Aging albacore tuna using epigenetic clocks

Chevrier Thomas[1,2], Sylvain Bonhommeau[2], Michael Thompson[3], Giorgia Del Vecchio[3], Anne-Elise Nieblas[1], Dominique Cowart[1], Julie Imbert Nguyen[2], Serge Bernard[4], Jessica Farley[5], Yann Guiguen[6], Cedric Cabau[7], Christophe Klopp[7], Joseph A. Zoller[8], Steve Horvath[8,9,10], Robert Brooke[11], Matteo Pellegrini[3]

[1]Company for Open Ocean Observations and Logging (COOOL), Saint-Leu, La Réunion, France

[2]Ifremer, DOI Délégation Océan Indien, F-97420 Le Port, La Réunion, France

[3]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA, USA

[4] LIRMM-CNRS, university of Montpellier, rue Ada, 34000 Montpellier, France

[5]CSIRO Environment, Hobart, TAS, Australia

[6]INRAE, LPGP, 35000, Rennes, France

[7]Sigenae, GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France.

[8]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, California, USA

[9]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA

[10]Altos Labs, San Diego, California, USA

[11]Epigenetic Clock Development Foundation, Torrance, CA, USA

Emails: t.chevrier.cooolresearch@gmail.com

I. Introduction

Accurately determining the age of individuals is a cornerstone of ecological and biological research. Age information underpin a wide range of applications, from assessing population dynamics to evaluating ecosystem health (Ono *et al.*, 2015). Reliable age estimates are essential for inferring key life-history traits such as growth rates, age at sexual maturity, and age-specific fecundity (Schaffer, 1974; Western, 1979; Frisk *et al.*, 2001). Consequently, the ability to estimate age enables more robust monitoring of demographic parameters, including population structure and reproductive potential. Despite its central role, chronological age remains difficult to determine in wild animal populations, particularly in species lacking clear aging markers in hard structures or with limited access to long-term observational data (Cailliet *et al.*, 2001; Campana, 2001).

The importance of chronological age for wild populations has prompted the development of many estimation methods (Campana, 2001). Morphological features of internal structures have been commonly used to provide accurate age estimation (e.g. mammalian teeth (Goren *et al.*, 1987)). For fish, age is traditionally determined by counting growth zones in a range of hard structures including otoliths, vertebrae, scales, and fin rays (Pannella, 1971; Secor *et al.*, 1995; Campana, 2001). Such techniques have however some shortcomings : they can be costly and time consuming (Helser et al., 2019), of low accuracy for some species, require inter-calibration and cross-validation between labs, and are necessarily lethal in the case of otolith and vertebra readings (Campana, 2001; Anastasiadi & Piferrer, 2019). As the demand for fish age composition data is increasing (Helser *et* al., 2019), developing a non-lethal method for the accurate estimation of chronological age is an important first step for furthering our understanding of aging in wild populations.

Biological aging is a widespread process across animal species and is typically accompanied by molecular modification (Boyd-Kirkup et al., 2013; Booth & Brunet, 2016). However, evidence of senescence in fish remains limited. Unlike most vertebrates, several fish species show no-age related telomere shortening, raising questions about how aging manifests in this taxa (Simide et al., 2016; Sauer et al., 2021). Among the molecular processes associated with aging, DNA methylation at cytosine-phosphate-guanine (CpG) sites is known to change with age (Horvath, 2013; Lu *et al.*, 2023). Methylation profiles have been

used to develop biomarkers of age known as epigenetic clocks, which predict chronological age with remarkable accuracy (Mayne *et al.*, 2021a; Arneson *et al.*, 2022; Lu *et al.*, 2023). This genetic method is promising for inferring health status as an indicator of biological age. Epigenetic clocks were first built to monitor human aging (Horvath, 2013), but their underlying principles appear to be evolutionarily conserved, as they have now been successfully developed for many mammalian species (Polanowski et al., 2014; Bors et al., 2021). More recently, a mammalian methylation array was developed by Arneson et al. (2022), which is a single custom array that measures up to 36k CpGs per species that are well conserved across many mammalian species (Lu et al., 2023). It is not yet known whether the CpGs on the mammalian methylation array lend themselves for measuring cytosine methylation levels in fish.

Only a handful of epigenetic clocks have been developed for laboratory-raised fishes, including for European sea bass (*Dicentrarchus labrax*; (Anastasiadi & Piferrer, 2019) and zebrafish (*Danio rerio*; (Mayne et al., 2020)). Subsequently, Mayne et al. 2021 developed epigenetic clocks for three species of threatened fishes (Australian lungfish, *Neoceratodus forsteri*; Murray cod, *Maccullochella peelii*; and Mary River cod, *Maccullochella mariensis*), using a combination of wild and laboratory-raised individuals. Weber et al., 2022 developed a novel epigenetic age method estimation for two wild-caught reef fish from the Gulf of Mexico (Red snapper: *Lutjanus campechanus* and red grouper: *Epinephelus morio*). More recently, Weber et al have advanced the development of epigenetic clocks in two marine species, focusing on both deepwater teleost and ray (Weber et al., 2024a, 2024b). In the Weber et al., 2024a study on the blackbelly rosefish (*Helicolenus dactylopterus*) with 61 samples, two single-tissue epigenetic clocks developed from fin clip and muscle tissues showed high correlation ($R^2 > 0.98$; MAE < 1 year) between epigenetic and chronological age. However, a multi-tissue clock with both tissues showed low performance, particularly for muscle samples (R^2=0.77; MAE=5.43 years), indicating tissue-specific differences in age-associated DNA methylation patterns for the targeted region in this species. In contrast, two single tissue (fin clip and whole blood) and multi-tissue epigenetic clocks were developed for the cownose ray (*Rhinoptera bonasus*), and as for the blackbelly rosefish, they achieved high accuracy (R^2=0.97-0.99; MAE < 1 year; Weber *et al.*, 2024b). Notably, CpGs sites used by the multi-tissue clock did not overlap with those in both single-tissue models.

Here we focus on albacore tuna (*Thunnus alalunga*) which is a commercially important species in all oceans. Given the importance of age data for its stock assessments, the development of epigenetic clocks represents an opportunity to improve estimates of life history parameters which is especially true in the Indian Ocean where some parameters are taken from those established in other oceans. The development of this clock would also enable the upscale of sample sizes, allowing for more robust estimates of the age structure of the stocks.

II. Materials and Methods

A. Samples collection, chronological age estimation

Albacore muscle samples (n = 101) were collected from three sites in the Western Indian Ocean (Reunion, Seychelles and South Africa). Most samples were obtained from the "Germon project" (Nikolic et al., 2015) and were collected in 2013 and 2014. An additional set of four larvae was sampled north of Reunion Island during a scientific larval survey conducted by Ifremer in January 2022. Decimal chronological age was determined using the count of opaque zones in the otolith of the sampled fish (Farley et al., 2019). The distribution of age estimates from otolith readings for the 105 albacore tuna specimens is presented in Figure 1.

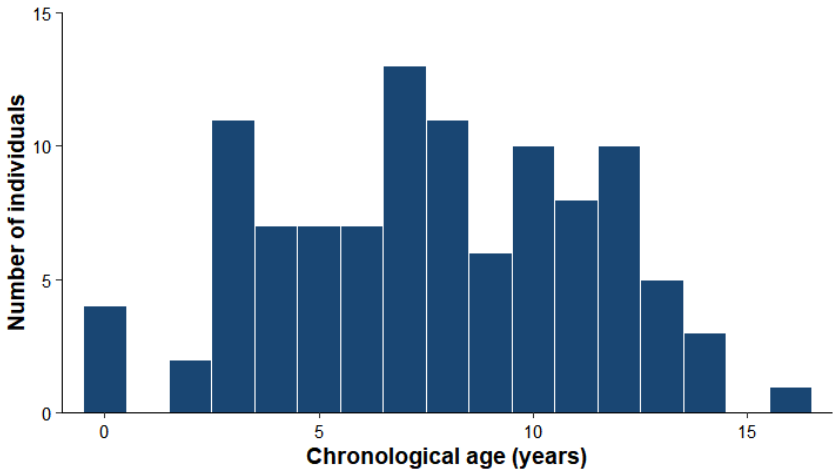

Figure 1: Total number of samples and age ranges used for Albacore clock

B. DNA extraction

Genomic DNA was extracted from muscle tissue for 101 individuals and from the eyes for four larvae (see Supplementary Information), using the DNeasy Blood & Tissue Kit (Qiagen), following the manufacturer's protocol. DNA concentrations were quantified with a Qubit™ 4 Fluorometer (Invitrogen™), and DNA purity was assessed using the 260/280 absorbance ratio measured with a NanoDrop Lite spectrophotometer (Thermo Scientific). Samples with quantity higher than 500 ng and a ratio between 1.8 and 2.0 were considered to contain high-quality DNA and usable for subsequent sequencing.

C. Larvae species identification

To determine the larvae species identity, a PCR amplification followed by 2% agarose gel electrophoresis was performed. These molecular analyses were essential because early developmental stages of tuna species exhibit highly similar morphological features, rendering visual identification unreliable. For amplification, species-specific primers developed by (Lee et al., 2022) were used with the following primer pair: forward primer -GTTTCGTGATCCTGCTAGTG- and reverse primer -CCTCCTAGTTTGTTGGAATAGAT-. The PCR was performed under the following thermal cycling conditions: an initial denaturation at 94 °C for 2 min, followed by 35 cycles of – denaturation at 94 °C for 30 s, annealing at 55 °C for 30 s, and extension at 72 °C for 30 s –, with a final extension step at 72 °C for 3 min, and a final hold at 4 °C. This genetic approach enabled accurate species identification, overcoming the limitations of morphological discrimination in early larval stages.

D. Library preparation and sequencing

For library preparation, 500 ng of genomic DNA was fragmented and subjected to end-repair, dA-tailing, and adapter ligation using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) with custom pre-methylated adapters (custom adapter plate, IDT). Pools of 16 purified libraries were hybridized to the biotinylated probe panel following the manufacturer's instructions (xGen Hybridization Capture Kit, IDT).

Captured DNA was treated with bisulfite using the EZ DNA Methylation-Gold Kit (Zymo Research), followed by PCR amplification with KAPA HiFi Uracil + ReadyMix (Roche) under the following conditions: initial denaturation at 98°C for 2 minutes; 14 cycles of 98°C for 20 seconds, 60°C for 30 seconds, and 72°C for 30 seconds; final extension at 72°C for 5 minutes; hold at 4°C.

Library quality was assessed using the High Sensitivity D1000 Assay on a 4200 Agilent TapeStation. Libraries were pooled (96 per pool) and sequenced as 150 bp paired-end reads on an Illumina NovaSeq 6000 (S1 flow cell).

E. Sequencing data processing

Raw bisulfite sequencing data were processed using *BSBolt* to generate high-quality methylation calls. Paired-end fastq files were first quality-checked, trimmed to remove adapters and low-quality bases using *fastp*. Prior to alignment, the reference genome was bisulfite-converted and indexed using *BSBolt*'s genome indexing module to enable accurate mapping of converted reads. Reads were then aligned to the bisulfite-indexed reference genome, which accounts for the C-to-T conversions characteristic of bisulfite treatment. After alignment, methylation calls were extracted using *BSBolt*'s methylation caller module, providing per-CpG site methylation ratios (methylated reads over total coverage). Sites with low coverage (<5×) were filtered out to ensure data reliability. The resulting data were merged across samples to generate a single methylation matrix, with rows corresponding to individual samples and columns representing common CpG sites. This matrix served as the input for downstream statistical analyses, including epigenetic clock construction and age prediction models.

F. Statistical modelling for age prediction from methylation profiles

To develop an epigenetic clock, we first imported the methylation matrix generated with *BSBolt* and harmonized sample identifiers to match age metadata. After quality control and matching, samples were split into a training set (85%) and a testing set (15%) using stratified partitioning to maintain age distribution. We used a LASSO regression approach implemented via the *glmnet* package in R, with 10-fold cross-validation on the training set to determine the optimal penalty parameter (lambda) minimizing prediction error. The final

model was then fitted to the training data and used to identify age-informative CpG sites (non-zero coefficients). Predicted ages were generated for both training and testing sets, and model performance was evaluated by calculating Pearson correlation coefficients and median absolute errors (MAE) between predicted and chronological ages. To further evaluate model robustness and reduce overfitting bias, a leave-one-out cross-validation (LOOCV) was performed on the filtered set of age-informative CpG sites, allowing age prediction for each individual while being left out from model training.

III. Results

A. Whole genome sequencing

Our genome assembly of the albacore tuna (*Thunnus alalunga*) represents the first chromosomal-scale reference available for this species. The initial assembly, generated using Hifiasm from PacBio HiFi long reads, resulted in 537 scaffolds, with a scaffold N50 of 3.6 Mb, reflecting a high level of sequence continuity. In a second step, we used Hi-C contact maps to scaffold and organize the assembly, resulting in a complete scaffolded genome comprising 38 scaffolds, with a markedly improved scaffold N50 of 34.1 Mb. Of these, 24 scaffolds correspond to chromosome-scale scaffolds, consistent with the expected haploid chromosome number (n = 24) previously reported in tunas (REF), into which Hi-C integration enabled anchoring of 99.88% of the 785.3 Mb genome sequence. Assembly completeness, assessed using BUSCO v4 (Actinopterygii lineage, protein mode), confirms the high quality of the genome, with 93.6% of genes detected as complete and single-copy genes.

B. Epigenetic clock

Across all individuals, we have recovered 95,191 unique CpGs sites with an average of 4,472,876 reads per individual, for an average sequencing depth of 585 reads/site. The final methylation matrix contains a total of 7,637 CpG sites shared by all samples, all with base and alignment quality scores > 10 and with a minimum site coverage of five reads. For each CpG in our methylation matrix, we computed a Pearson correlation between methylation values and chronological age of the fish samples.

For Albacore, using LASSO regression, 48 CpGs sites were used to calibrate the model, we found a high correlation between the chronological age and the predicted age in both the

training (Pearson's r = 0.973) and the testing dataset (Pearson's r = 0.946) (Figure 2). The median absolute error (MAE) in the training data set was 0.721 years and 1 years for the testing one.

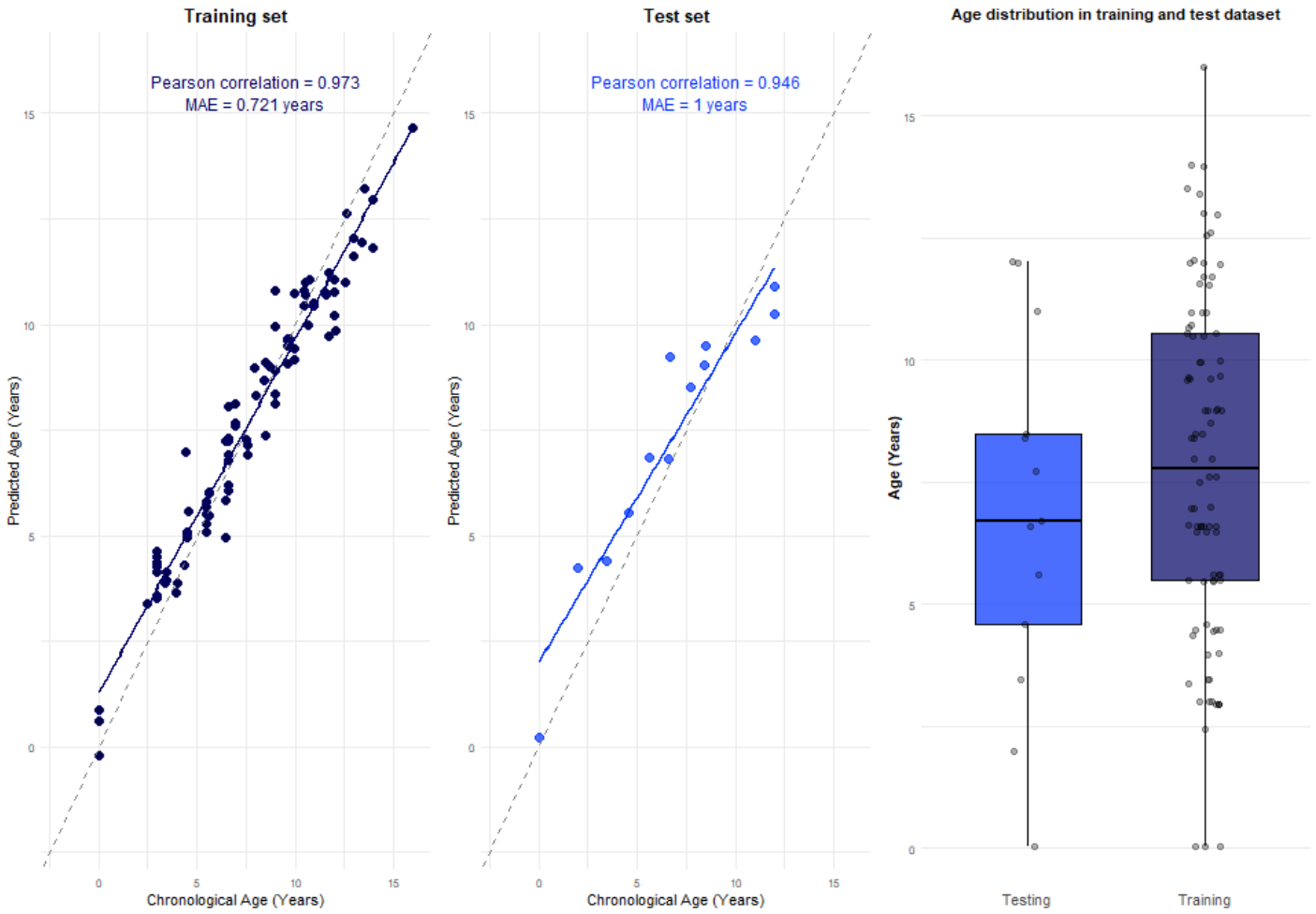

Figure 2: The Albacore clock based on 48 CpGs sites calibrated with chronological age. Correlation between otolith age and predicted age in (A) the training set and (B) in the testing set. (C) Samples repartition between training and testing data set.

The Leave-one-out cross-validation reveals a strong correlation between these 48 CpGs sites and chronological age (Pearson correlation = 0.955 and MAE = 0.795 years) (Figure 3).

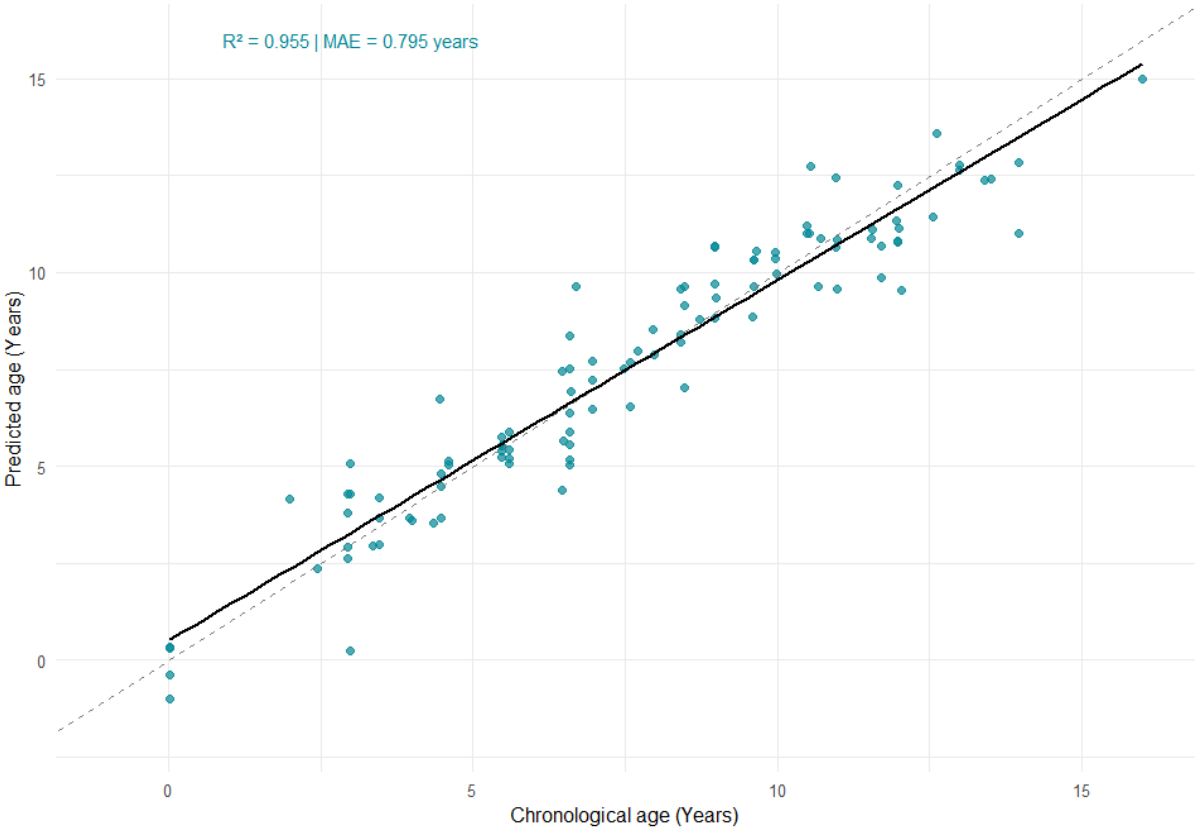

Figure 3: Leave-one-out cross-validation for Albacore clock based on 48 CpGs sites calibrated with chronological age

IV. Discussion

Otolith age readings do not fully cover the complete age range of the species. In particular, individuals under 3 years of age and those older than 13 years are underrepresented (Fig. 1), likely due to natural sampling challenges in the Indian Ocean. This limited representation may slightly reduce predictive accuracy for the oldest age classes, especially beyond 12 years. As the maximum lifespan of *T. alalunga* is estimated to be around 20 years (grey literature), expanding the dataset to better include juvenile and older individuals would enhance an already robust clock, improving its calibration and reliability across the species' full lifespan. Moreover, our models are based respectively on 105 samples. Increasing the sample size—particularly by targeting age classes currently underrepresented—would not only help capture the full range of age-associated methylation dynamics, but also improve the generalizability of the models. Ideally, as explained by Mayne et al., 2021b at least 70 and optimally 134 samples from a single tissue type are

recommended to develop a reliable epigenetic clock for a given species, in order to minimize prediction error rates.