Addendum: Detection of colinear blocks and synteny and evolutionary analyses based on utilization of MCScanX

Addendum to: *Nature Protocols* https://doi.org/10.1038/s41596-024-00968-2, published online 15 March 2024.

https://doi.org/10.1038/s41596-026-01380-8

Published online: 15 May 2026

Check for updates

Xi Zhang, Yupeng Wang, Paule V. Joseph, Andrew H. Paterson & David Roy Smith

The bioinformatics protocol by Wang et al.[1] outlines the steps for efficiently identifying colinear blocks in intra- and inter-species BLASTP outputs using the Multiple Colinearity Scan Toolkit Version X (MCScanX)[2]. Part 2 of the protocol provides steps for downloading data directly from NCBI and preparing the necessary .gff and .blast files yielded from blast all-vs-all analyses. While using MCScanX, we (X.Z. and D.R.S.) discovered that Part 2 lacks an essential pre-processing step—the step required for determining whether there are multiple isoforms derived from alternative splicing. Indeed, when analyzing data downloaded directly from NCBI, it can be crucial to have a process known as transcript filtering to identify the longest transcript as the primary protein sequence. Similar issues could happen when researchers fail to use primary assembly (one haploid set) in diploid genome assemblies, because the alleles could be mistakenly treated as gene duplicates.

To avoid misprediction of gene duplicates, especially for genome data from NCBI or other online resources, it can be helpful to include a transcript filtering step[3,4]. This is particularly true when analyzing datasets of species with large numbers of duplicated genes. Without this step, the number of duplicate genes will be overrepresented. In Fig. 1, we show how using a transcript filtering step can dramatically impact the results of the analysis carried out in the protocol by Wang et al. Indeed, by following the MCScanX protocol and comparing *Arabidopisis thaliana* with and without transcript filtering, we found as many as 25,776 protein-coding genes categorized as singleton duplications after employing transcript filtering, compared to only 3,086 protein-coding genes when not filtering. Similarly, there are 11,948 protein-coding genes categorized as dispersed duplications when filtering vs. 7,297 protein-coding genes when not. Finally, there are only 2,216 genes categorized as tandem duplications after transcript filtering compared to 30,101 genes without filtering. This suggests that in the absence of transcript filtering, multiple isoforms from the same gene can be misinterpreted as tandem duplication events.

Figure 2 shows the visualization of the synonymous substitution rate (Ks) distributions of colinear genes in *A. thaliana* and *Medicago truncatula* before and after transcript filtering. We are not questioning the reliability of running the MCScanX algorithms but want to highlight potential issues when using the protocol, particularly the potential challenges when preparing input files in Part 2 (Steps 7–20). Overall, MCScanX is a useful tool for efficiently identifying colinear blocks and downstream evolutionary analysis, but additional work is needed for preparing the input data and running the tool.

We provide the following useful tips for increasing the utility of the protocol: as noted by the authors in Steps 11–13 and the Troubleshooting section, we found that when generating the correct .gff file, it is better to offer alternative options similar to the 'mkGFF3.pl' program in the MCScanX_protocol package. This is because the downloaded .gff can have different formats and it is important to convert it to the one MCScanX can read. We have found that the 'gff2bed' script from BEDOPS v2.4.41 (ref. 5), AGAT v1.6.1 (ref. 6) and the custom processing script on the 'XX_feature_table.txt' can help yield the .gff file for MCScanX. In terms of generating the .blast file at Steps 14–20, we found it is not efficient to prepare the 'runBLASTP.sh', especially when an all-against-all BLASTP is needed for each reciprocal genome pair. We have provided custom scripts with the MCScanX_Assistant tool at GitHub (https://github.com/zx0223winner/MCScanX_Assistant) to iterate the genome all-against-all BLASTP processing, which greatly improves the preparation step (see also the Supplementary Text S1–S4 in ref. 7).

These comments were well-received by the MCScanX team (Y.W., P.V.J. and A.H.P.) and a notice has been added to the external link of the protocol (http://bdx-consulting.com/mcscanx-protocol/) stating the following: "…the current stage lacks a transcript filtering step for handling multiple alternative splice isoforms per locus, which may lead to confusion

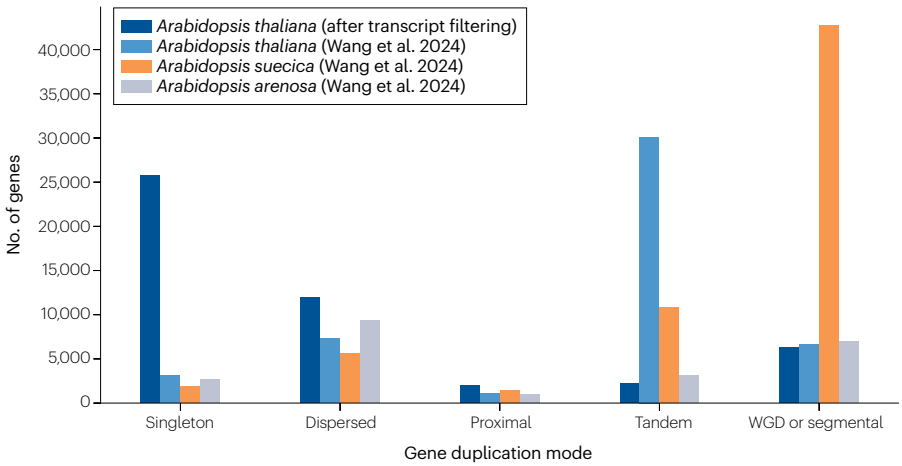

Fig. 1 | Comparison of gene duplication modes among closely related Arabidopsis taxa, with and without transcript filtering. This figure was adapted from Fig. 6 of the protocol[1], where transcript filtering was not used (*A. arenosa*, *A. suecica* and *A. thaliana* without filtering are shown in grey, in orange and in light blue, respectively, as in Fig. 6 of the protocol[1]). *A. thaliana* after transcript filtering is shown in dark blue. Strikingly, it appears that tandem duplications are less prevalent in *A. thaliana* than *A. arenosa*, and singleton duplications represent an overwhelming proportion of gene duplications in *A. thaliana* compared to the other two species, when transcript filtering is incorporated into the workflow.

among paralogous genes. To address this limitation, users are encouraged to utilize MCScanX_Assistant, which provides the necessary functionality".

The MCScanX team further addresses the protocol's lack of transcript filtering step as follows:

During the development of the original software, the MCScanX team recognized that alternative splicing could influence MCScanX results. To address this, the accompanying README file (https://github.com/wyp1125/MCScanX) explicitly states that "The xyz.bed file holds gene positions," and the included example uses *Arabidopsis thaliana* gene symbols (e.g., AT1G01010) rather than transcript identifiers (e.g., AT1G01010.1). This guidance clearly indicates that users should supply gene-level names and coordinates—not transcript-level entries—in the .bed file. Furthermore, the original publication[2] noted that "If a gene had more than one transcript, only the first transcript in the annotation was used." Although the MCScanX toolkit did not include a dedicated transcript-filtering utility, the expected use of gene-level information in the .bed file was unambiguous.

In the MCScanX team's subsequent research projects involving MCScanX, the .bed file was generated using custom scripts that performed transcript filtering, selecting either the first annotated transcript[8] or the longest transcript[9–12]. The choice of filtering rule was made heuristically based on the specific biological question. To our knowledge, no consensus has yet emerged within the scientific community regarding an optimal transcript-filtering strategy, as this topic has not been comprehensively evaluated. Consequently, studies examining the effects of transcript filtering on gene colinearity and paralogous gene detection are both valuable and timely.

In the 2024 protocol[1], the MCScanX team introduced a set of new automation and utility tools for MCScanX packaged separately as "MCScanX-protocol" (http://bdx-consulting.com/mcscanx-protocol/), including functions that enable direct processing of genome assembly

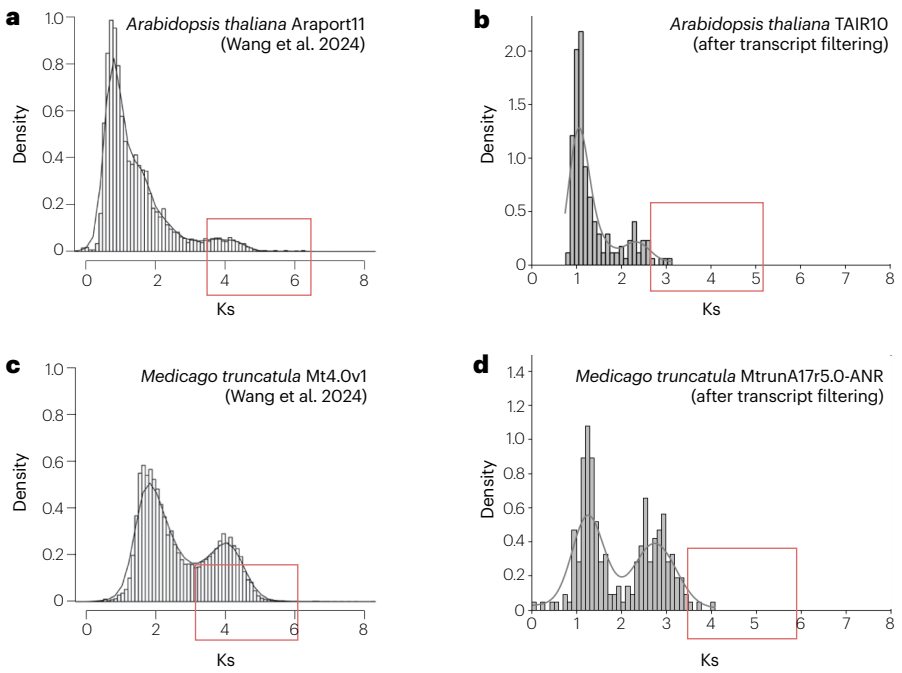

Fig. 2 | Visualization of the synonymous (Ks) substitution rate distributions of colinear genes in different genomes (*A. thaliana* and *M. truncatula*). Panels a and c are adapted from Fig. 5 of the protocol[1]. Red boxes highlight the differences in Ks before and after the transcript filtering. In Fig. 5 of the protocol, the authors used a different version of the *A. thaliana* genome (Araport11) as compared to the one demonstrated in the protocol (TAIR 10). Considering that the Araport11 version only has a few more protein-coding genes than the TAIR 10 version, we used the TAIR10 of *A. thaliana* to ensure data consistency throughout the protocol. Similarly, a high-quality, chromosome-scale reference genome for *M. truncatula* (MtrunA17r5.0-ANR) was used in this study, as the earlier Mt4.0v1 assembly was generated prior to the availability of long-read sequencing technologies.

files from NCBI. The current release of the mkGFF3.pl script within this package does not programmatically enforce transcript-filtering logic creating a potential pitfall for users unfamiliar with the original MCScanX documentation. The MCScanX team evaluated the consequences of this omission. Including all transcripts for a gene may lead to BLASTP matches among alternative isoforms of the same gene. Nonetheless, this has only limited impact on colinearity detection. As described in the original MCScanX publication[2] the algorithm mitigates inflated local colinearity signals by collapsing consecutive BLASTP matches that share a common gene and whose paired genes are separated by fewer than five loci, retaining only the match with the smallest *E*-value. This behavior aligns with the results shown here in Fig. 1, where the numbers of WGD and segmental genes are nearly identical regardless of whether primary-transcript filtering is applied. Thus, the colinearity and synteny analyses presented in the protocol remain robust.

The more substantial effect of omitting transcript filtering is an overestimation of tandem gene pairs, caused by BLASTP hits among transcripts belonging to the same gene. This leads to inaccuracies in the gene-type counts reported in Step 26B and Fig. 6 of the protocol[1]; however, the MCScanX team did not detect tandem relationships among isoforms of the same gene in the corresponding phylogenetic analyses.

The MCScanX team encourages users of the protocol to consult both MCScanX_Assistant (https://github.com/zx0223winner/MCScanX_Assistant) and this Addendum for guidance on preparing an appropriate .bed file. The MCScanX team recognizes that transcript filtering by either the first annotated or the longest isoform remains an arbitrary choice and may introduce bias into downstream analyses. The field still lacks a benchmark or gold-standard dataset for objectively assessing the performance of gene-colinearity detection and gene-duplication mode classification. Developing such resources would be valuable for future research. The MCScanX team appreciated Dr. Zhang and Dr. Smith for their examination of the MCScanX-protocol package and for identifying areas for improvement. Both parties are actively collaborating to maintain and enhance both the MCScanX protocol and the MCScanX toolkit.

Code availability

The Protocol uses SnakeMake pipeline and Conda to run and install dependencies. The distribution version is available online at GitHub: https://github.com/zx0223winner/MCScanX_Assistant and the archived version is at Zenodo: https://doi.org/10.5281/zenodo.19245923.

References

1. Wang, Y. et al. Detection of colinear blocks and synteny and evolutionary analyses based on utilization of MCScanX. *Nat. Protoc.* **19**, 2206–2229 (2024).
2. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
3. Zhang, X., Hu, Y., Cheng, Z. & Archibald, J. M. HSDecipher: A pipeline for comparative genomic analysis of highly similar duplicate genes in eukaryotic genomes. *STAR Protoc.* **4**, 102014 (2023).
4. Zhang, X., Hu, Y., Smith, D. R., Cheng, Z. & Archibald, J. M. HSDSnake: a user-friendly SnakeMake pipeline for analysis of duplicate genes in eukaryotic genomes. *Bioinformatics* **41**, btaf325 (2025).
5. Neph, S. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
6. Dainat, J. In *Plant and Animal Genome XXIX Conference* (8–12 January 2022).
7. Zhang, X. & Smith, D. R. How to prepare the input data and run MCScanX efficiently? Preprint at *bioRxiv* https://doi.org/10.1101/2025.07.29.666888 (2025).
8. Wang, Y., Wang, X., Lee, T. H., Mansoor, S. & Paterson, A. H. Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in *Oryza sativa* (rice). *New Phytol.* **198**, 274–283 (2013).
9. Wang, Y., Li, J. & Paterson, A. H. MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics* **29**, 1458–1460 (2013).
10. Wang, Y., Sun, Y. & Joseph, P. V. Diverse evolutionary rates and gene duplication patterns among families of functional olfactory receptor genes in humans. *PLoS One* **18**, e0282575 (2023).
11. Wang, Y., Sun, Y. & Joseph, P. V. Contrasting patterns of gene duplication, relocation, and selection among human taste genes. *Evol. Bioinform.* **17**, 11769343211035141 (2021).
12. Wang, Y., Tan, X. & Paterson, A. H. Different patterns of gene structure divergence following gene duplication in Arabidopsis. *BMC Genomics* **14**, 652 (2013).

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government and NIAAA Division of Intramural Research (Z01AA000135). The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the US Department of Health and Human Services.

Author contributions

X.Z. conceptualized the Addendum and analyzed associated data. The manuscript was written by X.Z. and edited by D.R.S. A.H.P. and P.V.J. conceived the original project and provided supervision and funding. Y.W. conducted data analysis.

Competing interests

The authors declare no competing interests.