When homology fails: lessons from liver-fluke phylogenies

Pilar Alda [*,1], Annia Alba [2] & Nicolás Bonel [1]

[1] Genética y Ecología Evolutiva, CERZOS, CONICET-UNS, Bahía Blanca, Buenos Aires, Argentina
[2] IHPE, UMR 5244 Université de Perpignan Via Domitia, CNRS, IFREMER, Université de Montpellier

* Correspondence: pilaralda@gmail.com (P. Alda)

Abstract
Misaligned sequences derail evolutionary inference. Datasets from GenBank require verification of positional homology and orientation before alignment and phylogenetic analysis. Liver-fluke case studies reveal how overlooked errors skew results, underscoring the need for rigorous checks in parasitology and all molecular research.

Keywords: Phylogenetic analysis; sequence alignment; GenBank; *Fasciola hepatica*; COI barcode; *nad1*.

Reconstructing the evolutionary history of organisms depends on comparing stretches of DNA that share a common origin—so-called homologous sequences [1]. In molecular phylogenetics, this concept of homology is foundational. If the DNA fragments being compared are not truly homologous, then any conclusions about evolutionary relationships rest on shaky ground.

Despite this, errors in identifying homologous regions are surprisingly common. These mistakes often arise when researchers use sequences from public databases like GenBank without carefully checking whether they actually match the same part of a gene—or whether they are even in the correct orientation. In theory, most scientists know that accurate alignment of homologous regions is essential for reliable phylogenetic analysis [2]. In practice, though, mismatches still happen. Non-homologous sequences may be inadvertently aligned and analyzed as if they were comparable, leading to false conclusions about how species are related. These errors can ripple outward, distorting estimates of genetic diversity, population structure, and even species boundaries.

The consequences of misaligning non-homologous sequences are not hypothetical—they're already evident in the literature. For example, in avian phylogenetics, a major study was recently shown to be flawed because it compared DNA regions that were not evolutionarily equivalent [3,4]. In another case, researchers working on the giant freshwater prawn—a species spread across Southeast Asia—discovered that two different regions of the same gene had long been mistaken for each other in genetic analyses [5]. Because they were treated as homologous, these unrelated sequences produced the illusion of deep population differences that didn't actually exist. Phylogenetic trees built from these misaligned data gave misleading views of the species' history and diversity.

We encountered a similar issue while examining sequence data for the liver fluke *Fasciola hepatica*—a parasitic flatworm that infects livestock and humans worldwide. In a recent study on the species' global genetic diversity, researchers downloaded hundreds of mitochondrial sequences from GenBank, focusing on two common markers: COI and *nad1* [6]. These markers are widely used in *Fasciola* studies and have been contributed by labs around the world over the past two decades. But a closer look revealed a problem: not all sequences labeled as COI or *nad1* were directly comparable.

The issue becomes clear when we look closely at the COI sequences. Although all of them were labeled as the same gene, they actually fall into two non-overlapping groups (Fig. 1A). One set, about 447 base pairs (bp) long, was generated by Bowles *et al.* [7] using a primer pair originally designed for another parasite genus, *Echinococcus*. These sequences cover a region near the end of the COI gene. The second set, roughly 493 bp long, was amplified with primers developed specifically for *Fasciola* by Itagaki *et al.* [8], targeting the beginning of the gene. Despite both being labeled "COI", these two sets do not overlap—they cover completely different parts of the gene. Aligning them side by side is like comparing apples to oranges. Yet this is exactly what happens in some phylogenetic analyses: non-overlapping fragments are aligned as if they were homologous, creating the illusion of shared ancestry where there is none (Fig. 1B–C).

The *nad1* sequences raise a different—but equally serious—problem: orientation. In this case, all sequences were amplified using the same primers, which should make them comparable. But not all were uploaded to GenBank in the same direction. DNA sequences are directional, and some entries were stored in reverse (e.g. [9]). If this orientation isn't corrected before alignment, the sequences are read backward. This produces mismatches that look like real differences, when in fact they are artifacts. The outcome: inflated estimates of genetic divergence and misleading phylogenetic trees.

These examples show how even well-intentioned studies can fall into serious methodological traps if basic checks on homology and sequence orientation are skipped. A key part of the problem is a common assumption: that all sequences labeled as “COI” or “nad1” must correspond to the same region of the gene and are therefore directly comparable. But this isn’t always true. In public databases like GenBank, entries are only lightly curated. Important details—such as the specific primers used to amplify a fragment or the direction in which the sequence is stored—are often missing. Without this information, researchers may unknowingly align fragments that do not overlap or compare sequences that are facing in opposite directions.

This problem is especially pronounced for species like *F. hepatica*, which have been studied by dozens of research groups across the world. Over time, different labs have used different protocols, resulting in a patchwork of sequence types. The more diverse the data sources, the greater the likelihood of inconsistencies. And for researchers who are new to molecular phylogenetics, there’s a strong temptation to rely on automated tools—assuming they will "catch" these problems. But no software can replace careful inspection. Without manual checks, even experienced researchers may be misled by errors baked into the sequence data.

How to avoid these pitfalls: five essential practices

To minimize errors in sequence-based phylogenetic analyses—especially when using public data—we recommend the following best practices:

1. Verify homology before alignment
Before aligning sequences, make sure they actually correspond to the same region of the gene. This may require consulting the original publications where the sequences were first reported, or identifying the primers used for amplification. Tools like the NCBI Multiple Sequence Alignment Viewer (MSA Viewer) can help visualize how sequences map along a reference gene. For large-scale datasets, tools such as PREQUAL [2] can flag suspicious or non-homologous regions before alignment begins.

2. Check and correct sequence orientation
DNA sequences are directional, and orientation errors—especially in mitochondrial genes—are common. If a sequence is stored in reverse (3′–5′ instead of 5′–3′), it can’t be aligned properly without correction. Tools like MAFFT include an option to automatically detect and adjust reversed sequences [10].

3. Manually inspect alignments
Even the best alignment software can make mistakes. Look for tell-tale signs of problems: long gaps, clusters of mismatches, or sequences that are much longer or shorter than others. Software like MEGA [11] or AliView [12] provides an intuitive interface for manual inspection. Both tools use color-coded bases, making it easier to spot mismatches and other alignment issues at a glance (Fig. 1B).

4. Watch for artifacts in downstream results
Misalignments can distort later analyses, producing misleading phylogenetic trees or haplotype networks. Unusually long branches, unstable topologies, or unexpected genetic clusters can all be signs of alignment errors (Fig. 1C).

5. Submit complete metadata
When depositing sequences into public databases like GenBank, include essential metadata: which primers were used, the exact region amplified, and the orientation of the sequence. This information improves transparency and helps future users avoid misinterpretation.

Public repositories like GenBank have revolutionized molecular research by making genetic data broadly accessible. They allow researchers to reuse and compare sequences across studies, species, and continents. But with this opportunity comes responsibility. Not all sequences in these databases are directly comparable, and failing to check their compatibility can lead to deeply flawed evolutionary conclusions.

Homology is not just a technical detail—it is the foundation of any phylogenetic analysis. Misaligned or misoriented sequences can mislead not only a single study, but entire research programs that build on flawed assumptions. For parasitologists, evolutionary biologists, and any researcher working with public sequence data, careful validation of homology and orientation is not optional—it is essential to the integrity of our science.

Acknowledgements

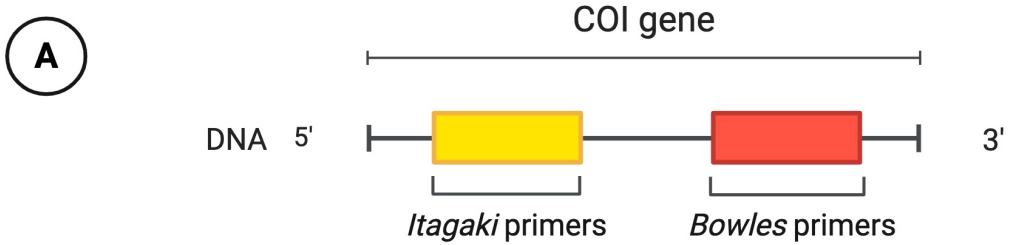
We thank the ECOS-SUD program and the Explore5 initiative from the University of Montpellier for providing the opportunity to meet, exchange ideas, and lay the groundwork for the development of this *Forum* article. We are also grateful to Micaela Müller for her thoughtful comments on the manuscript.

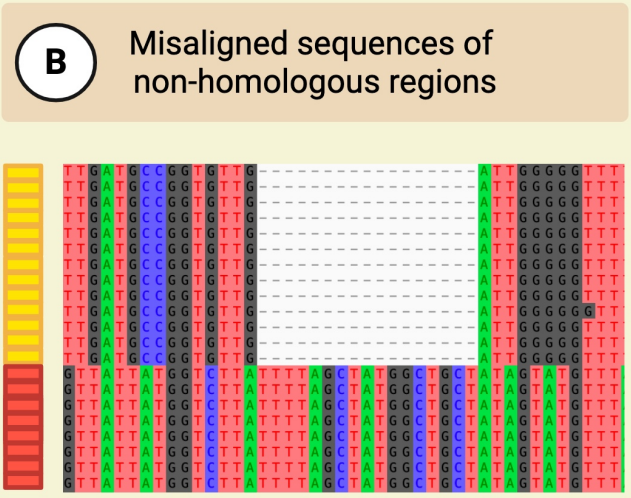
The authors declare no competing interests.

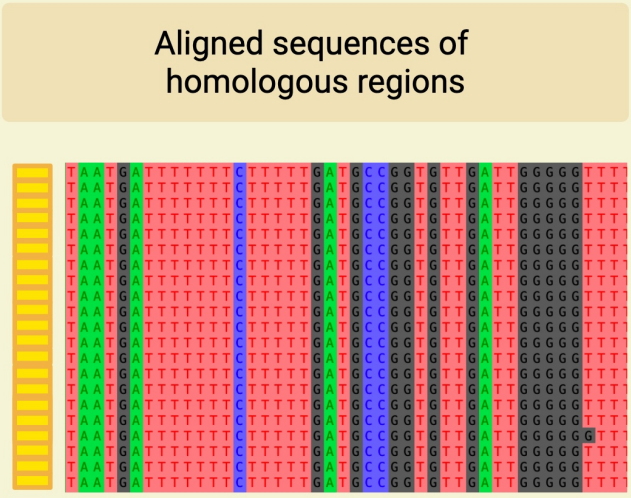
References

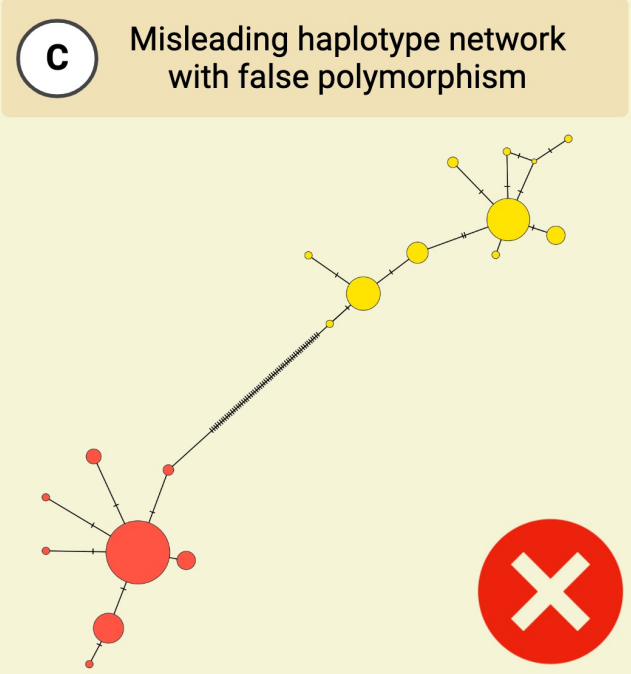
1. Lemey, P. *et al.*, eds. (2013) *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*, (2. ed, 6. print.), Cambridge University Press
2. Whelan, S. *et al.* (2018) PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* 34, 3929–3930
3. Springer, M.S. and Gatesy, J. (2024) A new phylogeny for Aves is compromised by pervasive misalignment and homology problems. *Proc. Natl. Acad. Sci. U.S.A.* 121, e2406494121
4. Wu, S. *et al.* (2024) Genomes, fossils, and the concurrent rise of modern birds and flowering plants in the Late Cretaceous. *Proc. Natl. Acad. Sci. U.S.A.* 121, e2319696121
5. Jose, D. and Harikrishnan, M. (2017) Non-homologous COI barcode regions: a serious concern in decapod molecular taxonomy. *Mitochondrial DNA Part A* 28, 482–492
6. Alvi, M.A. *et al.* (2023) Genetic variation and population structure of Fasciola hepatica: an in silico analysis. *Parasitol Res* 122, 2155–2173
7. Bowles, J. *et al.* (1992) Genetic variants within the genus Echinococcus identified by mitochondrial DNA sequencing. *Molecular and Biochemical Parasitology* 54, 165–173
8. Itagaki, T. *et al.* (2005) Genetic characterization of parthenogenic *Fasciola* sp. in Japan on the basis of the sequences of ribosomal and mitochondrial DNA. *Parasitology* 131, 679–685
9. Schwantes, J.B. *et al.* (2020) *Fasciola hepatica* in Brazil: genetic diversity provides insights into its origin and geographic dispersion. *J. Helminthol.* 94, e83
10. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30, 772–780
11. Kumar, S. *et al.* (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* 33, 1870–1874
12. Larsson, A. (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278

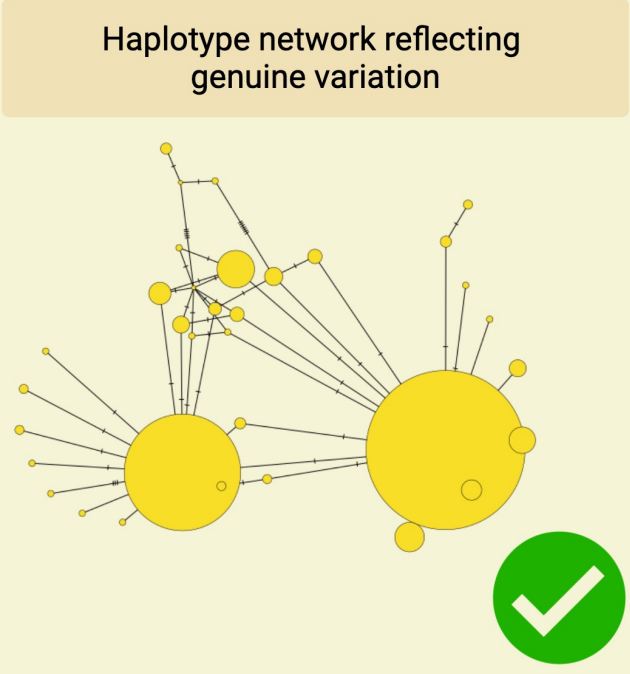
Figure 1. Misaligned and misleading: the hidden risks of using sequences from public databases without careful validation. (A) Diagram of the *Fasciola hepatica* COI gene showing two commonly amplified, non-overlapping regions—yellow (Itagaki primers) and red (Bowles primers). Although both regions belong to the same gene, they are non-homologous and should not be aligned together. (B) Comparison of multiple sequence alignments. Left: an incorrect alignment where COI fragments amplified with Bowles and Itagaki primers have been artificially aligned, producing numerous mismatches that appear as false polymorphisms. Right: a correct alignment including only homologous sequences from the Itagaki fragment. (C) Haplotype networks derived from the alignments in panel B. Each dash represents a single mutation. Left: the misaligned sequences generate a misleading network that overestimates genetic diversity due to false positional homology. Notably, the two groups of sequences are separated by over 40 mutational steps (indicated by dashes), all of which result from artificial mismatches introduced by aligning non-homologous regions. Right: the network based on homologous sequences provides an accurate view of genuine variation. Neglecting homology validation prior to alignment can lead to severely biased phylogenetic and population genetic inferences.


A
COI gene
DNA
5'
3'
Itagaki primers
Bowles primers


B
Misaligned sequences of non-homologous regions


Aligned sequences of homologous regions


C
Misleading haplotype network with false polymorphism


Haplotype network reflecting genuine variation