Dear Editor,

We appreciate the effort that some of the authors of the original paper by Ren *et al.* (2018) went through in debating and trying to resolve the issues raised in our "Rebuttal" letter. However, we are afraid that they have not really addressed the issues we brought up and that they in fact could not resolve any of our concerns. Moreover, some parts of the reply raise additional concerns regarding their original analyses and interpretations. Please allow us to explain and to argue why we feel that these issues are not minor.

Our main concern with the original paper of Ren *et al.* was that the authors failed to recognize the effect that small-scale duplications (SSDs) have on the inference of 'recent' whole genome duplication (WGD) events. We believe that in many cases, the newly-inferred recent WGDs proposed in Ren *et al.* cannot be distinguished from SSDs based on the analyses presented, and particularly without evidence from structural genomic data. In our rebuttal, we discussed several common interpretative pitfalls which we believe may have caused erroneous interpretations and evaluated the impact of these methodological hazards on the results presented by Ren *et al.*.

In their reply, Wang *et al.* argue that their analyses are sound and their conclusions reliable because the generally-recognized patterns of gene duplicate evolution do not apply to their data sets. In a nutshell, the authors claim that the hallmark evolutionary demographic feature of SSDs — the expected frequency of their ages (K_S values) showing an exponential distribution (often described as an 'L'-shape) — is generally not present nor expected to be present in their data due to careful removal of transcriptional isoforms, in both genomic and transcriptomic data sets. This claim is very surprising and it seems that the authors are unaware of the radical implications of such a statement. The expected exponential decay or 'L'-shape of the frequency of duplications as a function of age is the natural outcome of a model of continuous small-scale duplication and eventual loss of duplicate genes not under selection by non-functionalization (e.g., through frameshift, missense, nonsense or regulatory mutations) (Lynch 2007; Figure 1 in this reply). This model relies solely on "true" gene duplicates, and the resulting exponential distribution is *not* caused by transcriptional isoforms (or other "false paralogs" like allelic variants). Thus, unlike what Wang and co-authors seem to imply, their removal is not expected to result in a change of the 'L'-shape of K_S distributions of paralogs (we do of course recognize that transcriptional isoforms do, however, spuriously amplify the 'L'-shape). For further information, we would like to point to the many studies (both theoretical and based on empirical data, not only in plants, but also in animals and fungi) that have been published on the demography of duplicate genes and properties of K_S age distributions (Lynch & Conery 2000, Lynch & Conery 2003, Blanc & Wolfe 2004 (e.g., Figure 1), Maere *et al.* 2005, Cui *et al.* 2006, Lynch 2007, Vanneste *et al.* 2013). We acknowledge that the absence of a (strong) 'L'-shape of the exponential distribution can of course also be explained in terms of this basic model, but both simulation and empirical data suggest this case to be rare (a uniform distribution can be obtained as a limit). Thus, we believe that any clear *absence* of such a signature, as claimed by the authors to be displayed in their data (but see further), would in fact raise additional concerns and would need to be explained in biological terms[1]. Of course, there are well-known, but rare cases where the apparent absence of the 'L'-shape is due to a high retention of paralogs from a very recent WGD event, where the exponential distribution of SSD-derived paralogs is overshadowed by a narrow and tall peak from a huge number of WGD-derived paralogs (e.g., as in *Glycine max* (Wang *et al.* Figure 1C and D) or, presumably, *Panicum virgatum* (Wang *et*

[1] For example, the small SSD peak below $K_S < 0.5$ evident in *A. thaliana* (also described by Wang *et al.*) has been shown by Blanc & Wolfe (2004) to originate from a temporal increase in the number of tandem duplications and possibly selection for genome downsizing in the recent evolutionary past of *Arabidopsis.* Note however, that besides this secondary SSD peak, a clear initial peak of an exponential distribution is still present in the distributions shown by Blanc & Wolfe (and others). We would also like to stress that the presence of SSDs in secondary peaks, as acknowledged by Wang *et al.* and explained by Blanc & Wolfe (2004), of course only causes *more* trouble for the reliable inference of WGDs in low-K_S regions, as we discussed in our rebuttal to Ren *et al.*.

al. Figures 2E and 3C)). However, in such (extreme) cases, structural genomic evidence has been used and is required to unequivocally confirm the presence of a WGD instead of simply relying on arbitrary thresholds for the number of gene duplicates.

Furthermore, and contrary to the claims of Wang *et al.*, it is our strong opinion that at least five of the distributions shown by the authors in Figure 2 and 3 in fact *do show* an exponential decay (*R. communis, V.vinifera, O. sativa, A. trichopoda* and *C. caroliniana*). Only if one solely considers the kernel density estimate (KDE) and blindly ignores the actual distribution in the histograms, is the 'L'-shape indeed not observed. However, as we emphasized in our rebuttal paper, the KDE is strongly affected by boundary effects (this point was completely ignored by Wang and co-authors in their reply), and this should be obvious from inspecting the histograms (especially clear in, e.g., Wang *et al.* Figures 2C and F, and 3A). We also note that empirical distributions are of course expected to deviate to a certain extent from the perfect exponential expectation due to stochastic effects (see Figure 1; left panel), and more strongly so if only low numbers of gene duplicates are involved. Lastly, the lack of a clear exponential distribution might also signal technical issues, such as too aggressive filtering, for example when distinguishing very recent gene duplicates from transcriptional isoforms, which could lead to fewer identified paralogs and stronger stochastic effects. Based on the number of gene duplicates in the distributions presented by Wang *et al.,* we suspect this could also play a role (compare Figure 2A and 2C in Wang *et al.* with Figure 2 enclosed here for an example).

In conclusion, our original concerns have not been resolved or fully addressed. The reply by Wang *et al.* does not provide any arguments as to why their basic results could be regarded trustworthy, except for saying that their K_S distributions do not show an exponential decay — which, as stated above, disagrees with generally accepted assumptions and does not hold true when carefully inspecting the presented distributions — and that therefore "peaks", even in the second bin, are regarded as evidence for WGDs (see, e.g., Wang *et al.* Figure 3A and E). In a few cases, Wang and co-authors claim that we have misunderstood methods (we have not, note our original comment on GMM usage in our Supplementary Information) or on aspects of their analyses that we have not actually called into question, such as the problem of distinguishing transcriptional isoforms from recent paralogs, which we (until now) assumed was handled correctly in the analyses by Ren *et al.*. With regard to the "Calibrated WGDs" that we deemed unreliable based on their analyses (listed in our Supplementary Table), Wang *et al.* essentially recite their literature sources without addressing our specific concerns. Our main issue here is not that these literature references are misplaced (although some are), but rather that the presence of many of these WGDs at the given K_S values (and thus absolute age) cannot be inferred from the data presented by Ren *et al.,* because of the issues we discuss in our rebuttal (we here refrain from a detailed itemized reply to their list, but would be happy to provide one if requested). As the inference of "newly identified WGDs" relies on these "Calibrated WGDs" and their inferred ages, we believe the reliability of these WGD events is of crucial importance for their study. Similarly, in our rebuttal we questioned the reliability of their absolute dating analysis and the associated speculations about the evolutionary causes and consequences of their inferred WGDs (which are some of the main conclusions of the original paper). Instead of addressing our concerns the authors replied with simply reiterating their methods section.

In summary, we remain convinced that the inference of many of the WGDs in Ren *et al.* is highly problematic, and that as a result, any further speculation on the timing and macro-evolutionary roles of these events in Ren *et al.* is unjustified. We stressed this in our rebuttal paper, and stress this again after reading the reply by Wang *et al.*. Also, we think the issues discussed here cannot be called 'minor', as they severely impact the reliability of the results and general conclusions presented by Ren *et al.,* and we would like to urge the editor to take our concerns seriously. We believe that our original rebuttal provides a clear and constructive review of the data and methods presented in Ren *et al.*. We also believe that our concerns highlighted therein are of general relevance to the field, and we hope that our contribution will help to prevent potential further misinterpretations in other such studies in the future.

Sincerely,

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

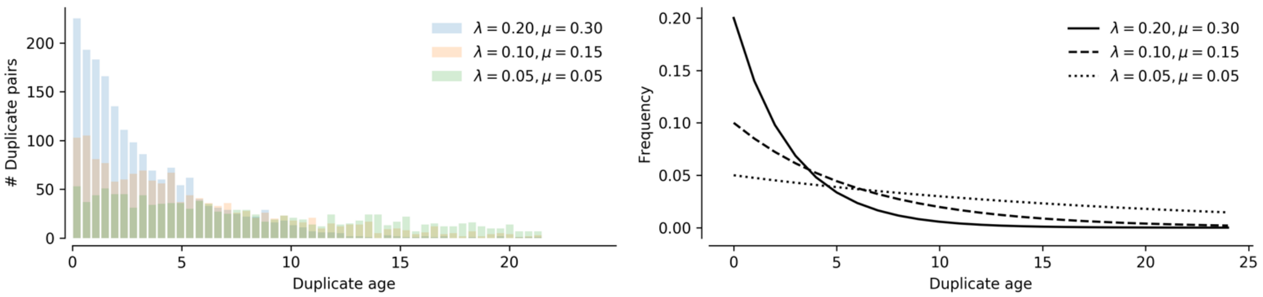

Figure 1: Left: stochastic simulation of the effect of continuous duplication (rate λ) and loss (rate μ) on the age distributions of duplicate genes. Right: Analytically derived expectations for the same parameter settings as used in the simulations, after Lynch (2007). For details as well as simulation code we refer to https://github.com/arzwa/wgd/blob/master/example/dlsim.ipynb

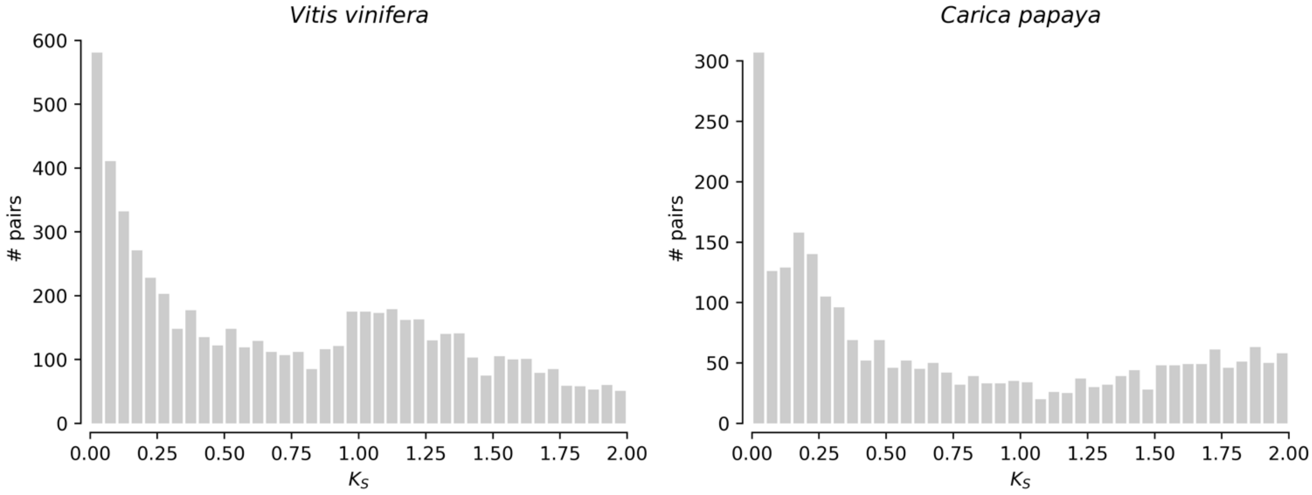

Figure 2: K_S distributions for *Vitis vinifera* and *Carica papaya*, for comparison with the additional data presented in the reply by Wang *et al.* We explicitly note that these distributions excluded transcriptional isoforms (only longest transcripts for every gene locus were used). Bin-width is 0.05 K_S units as in the reply by Wang *et al.,* enabling direct comparison of the number of duplicate pairs (*y*-axis). Methods used were identical to those described in the supplementary material of our rebuttal paper.

References

- 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.
- 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.
- Lynch, M. (2007). The origins of genome architecture (Sunderland, Mass.: Sinauer Associates).
- Lynch, M., and Conery, J.S. (2003). The evolutionary demography of duplicate genes. In Genome Evolution, (Springer, Dordrecht), pp. 35–44.
- Lynch, M., and John S. Conery (2000). The Evolutionary Fate and Consequences of Duplicate Genes. Science *290*, 1151–1155.
- 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. Proceedings of the National Academy of Sciences *102*, 5454–5459.
- 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. Molecular Plant *11*, 414–428.
- Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. Am. J. Bot. *96*, 336–348.
- 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.