors fail to recognize or discuss the impact of taxon sampling on their results. It is however obvious that the criterion that at least 1000 GDs should map on a branch to infer a WGD (criterion 2 on p. 425 in Ren et *al.*) is highly sensitive to the taxon sampling at hand. As an example we note that, if the authors would not have used the dense sampling they have within the Brassicaceae, and would for example not have sampled *Arabidopsis lyrata* and *Capsella rubella*, they would have been forced to conclude a WGD happened in the branch leading to *Arabidopsis thaliana* after divergence from *Brassica* based on this criterion (it may fail to meet another criterion however). As a result we suspect that if some clades would have been more thoroughly sampled (breaking up long terminal branches), Ren *et al.* would not have inferred some WGDs using the same identification criteria. We do acknowledge it is unfeasible to gather a more dense taxon sampling in many cases, but still think a reflection on these issues is appropriate.

Supplementary Methods

K_S distribution construction

All CDS sequence data was downloaded from PLAZA Dicots 4.0 (Van Bel et al., 2018). K_S distributions were constructed using the '`wgd`' package (Zwaenepoel & Van de Peer, 2019) following an approach based on Vanneste et al., 2013. In brief, within species paralogous gene family delineation was performed using all-versus-all Blastp (Altschul et al., 1997) with an *e*-value cut-off of 10^{-10} and Markov Clustering (MCL) with the `mcl` program (v10-201) (van Dongen, 2000). The specific command in the version of `wgd` used was (example for *Selaginella moellendorffii*):

```
wgd mcl --mcl --cds -s cds.smo.fasta
```

Multiple sequence alignments (MSAs) of paralogous protein sequences were constructed using MUSCLE (v3.8.31) (Edgar, 2004) with default parameters. This MSA is then used as a guide for a nucleotide level MSA. Subsequently, all pairwise K_S estimates for sequence pairs with pairwise alignment length exceeding 300 nucleotides were estimated by maximum-likelihood (ML) using the basic model of Yang & Nielsen (1998) as implemented in the `codeml` program (Goldman and Yang, 1994) from the PAML package (v4.4c) Yang (2007). Codon frequencies were determined using the F3X4 method based on the average nucleotide frequencies at the three codon positions. Codon model 0 was used for pairwise K_S, K_A and ω estimation, assuming a constant ω across sites and branches. This option is selected since one pair of sequences is usually not enough to detect selective pressure (Vanneste et al., 2013). Subsequently, an approximate maximum likelihood tree for the paralogous gene family is computed using FastTree (Price et al., 2010) and rooted using midpoint rooting. Weights for K_S values are determined using a post-order traversal where at each joining node the m pairwise K_S estimates between leaves of the left and right subtree are added with a weight $1/m$ to the K_S distribution. Using this approach, the weights for each duplication event sum up to one and contribute accordingly to the empirical distribution. This procedure was performed using the following command with `wgd (the `--pairwise` option was used as this is likely more similar to the analysis of Ren et al. than the default)`:

```
wgd ksd smo.mcl cds.smo.fasta --preserve --pairwise
```

Kernel density estimation

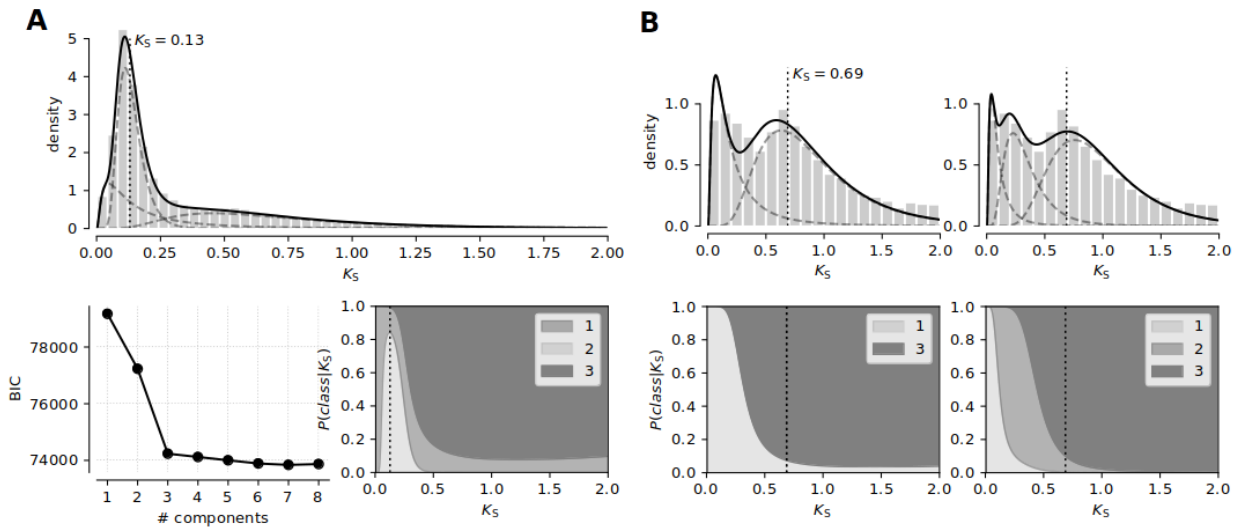
Kernel density estimation was performed on the node-averaged distributions using the seaborn library (v0.8.1) which uses the `` `gaussian_kde` `` function from the SciPy library (v1.1.0) in Python (3.6.2). We used the default bandwidth chosen by Scott's 'rule of thumb' (Scott, 1992) as we considered this would likely be similar to what Ren *et al.* used. We filtered out all K_S values < 0.005 and K_S > 2. KDEs were fitted both for the K_S distribution as such as well as for a distribution with the same K_S distribution reflected around zero added to it. This procedure amounts to mitigating the boundary effect induced by KDEs. We note that the implied assumption in this approach is that $f'(K_S = 0) = 0$.

Gaussian mixture modeling

For fitting Gaussian mixture models (GMMs) to node-averaged K_S distributions, we used the `` `GaussianMixture` `` class as implemented in the `mixture` utilities from the `` `sklearn` `` library (v0.19.1) in Python (3.6.2) (Pedregosa et al., 2011). Again we filtered the K_S distribution such that $0.005 < K_S < 2$. The Expectation-Maximization (EM) algorithm was initialized using *k*-means and was run until convergence (using a threshold of 10^{-3} on the lower bound of the average gain). Up to 8 components were fitted. We used the same

methods to fit normal and log-normal mixtures by fitting normal components to the log-transformed data and subsequent back-transformation to the original scale. We used both the Akaike and Bayesian information criteria (AIC & BIC) to assess relative model fit and report the BIC in our Figure 1. Python code for all analyses and plots can be obtained from a Jupyter notebook available at (https://github.com/arzwa/ksnotebooks).

Supplementary Figures



Supplementary Figure 1: Modeling K_S distributions for *bona fide* WGDs and the contribution of small scale duplications. (A) A K_S distribution for *Glycine max*, a species with a well-characterized recent WGD, dated to have occurred about 14 million years ago (Vanneste et al. 2014). A three-component mixture model is shown, which corresponds to the obvious 'knee' in the BIC plot. The bottom right plot shows the posterior probability to belong to each component for different pairwise K_S estimates. This plot clearly shows how a significant fraction (~18%) of duplicate pairs at $K_S = 0.13$ (the median value Ren *et al.* chose to represent this WGD) stems from SSDs (given the fitted GMM). Such an issue can be avoided by using syntenic structural information to select a subset of paralogs that stem from the WGDs (e.g., as in Vanneste et al. 2014). (B) Similar to (A) but for *Solanum lycopersicum*, again the peak indicated at $K_S = 0.69$ reflects the value reported by Ren *et al.* Plots are shown for a two- and three-component log-normal GMM (the latter obviously giving a better fit even though the second peak does not correspond to any known biological features, this peak is for example not detected in K_S distributions based on syntenic information (Vanneste et al. 2014)). Both panel A and B illustrate that, if we assume that these mixture models reliably reflect the underlying theoretical distribution, SSDs contribute significantly to peaks induced by WGDs. This effect becomes stronger the more recent the WGD, where reliable inference of WGDs from K_S distributions eventually becomes impossible. We note that these plots can underestimate this effect for higher K_S regions, as the fitted SSD decay peak is unable to model those SSDs that get fixed in the genome.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J., and Rieseberg, L.H. (2008). Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years. Mol. Biol. Evol. *25*, 2445–2455.

van Dongen, S. (2000). Graph Clustering by Flow Simulation. University of Utrecht.

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

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. *11*, 725–736.

Lynch, M., and Conery, J.S. (2003). The evolutionary demography of duplicate genes. In Genome Evolution, (Springer, Dordrecht), pp. 35–44.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. *102*, 5454–5459.

Morrison, D.A. (2008). How to summarize estimates of ancestral divergence times., How to Summarize Estimates of Ancestral Divergence Times. Evol. Bioinforma. Online Evol. Bioinforma. Online *4*, *4*, 75, 75–95.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE *5*, e9490.

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., Ma, H., and Qi, J. (2018). Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. Mol. Plant *11*, 414–428.

Scott, D., W. (1992). Kernel Density Estimators. In Multivariate Density Estimation, (Wiley-Blackwell), pp. 125–193.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. *46*, D1190–D1196.

Vanneste, K., Van de Peer, Y., and Maere, S. (2013). Inference of Genome Duplications from Age Distributions Revisited. Mol. Biol. Evol. *30*, 177–190.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. Genome Research *24*, 1334–1347.

Zwaenepoel, A., and Van de Peer, Y. (2019). wgd - simple command line tools for the analysis of ancient whole genome duplications. Bioinformatics, in press (doi: 10.1093/bioinformatics/bty915).

Supplementary information - Finding evidence for whole genome duplications: a reappraisal

Arthur Zwaenepoel, Zhen Li, Rolf Lohaus & Yves Van de Peer

Table 1: WGDs used by Ren *et al.* (2018) for the calibration of the survivalship curve for transciptomic data. The species included in this table are those referred to by Ren *et al.* as species with 'Calibrated' WGDs (purple dots in their Figure 2 and 3). When we comment on evidence, we specifically mean evidence as can be judged from the analyses in Ren *et al.* Where appropriate, we mention additional references that agree or disagree with the results presented by Ren *et al.*

Species	Median K_S	Comment
Olea europaea	0.13	KDE artifact (median K_S reflects SSD peak); K_S distribution differs strongly from Unver *et al.* (2017); Possibly in agreement with Julca *et al.* (2018)
Lactuca sativa	0.91	Some evidence for WGD; Median K_S affected by SSD peak
Flourensia thurifera	0.20	Strong evidence for WGD; Median K_S may be affected by SSD peak
Asteraceaea	1.28	No clear evidence for WGD
Actinidia arguta	0.09	KDE artifact (median K_S reflects SSD peak)
Hydrangea macrophylla	0.16	Some evidence for WGD; Median K_S affected by SSD peak
Eschscholzia californica	0.18	Missing from supplementary material; Cui *et al.* (2006) reported a peak at $K_S \approx 0.65$
Chloranthales	0.89	No clear evidence for WGD; Reference in Supplementary Table 3 does not report ancient WGD but neopolyploidy
Panicum virgatum	0.05	KDE artifact (median K_S reflects SSD peak)
Yucca filamentosa	0.08	KDE artifact (median K_S reflects SSD peak)
Asparagus officinalis	0.53	Some evidence for WGD; Median K_S affected by SSD peak; In agreement with Harkess *et al.* (2017)
Asparagales	1.09	Missing from supplementary material; Zhang *et al.* (2018) suggest no shared WGD for Asparagales
Pinellia ternata	0.10	KDE artifact (median K_S reflects SSD peak)
Acorus calamus	0.10	KDE artifact (median K_S reflects SSD peak); Cui *et al.* (2006) reported a peak (or possibly two) at $K_S \approx 0.5$ for *Acorus americanus*
Cabomba caroliniana	0.16	KDE artifact (median K_S reflects SSD peak); References in Supplementary Table 3 do not report ancient WGD but neopolyploidy

References

- Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., et al. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. 16, 738–749.
- Harkess, A., Zhou, J., Xu, C., Bowers, J.E., Hulst, R., Ayyampalayam, S., Mercati, F., Riccardi, P., McKain, M.R., Kakrana, A., et al. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. Nature Communications 8, 1279.

- Julca, I., Marcet-Houben, M., Vargas, P., and Gabaldón, T. (2018). Phylogenomics of the olive tree (*Olea europaea*) reveals the relative contribution of ancient allo- and autopolyploidization events. BMC Biology 16.
- Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., Yang, M., He, L., Deng, T., Escalante, F.J., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. Proceedings of the National Academy of Sciences 114, E9413–E9422.
- Zhang, G.-Q., Liu, K.-W., Li, Z., Lohaus, R., Hsiao, Y.-Y., Niu, S.-C., Wang, J.-Y., Lin, Y.-C., Xu, Q., Chen, L.-J., et al. (2017). The *Apostasia* genome and the evolution of orchids. Nature 549, 379–383.