

A curated global synthesis of mitochondrial diversity in milkfish (*Chanos chanos*): specific marker resolution and geographic gaps in public sequence data

Ziwei Wang

University of Shanghai for Science and Technology

Biyuan Liu

52303904041@stu.ecnu.edu.cn

East China Normal University

Research Article

Keywords: *Chanos chanos*, mitochondrial DNA, D-loop, cytochrome b, COI, haplotype diversity

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11251614/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A curated global synthesis of mitochondrial diversity in milkfish (*Chanos chanos*): specific marker resolution and geographic gaps in public sequence data

Ziwei Wang¹, Biyuan Liu^{2,3*}

1. Business School, University of Shanghai for Science and Technology, Yangpu District, Shanghai, China
2. State Key Laboratory of Estuarine and Coastal Research, East China Normal University, Shanghai, China
3. East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Yangpu District, Shanghai, China

Corresponding author

Biyuan Liu

Email: 52303904041@stu.ecnu.edu.cn

Abstract

Milkfish (*Chanos chanos*), the only living species of the family Chanidae, is widely distributed across the Indo-Pacific. Although mitochondrial sequences have accumulated from regional studies, their marker specific resolution and geographic comparability remain unevaluated under a common quality control framework. Here, we integrated public GenBank sequences for three mitochondrial markers, the control region (D-loop), cytochrome b (cytb), and cytochrome c oxidase subunit I (COI), and screened them for geographic provenance, sequence integrity, and sequence level outliers. The final datasets contained 676 D-loop, 114 cytb, and 257 COI sequences. D-loop showed the highest resolution (396 haplotypes, $\pi = 0.04505$), whereas cytb and COI contained 22 and 30 haplotypes ($\pi = 0.00201$ and 0.00239). Negative Tajima's D and Fu's Fs values and the mismatch distributions were compatible with historical expansion, establishing a baseline hypothesis for future calibrated dating. Most sequence distance variance occurred within geographic groups (>95% for every marker). The French Polynesia–India cytb comparison had the highest between group component (4.86%; $p = 0.045$), based mainly on Niau, Kauehi, and India NBFGR; this contrast provides a sequence level reference spanning both ends of the Indo-Pacific gradient and a testable starting point for future barrier studies. D-loop was most informative for fine scale variation, COI provided the broadest reference coverage, and cytb supplied complementary coding region information. The curated panel therefore delivers a reproducible mitochondrial baseline for *C. chanos* across its best sampled regions, offering a practical foundation for extending coverage into currently sparse areas.

Keywords: *Chanos chanos*; mitochondrial DNA; D-loop; cytochrome b; COI; haplotype diversity

Introduction

Milkfish, *Chanos chanos* (Forsskål, 1775), is the sole living representative of an ancient gonorynchiform lineage and one of the most widely distributed Indo-Pacific fishes supporting fisheries and aquaculture (Bagarinao, 1994). The species belongs to the order Gonorynchiformes and the family Chanidae; both the genus *Chanos* and the family Chanidae are monotypic among living taxa, making *C. chanos* the only extant chanid (Grande and Poyato-Ariza, 2010; WoRMS, 2025). A lineage with this combination of taxonomic isolation, broad distribution, coastal life history, and long history of aquaculture provides an unusually clear system for examining how mitochondrial markers describe intraspecific genetic diversity across regions.

Previous genetic studies have examined regional milkfish populations, Philippine wild populations, broodstock, and aquaculture stocks using mitochondrial and nuclear markers (Ravago-Gotanco and Juinio-Meñez, 2004; Adiputra et al., 2011; Romana-Eguia et al., 2018; Santos et al., 2019). These studies provide important regional evidence, but the sequences were generated for different purposes and are distributed across multiple database records, marker types, sampling designs, and annotation standards. For a taxonomically distinctive, widely distributed species such as milkfish, a central methodological question is whether such public records can be combined into a reliable species level genetic reference panel. A marker resolved synthesis is thus needed before broad geographic patterns can be evaluated. Such a synthesis must distinguish genuine biological variation from differences in sequence length, missing data, marker coverage, and metadata quality.

Mitochondrial control region sequences can provide high resolution for recent intraspecific variation, whereas protein coding markers such as cytochrome b and COI generally provide more conserved and cross study comparable sequence data (Parida et al., 2024). Large scale comparisons of marine fishes also show that genetic diversity is shaped by interacting biological and geographic factors, including life history, range size, and sampling history (Manel et al., 2020; Clark and Pinsky, 2024). The relative suitability of these markers for milkfish should thus be evaluated from their observed diversity, alignment performance, and geographic representation rather than assumed in advance.

This study systematically integrates publicly available GenBank sequences of *C. chanos* for three mitochondrial markers: D-loop, cytb, and COI. We apply a common screening and quality control framework, compare marker specific diversity, and examine the distribution of haplotypes using minimum spanning networks and maximum-likelihood trees. We specifically ask whether public sequence data can provide a reliable species level mitochondrial panel for this taxonomically distinctive fish, how much sequence resolution each marker provides, and how uneven public-data coverage affects the interpretation of a global mitochondrial dataset.

Materials and Methods

GenBank data retrieval and screening

Nucleotide records assigned to *C. chanos* were retrieved from the NCBI GenBank nucleotide database using the organism field. Three mitochondrial markers were targeted: the control region (D-loop), cytochrome b (cytb), and cytochrome c oxidase subunit I (COI). Marker specific searches used gene or annotation terms for each locus. Records annotated as complete or nearly complete mitochondrial genomes were excluded from Marker specific datasets to avoid counting the same mitochondrial genome as a population sequence for multiple loci. One complete mitochondrial genome was retained separately as a reference sequence.

The initial marker datasets contained 1,249 D-loop, 206 cytb, and 279 COI records after exclusion of complete and nearly complete mitochondrial genomes. GenBank flat files were retained together with FASTA files so that accession identifiers, sequence annotations, isolate names, geographic qualifiers, and bibliographic information could be traced to the original records.

Only records with geographic provenance were considered for the geographic analyses. Geographic information was obtained from the GenBank /geolocname or legacy /country qualifiers and, where explicitly documented, from traceable isolate or publication information. Records without geographic information were not assigned coordinates by inference and were excluded from the geographic dataset. The resulting geo only datasets contained 680 D-loop, 119 cytb, and 271 COI sequences.

Sequence quality control

Quality control was conducted at three levels. First, each record was checked for a valid accession identifier, a non-empty nucleotide sequence, and an admissible DNA character set. Second, alignment quality was assessed using sequence occupancy, the proportion of missing or ambiguous characters, and identity relative to a Marker specific majority rule consensus. Third, a sequence level outlier score was calculated from the distance to the nearest sequence in the same marker alignment. Robust scaling used the median, with a fallback scale based on the interquartile range or standard deviation when the median absolute deviation was zero. Hamming distance and matrix operations for sequence level QC were implemented with SciPy (Virtanen et al., 2020). Sequences with more than 10% missing or ambiguous characters were excluded from the high-quality analysis set. Sequences with lower identity or unusual nearest-neighbour distances but without a hard missing data failure were classified as review candidates and were not retained in the primary high-quality dataset. This procedure produced final high-quality datasets of 676 D-loop, 114 cytb, and 257 COI sequences. The original geo only files and the recordlevel QC decisions were retained separately from the high-quality files.

Because the three markers differ in evolutionary rate and annotation history, quality control was conducted separately for each marker. No sequence was removed solely because it belonged to a geographically distinct population.

Multiple sequence alignment and trimming

Sequences were aligned separately by marker using MAFFT version 7.526 with the auto strategy (Katoh and Standley, 2013). Alignment columns were filtered using a 90% valid-character coverage threshold: columns containing gaps or ambiguous characters in more than 10% of the sequences were removed, whereas the remaining columns were retained regardless of whether they formed one uninterrupted block. This column-wise procedure prevents sparse insertion columns from fragmenting an otherwise usable alignment.

The final alignments contained 470 bp for D-loop, 744 bp for *cytb*, and 558 bp for *COI*. D-loop records were handled separately because the public dataset contains control-region records with different lengths and flanking annotations. The final D-loop alignment was generated from the high-quality geo-only set after removal of records identified as misaligned during the dedicated D-loop QC procedure.

Haplotype definition and diversity statistics

For each marker, identical aligned nucleotide sequences were collapsed into haplotypes using exact sequence matching. Haplotype number (H), segregating sites (S), the mean number of pairwise differences (k), nucleotide diversity (π), and haplotype diversity (hD) were calculated separately for each marker and for each sampling population when sample size permitted. Haplotype diversity was estimated with the small sample correction of Nei (1987), and nucleotide diversity as the mean pairwise nucleotide difference divided by the aligned sequence length. Uncertainty estimates for both statistics were obtained from 500 sequence level bootstrap resamples, with the random seed fixed at 20260902 for reproducibility.

Haplotype networks

A minimum-spanning network (not a median-joining network) was constructed independently for each marker using the haploNet implementation in the R package pegas (Paradis, 2010). Pairwise distances among haplotypes were calculated from nucleotide differences with pairwise deletion for positions unresolved in either sequence, and the network was constructed from the complete haplotype distance matrix. Each node represents a haplotype, with node size proportional to haplotype frequency and node sectors indicating sampling population composition. Network figures were prepared with smaller nodes and reduced label density to improve readability in the high diversity D-loop dataset. Edge lengths in the displayed layout are layout artefacts and are not treated as a scale of evolutionary distance.

Maximum-likelihood phylogenetic trees

Maximum-likelihood trees were inferred independently for the three high-quality marker alignments using IQ-TREE version 2 (Minh et al., 2020). ModelFinder (Kalyaanamoorthy et al., 2017) selected TPM3u+F+I+G4 for D-loop, TN+F for *cytb*, and HKY for *COI* according to the Bayesian information criterion. Each analysis included 1,000 ultrafast bootstrap replicates and 1,000 SH-aLRT replicates (Hoang et

al., 2018). Trees were used as a complementary diagnostic of haplotype structure rather than as a basis for dated inference.

Neutrality tests, population diversity, and exploratory AMOVA

Tajima's D was calculated using the `tajima.test` function in `pegas` (Paradis, 2010). Fu's F_s was calculated using the Fu (1997) formulation in log space, using the same sample size, haplotype number, and mean pairwise difference as in the marker level diversity table. The 500 neutral, no-recombination Kingman coalescent simulations were executed in parallel with `joblib` (joblib developers, 2025), with a fixed seed of 20260902. Empirical left tail probabilities were estimated from these simulations; because 500 replicates resolve tail probabilities only to approximately 0.002, these probabilities should be regarded as preliminary. R, based diversity, neutrality, network, and AMOVA analyses were performed in R (R Core Team, 2025) using `ape` (Paradis and Schliep, 2019) and `pegas` (Paradis, 2010).

Populationlevel diversity statistics were summarized for sampling groups with sufficient sample size and were used to describe Marker specific patterns. An exploratory countrylevel analysis of molecular variance (AMOVA) was conducted separately for each marker using the distance based implementation in `pegas` (Paradis, 2010), following the framework of Excoffier et al. (1992). Only country groups represented by at least five high-quality sequences were included. AMOVA was used to describe the allocation of sequence distance variance among broad geographic groups; it was not used as a substitute for balanced sampling or precise geographic coordinates.

Historical dynamics and relative lineage depth

Historical demographic dynamics were evaluated using pairwise mismatch distributions. For each marker and each sampling group with at least five sequences, pairwise nucleotide differences were calculated with pairwise deletion, and the distribution peak, mean pairwise difference, and Harpending's raggedness index were recorded. The raggedness index was calculated as the summed squared deviation between observed mismatch class frequencies and those expected under a fitted unimodal expansion model (Harpending et al., 1994). Distributions were plotted with `Matplotlib` (Hunter, 2007). Mismatch distributions were interpreted together with Tajima's D, Fu's F_s , haplotype network structure, and the distribution of haplotype frequencies; a single unimodal distribution or a negative neutrality statistic was not treated as sufficient evidence of demographic expansion.

Relative lineage depth was summarized from the terminal and internal branch lengths of the IQ-TREE maximum-likelihood trees. Absolute divergence or expansion times were not estimated because the present dataset lacks a reliable *C.chanos* specific

calibration point, a validated mitochondrial substitution rate, and a consistent temporal sampling design. The branch length analysis therefore describes relative genetic depth among Marker specific lineages rather than dates in years.

Results

Data curation and public data coverage

Of the 1,734 retrieved marker records, 1,070 (61.7%) carried usable geographic provenance; the remainder either lacked geographic annotation or were hatchery coded records that were not treated as wild geographic samples. Geographic screening yielded 680 D-loop, 119 cytb, and 271 COI records, and Marker specific quality control retained 676, 114, and 257 of these, corresponding to retention rates of 99.4%, 95.8%, and 94.8%, respectively. The curated panel therefore contains 1,047 Marker specific records. Standardised alignments spanned 470 bp (D-loop), 744 bp (cytb), and 558 bp (COI); the D-loop alignment was the shortest but by far the most variable, whereas cytb and COI provided longer and more conserved coding region alignments. Because the same individual was generally not sequenced at all three loci and the public records provide no consistent multilocus sampling design, the three markers were analysed independently.

Regional coverage was highly uneven: of the 1,070 geographically annotated records, 722 came from Southeast Asia, 214 from South Asia, and 90 from French Polynesia, with the remaining 44 distributed among other regions. Several portions of the species range were represented sparsely or only by isolated records. The panel thus reflects the quality and geographic structure of the public archive rather than a spatially balanced sampling survey. In comparing markers or regions, the large D-loop contribution from Philippine and Indonesian wild population studies, the concentration of cytb records in French Polynesia and India, and the strong COI representation of India should be borne in mind. Accordingly, the database supports comparative analyses of the sampled regions and of marker performance, but does not support the inference that unsampled regions have low diversity or lack population structure.

Marker diversity and neutrality statistics

The three markers differed by roughly an order of magnitude in the amount of sequence variation they retained (**Table 1.**). D-loop showed the highest markerlevel diversity, with 396 haplotypes among 676 sequences, 215 segregating sites, a mean pairwise difference of 21.17 substitutions, and nucleotide diversity of 0.04505 ($hD = 0.99671$). The coding markers were less variable: cytb contained 22 haplotypes among 114 sequences, 27 segregating sites, $k = 1.50$, and $\pi = 0.00201$ ($hD = 0.68468$), whereas COI contained 30 haplotypes among 257 sequences, 41 segregating sites, $k = 1.34$, and $\pi = 0.00239$ ($hD = 0.77432$). D-loop therefore recovered approximately eighteen times as many haplotypes as cytb from fewer than six times as many sequences, a difference reflected in segregating sites and mean pairwise differences rather than in sample size alone. Tajima's D was negative for all three markers but significant only for cytb and

COI; Fu's F_s was strongly negative throughout, although its probabilities are bounded at approximately 0.002 by the 500 replicate resolutions.

Distribution of haplotypes among major sampling groups

D-loop sampling was concentrated in 22 fine scale groups. Six groups — CRwSarangani, CRwBohol, CRwCamiguin, CRwAurora, CRwMedan, and CRwAntique — contributed 50 sequences each. Other groups included CRwSulawesi, CRwClaveria, CRwCurrima, CRwCamarinesSur, CRwDavao, CRwGeneralSantos, CRwZamboanga, and the Pangasinan groups Bonuan, SanFabian, and Binmaley. This dataset therefore provided comparatively detailed within region sampling for Philippine and Indonesian wild populations.

The cytb dataset was dominated by two French Polynesian sampling localities, Niau (53 sequences) and Kauehi (35), and by the India NBFGR group (20). These three groups provided the clearest regional comparison available for cytb, but they represented two French Polynesian localities and one Indian reference group rather than complete Pacific and Indian Ocean populations. The remaining cytb records represented small groups from the South China Sea, Indonesia, Taiwan, Pakistan, and Pangasinan.

COI was dominated by the India Indian coast group, which contributed 180 high-quality sequences. Other relatively well represented groups included Taiwan ASIZP, Qatar, ASIZP, and Batangas; several additional groups were represented by only one to three sequences. This unequal distribution of records among groups was taken into account when interpreting populationlevel statistics.

Population diversity

D-loop diversity was high in both major countrylevel aggregates. The Philippines contained 595 high-quality D-loop records, 370 haplotypes, and nucleotide diversity of 0.04498 ($hD = 0.99672$); Indonesia contained 81 records, 47 haplotypes, and nucleotide diversity of 0.04517 ($hD = 0.97809$). The Philippine and Indonesian haplotype counts together exceed the 396 haplotypes in the full D-loop dataset, indicating that a subset of haplotypes is shared between the two countries.

For cytb, French Polynesia contained 88 sequences, 9 haplotypes, and nucleotide diversity of 0.00193, whereas India contained 20 sequences, 10 haplotypes, and nucleotide diversity of 0.00220. The Indonesian cytb group contained only five sequences and yielded lower estimated diversity, so its statistics should be interpreted cautiously.

For COI, India contributed 190 sequences in total, of which 180 came from the India Indian coast group; the countrylevel aggregate contained 18 haplotypes and nucleotide diversity of 0.00223. Among the better represented groups, Vietnam showed the highest nucleotide diversity (12 sequences, 11 haplotypes, $\pi = 0.00434$), although this group is small and the estimate is correspondingly unstable. The Philippines and Taiwan were represented by 13 and 10 sequences, respectively. These comparisons are descriptive, because population sample sizes and geographic coverage were not balanced across markers.

Exploratory countrylevel AMOVA

Exploratory countrylevel AMOVA showed that most sequence distance variance occurred within groups (**Table 2**). For D-loop, the Indonesia–Philippines grouping assigned 0.38% of the variance to between group differences and 99.62% to within group differences ($p = 0.120$). For cytb, the French Polynesia–India grouping assigned 4.86% to between group differences and 95.14% to within group differences ($p = 0.045$) — the highest between group variance proportion among the three markers. For COI, 2.73% of the variance was between the six country groups meeting the sample size criterion and 97.27% was within groups ($p = 0.195$).

The cytb comparison therefore provided the strongest broad regional signal in the present AMOVA, but it remains a wide scale regional comparison rather than a direct test of the Indo-Pacific Barrier: Niau and Kauehi represent two French Polynesian sampling localities rather than the entire Pacific population, and India NBFGR does not represent the entire Indian Ocean population. Because the groupings were based on uneven public data coverage and were not sampled under a common spatial design, the AMOVA results should be interpreted as exploratory evidence of geographic variance rather than as estimates of complete population structure.

Haplotype networks and maximum-likelihood trees

The D-loop network contained 396 haplotypes and was characterized by many low frequency haplotypes surrounding more frequent nodes. No single country specific cluster encompassed all records from a region, and the high number of singleton and low frequency haplotypes produced a diffuse network structure, indicating extensive haplotype diversity within the Philippine and Indonesian sampling groups (**Fig.1 A**).

The cytb and COI networks were more compact because they contained fewer haplotypes relative to their sample sizes — 22 and 30 haplotypes, respectively. Shared or closely connected haplotypes occurred across multiple sampling groups, consistent with weak overall differentiation in the public dataset (**Fig. 1 B, C**).

The maximum-likelihood trees showed the same broad marker contrast (**Fig.2**). The D-loop tree contained many short terminal branches and a large number of low frequency tips, reflecting its high sequence resolution; the cytb and COI trees were more compact and contained fewer distinct sequence states. The trees did not provide uniform, strongly supported geographic separation across all sampling groups, so their interpretation was based on the combined patterns of branch support, haplotype composition, and Marker specific diversity rather than on any single topology.

Mismatch distributions and relative lineage depth

Pairwise mismatch distributions were strongly marker dependent (**Fig. 3**). The pooled D-loop dataset showed a mean pairwise difference of 21.17 substitutions and a raggedness index of 0.000589, with a distribution peak at 22 differences. The pooled cytb dataset showed a mean of 1.50 differences and a raggedness index of 0.0636, with a peak at one difference; the pooled COI dataset showed a mean of 1.34 differences and a raggedness index of 0.0940, also with a peak at one difference. At the level of the

better represented D-loop groups, raggedness values were low for most Philippine and Indonesian groups, whereas the smaller Pangasinan groups showed less smooth distributions. These patterns are compatible with recent expansion signals in some groups, but the uneven public sampling prevents a uniform demographic history from being assigned to the species.

The relative branch-length analysis provided a comparative rather than temporal result. Maximum observed branch lengths were 0.0366 for D-loop, 0.00544 for *cytb*, and 0.0258 for COI. These values are model and alignment dependent substitutions per site quantities and should not be interpreted as divergence times without calibration.

Discussion

Taxonomic context and marker suitability

Milkfish occupies a distinctive taxonomic position among teleosts. It is classified as *C. chanos* (Forsskål, 1775), in the genus *Chanos*, family *Chanidae*, order *Gonorynchiformes*, and class *Actinopterygii*. *Chanidae* is a monotypic extant family, and *Chanos* is a monotypic extant genus containing the only living chanid species. This taxonomic distinctiveness, together with broad Indo-Pacific distribution and aquaculture importance, provides a clear biological context for a marker resolved specieslevel synthesis.

The marker suitability assessment follows from the results of this study rather than from a priori preference. D-loop produced the largest number of haplotypes and the highest nucleotide diversity, indicating the highest resolution for fine scale intraspecific variation in the present dataset. Its higher resolution was accompanied by greater sensitivity to alignment strategy, length variation, and repeat associated indels, consistent with the control-region structure reported previously in milkfish (Ravago et al., 2002). This combination of high information content and greater analytical sensitivity is important: D-loop can reveal variation that coding markers do not resolve, and the additional variation can be interpreted reliably once alignment and quality control decisions have been documented — an advantage the present framework makes explicit and transferable to future studies.

COI was represented by more records than *cytb* and showed a more conserved sequence profile. This combination supports broad public data integration and specimen identification, although lower variation reduces its ability to distinguish very recent population differences. *Cytb* provided an additional coding region marker with intermediate practical value in the present dataset. Thus, the results support a target dependent conclusion: D-loop is the most informative marker for fine scale diversity discovery, COI is the most practical marker for broad reference data integration, and *cytb* is a complementary coding region marker. This ranking is specific to the present dataset and objectives, not a universal ranking for all *C. chanos* studies; it nevertheless gives researchers working on this species an evidence-based basis for choosing markers according to their objectives. The comparison also illustrates why marine fish genetic diversity studies benefit from reporting marker choice explicitly rather than treating

mitochondrial loci as interchangeable measures of diversity (Manel et al., 2020; Clark and Pinsky, 2024).

Haplotype sharing and geographic structure

The haplotype networks and maximum-likelihood trees did not show uniform separation of samples according to their geographic labels. Closely connected haplotypes occurred across multiple sampling groups, while some groups contained several low frequency or private haplotypes. This pattern is consistent with a combination of shared mitochondrial variation and local frequency differences. Although it does not by itself distinguish among natural dispersal, historical demographic processes, and human mediated movement of aquaculture seedstock, it defines a set of competing hypotheses that nuclear markers and denser sampling can now test directly.

The cytb dataset provided a focused comparison between French Polynesia and India. The French Polynesian records were separated into the Niau ($n = 53$) and Kauehi ($n = 35$) localities, whereas the Indian records comprised the India NBFGR group ($n = 20$). In the exploratory country level AMOVA, 4.86% of cytb sequence distance variance was assigned to the difference between French Polynesia and India ($p = 0.045$). This result indicates a detectable broad geographic component in the cytb dataset. The comparison spans the two ends of the Indo-Pacific gradient and therefore provides a sequencelevel reference for future barrier studies, even though it is not itself a direct test of the Indo-Pacific Barrier: it includes two localities from French Polynesia, one Indian reference group, and unequal sampling histories. A dedicated barrier test would require balanced sampling on both sides of a defined geographic boundary, and the present panel identifies the intervening regions where that additional sampling would be most informative.

Because the geographic groupings were defined by uneven public data coverage rather than by a common spatial design, geographic structure is interpreted here from haplotype composition, network connectivity, tree topology, and descriptive population diversity. These complementary lines of evidence describe the sampled regions well and provide a baseline against which future balanced surveys can be compared, even though they are not a substitute for a sampling design with reliable coordinates and comparable sample sizes.

The contrast with earlier broad scale results is instructive. Ravago-Gotanco and Juinio-Meñez (2004), using RFLP analysis of the control region, recovered approximately 11% of molecular variance among three broad groups Indian Ocean, west Pacific, and north-central Pacific and interpreted this as limited gene flow among disjunct populations. The present sequencelevel analysis assigns only 4.86% of cytb variance to the French Polynesia–India contrast. Several non-exclusive explanations exist: differing marker resolution, the inclusion here of only two French Polynesian localities rather than a north-central Pacific sample, and hatchery mediated admixture that would dilute an historical signal. Testing among them requires the north-central Pacific populations included in the earlier work.

Human mediated seedstock movement as a third mechanism

This explanation deserves particular weight in milkfish, which has been farmed for centuries across Southeast Asia and Taiwan (Bagarinao, 1994). The industry has depended on seedstock movement throughout that history: collection grounds, nurseries, and grow out facilities are often located in different provinces or countries, and the Philippines where milkfish is the national fish — has relied on imports from Indonesia and Taiwan, with peak season imports estimated at up to 360 million fry (Salayo et al., 2021). Seedstock circulation is thus a structural feature of how this taxon occupies its range, not a recent perturbation.

Two independent lines of genetic evidence support treating this as a mechanism rather than a caveat. Santos et al. (2019), analysing control-region and *cytb* sequences from 631 Philippine and Indonesian fish — largely the same underlying records examined here — attributed weak structure and high gene flow partly to the influx of haplotypes brought about by importation. Romana-Eguia et al. (2018) reached a comparable conclusion from short tandem repeats, noting that stock transfers and translocations could account for negligible differentiation among Philippine stocks and that hatchery escapees may subsequently spawn in the wild. Both processes would homogenise mitochondrial variation regionally while leaving underlying structure intact but progressively harder to detect.

This distinction matters for interpretation. In most Indo-Pacific phylogeographic systems, mitochondrial structure reflects a balance between historical vicariance and contemporary dispersal; for milkfish a third term must be added. Extensive haplotype sharing cannot be read as evidence of unrestricted natural gene flow, because anthropogenic mixing generates the same pattern; equally, the weak structure reported here should not be taken as evidence against historical barriers, since admixture would dilute rather than erase their signal. The two hypotheses make different, testable predictions: natural dispersal should generate isolation by distance, whereas human mediated movement should produce scattered, distance independent sharing concentrated around fry collection grounds, hatchery centres, and grow out regions. Distinguishing them requires nuclear markers and, ideally, archived samples predating the expansion of hatchery-based seed supply; the present panel provides a mitochondrial baseline for such comparisons.

Marker specific representation of the major sampled regions

The Philippines is the best represented region for D-loop and provides the greatest within region resolution in the current dataset; within this region, the panel is sufficiently dense to characterise local haplotype composition in detail. The Indonesian D-loop records include Medan and Sulawesi and provide a useful comparison with Philippine wild groups, although they remain less geographically extensive than the Philippine sampling; the D-loop comparison between the two countries should therefore be understood as a comparison between the available sampled groups rather than a complete country wide estimate.

India is strongly represented in COI and is also present in *cytb*. The Indian COI dataset contributes substantial information about within region variation, although Indian coast combines records for which the public metadata do not consistently provide a

single fine scale sampling locality and should not be interpreted as one precisely located population. French Polynesia is most strongly represented in *cytb*, particularly by Niau and Kauehi, which allows a clearer comparison between two named localities, although it does not provide equivalent coverage of the wider Pacific distribution.

These Marker specific geographic concentrations explain why the three markers cannot be treated as interchangeable global surveys: each emphasizes a different combination of sampling history, geographic region, and evolutionary resolution. Read jointly, however, the three markers cover the species range more broadly than any single marker would, so that their combination offers wider geographic coverage than has previously been available for *C. chanos*.

Advantages and constraints of public mitochondrial data

Public sequence data provide an efficient way to assemble a cross-study reference panel for a widely distributed aquaculture species. For a taxonomically distinctive species with a broad Indo-Pacific range, this approach makes it possible to compare regional records that would otherwise remain separated by study specific sampling and annotation conventions, allowing accessions generated by multiple projects to be compared under a common alignment and quality-control framework. The same dataset can support Marker specific diversity estimates, haplotype networks, phylogenetic trees, and an explicit assessment of geographic sampling imbalance.

The same integration also introduces constraints. Metadata quality varies among studies; sample sizes are highly uneven; some localities are represented only at country level; market or trade associated records may not represent natural populations; and different studies may target different portions of the D-loop. The public records also generally do not provide the original chromatograms or sufficient information to evaluate within individual heteroplasmy directly. These factors define the resolution at which geographic and demographic inference is currently achievable, and they indicate where targeted new sampling would most efficiently strengthen inference. Curated mitochondrial reference resources have adopted a comparable logic, combining broad database coverage with explicit quality control and annotation procedures (Kasmi et al., 2023).

The results therefore support a marker resolved description of mitochondrial diversity and an exploratory assessment of broad geographic variance in the available *C. chanos* records. The AMOVA results provide a first quantitative comparison of broad groups and a reference point for future work, even though a balanced sampling design would be needed before population boundaries can be defined. Within the scope of the NCBI GenBank nucleotide records and Marker specific searches used here, this study provides a first systematic integration of publicly available global D-loop, *cytb*, and COI sequences for *C. chanos* under a common screening, alignment, and quality control framework. This statement refers to the scope of the present database integration and does not imply that previous regional studies or single marker analyses did not exist. The value of the dataset lies in the combination of a standardized high-quality reference panel, comparison of coding and control region resolution, and explicit documentation

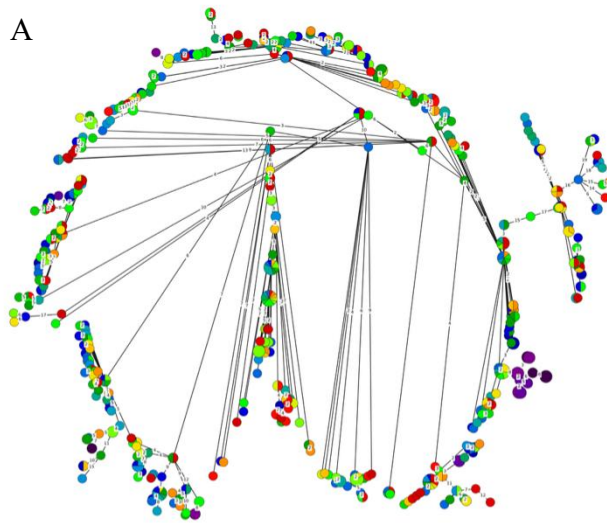
of where public data are concentrated and where they remain sparse — information that makes the panel directly usable for designing the next generation of sampling.

Conclusion

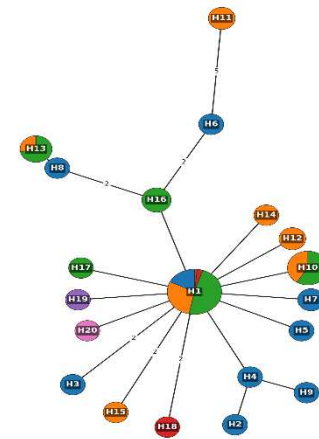
A curated public data panel of *C. chanos* mitochondrial sequences revealed clear differences among D-loop, cytb, and COI. D-loop provided the highest intraspecific resolution, whereas COI and cytb provided more conserved coding region comparisons. Negative neutrality statistics and the mismatch distributions were compatible with historical expansion signals, establishing a baseline hypothesis that subsequent work can test against calibrated models. The uneven geographic distribution of GenBank records shows that the present dataset is a systematic global public data synthesis rather than a spatially balanced survey of the species range, while providing representative coverage of the best sampled regions. This panel therefore offers a reproducible baseline for *C. chanos* mitochondrial diversity and a practical starting point for expanding coverage into currently sparse areas; a time calibrated demographic analysis will benefit from additional sampling and a defensible mutation rate or fossil calibration.

Figures

A



B



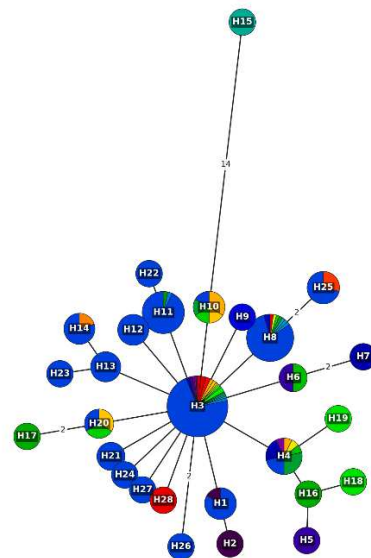
Populations

Bonuan (Philippines)	GeneralSantos (Philippines)	Davao (Philippines)	Claviera (Philippines)
Sarabian (Philippines)	GeneralSantos (Philippines)	Aurora (Philippines)	Claviera (Philippines)
Bimray (Philippines)	Zamboanga (Philippines)	Medan (Indonesia)	CamarinesSur (Philippines)
Sulawesi (Indonesia)	Zamboanga (Philippines)	Curimas (Philippines)	CamarinesSur (Philippines)
Sanagani (Philippines)	Campung (Philippines)	Curimas (Philippines)	Antique (Philippines)
Bohol (Philippines)	Davao (Philippines)		

Populations

- NDFGR (India)
- Kauehi (French Polynesia)
- Niau (French Polynesia)
- South China Sea (open sea)
- Lampung (Indonesia)
- Pangasinan (Philippines)
- Cilacap (Indonesia)

C



- ZSI SRC (India)
- Singapore (Singapore)
- Gombong (Indonesia)
- FGB (India)
- Philippines
- ASIZP (Taiwan, China)
- Kochi (Japan)

- Milkfish (Saudi Arabia)
- Indian coast (India)
- Cebu (Philippines)
- National Capital Region (Philippines)
- KACH (India)
- MZ (Mexico)

Populations

■	NBFG:86 (India)	■
■	India other (India)	■
■	Indonesia	■
■	Qatar (Qatar)	■
■	Moheshkhali (Bangladesh)	■
■	KhanhHoa (Vietnam)	■

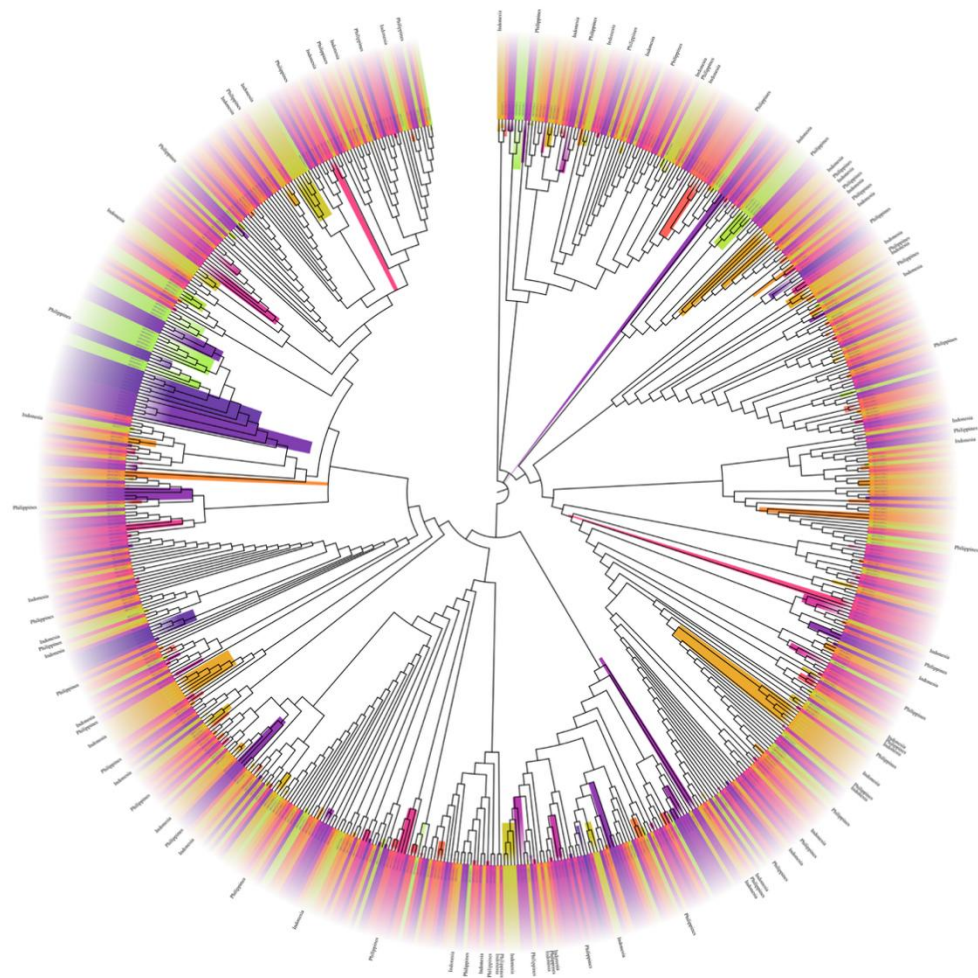
ons

PhuYen (Vietnam)	A417c (Indonesia)
BinhDinh (Vietnam)	A417b (Indonesia)
QuangBinh (Vietnam)	A417a (Indonesia)
NgheAn (Vietnam)	CES (Australia)
ANFB-2016 (India)	A51PZ (Taiwan, China)
Liverpool (UK)	Batangas (Philippines)

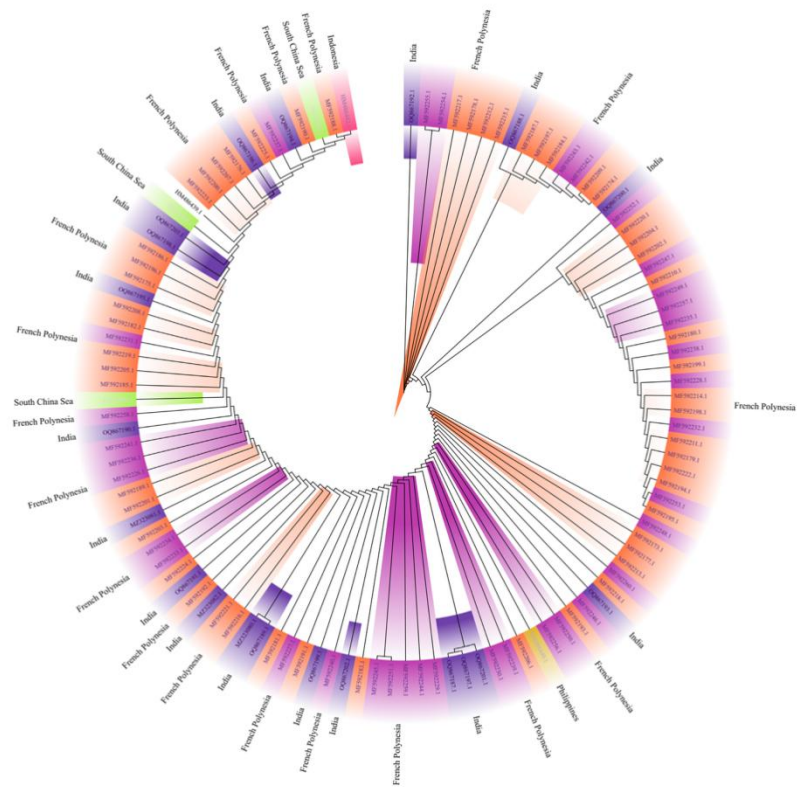
-  Bushehr (Iran)
- Kwazulu Natal (South Africa)
-  Jawa Tengah (Indonesia)
-  Society Islands (French Polynesia)
-  South China Sea (open sea)
-  Queensland (Australia)

Figure 1. Haplotype networks of *Chanos chanos* mitochondrial markers. Minimum-spanning networks are shown for (A) D-loop, (B) *cytb*, and (C) COI. Each node represents a haplotype, with node size proportional to haplotype frequency and node sectors indicating the sampling-population composition of that haplotype. Networks were constructed from high-quality sequences with geographic provenance: the D-loop panel contains 676 sequences and 396 haplotypes, the *cytb* panel 114 sequences and 22 haplotypes, and the COI panel 257 sequences and 30 haplotypes. Connecting lines represent mutational steps but are drawn to a uniform length; the displayed layout is not a proportional evolutionary-distance scale. In the D-loop panel, most haplotypes occur at low frequency, producing a diffuse structure dominated by rare haplotypes.

A



B



C

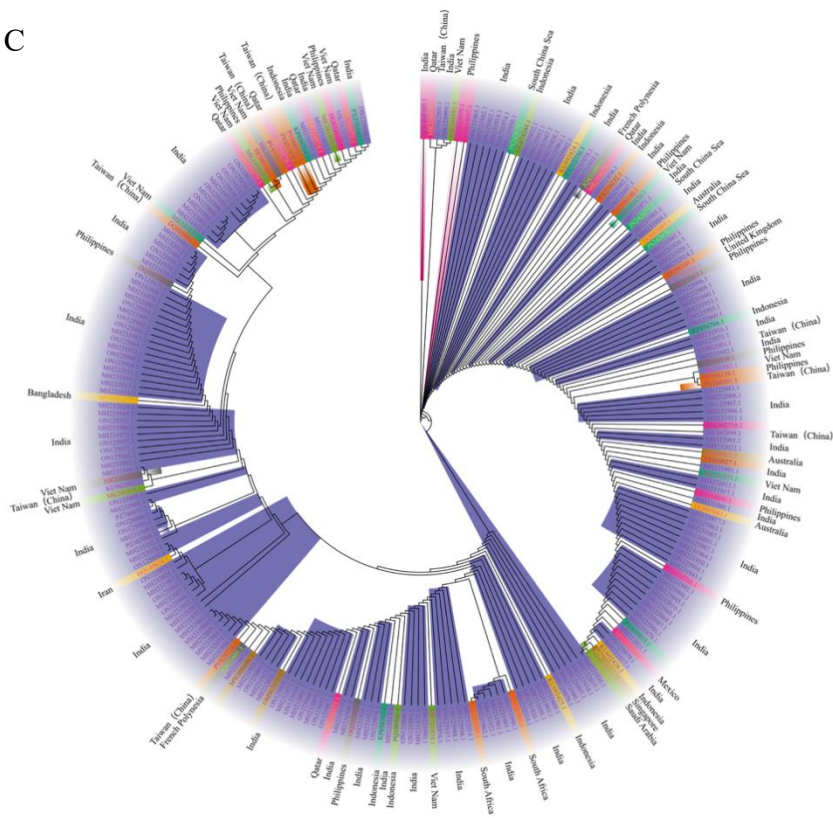


Figure 2. Maximum-likelihood trees for *Chanos chanos* mitochondrial markers. Trees are shown for (A) D-loop, (B) *cytb*, and (C) COI. Trees were inferred in IQ-TREE 2 under marker-specific models selected by ModelFinder: TPM3u+F+I+G4 for D-loop, TN+F for *cytb*, and HKY for COI. Branch support was assessed with 1,000 ultrafast bootstrap and 1,000 SH-aLRT replicates.

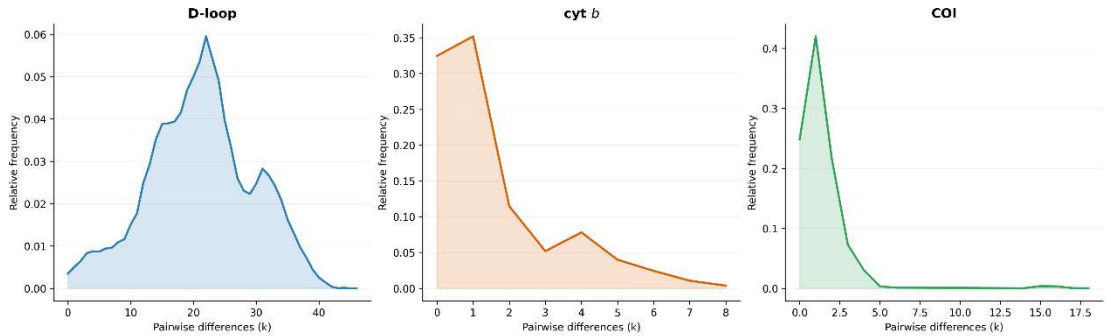


Figure 3. Pairwise mismatch distributions. Relative-frequency distributions of pairwise nucleotide differences are shown for the pooled D-loop, *cytb*, and COI datasets. The distributions were interpreted together with neutrality statistics and haplotype-network structure.

Tables

Table 1. Marker-level mitochondrial diversity and neutrality statistics.

Marker	n	H	S	$\pi \pm \text{SD}$	$hD \pm \text{SD}$	k	Tajima's D	p	Fu's Fs	p
D-loop	676	396	215	0.04505 ± 0.00069	0.99671 ± 0.00037	21.1746	-0.8836	0.3769	-550.6140	0.002
<i>cytb</i>	114	22	27	0.00201 ± 0.00029	0.68468 ± 0.04301	1.4988	-2.0894	0.0367	-16.5149	0.002
COI	257	30	41	0.00239 ± 0.00025	0.77432 ± 0.02293	1.3355	-2.2898	0.022	-27.5947	0.002

Table 2. Exploratory country-level analysis of molecular variance (AMOVA) for each mitochondrial marker.

Marker	Grouping	Populations	n	n_pop	% between	% within	p	Permutations
D-loop	country	Indonesia; Philippines	676	2	0.38	99.62	0.12	199
<i>cytb</i>	country	French Polynesia; India	108	2	4.86	95.14	0.045	199
COI	country	India;Indonesia;Philippines; Qatar;Taiwan;Viet Nam	241	6	2.73	97.27	0.195	199

Declarations

Authors' Contribution

Conceptualization: [Biyuan Liu]; Data curation: [Biyuan Liu, Ziwei Wang]; Formal Analysis: [Ziwei Wang, Biyuan Liu]; Investigation: [Ziwei Wang, Biyuan Liu]; Methodology: [Ziwei Wang, Biyuan Liu]; Project administration: [Biyuan Liu]; Resources: [Biyuan Liu, Ziwei Wang]; Software: [Ziwei Wang, Biyuan Liu]; Supervision: [Biyuan Liu]; Validation: [Biyuan Liu, Ziwei Wang]; Visualization: [Biyuan Liu, Ziwei Wang]; Writing – original draft: [Ziwei Wang]; Writing – review & editing: [Ziwei Wang, Biyuan Liu]

Funding information

No funding was received for this research.

Competing of Interest

The authors declare no conflicts of interest.

Ethical Conduct Approval – IACUC

Not applicable. This study did not involve live animals or animal experiments; therefore, IACUC approval was not required.

Data availability statements

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request.

References

- Adiputra, Y. T., Hsu, T. H., and Gwo, J. C. (2011). Genetic relationship of milkfish (*Chanos chanos*) from Indonesia, the Philippines and Taiwan using mitochondrial cytochrome b sequences. *Journal of Applied Ichthyology*, 27, 1100–1103. <https://doi.org/10.1111/j.1439-0426.2010.01629.x>
- Bagarinao, T. (1994). Systematics, distribution, genetics and life history of milkfish, *Chanos chanos*. *Environmental Biology of Fishes*, 39, 23–41.
- Clark, R. D., and Pinsky, M. L. (2024). Global patterns of nuclear and mitochondrial genetic diversity in marine fishes. *Ecology and Evolution*, 14, e11365. <https://doi.org/10.1002/ece3.11365>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9, 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131, 479–491.
- Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking, and background selection. *Genetics*, 147, 915–925.
- Grande, T., and Poyato-Ariza, F. J. (2010). Phylogenetic relationships of fossil and Recent gonorynchiform fishes (Teleostei: Ostariophysi). *Zoological Journal of the Linnean Society*, 159, 542–582.
- Harpending, H. C., et al. (1994). Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology*, 66, 591–600.
- Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35, 518–522.
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14, 587–589. <https://doi.org/10.1038/nmeth.4285>
- Kasmi, S., Eschbach, E., and Hanel, R. (2023). Mare-MAGE curated reference database of fish mitochondrial genes. *BMC Genomic Data*, 24, 18. <https://doi.org/10.1186/s12863-023-01119-4>

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. <https://doi.org/10.1093/molbev/mst010>

Manel, S., Guerin, P.-E., Mouillot, D., Blanchet, S., Velez, L., Albouy, C., and Pellissier, L. (2020). Global determinants of freshwater and marine fish genetic diversity. *Nature Communications*, 11, 692. <https://doi.org/10.1038/s41467-020-14409-7>

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37, 1530–1534. <https://doi.org/10.1093/molbev/msaa015>

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Paradis, E. (2010). PEGAS: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, 26, 419–420. <https://doi.org/10.1093/bioinformatics/btp696>

Paradis, E., and Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35, 526–528. <https://doi.org/10.1093/bioinformatics/bty633>

Parida, S., Parida, A., Rout, J. R., Kumar, S., Dhar, S., Bisai, S., Behera, P., and Behera, A. (2024). Mitochondrial genes and their application in fish diversity studies. *In Current Trends in Fisheries Biotechnology*. https://doi.org/10.1007/978-981-97-3165-7_2

Ravago, R. G., Monje, V. D., and Juinio-Ménez, M. A. (2002). Length and sequence variability in mitochondrial control region of the milkfish, *Chanos chanos*. *Marine Biotechnology*, 4, 40–50. <https://doi.org/10.1007/s10126-001-0076-4>

Ravago-Gotanco, R. G., and Juinio-Meñez, M. A. (2004). Population genetic structure of the milkfish, *Chanos chanos*, based on PCR-RFLP analysis of the mitochondrial control region. *Marine Biology*, 145, 789–801. <https://doi.org/10.1007/s00227-004-1372-0>

Romana-Eguia, M. R. R., Santos, B. S., Ikeda, M., Basiao, Z. U., and Kijima, A. (2018). Genetic assessment of milkfish (*Chanos chanos* Forsskal) stocks based on novel short tandem repeats for marker-aided broodstock management. *Aquaculture Research*, 49, 1557–1568. <https://doi.org/10.1111/are.13610>

Salayo, N. D., Marte, C. L., Toledo, J. D., Gaitan, A. G., & Agbayani, R. F. (2021). Developing a self-sufficient Philippine milkfish industry through value chain analysis. *Ocean & Coastal Management*, 201, 105426. <https://doi.org/10.1016/j.ocecoaman.2020.105426>

Santos, B. S., Basiao, Z. U., and Quilang, J. P. (2019). Genetic diversity and patterns of demographic expansion in natural populations of milkfish, *Chanos chanos* (Forsskal, 1775), in

the Philippines. *Mitochondrial DNA Part A*, 30, 312–324.
<https://doi.org/10.1080/24701394.2018.1504931>

R Core Team. (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, 261–272.
<https://doi.org/10.1038/s41592-019-0686-2>

World Register of Marine Species. (2025). *Chanos chanos* (Fabricius, 1775). Retrieved September 2026, from <https://www.marinespecies.org/aphia.php?p=taxdetails&id=217625>

Joblib developers. (2025). joblib: Computing with Python functions. Zenodo.
<https://doi.org/10.5281/zenodo.15496554>