Marine Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Method paper

Selective pressure on the protein-coding genes of the pufferfish is correlated with phenotypic traits

Hyeonju Ahn[a], Chul Lee[d], Bo-Hye Nam[b], Eun Bae Kim[c], Kelsey Caetano-Anolles[a], Heebal Kim[a,*]

[a] Department of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Republic of Korea
[b] Biotechnology Research Division, National Fisheries Research & Development Institute, Busan 619-705, Republic of Korea
[c] Department of Animal Life Science, Kangwon National University, Chuncheon 200-701, Republic of Korea
[d] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Republic of Korea

ARTICLE INFO

Keywords:
Pufferfish
Takifugu rubripes
Tetraodon nigroviridis
Evolution
Protein-coding gene
Tetrodotoxin

ABSTRACT

The pufferfish accumulates neurotoxic tetrodotoxin in its body and inflates by filling its stomach with water. These traits are unique to this species, and may be a result of adaptation post-divergence of Tetraodontidae. However, evolution of the protein-coding genes in the pufferfish has not yet been well elucidated. Detection of positive selection on these genes can help us understand the mechanisms associated with functional evolution. We downloaded well-annotated gene information of two pufferfish species, *Takifugu rubripes* and *Tetraodon nigroviridis*, from the public ENSEMBL database. In order to detect selective pressure on protein-coding sequences, we performed dN/dS estimation using codeml within the PAML software package. We selected one to one orthologous genes among seven fish species (*Gasterosteus aculeatus*, *Oryzias latipes*, *Poecilia formosa*, *Takifugu rubripes*, *Tetraodon nigroviridis*, and *Xiphophorus maculatus*). Results of dN/dS analysis on orthologous genes indicate that pufferfish showed high non-synonymous substitution rate for positively selected genes, and the evolutionary rate was faster during the diversification of two pufferfishes after divergence. Additionally, a candidate mechanism for regulation of neuro-toxicity of tetrodotoxin was identified from functional annotation of positively selected genes. These results support positive selection on protein-coding genes of the pufferfish with the acquisition of specific phenotypic traits.

1. Introduction

The pufferfish display some species-specific characteristics, including the accumulation of tetrodotoxin (TTX) in their tissues. It has been reported that many species of pufferfish, including *Takifugu rubripes* and *Tetraodon nigroviridis,* accumulate TTX in their tissue through the marine food chain, which begins with TTX-producing bacteria (Noguchi et al., 2006). TTX is a toxin which induces abnormalities in nerve excitability by blocking Na + channels (Kiernan et al., 2005). Additionally, the pufferfish pumps water into its stomach and inflates its body volume in order to defend itself from predators. As the pufferfish's stomach specializes in inflation, it shows high extensibility and unique digestive function (Brainerd, 1994).

Evolutionary study of pufferfish coding genes, along with other fish orthologous genes, may allow for better understanding of the genetic mechanisms behind pufferfish-specific phenotypes. A study was previously performed investigating the evolutionary patterns of *Takifugu rubripes* and *Tetraodon nigroviridis* and their protein-coding genes using pairwise dN/dS analysis (Montoya-Burgos, 2011). Both pufferfishes showed similar strength of purifying selection when compared to murids, and the strength in hominids was stronger with their smaller population size. However, a comparative evolutionary study of protein-coding genes in pufferfish has yet to be performed fully.

dN/dS represents the ratio of non-synonymous substitution rate to synonymous substitution rate, which is calculated using the aligned orthologous sequences of protein-coding genes (Yang et al., 2000). It is used as an estimator to measure selective pressure on genes, and the significance of positive selection is decided using a likelihood test between the null (neutral selection) and alternative hypotheses (positive selection) (Yang, 2007). This estimation enables investigation of single gene evolution as well as evolution of several genes at once, from which biological functions can be assumed through functional annotation of selected genes using a public database.

In the present study, we utilized dN/dS estimation and a branch-site model to identify selective pressure on pufferfish coding genes. We selected seven fish species from the ENSEMBL genome database,

* Corresponding author at: Department of Agricultural Biotechnology, Animal Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Republic of Korea.
E-mail address: heebal@snu.ac.kr (H. Kim).

https://doi.org/10.1016/j.margen.2017.11.015
Received 28 June 2017; Received in revised form 4 November 2017; Accepted 29 November 2017

considering the problem of substitution saturation which arises from a long divergence time. The seven species, *Gasterosteus aculeatus*, *Oryzias latipes*, *Oreochromis niloticus*, *Poecilia formosa*, *Takifugu rubripes*, *Tetraodon nigroviridis*, and *Xiphophorus maculatus*, diverged from Percomorphaceae 117 mya (Supplementary Fig. 1, the divergence time from TIMETREE) (Hedges et al., 2015). Percomorphs are a large group of spiny-finned fishes which contains around 55% (> 17,000 species) of teleost diversity (Sanciangco et al., 2016). Because the resolution of the order lacks in density, studies aiming to classify phylogenetic placement and give a concrete shape have been consistently performed. In the meantime, comparison between these seven fish species may suggest direction for future functional evolutionary studies in Percomorphs. Results of this study may also help to elucidate rapid evolution of the protein-coding genes of pufferfish and suggest possible underlying genomic mechanisms behind pufferfish-specific characteristics.

2. Materials and method

2.1. Selection of fish species for dN/dS analysis

We selected seven out of the eleven fish species whose data is available in the ENSEMBL genome database- *Gasterosteus aculeatus* (stickleback), *Oryzias latipes* (medaka), *Oreochromis niloticus* (tilapia), *Poecilia Formosa* (amazon molly), *Takifugu rubripes* (fugu), *Tetraodon nigroviridis* (tetraodon), and *Xiphophorus maculatus* (platyfish). Although saturation of synonymous substitutions weakens the power of the branch-site test in dN/dS estimation (Gharib and Robinson-Rechavi, 2013), the relation of divergence time and saturation of synonymous substitution in the whole genome of teleost fish has not yet been studied sufficiently. For the present study, we chose species who diverged no > 200 mya of each other based on information taken from TIMETREE (Supplementary Fig. 1) (Hedges et al., 2015). The genome assembly version used for each of the seven species was TETRAODON8.0 (*Tetraodon nigroviridis*), FUGU4.0 (*Takifugu rubripes*), HdrR (*Oryzias latipes*). Orenil1.0 (*Oreochromis niloticus*), BROADS1 (*Gasterosteus aculeatus*), PoeFor_5.1.2 (*Poecilia formosa*), and Xipmac4.4.2 (*Xiphophorus maculatus*).

2.2. Data preparation

First, we downloaded text data files of cDNA and peptide sequences of all transcripts via ENSEMBL Biomart. We retained the longest transcripts of each gene to use for our analysis, as longer sequences provide more genetic information and are more advantageous. Only genes that were 1:1 orthologous between our selected seven species were used for dN/dS estimation. Gene homology information was collected from the ENSEMBL database. Multiple alignment of peptide sequences was performed using ClustalO (Sievers et al., 2011), and the sequences were converted into corresponding cDNA sequences using PAL2NAL (Suyama et al., 2006). Poorly aligned transcripts were eliminated using Gblocks (Castresana, 2000). After all of these filtering steps were performed, a total of 10,136 orthologs remained.

2.3. dN/dS analysis of protein coding sequences for detecting positive selection

Codeml within the PAML package was used to estimate dN/dS and the likelihood of positive selection (Yang, 2007). Using the branch-site model, we set three cases of foreground branches which were considered as their few sites under positive selection (Fig. 1). Fig. 1 illustrates three phylogenetic trees with foreground branches marked by red lines, representing positive selection in Tetraodontiformes, *Takifugu rubripes*, and *Tetraodon nigroviridis*. Foreground branches represent the particular lineages that positive selection affects few sites of them in branch-site models (Yang, 2007). Codeml control files (.ctl) of branch-site model set model = 2, NSsites = 2, fix_omega = 1, and omega = 1 in the null model and model = 2, NSsites = 2, and fix_omega = 0 in the alternative model according to the PAML manual. From the output, the p-value of each gene was calculated from $2\Delta l$ (where l is likelihood of a model) and a chi-square distribution. FDR (False-Discovery rate) for all p-values was calculated using the statistical computing program R (Team, R.C., 2013). 982 genes in Tetraodontiformes, 1526 genes in *Takifugu rubripes*, and 2789 in *Tetraodon nigroviridis* were identified as significant at a cutoff of FDR < 0.1. An ANOVA and Tukey's HSD test for dN/dS was also performed using R (Supplementary Tables 1and 2). Additionally, in order to compare the dN/dS values of pufferfish and other fish species, the same dN/dS analysis was performed on *Poecilia Formosa* and *Xiphophorus maculatus* with similar lineage structures.

2.4. Functional annotation of genes under positive selection

Functional annotation was performed on identified genes using GOEAST (Gene Ontology Enrichment Analysis Software Toolkit) (Zheng and Wang, 2008). First, we downloaded a list of GO term annotations from ENSEMBL Biomart. Annotation lists of *Takifugu rubripes* and *Tetraodon nigroviridis* were prepared in order to compare the results of the enrichment tests. The annotation data was modified into appropriate formatting according to the GOEAST web analyzer requirements. Modified data was used as a background file in GOEAST Customized Result Analysis. We uploaded ENSEMBL ids of selected genes (FDR < 0.1) to GOEAST; the result was outputted as sorted list based on p-values and represented as hierarchical figures (Fig. 2, Supplementary Figs. 2–5). Analyses were run using GOEAST default options (FDR < 0.1, Benjamini–Hochberg–Yekutieli procedure).

3. Results

3.1. Evolutionary pressure on coding sequences of Tetraodontiformes, Takifugu rubripes and Tetraodon nigroviridis

We estimated dN/dS values of 10,136 orthologous genes for foreground branches and background branches using the codeml package within PAML 4. When the foreground branch was set as Tetraodontiformes, the median dN/dS value was 0.1302 in background branches under null hypothesis, 0.1649 in foreground branches under null hypothesis, 0.1308 in background branches under alternative hypothesis, and 0.1886 in foreground branches under alternative hypothesis (Table 1). In the branch-site model, the null hypothesis supposes that positive selection has not affected any sites of genes of foreground branches or background branches ($w_2 = 1$ fixed by fix_omega = 1 and omega = 1), while the alternative hypothesis assumes positive selection on foreground branches (Yang, 2007). *Takifugu rubripes* as a foreground lineage showed a higher median dN/dS value than Tetraodontiformes, which showed 0.1398 in background branches under null hypothesis, 0.1674 in foreground branches under null hypothesis, 0.1396 in background branches under alternative hypothesis, and 0.1956 in foreground branches under alternative hypothesis. The median dN/dS value in *Tetraodon nigroviridis* as a foreground lineage was shown to be 0.1372 in background branches under null hypothesis, 0.1737 in foreground branches under null hypothesis, 0.1370 in background branches under alternative hypothesis, and 0.2300 in foreground branches under alternative hypothesis. Analysis of only positively selected genes (PSGs, fdr < 0.1 in the likelihood test) resulted in an median dN/dS value of 0.8598 in Tetraodontiformes, 4.6688 in *Takifugu rubripes* and 4.0973 in *Tetraodon nigroviridis* as alternative foreground branches.

dN/dS values significantly ($p < 0.001$, ANOVA) differed among Tetraodontiformes, *Takifugu rubripes* and *Tetraodon nigroviridis* in all hypotheses (null background, null foreground, alternative background, alternative foreground and alternative foreground (fdr < 0.1)) (Supplementary Table 1). Results of the Tukey's HSD Post-hoc comparison test showed a significant ($p < 0.001$) difference between