DISPATCH

Phylogenomics: Is less more when using large-scale datasets?

Davide Pisani[1,2], Eleonora Rossi[2], Ferdinand Marlétaz[3] and Roberto Feuda[4]

E-mail: Davide.pisani@bristol.ac.uk;

1 Palaeobiology Research Group, School of Biological Sciences, University of Bristol, UK.
2 Palaeobiology Research Group, School of Earth Sciences, University of Bristol, UK.
3 Centre for Life's Origin & Evolution, Department of Genetics, Evolution & Environment, University College London, UK.
4 Department of Genetics and Genome Biology, University of Leicester, Leicester, UK.

Phylogenetic studies have traditionally placed the simple Xenoacoelomorph worms as the sister group of all other animals with bilateral body symmetry. A new study shows that misidentification of orthologous genes might have been the source of at the least some support for this placement.

Resolving evolutionary (i.e. phylogenetic) relationships among species or more inclusive taxa such as families and phyla is notoriously difficult. It has thus become common for phylogenetic studies to use datasets consisting of hundreds, sometimes thousands, of genes sampled from complete genomes[1–4]. This approach is rooted in a series of seminal papers that first attempted to "concatenate" (join) individual gene alignments, which are usually not very informative because they include few sites, into a single "supermatrix" to increase signal and reduce the effect of stochastic errors[5,6]. Early supermatrices including tens of genes[7] were considered large, but the datasets available to modern phylogenetic studies have swollen enormously[4]. In the early days of molecular phylogenetics, scientists struggled to sequence the genes to include in their datasets. Today, the trend is for genomic data to be generated by large scale initiatives such as the Darwin Tree of Life[8], and phylogeneticists mostly engage instead in the development of computational pipelines to subsample the data and identify the genes that are most appropriate to attempt to accurately resolve specific phylogenetic problems. While it would be tempting to assume that the rule of thumb should be that supermatrices should include as many genes as possible, maybe even all the genes in the genomes of the species under study, this is both unrealistic and problematic. It is unrealistic because the analysis of massive datasets continue to be extremely time consuming[4] and has a large, associated carbon footprint. It is problematic,

because when assembling increasingly large datasets, more genes with complex evolutionary histories (i.e. genes that underwent many duplications and deletions) will tend to be included in the supermatrix. For such genes, it can be impossible to distinguish orthologs (genes separated by speciation) from paralogs (genes separated by duplication), and this can potentially lead to the inference of incorrect phylogenies (Figure 1). In a recent paper in *Current Biology*, Peter Mulhair, Mary O'Connell and colleagues[9] address the problem of filtering collections of single gene alignments to reduce the potential negative effect of 'hidden paralogy', that is the inclusion of sequences related by paralogy, rather than orthology, in a phylogenetic data set. They found that filtering the data to try to minimise the presence of gene families affected by hidden paralogy can change the results of phylogenetic analyses aimed at resolving relationships at the root of animals with bilateral body symmetry (Bilateria).

When subsampling gene sequences to assemble phylogenetic datasets, the key problem is identifying, for each considered gene, only orthologous rather than a mixture of paralogous and orthologous sequences (Figure 1). This is trivial for genes that do not have a history of gene duplication (e.g. some ribosomal proteins that evolved as in Figure 1A), but becomes progressively more complex (Figure 1B), and can become impossible (Figure 1C), when genes with more complex histories of duplications and deletions are considered. The implication is that not all genes can be expected to be good phylogenetic markers, and if the sequences sampled for a given gene include paralogs, the inferred phylogeny will reflect the duplication history of the gene (Figure 1C), rather than the history of the species from which the genes were sampled. Many tools have been developed to identify sets of orthologs from gene families that underwent duplications and deletions[10,11], but none of these tools is fool-proof. In part, this is because these methods frequently rely on the topologies of gene trees (the accuracy of which can be poor), and this is in part because the identification of a set of orthologs might simply be impossible (Figure 1C). Mulhair and colleagues[9] address the problem of filtering single gene alignments to reduce the potentially negative effect of "hidden paralogy", using an approach they aptly named "Clan Check"[12]. Clan Check uses prior knowledge of phylogenetic relationships to identify uncontroversial groups and identify genes that fail to support them. Uncontroversial groups are clusters of species whose validity is out of question (examples could be groups such as mammals, birds or — as in the case of Mulhair and colleagues[9] — sponges, jellyfishes and corals, etc). Such groups are expected to be characterised by a clear and unambiguous signal, and the Clan Check approach uses them as benchmarks to filter individual genes because, the authors reason, the homology relationships of sets of sequences that fail to support these groups must be dubious.

Mulhair and colleagues[9] test their approach on three datasets[1-3] addressing one of the most topical current problems in animal phylogenetics, the placement of the Xenoacoelomorpha (a clade including the enigmatic Xenoturbella[13] and the acoel worms). Xenoacoelomorpha has been previously recovered as the sister of the remaining animals with bilateral body symmetry (the Nephrozoa hypothesis[1,2]), or as the sister of Ambulacraria, a group of deuterostomes including the hemichordates (acorn worms) and echinoderms (sea urchins, sea stars and their allies) the Xenoambulacraria hypothesis[3]. Distinguishing between these hypotheses has significant implications for the understanding of the morphological complexity of the last common ancestor of the Bilateria[1], which is frequently assumed to have been a small, acoel-like worm. Mulhair and colleagues[9] defined ten incontrovertible groups for the datasets they tested and found that these groups were frequently violated, in some cases by more than 60% of the genes. Accordingly, Mulhair and colleagues[9] attempted to reduce the potential negative effect of hidden paralogs on phylogeny estimation by excluding all the genes that violated more than three to five (depending on the dataset) incontrovertible groups, and found that this had an impact on their results. Support for the Xenoambulacraria hypothesis increased, while support for the Nephrozoa (Figure 2) decreased. More precisely, of the three datasets that Mulhair and colleagues[9] tested, the only one that originally supported Xenoambulacraria[3], continued to do so after filtering with Clan Check. By contrast, of the other two datasets (which originally supported Nephrozoa), one[1] switched to support Xenoambulacraria (albeit not very strongly), while the other[2] become almost inconclusive, with the precise extent of support lost for Nephrozoa depending on the choice of outgroup.

The results of Mulhair and colleagues[9] are intriguing because they suggest that the effect of hidden paralogy, which is rarely tested for (see Philippe[3] for an exception), can impact the results of phylogenomic analyses. However, it is important to stress, as implied also by Mulhair and colleagues[9], that there is no clear proof yet that the genes excluded by Clan Check are exclusively enriched in hidden paralogs, as other problems might affect the capacity of individual genes to recover incontrovertible groups, e.g. contaminants, problems of compositional heterogeneity, and variations in the rate of substitution. While a better characterisation of Clan Check is necessary, we are convinced that its application should already be strongly encouraged. This is because, irrespective of what might be causing a gene to fail a Clan Check test, if a gene does not recover a set of incontrovertible groups, what hope can we have that it will reliably resolve tricker sets of relationships?

The most important take-home message from the study of Mulhair and colleagues[9] is that less can be more in phylogenomics[14], and that dataset size does not equal dataset quality. Irrespective of whether Nephrozoa or Xenoambulacraria will ultimately be proven

correct, Clan Check and other similar tools[3] may help define better curated datasets, that can be expected to achieve more accurate phylogenies. These are the datasets we should aim to generate in modern phylogenomic research, and it would be interesting to see how software like Clan Check might impact other current controversies, such as the placement of the comb jellies in the animal phylogeny.

Declaration of Interests

The authors declare no competing interests.

Figure 1. Sampling orthologs from gene families with complex evolutionary histories can be impossible.
(A) a gene family that did not undergo processes of gene duplication and losses, all sequences are related by the process of speciation and are thus orthologs. When all sequences are orthologous, the gene tree and the species tree correspond making ortholog sequence selection trivial. (B) a gene family that underwent one duplication and no deletions. The duplication (the node represented by a square) results in the emergence of two genes (B1 and B2) that are not related by speciation and are referred to as 'paralogs'. At speciation both paralogs are independently passed to descending species. Sequences in the B1 or the B2 side of the gene tree are in an orthologous relationship with each other, but B1 and B2 copies are paralogous. Accordingly, if sequences are sampled (for each species) that belong exclusively to gene B1 or B2, the history of the sampled gene sequences will reflect the correct species phylogeny. (C) a gene family that underwent one duplication (resulting in the paralogs C1 and C2), and several lineage-specific deletions that caused the loss of C1 in Species 2 and of C2 in Species 1, 3, 4 and 5. For this gene, a full set (for Species 1 to 5) of orthologs cannot be sampled, and the correct species tree cannot be recovered. Circles represent speciations. Squares at nodes represent gene duplications in panels B and C. Dotted lines indicate gene losses in panel C.

Figure 2. A schematic overview of the Clan check approach. As genes with hidden paralogs are excluded some support for Xenoambulacraria emerges whilst some support for Nephrozoa disappears (*Branchiostomidae* silhouette by Michelle Site (CC by 3.0)).

References [AU please ensure references are in CB style]

1. Rouse, G.W., Wilson, N.G., Carvajal, J.I., and Vrijenhoek, R.C. (2016). New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. Nature 530, 94–97.
2. Cannon, J.T., Vellutini, B.C., Smith, J.3rd, Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. Nature 530, 89–93.
3. Philippe, H., Poustka, A.J., Chiodin, M., Hoff, K.J., Dessimoz, C., Tomiczek, B., Schiffer, P.H., Müller, S., Domman, D., Horn, M., et al. (2019). Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. Curr. Biol. 29, 1818–1826.e6.
4. Tarver, J.E., dos Reis, M., Mirarab, S., Moran, R.J., Parker, S., O'Reilly, J.E., King, B.L., O'Connell, M.J., Asher, R.J., Warnow, T., Peterson, K.J., Donoghue, P.C.J., and Pisani, D. (2016). The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. Genome Biol. Evol. 8:330-344.
5. Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? Trends Genet. 22:225-31.
6. Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. Annu. Rev. Ecol. Syst. 36:541-562.
7. Rokas, A., Williams, B.L., King, N., and Carroll, S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798-804.
8. https://www.darwintreeoflife.org/
9. Mulhair, P.O., McCarthy, C.G.P., Siu Ting, K., Creevey, C.J., and O'Connell, M.J. (2022). Filtering artefactual signal increases support for Xenacoelomorpha and Ambulacraria sister relationship in the Animal tree of life. Curr. Biol, 32, XXXXXXX.
10. Kocot, K.M., Citarella, M.R., Moroz, L.L., and Halanych, K.M. (2013). PhyloTreePruner: A Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics. Evol Bioinform Online. 9:429-35.
11. Ballesteros, J.A., and Hormiga G. (2016). A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. Mol Biol Evol. 33:2117-34.
12. Siu-Ting, K., Torres-Sánchez, M., San Mauro, D., Wilcockson, D., Wilkinson, M., Pisani, D., O'Connell, M.J., and Creevey, C.J. (2019) Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics. Mol Biol Evol. 36:1344-1356.
13. Bourlat, S.J., Nielsen, C., Lockyer, A.E., Littlewood, D.T., and Telford M.J. (2003). Xenoturbella is a deuterostome that eats molluscs. Nature. 424(6951):925-8.

14. Tihelka, E., Cai, C., Giacomelli, M., Lozano-Fernandez, J., Rota-Stabelli, O., Huang, D., Engel, M.S., Donoghue, P.C.J. and Pisani D. (2021). The evolution of insect biodiversity. Current biology 31, R1299–R1311.
15. Li, Y., Shen, X.X., Evans, B., Dunn, C.W., and Rokas A. (2021) Rooting the Animal Tree of Life, Mol Biol Evol 38:4322–4333.
16. Giacomelli, M., Rossi, E.M., Lozano-Fernandez, J., Feuda, R., and Pisani, D. (2022). Resolving tricky nodes in the tree of life through amino acid recoding. Iscience https://doi.org/10.1016/j.isci.2022.105594.