Evolution of Conserved Non-coding Elements in Teleost Fish

November 2021

by

Elisavet Iliopoulou

A thesis submitted to School of Medicine, University of Crete

for the degree of M.Sc. in Bioinformatics

Examination Committee:

Associate Professor, Dr. Ioannis Iliopoulos[2]

Researcher B', Dr. Pavlos Pavlidis[3]

Researcher C', Dr. Tereza Manousaki[1]

1 Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Greece

2 School of Medicine, University of Crete

3 Foundation for Research and Technology - Hellas, FORTH

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του Διιδρυματικού Προγράμματος Μεταπτυχιακών Σπουδών στη Βιοπληροφορική της Ιατρικής Σχολής του Πανεπιστημίου Κρήτης και του Ιδρύματος Τεχνολογίας & Έρευνας (ΙΤΕ).

Acknowledgments

First and foremost I am extremely grateful to my supervisor, Tereza Manousaki, for her invaluable support and guidance during my thesis. Although the situation over the last year has been unprecedented for everyone, due to Covid 19 and quarantine, she was there making sure that "Genome nerds" will remain united and motivated.

I also very much appreciate my esteemed second supervisor, Costas Tsiggenopoulos, whose experience and companionable attitude was valuable.

And of course, I'm extremely grateful to meet all members of our team. Our postdoc Vasilis Papadogiannis, who I am really lucky to have met and helped me a lot with my thesis and kept me constantly motivated. Who always finds a way to get it done, and done well! Klara Eleftheriadi, whose very friendly and clubby mood brightened up our lab atmosphere. Thodoris Danis for the tech-knowledge and advice when starting my thesis. Nelina Angelova for her company, discussions and encouragement. Of course Christos Kitsoulis and Andromachi Papadopoulou, the new members of our lab, for their patience hearing me moaning through the last month of my thesis. And Harris Zafeiropoulos and Savvas Paragkamian, for the great support and friendly atmosphere when we were sharing the lab.

I would also like to thank our High Performance Computing cluster 'zorbas' admins, Antonis Potirakis and Stelios Ninidakis, for their patience, support and troubleshooting, but also the people who created it Evaggelos Pafilis and Dimitrios Sidirokastritis.

A special thank you, to my parents and my sister whose support and understanding is always important and valuable.

CONTENTS

Περίληψη

Πολλά γονίδια που σχετίζονται με εξελικτικά συντηρημένες λειτουργίες βρίσκονται κοντά σε εξαιρετικά συντηρημένα μη κωδικα στοιχεία (CNE), τα οποία συχνά δρουν ως ενισχυτές των γειτονικών τους γονιδίων. Ένας μικρός αριθμός μελετών έχει υποδείξει την απώλεια και την ταχεία εξελικτική απόκλιση των CNE στους τελεόστεους. Εδώ, επικεντρωθήκαμε στον εντοπισμό και τη μελέτη του συνόλου των CNE των τελεόστεων, εκμεταλλευόμενοι την πρόσφατη αύξηση των γονιδιωμάτων υψηλής ποιότητας. Σχεδιάσαμε μια διαδικασία συγκριτικής γονιδιωματικής για την ευθυγράμμιση των γονιδιωμάτων και την ανίχνευση των CNE. Επιπλέον, κατασκευάσαμε φυλογενωμικά δέντρα με βάση τα CNE, για να ελέγξουμε τις δυνατότητές τους στην επίλυση των φυλογενετικών τοποθετήσεων των τελεόστεων και να μελετήσουμε την απώλεια, την αύξηση και την εξελικτική απόκλιση των CNE μεταξύ των τελεόστεων. Τέλος, προσπαθήσαμε να αποκτήσουμε πληροφορίες σχετικά με τη δυνατότητα των εντοπισμένων CNE ως αναπτυξιακών ενισχυτών, συσχετίζοντάς τα με κοντινά αναπτυξιακά γονίδια. Η μελέτη αυτή παρείχε πληροφορίες σχετικά με την εξέλιξη των CNE στους τελεόστεους και ένα σύνολο πιθανά συντηρημένων αναπτυξιακών ρυθμιστικών στοιχείων που μπορούν να χρησιμοποιηθούν σε περαιτέρω πειραματικές δοκιμές.

Λέξεις-κλειδιά: *συντήρηση, συντηρημένα μη κωδικά στοιχεία, ολόκληρος διπλασιασμός του γονιδιώματος, ενισχυτές, τελεόστεοι, σπονδυλωτά, φυλογενωμική ανάλυση*

Abstract

Many genes associated with evolutionarily conserved functions are proximal to highly Conserved Non-coding Elements (CNE), which often act as enhancers of their neighboring genes. A small number of studies have suggested loss and rapid evolutionary divergence of CNE in teleost fish. Here, we focused on identifying and studying the teleost CNE repertoire, taking advantage of the recent increase in high-quality genomes. We developed a comparative genomics pipeline to align genomes and detect CNEs. Moreover, we built CNE-based phylogenomic trees, to test their potential in resolving teleost phylogenetic placements and study CNE loss, gain and evolutionary divergence across teleosts. Finally, we attempted to gain information on the potential of identified CNE as developmental enhancers, by associating them with proximal developmental genes. This study provided insight on CNE evolution in teleost fish and a set of putative conserved developmental cis-regulatory elements that show promise for further experimental testing.

Keywords: *Conserved non-coding elements, WGD, enhancers, teleosts, vertebrates, conservation, pairwise whole genome alignment, phylogenomic analysis*

1. Introduction

Conserved non-coding regions of the genome have been found to regulate genes associated with evolutionarily conserved functions, such as developmental processes (Woolfe et al., 2004). During embryonic development, gene expression must be controlled with precision both spatially and temporally. For this purpose, there is a combinatorial interaction of Transcription Factors (TFs) with cis-regulatory elements, which are mainly located in non protein-coding genomic sequence (Davidson et al., 2006). Conserved Non-coding Elements (CNE), often act as enhancers of their neighboring target genes, playing an important role in the spatiotemporal regulation. In vertebrates, gene expression is regulated by many cis-regulatory elements, such as promoters, with some of them located near the transcription start sites (TSS) of genes. Promoters direct the process of transcription, but sometimes have low basal activity (Haberle V, Stark A , 2018). Some important cis-regulatory modules are located in regions of the genome far from the TSS , including enhancers and silencers (Verheul TC, et al., 2020). Therefore, enhancers and their associated TFs have a leading role in the regulation of gene expression (Spitz F, Furlong EE, 2012). Experiments have shown that only 25% of the enhancer–promoter interacting fragments were within 50 kB and about 57% of the contacts spanned more than 100 kB (Yao, L. et al, 2013).

Conserved sequences within the non-coding parts of genomes were initially identified almost five decades ago (Comings, 1972, Ohno S., 1972). Such elements had been shown to maintain >70% sequence identity for over 400 millions of years of evolution that, in many cases, exhibit the percentage of conservation in protein-coding genes (Polychronopoulos et al., 2017). In the past few years, many studies have searched for CNEs, each using different criteria and species.

It has long been known that vertebrates have undergone two rounds of Whole Genome Duplication (WGD), and it has been suggested that these events have contributed to their diversification (Ohno S., 1970). Within vertebrates, the group of teleosts form the most species-rich group of vertebrates, with unprecedented diversity. It is hypothesized that one of the main drivers of this adaptation is the teleost specific (3R) WGD (Glasauer and Neuhauss, 2014). We investigated the fate of CNEs and focused on identifying and studying the teleost CNE repertoire. In teleost fish many CNE have been

suggested to have been lost or rapidly diverged following the 3R WGD event (Glasauer and Neuhauss, 2014), but this is based on few model species.

We developed a comparative genomics pipeline to align genomes and detect CNEs across the teleost phylogeny. With that dataset in hand we opted to answer three important questions. First, whether CNEs are a better proxy for resolving phylogenetic relationships compared to protein-coding sequences, second which are the ancestral teleost CNEs, and finally how many of the known vertebrate CNEs were lost/kept in the teleost lineage.

To answer those questions, we first identified CNEs that are shared among all studied teleosts. We then used this set to build the first thorough CNE-based teleost phylogeny and assess their potential in resolving phylogenetic placements. Further, we focused on identifying the set of CNE shared by all main teleost clades, attempting to recover the ancestral teleost CNE repertoire. We also hypothesised that such highly conserved ancestral elements are likely to be functional. Finally, we analysed our data to gain information on the potential action of identified CNEs as developmental cis-regulatory elements in the zebrafish and human, by associating them with proximal genes. This work represents the first CNE analysis on a large number of teleosts genomes, and as such it provides a unique dataset for understanding the evolution of the functional and conserved noncoding genome.

2. Materials and Methods

All the analyses were carried out on the ‘zorba’ IMBBC HPC cluster, HCMR, Heraklion, Greece (Zafeiropoulos et al. 2021).

To identify teleost CNEs multiple steps had to be implemented (Figure 1), as we describe below in section 2.1. Briefly, we first downloaded 31 high quality genomes from the Ensembl database spanning the entire teleost phylogeny. The selection of genomes was based on two important scores N50 > 8Mb and Busco > 95%. The second step was to conduct pairwise whole genome alignments in two chosen pairs of teleosts that represent the core dataset of the analysis. Then, the pairwise alignments had to be chained/netted, in order to proceed with the main step of CNE identification and the search of the identified CNEs in all other teleost species.

2.1 CNE Identification pipeline

2.1.1 Pairwise whole-genome alignment

We performed pairwise whole-genome alignments using LASTZ v1.04.03 (Harris R.S.,2007) with four genomes downloaded from the Ensembl database (Howe et al., 2021). All genomes were downloaded hard masked, i.e. interspersed repeats and low complexity regions were detected with the RepeatMasker tool (Chen, N. ,2004) and masked by replacing repeats with 'N's. *Danio rerio* (DanRer11)*, Astyanax mexicanus* (Astyanax_mexicanus-2.0)*, Takifugu rubripes* (fTakRub1.2) and *Sparus aurata* (fSpaAur1.2) were used as the representative teleosts and aligned in pairs as they span a large evolutionary distance amongst teleosts. Based on the evolutionary relations among species (Figure 6), *D.rerio (*Zebrafish) was paired with *A.mexicanus* (Mexican tetra) and *T.rubripes* (Fugu) with *S.aurata* (Gilthead seabream), carrying out reciprocal alignments within pairs. Any Zebrafish or Fugu sequences larger than 10,100,000 bases were split into chunks of 10,100,000 bases overlapping by 100,000 bases for alignment. A similar process was followed for Mexican tetra and Gilthead seabream, but using non-overlapping chunks of 10,000,000 bp. This process was implemented using utilities found in UCSC, because aligners fail to align very large sequences.

Two rounds of pairwise whole-genome alignments were run for each pair as previously suggested (Hiller et al.,2013). In the first round, we used sensitive parameters, originally recommended for distantly related species (> 100 Mya). Parameters used were the following: M=50 E=30 H=2000 K=2200 L=6000 O=400 T=1 Y=3400 Q=HoxD55.q (Tan et al., 2019). To achieve high sensitivity and identify all conserved elements, even with a length less than 100 bases we carried a second round of alignment, after masking aligned regions from the first alignment round. Again, any Zebrafish or Fugu sequences larger than 10,100,000 bases were split into chunks with the same procedure as before. A similar process was followed for Mexican tetra and Gilthead seabream.(Kent et al., 2002). During this round, parameters for medium sensitivity were used, more specifically K=1500, L=2300, M=0 and W=5. (Hiller et al., 2013). Alignment results from LASTZ were parsed with the *alignment_processing.py* custom python script to produce the final alignment input for downstream processing (see CODE AVAILABILITY).

2.1.2 Chaining - Netting

Aligned regions found by both rounds of pairwise whole-genome alignment steps were chained using CNEr v1.8.3 Bioconductor wrapper functions (Tan et al., 2019) inside R environment (using utilities found in UCSC Browser) (Kent, 2002). During chaining, if two matching alignments are close enough, they are joined into one bigger fragment. Next, we kept only the longest fragments and formed netted alignments in net Axt format.

2.1.3 CNE Identification in focal species

In order to carry out CNE detection, alignment chain nets produced as described in sections 2.1.1 & 2.1.2 were used (Figure 1). Zebrafish and Fugu were selected as reference genomes for each pair of alignments detailed in the section "*Pairwise whole-genome alignment"* as they are relatively distantly related within teleosts (206 - 252 MYA) (Kumar et al., 2017), and suitable teleost species to study conservation across the teleost lineage. Therefore, from this comparison we obtained an ancient CNE dataset shared across the teleost phylogenomic tree. We filtered the chained alignments that were previously detailed using annotation information for exons, to keep only the non-exonic regions of each of the four genomes. All annotation files were downloaded from the Ensembl database,

*Takifugu_rubripes.fTakRub1.2.104 (*Takifugu rubripes), *GCF_900880675.1_fSpaAur1.1_genomic (*Sparus aurata), *Astyanax_mexicanus-2.0.104* (Astyanax mexicanus) and *Danio_rerio.GRCz11.104 (*Danio rerio). CNEr v1.28.0 Bioconductor (Tan et al., 2019) package was used for the filtering procedure and for CNE detection, inside the R environment. Conserved regions were filtered based on three thresholds for each pair of species. First, the maximum number of hits per element (cutoffs = 4), i.e. how many times we expect to see an element. Second, the minimum identity of aligned regions (Identities = 70pc), i.e. the minimum percentage of matches in a single alignment. And finally, the sliding window size of the detection (windows = 100).

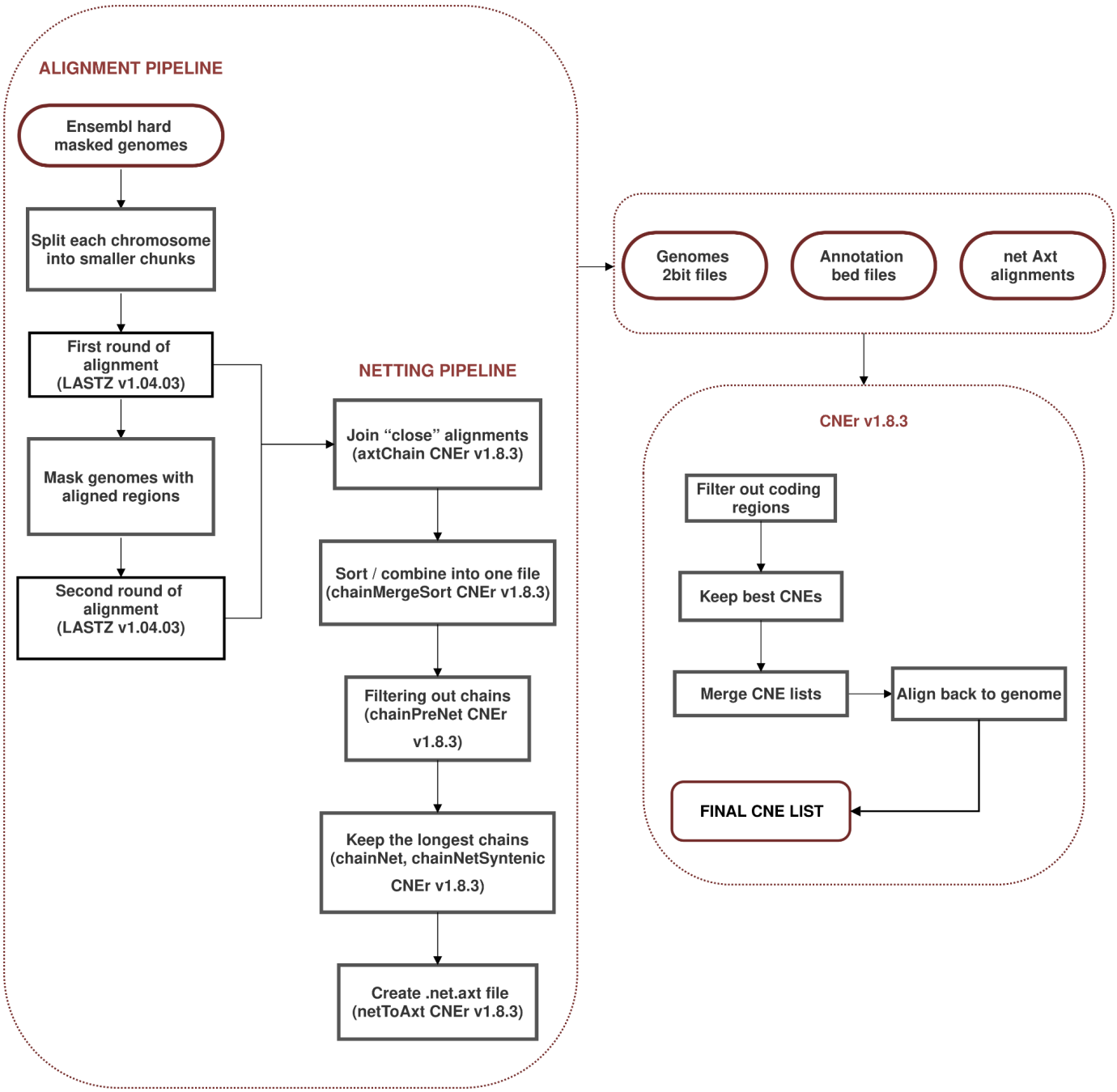

Figure 1. ***Alignment pipeline****. Two rounds of pairwise whole-genome alignments using masked genomes.* ***Netting pipeline****. Join close alignments and form chains, then select the longest chains.* ***CNEr v1.8.3.*** *CNE Identification pipeline using genomes in 2bit format, annotation bed files and the chained alignments as input. The output of this pipeline is a list with all identified CNEs.*

2.1.4 Positive control and comparison with published data

To confirm our method is able to recover at least highly conserved ancestral CNE shared among teleosts and other vertebrates, we compared the Zebrafish-*A. mexicanus* chained alignments against publicly available Zebrafish-Human pairwise genome alignment chains from the UCSC Genome Browser Downloads (Kent, 2002). As Zebrafish and *A. mexicanus* are relatively closely related, giving a high rate of genome alignment, we expected that we would be able to recover a high fraction of such highly conserved ancestral areas. Hence we used this comparison to optimise alignment parameters until a very high level of Zebrafish-Human conserved blocks were recovered through our method (at least 95%). A higher threshold was not sought, as remaining discrepancies could be explained due to some level of divergence of Zebrafish and *A. mexicanus*. To do so, we kept the coordinates for both alignments and also filtered the Zebrafish-Human chains for repeats inside UCSC Table Browser (Karolchik et al., 2004). We then used BEDtools (Quinlan and Hall, 2010) to check the coverage of the two alignments and test the completeness of our dataset.

In addition, we used a previously published Zebrafish CNE dataset (Hiller et al., 2013) using an older genome version (danRer7). This set was derived from the comparison of Zebrafish with other 15 vertebrate genomes, including four teleosts. Finally, we used one more CNE dataset for the Zebrafish genome, included in the ANCORA database (Engström et al., 2008), which was also produced with older genome versions (danRer10).

2.2 CNE search in other teleost species

Using the resulting datasets from our CNE identification, those of Zebrafish (*zCNEs*) and Fugu (*fCNEs*), we made similarity searches using BLAST v2.10.0+ (Altschul et al., 1990) against 30 teleost fish genomes, which were pre-masked for repeats. We used a maximum e-value for the alignments (evalue=1e-6), a word size for wordfinder algorithm (length of best perfect match) (word size=6), the maximum number of aligned sequences to keep (max target seqs=1), and the maximum number of hits per subject sequence to save for each query (max hsps=1). This comparative dataset (Table 2) was used as the basis for studying CNE conservation, gain and loss, as well as identifying

the teleost specific CNE core dataset shared among all 31 teleost fish, which we used for the Phylogenomic analysis as described in section 2.3.

2.3 CNE-based phylogenomic analysis

To assess the potential of CNE for phylogenomic studies and form the basis for the rest of our study on teleost evolution, we built a CNE-based phylogenomic tree with the richest (in terms of total elements) and representative (deriving from a well studied genome) set, *zCNEs*. We also included an outgroup organism for better comparison, which in our case is *Lepisosteus oculatus* (Spotted gar), belonging to the holostei infraclass of ray-finned bony fish, which diverged before the teleost WGD (3R) occurred (Braasch et al., 2016). For this study, we kept only CNE found in a single copy in all *teleostei* and *holostei* species used, to avoid ambiguities related to paralog resolution. Thus, we kept only the shared and unique CNEs for each species and made local alignments against the Spotted gar genome using the shared teleost CNEs. We then extended all sequences identified via BLAST, to match the length of the respective query CNE (zCNE), as well as a further 50bp in each side, to try and capture potential alignable sequence that may be shared by some of the species used, but not in with the initial query. When any two elements overlapped with each other, we adjusted the coordinates accordingly to split the overlap. This produced a collection of shared CNE from each species. All CNE dataset processing was carried out with the *CNEs_processing.py* custom python script.

Next, we aligned each CNE from all 32 species. We used MAFFT v7.407 (Katoh et al.,2002) for multiple sequence alignment based on fast Fourier transform (FFT), which allows rapid detection of homologous segments. By default, MAFFT will determine the best algorithm based on the number of sequences. The progressive method FFT-NS-2 was automatically selected. Individual CNE alignments were then collated to produce a single supermatrix of all CNE from across all 32 species (including Spotted gar).

After running an alignment, it is necessary to trim it to remove sites with erroneous alignment and gappy regions. To do that, we used trimAl v1.4.rev15 (Capella-Gutiérrez et al.,2009) with a gap threshold of 50 percent of the species for each alignment column (gt 0.5).

Finally, IQ-TREE multicore v1.6.12 (Minh et al.,2020) was used to build a CNE-based tree construction with extended model selection followed by tree inference and ultrafast bootstrap with 1000 replicates (m=MFP, bb=1000). The best-fit model for alignment was chosen (TVM+F+I+G4) using the Bayesian information criterion (BIC). For visualisation purposes FigTree v1.4.4 (Rambaut A., 2020) was used.

For evaluation purposes, a gene-based species tree was also included in our study provided by Papadogiannis et al (personal communication; unpublished data). The procedure was similar, but this time starting with single copy orthologous proteins for all 32 species, obtained by running OrthoFinder2 v2.5.4 (Emms and Kelly, 2018) on the proteomes of these species.

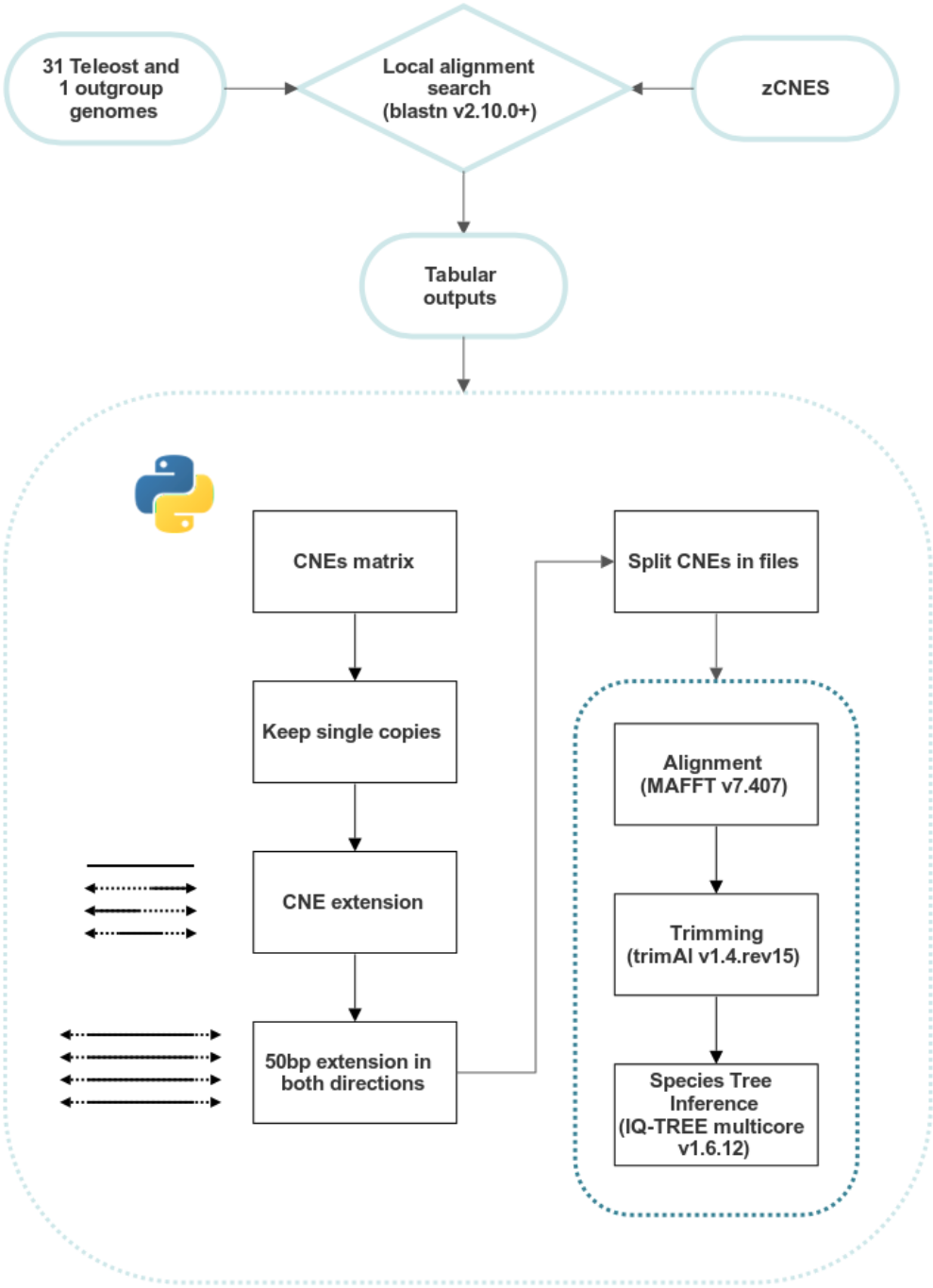

Figure 2. *CNE processing pipeline and CNE-based tree construction using single copy CNE orthologs*.

2.4 Identifying the vertebrate CNE set

After constructing the teleost phylogeny and visualizing the teleost clades evolution we opted to gain information about the evolution of CNEs through the transition to the 3R. To do so, we conducted an independent search of known vertebrate CNEs against the teleost genomes.

For this purpose *Homo sapiens* (Human), *Callorhinchus milii* (Elephant shark) and *Lepisosteus oculatus* (Spotted gar) (Figure 3) CNE datasets (with Human as the reference) were downloaded from the ANCORA database and were checked for any remaining coding regions with the *hg38.knownGene* annotation file (Gene predictions based on data from RefSeq, Genbank, CCDS and UniProt, from the UCSC KnownGene track). We selected sequences with >70% identity over a window of 50 bases, using a bash script. The union of these sets represented the core CNE set of the vertebrate subphylum.

To see which of these CNEs are present and conserved in the teleost genomes, we conducted a similarity search of this set using BLAST (evalue=1e-6, word size=6, max target seqs=1, max hsps=1) against all 31 teleost fish (including Zebrafish) genomes to detect gains and losses of CNEs.

2.5 Identifying Ancestral teleost CNEs

To find the *ancestral teleost CNE set,* we kept those elements that were present in the two initial (largest) teleost clades of both trees (Figures 5-6). Those are, *Clade 1* from *Siluriformes (P.hypophthalmus)* to *Cypriniformes (D.rerio),* and *Clade 2* from *Cyprinodontiformes (K.marmoratus)* to *Tetraodontiformes (M.mola)*. To do so, using the two CNE datasets, zCNEs and fCNEs, we scanned through the phylogenomic tree and kept only those elements that were present in at least one of the species of the opposite clade. Clade 1 was represented by zCNEs and Clade 2 by fCNEs. By concatenating these two sets, we got the *Ancestral teleost CNE set.*

We then conducted further similarity searches of this dataset with BLAST (evalue=1e-6, word size=6, max target seqs=1, max hsps=1) against the Human, Elephant shark and Spotted gar set derived from the ANCORA database. Doing so, we found the CNEs shared among teleost fishes and

the rest of vertebrates (*shared Ancestral teleost CNEs*), and further analysed our output to gain information about any elements that had no hit and thus represented those possibly gained in teleost lineage.

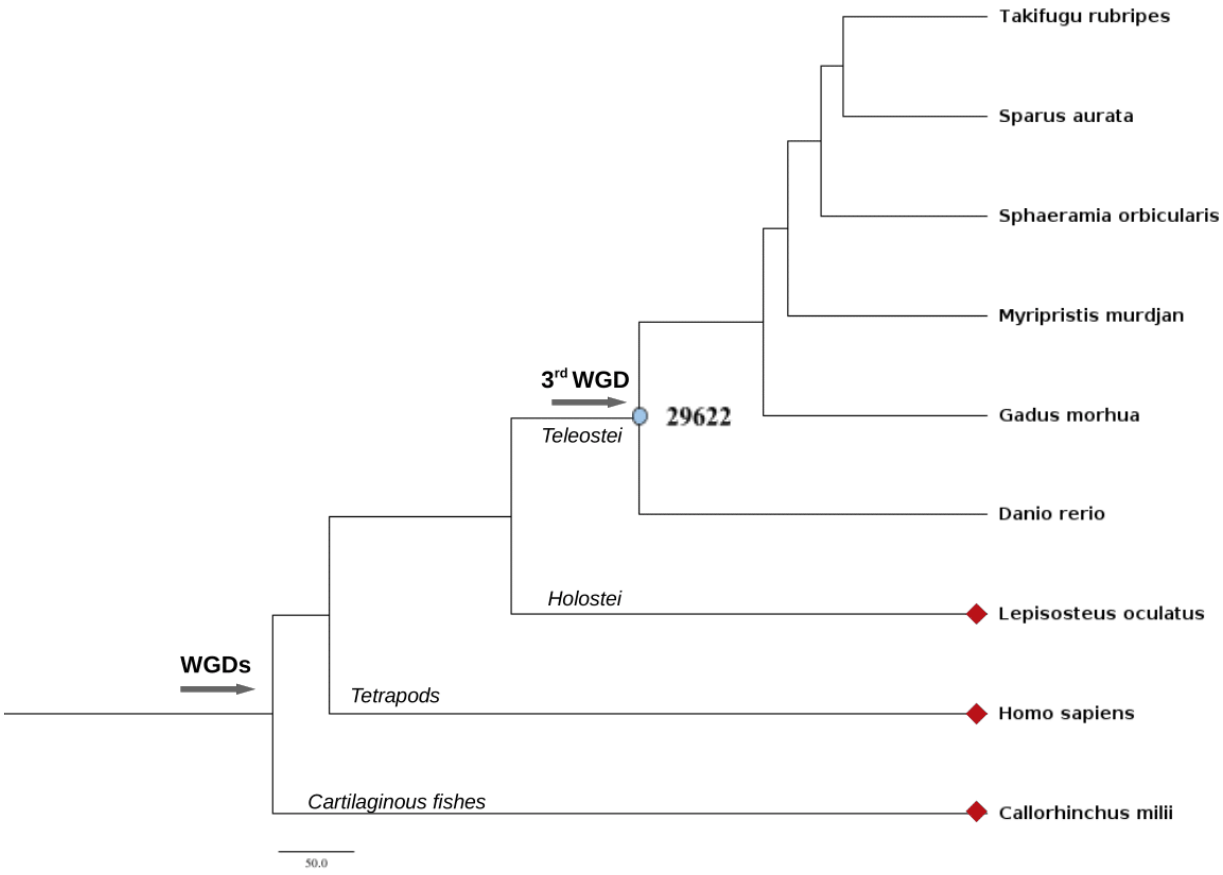

Figure 3. *Tree showing phylogenetic relationships among species used in the present study. Red diamond shape marks three vertebrate outgroup species (Homo sapiens (Human), Callorhinchus milii (Elephant shark) and Lepisosteus oculatus (Spotted gar)) used for the analysis. "**WGDs**" indicates the two previous Whole Genome Duplication events that are shared between all vertebrates.*

The selection of the reference teleost species (Zebrafish and Fugu) for the alignments, along with their pairs, is based on the evolutionary distance of these species.

2.6 Gene association

To be sure that many of those highly conserved sequences are also functional, we searched for genes nearby. We used ChIPseeker v1.31.0 Bioconductor package (Yu et al., 2015) to annotate our CNEs and associate them with proximal genes, and also to visualize CNE coverage across chromosomes.

We searched for gene associations, for both datasets, those that were gained or those lost in the transition to beginning of teleost evolution. To find any possibly lost CNE we kept unmatched CNEs resulting from the BLAST outputs of vertebrate comparisons with all 31 teleosts. In parallel, to find CNEs gained in teleosts, we kept any unmatched elements derived from the BLAST outputs of ancestral teleost CNEs against Human, Elephant shark, and Spotted gar (see section 2.5).

Within ChIPseeker, we provided the CNE datasets in the function *readPeakFile (https://laderast.github.io/surrogateMutation/reference/readPeakFile.html)*. After loading the coordinates, we assessed how CNEs were distributed across the genome. The *covplot* function *(https://github.com/YuLab-SMU/ChIPseeker/blob/master/R/covplot.R)* was used to calculate the coverage of loaded regions over chromosomes and generate a plot for visualization (Figure 7). Next, *annotatePeak (https://github.com/YuLab-SMU/ChIPseeker/blob/master/R/annotatePeak.R)* was used for CNE annotation with their neighboring genes. We extended the neighboring CNE searching window from -100kb to +100kb, as enhancers may be located far away from the target (Yao, L. et al., 2015). All functions were provided by Bioconductor (Huber et al., 2015)(Gentleman et al., 2004) to process the results of this analysis. We also used *TxDb.Hsapiens.UCSC.hg38.knownGene* for Human and *TxDb.Drerio.UCSC.danRer11.refGene* (Team BC, Maintainer BP, 2019) for Zebrafish after converting our data to UCSC format (from Ensembl format) using *cvbio UpdateContigNames* tool *(*https://github.com/dpryan79/ChromosomeMappings*)*. We used annotation information for 5' UTR, 3' UTR, Intron, Downstream and Intergenic. The distance to the nearest gene was calculated.

2.6.1 Gene association comparisons of Teleost CNE gains and losses

To assess if CNE losses and gains during the evolution of teleosts were associated with the same or different sets of genes, we ran an orthology analysis. Taking the genes associated with Human CNEs that have been lost in teleosts, and also those associated with Zebrafish CNEs, we compared them with ortholog information from the Ensembl BioMart data mining tool (Smedley et al., 2009). Thus, we used the *gene_association_orthology.py* custom python script to compare the above datasets (see sections 2.4, 2.5 for data acquisition details) using gene association information (see section 2.6 for details), as well as information for Zebrafish - Human orthologous genes from Ensembl BioMart, to test if CNEs from both lost and gained lists are associated with orthologous loci.

3. Results

3.1 CNE Identification pipeline

First, we carried out extensive Whole-Genome alignment-based searches, based on two pairs of focal reference species, as described in section 2.1. The alignment section of our extensive pipeline, detailed in Figure 1 resulted in a set of 298743 chains between Zebrafish and *A.mexicanus*, and 372053 between Fugu and *S.aurata*.

The pipeline yielded 63023 Zebrafish CNEs (zCNE) and 39532 Fugu CNEs (fCNE). These two sets span the entire teleost conservation information.

3.1.1 Positive controls

To make sure that CNE identification was successful, we compared our chains and CNE sets with published information as positive control. As mentioned above, 3 sets were used for this step. Zebrafish - Human chain sets downloaded from the UCSC genome browser, an ANCORA zCNE dataset, and a published zCNE set from another study in vertebrates (Hiller et al., 2013).

By comparing Zebrafish- *A.mexicanus* with Zebrafish-Human chains, we retrieved 97,1% which is a way higher percentage of the already strict one that we defined in section 2.1.4.

Previous CNE datasets for the Zebrafish genome, also included in the ANCORA database, were produced with older genome versions (*danRer10*) and include some small CNE fragments (maximum 50 bases window), while a previous study (Hiller et al., 2013) identified 54533 elements conserved between Zebrafish and 15 other vertebrates including 4 teleost fish. Thus, our extensive search here provides an updated set including a larger number of elements in Zebrafish.

3.2 Phylogenomic analyses

To meet our goal of testing whether CNEs are a good proxy for resolving phylogenetic relationships compared to protein-coding sequences, we carried out a CNE-based phylogenomic analysis as

detailed in section 2.3. To obtain an informative CNE alignment for this analysis, we used trimAl to remove non-homologous sites and gappy regions by using a threshold to filter regions with more than 50% gaps, aiming to retain maximal phylogenetic signal at the same time. As shown in Figure 4, conservation of filtered positions still remained high.

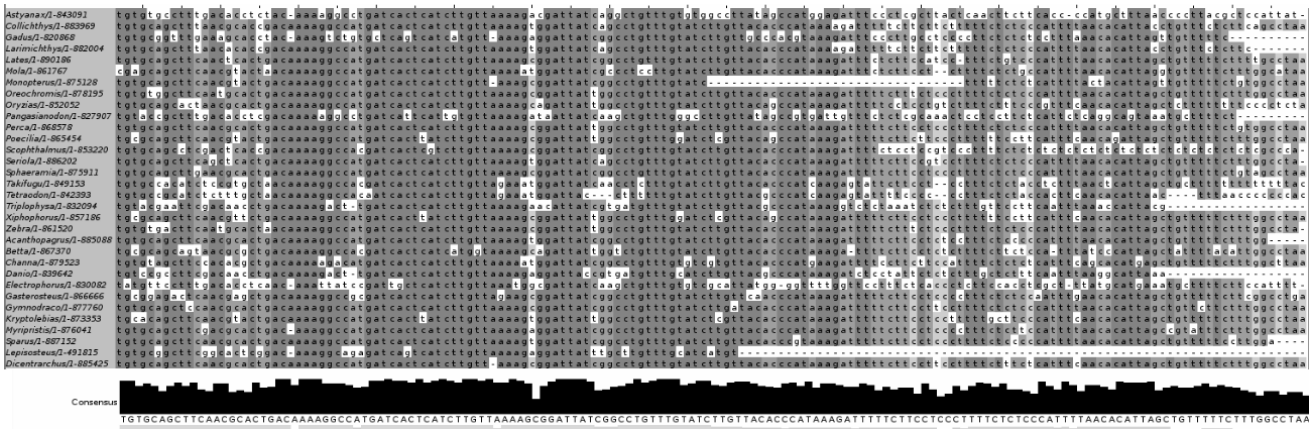

Figure 4. *Trimmed alignment inside the Jalview v2.11.1.4 software, which is used for multiple sequence alignment editing, visualisation and analysis.. Grey indicates different conservation levels. Black bars on the lower part of the figure show the percentage of conservation in each column.*

The single copy CNE orthologs Python code described in section 2.3, returned 2668 elements shared among teleosts and Spotted gar. The tree inference pipeline used in this study (Figure 2), returned the CNE-based tree shown in Figure 5.

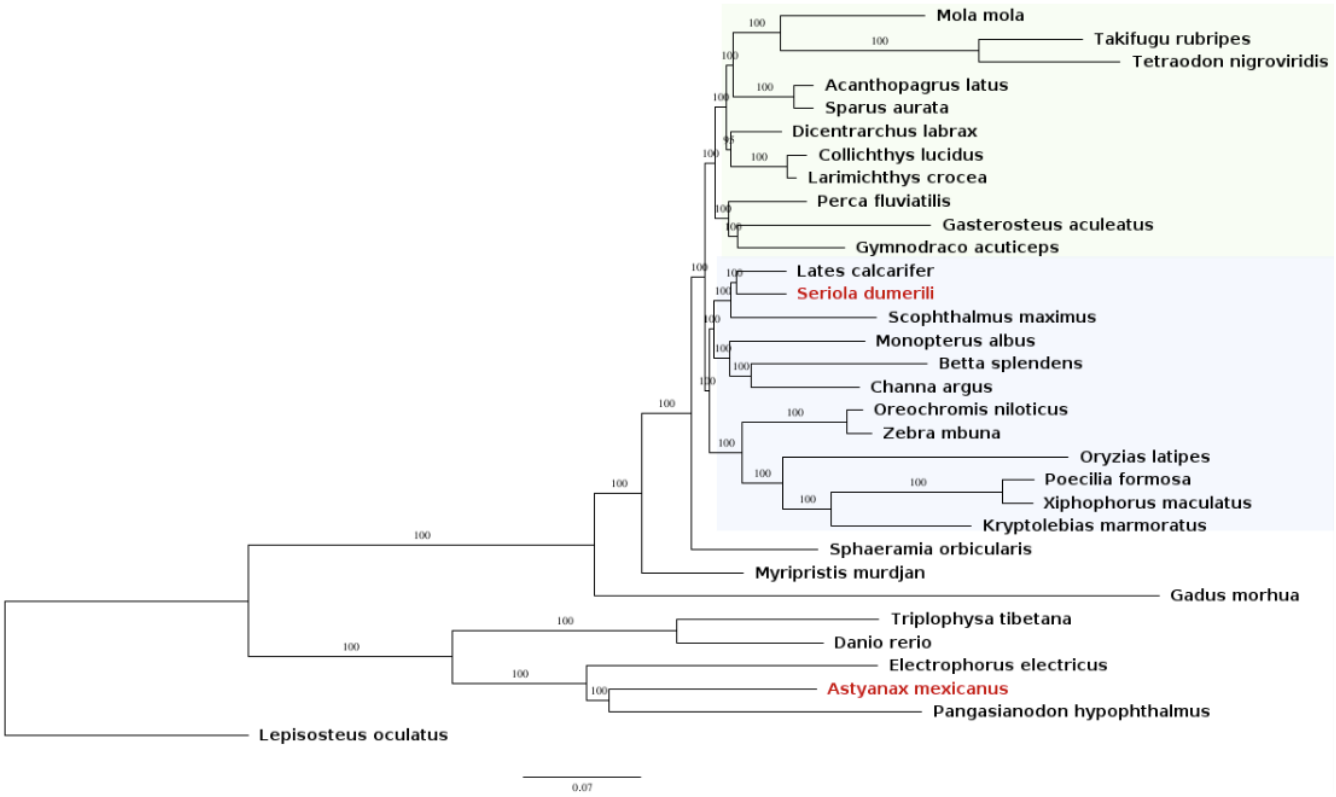

Figure 5. *CNE-based tree using 2668 single copy orthologs shared between teleosts and Spotted gar, derived from Supermatrix method and our tree inference pipeline. Red colour indicates species that were expected to be in a different position, and frames represent the different division of the clades, due to high conservation and less informative sequence alignments.*

The resulting tree holds some differences compared to the one retrieved by using protein-coding genes. As we can see from the frames, the species contained in the blue frame in Figure 5, are grouped in two different clades (*blue* and *red* frame) in Figure 6, with the blue framed clade grouping more closely with the green framed clade in Figure 6.

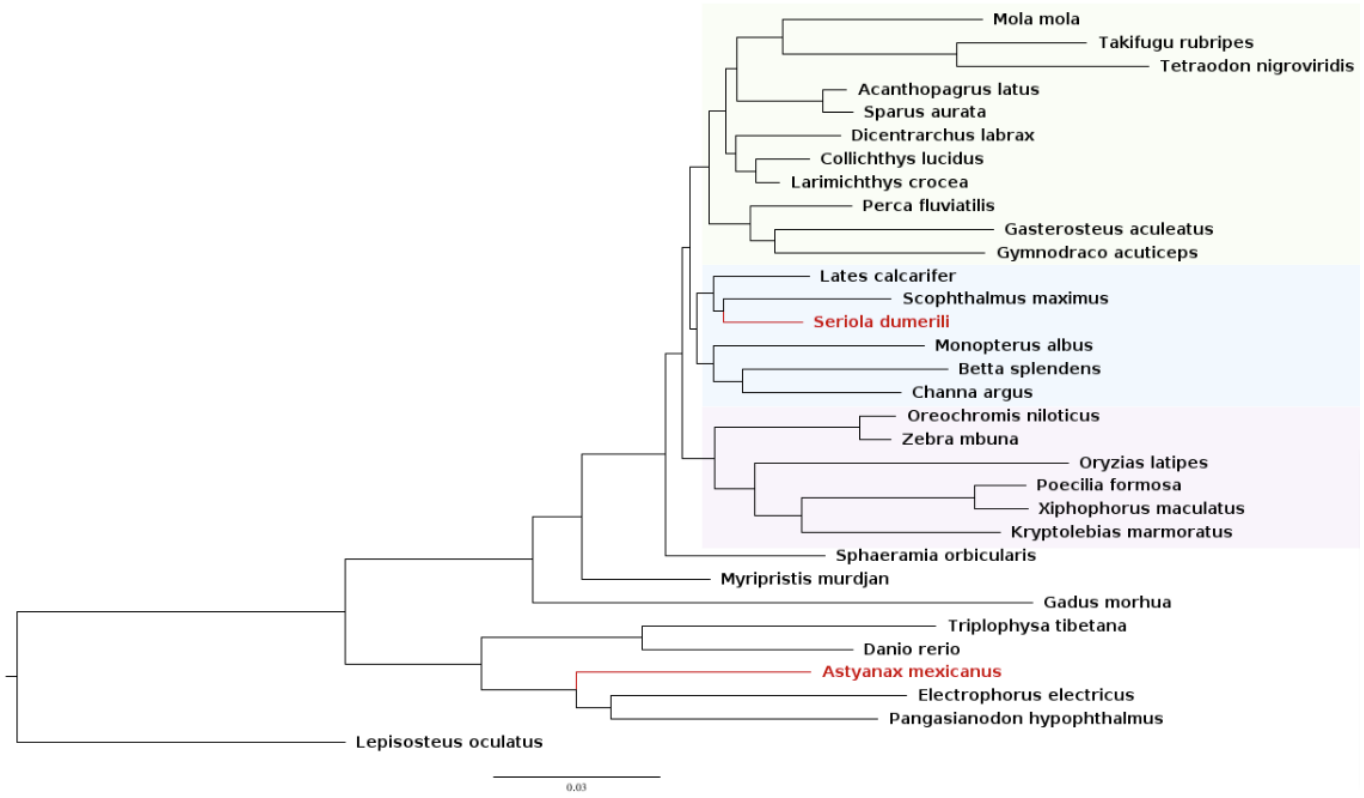

Figure 6. *Gene based tree using single copy orthologs, derived from a tree inference pipeline of our lab. Red colour indicates the differences in the position compared to CNE-based tree, and frames represent the right division of the clades.*

3.3 Identifying the ancestral teleost CNE repertoire

To capture the ancestral teleost CNE repertoire we opted to find the pan-teleost set of CNE. Whereas previous studies have suggested large CNE loss in early teleost evolution (Lee et al., 2011), we wanted to study CNE gain, loss or divergence across teleost species. When comparing the presence of Fugu specific elements with those conserved in Clade 1 as described in section 2.5 *of Materials and Methods*, we found only 6563 (16.6% of fCNEs). Making the same bidirectional comparison, 23059 Zebrafish specific CNEs (36.5% of zCNEs) were found in Clade 2. This Ancestral teleost

CNE set (Figure 3) consisted of 29622 elements. At the same time, only 3210 (5.1%) of the zebrafish CNEs were shared among all teleosts.

3.4 CNE gain and loss during the transition from vertebrates to teleost fishes

Taking advantage of published datasets and using our results derived from previous steps, we further extended our study on investigating possible gains and losses during the transition from vertebrates to teleost fishes. We ran reciprocal similarity searches which gave us the information described below, about those studied CNEs following the 3R WGD.

First, the analyses using similarity searches with our Ancestral teleost CNE set as queries against the ANCORA -Human, Elephant shark, Spotted gar- union set yielded 25007 unmatched elements that probably have been gained just at the formation of the teleost lineage. Moreover, similarity searches with ANCORA -Human, Elephant shark, Spotted gar- union set as query against the 31 teleost genomes, showed that a number of 18321 out of 26198 total elements had no hit and possibly have been lost or diverged during the evolution of the early teleost ancestor.

3.5 Gene association

Next we investigated the functionality of those conserved elements. First, we visualized how they were distributed in each chromosome (Figure 7). The possibly 3R-gained elements using the Zebrafish genome as reference for the 3R (left), and the lost or diverged taking Human as reference of the 2R (right).

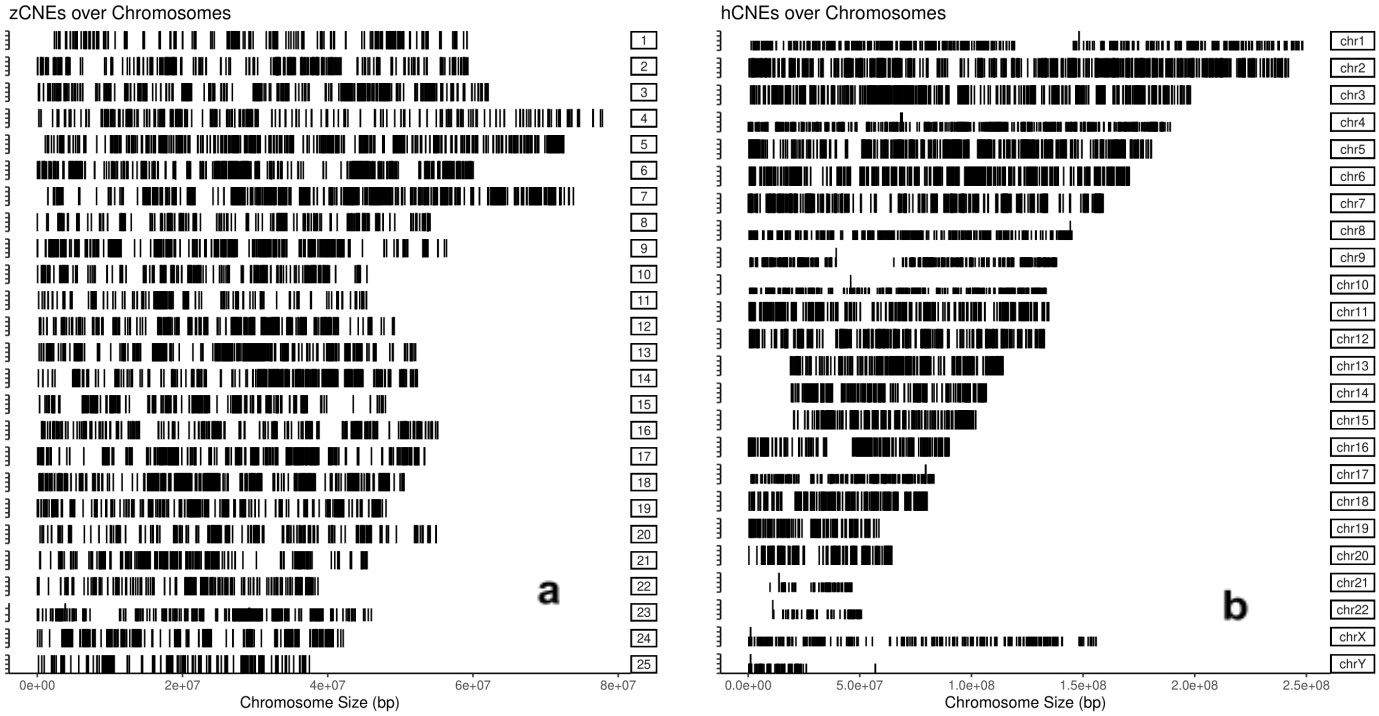

Figure 7. *Coverage plot for visualization of elements on each chromosome. Chromosome size is represented on the x-axis, while chromosome numbers on the y-axis.* ***a.*** *Gained in 3R zCNEs* ***b.*** *Lost in 3R hCNEs. Sex chromosomes were not excluded from the plot.*

3.5.1 Gene association comparisons of CNEs gained and lost at the teleost ancestor

A large number of CNE gains and losses were identified at the base of the teleost phylogeny. This pattern led us to investigate whether lost and gained CNE may have been neighbouring the same genes, or whether gained CNE evolved in unrelated loci to those with lost CNE. To search this, we used the human genome to associate CNEs lost in teleosts with their neighboring genes, which yielded 18316 associations with a total of 4568 genes. This is coherent with previous studies, as many genes own a set of multiple CNEs hidden inside introns, or located far upstream or downstream their locus (Woolfe et al., 2004). The same process was carried out in Zebrafish for teleost "gained" CNEs, yielding 19183 (FIgure 8) associations with a total of 3545 genes. We then used

Human-Zebrafish orthologue gene information from Ensembl as described in section 2.6.1, to compare these two datasets.

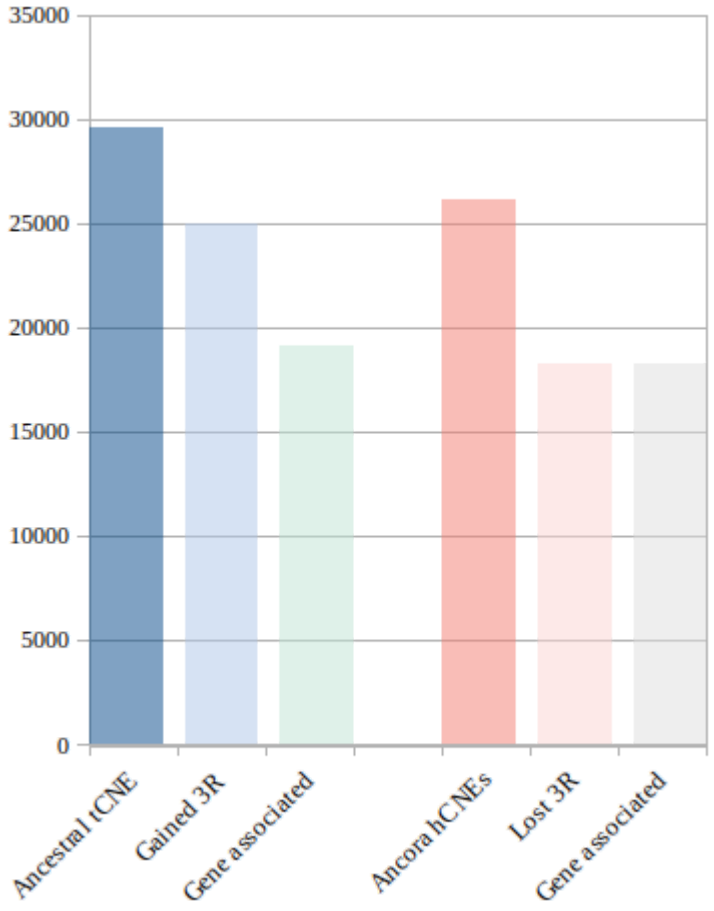

Figure 8. *Results from the analysis using ANCORA elements derived from the Human, Elephant shark and Spotted gar genomes with reciprocal local alignments against ancestral teleost CNEs. Blue bars indicate that most of the ancestral teleost CNEs (tCNE) were gained with the creation of teleost infraclass and the 3R event. Pink bars show that most of the 2R vertebrate (ANCORA) CNEs were lost or diverged following the 3R event.*

The orthology analysis showed that 2800 of the human genes associated with CNEs lost in teleosts have an orthologous Zebrafish gene and 2851 of the genes associated with Zebrafish elements that

seemed to have been gained early in teleost evolution have an orthologous Human gene. Comparing these two subsets revealed a total of 846 orthologous loci between human and zebrafish that are associated with both a gain and a loss of CNE (Table 1).

Table 1: Gene association results

	CNE Associated Genes	**Orthologous Associated Genes**	**Shared**
Gained in 3R	3545	2851	
Lost in 3R	4568	2800	**846**

4. Discussion

During the past few years more studies are discussing, deciphering and investigating the mystery behind non-coding parts of the genome. We already have evidence that some of these non-coding regions are functional and act in cis with genes, especially developmental ones (Sandelin et al., 2004).

Alignment quality and robustness

While there are many obstacles in identifying CNE in non-model genomes, due to the computational cost and the time necessary to properly align genomes, in this study we used a strategy to minimize this cost and obtain the best information possible by carrying out extensive alignments in two focal key points across teleost phylogeny.

When dealing with large genomes and increasing complexity of assemblies, the computational resources are limited and worthy of attention, and of course the results must be evaluated. Therefore, we focused on building our pipeline to yield as much information as possible, by continuously evaluating the process and making positive controls in all steps.

The initial set of CNE from the focal species was used to search 30 more teleost species via BLAST and produce a large database of conserved CNE across various teleost clades. This database was then used to test the potential of conserved elements for performing teleosts phylogenomic studies, as well as study CNE evolution during the vertebrate to teleost transition and associate elements with candidate genes. Moreover, datasets of this study can be used to scan many more other species.

CNEs as a tool for Phylogenomic analyses

The species tree based on CNEs that was constructed in our analysis, had a few differences from the gene-based species tree. This is a fact worth discussing, as the differences between the two trees were likely caused by the extreme conservation found in CNEs across species. When comparing species of

the same infraclass (inside the class of Actinopterygii, the ray-finned fishes), we expect to have low information regarding the sequence diversity, and therefore the difficulty of the tree inference algorithm in distinguishing some key subdivisions in the clades.

CNEs are extremely conserved sequences with unknown sequence evolution patterns. Overall, the conservation is much higher than the protein coding genes selected for the gene-based phylogeny, which has resulted in fewer informative sites. As a result the phylogeny produced, although very similar to the one produced by the protein coding moiety of the genome, has slight discrepancies which shows that CNEs are less reliable than coding regions for phylogenetic reconstruction. To further test this information in the future, this could be potentially resolved via a combinatorial approach that combines CNE and gene based phylogenies.

CNE evolution in teleosts

There is a large set of conserved elements found to be ancestral in teleosts. However, among species of the same infraclass with 206 - 252 MYA divergence time, many developmental cis-regulatory elements have diverged. The most representative example is the family of Tetraodontidae (e.g. *T.rubripes*, *T.nigroviridis*) of the Tetraodontiformes order which share with the family of Cyprinidae (e.g. *D.rerio*, *A.mexicanus*) a relatively small amount of CNEs, compared to other families with the same divergence time such as Holocentridae (e.g. *M.murdjan*) (see Table 2). Evolutionary rates may be the key to deciphering this "imbalance", as for example Tetraodontidae is a well-known family for it's fast evolutionary rate (Danis et al. in press).

Another key point of this study is how massive both gain and loss is, in the transition to teleosts. A small fraction of genes associated with multiple CNEs, that seemed to have been lost or gained in this transition, were found to be related. Nevertheless, these genes are particularly interesting, as there has to be an explanation on how they seemed to have lost or gained following the transition. Following this finding, we examined two hypotheses (Figure 9). The first (*Case 1)* is the hypothesis that those elements were simply diverged in the common ancestor until the sequence has changed and was inherited in the descendants. The second (Case 2), is that the ancestral CNE has been "lost" and replaced by another *de novo* element. It is also possible that both of these scenarios could have taken

place in different sets of CNEs. Based on the results of the ortholog analysis in which some genes associated with multiple lost or gained CNEs are related, we assume that at least for the 846 orthologous loci between Human and Zebrafish Case 1 holds, while for the rest Case 2 seems to be the applicable. Thus, many CNEs have actually diverged in the long branch of the teleost ancestor, and many others seem to have evolved right after the loss of the respective CNEs inherited from vertebrates.

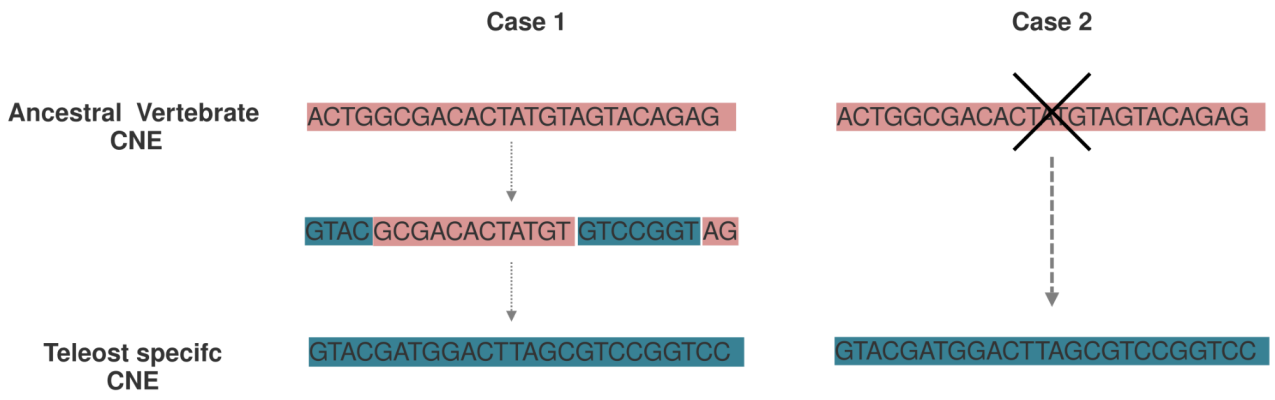

Figure 9. *Testing two possible cases for the evolution of "unmatched" elements. Case 1. Divergence in the common ancestor. Case 2. De novo CNE birth.*

5. Conclusions

In the past decades conserved non-coding sequences of the genome have been revolutionizing elements to study cis-regulation. In this study we used a carefully designed strategy which yielded an extensive teleost CNE dataset, compared ancestral teleost CNEs with those derived from other vertebrate species and performed a gene association analysis with an important subset of our CNE datasets. Also, we identified a set of related orthologous loci between Zebrafish and Human, which in previous steps seemed to have been lost. This is the first thorough analysis following the explosive increase in the availability of vertebrate genomes. Future analysis on many more species will shed light on the evolution of the non-coding genomes within teleosts and all other organisms.

CODE AVAILABILITY

All the custom scripts that were used in this project are available in the GitHub repository:

https://github.com/genomenerds/CNE_analysis

Table 2. BLAST results

Species	***# of zCNEs***	***# of fCNEs***	***# of hCNEs***
A. latus	*18255*	*34899*	*4653*
A. mexicanus	*57551*	*4532*	*5350*
B. splendens	*15020*	*22667*	*4553*
C. argus	*16455*	*26226*	*4562*
C. lucidus	*18029*	*32927*	*4768*
D. rerio	*63023*	*5074*	*4932*
D. labrax	*18471*	*33585*	*5340*
E. electricus	*40392*	*3937*	*4806*
G. morhua	*10032*	*8335*	*3170*
G. aculeatus	*14497*	*23361*	*4420*
G. acuticeps	*16538*	*28692*	*4708*
K. marmoratus	*14563*	*21202*	*4667*
L. crocea	*18187*	*32965*	*5321*
L. calcarifer	*18420*	*32046*	*5410*
M. mola	*14729*	*30942*	*4700*
M. albus	*15834*	*26227*	*4995*
M. murdjan	*20044*	*25670*	*5057*

O. niloticus	*16856*	*27584*	*5075*
O. latipes	*12454*	*16772*	*4275*
P. hypophthalmus	*40328*	*4625*	*4508*
P. fluviatilis	*16727*	*32237*	*4884*
P. formosa	*13910*	*19958*	*4535*
S. maximus	*14241*	*27135*	*4498*
S. dumerili	*18296*	*32193*	*5362*
S. aurata	*18476*	*37371*	*5202*
S. orbicularis	*16413*	*28605*	*4940*
T. rubripes	*12526*	*39532*	*4329*
T. nigroviridis	*11469*	*32867*	*3955*
T. tibetana	*55660*	*4967*	*4789*
X. maculatus	*13764*	*19640*	*4554*
Z. mbuna	*15297*	*27592*	*4672*

References

Aaron R. Quinlan, Ira M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, Volume 26, Issue 6, 15 March 2010, Pages 841–842, https://doi.org/10.1093/bioinformatics/btq033

Alison P. Lee, Sze Yen Kerk, Yue Ying Tan, Sydney Brenner, Byrappa Venkatesh, Ancient Vertebrate Conserved Noncoding Elements Have Been Evolving Rapidly in Teleost Fishes, Molecular Biology and Evolution, Volume 28, Issue 3, March 2011, Pages 1205–1215, https://doi.org/10.1093/molbev/msq304

Armstrong, J., Fiddes, I. T., Diekhans, M., & Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. Annual Review of Animal Biosciences, 7, 41–64. https://doi.org/10.1146/annurev-animal-020518-115005

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402. PubMed

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A. M., Campbell, M. S., Barrell, D., Martin, K. J., Mulley, J. F., Ravi, V., Lee, A. P., Nakamura, T., Chalopin, D., … Postlethwait, J. H. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics, 48(4), 427–437. https://doi.org/10.1038/ng.3526

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V., & Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Molecular Biology and Evolution, 23(9), 1808–1816. https://doi.org/10.1093/molbev/msl049

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V., & Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication

in teleost fishes. Molecular Biology and Evolution, 23(9), 1808–1816. https://doi.org/10.1093/molbev/msl049

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Casane, D., & Rétaux, S. (2016). Evolutionary Genetics of the Cavefish Astyanax mexicanus. Advances in Genetics, 95, 117–159. https://doi.org/10.1016/bs.adgen.2016.03.001

Comings D. E., 1972. The genetic organization of chromosomes. Adv. Hum. Genet. 3: 237–431.

Davies, K. T. J., Tsagkogeorga, G., & Rossiter, S. J. (2014). Divergent evolutionary rates in vertebrate and mammalian specific conserved non-coding elements (CNEs) in echolocating mammals. BMC Evolutionary Biology, 14(1), 1–19. https://doi.org/10.1186/s12862-014-0261-5

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biology, 3(10). https://doi.org/10.1371/journal.pbio.0030314

Dna, F. (n.d.). Analysis workflow for IQ-TREE 2 paper Non-reversible models Mixture models. 0, 0–4.

Earl, D., Nguyen, N., Hickey, G., Harris, R. S., Fitzgerald, S., Beal, K., Seledtsov, I., Molodtsov, V., Raney, B. J., Clawson, H., Kim, J., Kemena, C., Chang, J. M., Erb, I., Poliakov, A., Hou, M., Herrero, J., Kent, W. J., Solovyev, V., … Paten, B. (2014). Alignathon: A competitive assessment of whole-genome alignment methods. Genome Research, 24(12), 2077–2089. https://doi.org/10.1101/gr.174920.114

Engström, P. G., Fredman, D., & Lenhard, B. (2008). Ancora: A web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. Genome Biology, 9(2), 8–11. https://doi.org/10.1186/gb-2008-9-2-r34

Frith, M. C., Hamada, M., & Horton, P. (2010). Parameters for accurate genome alignment. BMC Bioinformatics, 11. https://doi.org/10.1186/1471-2105-11-80

Glasauer, S. M. K., & Neuhauss, S. C. F. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. Molecular Genetics and Genomics, 289(6), 1045–1060. https://doi.org/10.1007/s00438-014-0889-2

Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., Forczek, E., Joly-Lopez, Z., Steffen, J. G., Hazzouri, K. M., Dewar, K., Stinchcombe, J. R., Schoen, D. J., Wang, X., Schmutz, J., Town, C. D., Edger, P. P., Pires, J. C., Schumaker, K. S., … Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature Genetics, 45(8), 891–898. https://doi.org/10.1038/ng.2684

Hiller, M., Agarwal, S., Notwell, J. H., Parikh, R., Guturu, H., Wenger, A. M., & Bejerano, G. (2013). Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: Application to zebrafish. Nucleic Acids Research, 41(15). https://doi.org/10.1093/nar/gkt557

Hiller, M., Schaar, B. T., & Bejerano, G. (2012). Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Research, 40(22), 11463–11476. https://doi.org/10.1093/nar/gks905

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Ridwan Amode, M., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. Nucleic Acids Research, 49(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Inoue, J., & Saitou, N. (2020). dbCNS: A New Database for Conserved Noncoding Sequences. Molecular Biology and Evolution, 1–12. https://doi.org/10.1093/molbev/msaa296

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D493-6.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Kimmel, C. B., Hatta, K., & Eisen, J. S. (1991). Genetic control of primary neuronal development in zebrafish. Development, 113(SUPPL. 2), 47–57.

Leimeister, C. A., Dencker, T., & Morgenstern, B. (2019). Accurate multiple alignment of distantly related genome sequences using filtered spaced word matches as anchor points. Bioinformatics, 35(2), 211–218. https://doi.org/10.1093/bioinformatics/bty592

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Moghadam, H. K., Ferguson, M. M., & Danzmann, R. G. (2009). Comparative genomics and evolution of conserved noncoding elements (CNE) in rainbow trout. BMC Genomics, 10, 1–13. https://doi.org/10.1186/1471-2164-10-278

Naval-Sanchez, M., Potier, D., Hulselmans, G., Christiaens, V., & Aerts, S. (2015). Identification of lineage-specific cis-regulatory modules associated with variation in transcription factor binding and chromatin activity using ornstein-uhlenbeck models. Molecular Biology and Evolution, 32(9), 2441–2455. https://doi.org/10.1093/molbev/msv107

Ochoa, L. E., Datovo, A., DoNascimiento, C., Roxo, F. F., Sabaj, M. H., Chang, J., Melo, B. F., Silva, G. S. C., Foresti, F., Alfaro, M., & Oliveira, C. (2020). Phylogenomic analysis of trichomycterid catfishes (Teleostei: Siluriformes) inferred from ultraconserved elements. Scientific Reports, 10(1), 1–15. https://doi.org/10.1038/s41598-020-59519-w

Polychronopoulos, D., King, J. W. D., Nash, A. J., Tan, G., & Lenhard, B. (2017). Conserved non-coding elements: Developmental gene regulation meets genome organization. Nucleic Acids Research, 45(22), 12611–12624. https://doi.org/10.1093/nar/gkx1074

Ren, F., Tanaka, H., & Yang, Z. (2009). A likelihood look at the supermatrix-supertree controversy. Gene, 441(1–2), 119–125. https://doi.org/10.1016/j.gene.2008.04.002

Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., & Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics, 5, 1–9. https://doi.org/10.1186/1471-2164-5-99

Schmidt, F., Kern, F., & Schulz, M. H. (2020). Integrative prediction of gene expression with chromatin accessibility and conformation data. Epigenetics and Chromatin, 13(1), 1–17. https://doi.org/10.1186/s13072-020-0327-0

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., & Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Research, 13(1), 103–107. https://doi.org/10.1101/gr.809403

Sharma, V., & Hiller, M. (2017). Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. Nucleic Acids Research, 45(14), 8369–8377. https://doi.org/10.1093/nar/gkx554

Tan, G., Polychronopoulos, D., & Lenhard, B. (2019). CNER: A toolkit for exploring extreme noncoding conservation. PLoS Computational Biology, 15(8), 1–16. https://doi.org/10.1371/journal.pcbi.1006940

Theodoros Danis, Alexandros Tsakogiannis, Jon B. Kristoffersen, Daniel Golani, Dimitris Tsaparis, Panagiotis Kasapidis, Georgios Kotoulas, Antonios Magoulas, Costas S. Tsigenopoulos, Tereza Manousaki. Building a high-quality reference genome assembly for the the eastern Mediterranean Sea invasive sprinter Lagocephalus sceleratus (Tetraodontiformes, Tetraodontidae) bioRxiv 2020.02.17.952580; doi: https://doi.org/10.1101/2020.02.17.952580

Venkatesh, B., Kirkness, E. F., Loh, Y. H., Halpern, A. L., Lee, A. P., Johnson, J., Dandona, N., Viswanathan, L. D., Tay, A., Venter, J. C., Strausberg, R. L., & Brenner, S. (2007). Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome. PLoS Biology, 5(4), 932–944. https://doi.org/10.1371/journal.pbio.0050101

Voldoire, E., Brunet, F., Naville, M., Volff, J. N., & Galiana, D. (2017). Expansion by whole genome duplication and evolution of the sox gene family in teleost fish. PLoS ONE, 12(7), 1–20. https://doi.org/10.1371/journal.pone.0180936

Volff, J. N. (2005). Genome evolution and biodiversity in teleost fish. Heredity, 94(3), 280–294. https://doi.org/10.1038/sj.hdy.6800635

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., & Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biology, 3(1). https://doi.org/10.1371/journal.pbio.0030007

Yao, L., Berman, B. P., & Farnham, P. J. (2015). Demystifying the secret mission of enhancers: Linking distal regulatory elements to target genes. Critical Reviews in Biochemistry and Molecular Biology, 50(6), 550–573. https://doi.org/10.3109/10409238.2015.1087961

Yue, J. X., Kozmikova, I., Ono, H., Nossa, C. W., Kozmik, Z., Putnam, N. H., Yu, J. K., & Holland, L. Z. (2016). Conserved noncoding elements in the most distant genera of cephalochordates: The goldilocks principle. Genome Biology and Evolution, 8(8), 2387–2405. https://doi.org/10.1093/gbe/evw158

Zafeiropoulos, H., Gioti, A., Ninidakis, S., Potirakis, A., Paragkamian, S., Angelova, N., Antoniou, A., Danis, T., Kaitetzidou, E., Kasapidis, P., Kristoffersen, J. B., Papadogiannis, V., Pavloudi, C., Ha, Q. V., Lagnel, J., Pattakos, N., Perantinos, G., Sidirokastritis, D., Vavilis, P., … Pafilis, E. (2021). 0s and 1s in marine molecular research: A regional HPC perspective. GigaScience, 10(8), 1–12. https://doi.org/10.1093/gigascience/giab053

Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., & Flicek, P. (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience, 6(7), 1–8. https://doi.org/10.1093/gigascience/gix038

Zhou, Y., Xiao, S., Lin, G., Chen, D., Cen, W., Xue, T., Liu, Z., Zhong, J., Chen, Y., Xiao, Y., Chen, J., Guo, Y., Chen, Y., Zhang, Y., Hu, X., & Huang, Z. (2019). Chromosome genome assembly and annotation of the yellowbelly pufferfish with PacBio and Hi-C sequencing data. Scientific Data, 6(1), 1–8. https://doi.org/10.1038/s41597-019-0279-z

Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626 (2012).

Haberle, V., Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol 19, 621–637 (2018). https://doi.org/10.1038/s41580-018-0028-8

Chen, N. (2004), Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics, 5: 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s05

Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D493-6.

Ohno S., 1972. So much “junk” DNA in our genome, pp. 366–370 in Evolution of Genetic Systems, edited by Smith H. H. Gordon & Breach, New York.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. Track Data Hubs enable visualization of user-defined genome-wide

annotations on the UCSC Genome Browser. Bioinformatics. 2014 Apr 1;30(7):1003-5. Epub 2013 Nov 13.

S. Kumar, G. Stecher, M. Suleski, and S.B. Hedges, 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Molecular Biology and Evolution 34: 1812-1819, DOI: 10.1093/molbev/msx116.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

Rambaut A FigTree v1.4.4. http://tree.bio.ed.ac.uk/software/figtree/. Accessed 3 December 2020

G Yu, LG Wang, QY He. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015, 31(14):2382-2383. doi:[10.1093/bioinformatics/btv145]

Huber, W., Carey, V., Gentleman, R. et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12, 115–121 (2015). https://doi.org/10.1038/nmeth.3252

G Yu*, LG Wang, and QY He*. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015, 31(14):2382-2383. doi: 10.1093/bioinformatics/btv145

Gish, W. & States, D.J. (1993) "Identification of protein coding regions by database similarity search." Nature Genet. 3:266-272. PubMed

Madden, T.L., Tatusov, R.L. & Zhang, J. (1996) "Applications of network BLAST server" Meth. Enzymol. 266:131-141. PubMed

Zhang Z., Schwartz S., Wagner L., & Miller W. (2000), "A greedy algorithm for aligning DNA sequences" J Comput Biol 2000; 7(1-2):203-14. PubMed