Received: 19 August 2020 | Revised: 24 January 2021 | Accepted: 12 March 2021

DOI: 10.1002/ece3.7507

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

Genome-wide analysis of European sea bass provides insights into the evolution and functions of single-exon genes

Mbaye Tine[1,2] | Heiner Kuhl[3] | Peter R. Teske[4] | Richard Reinhardt[2]

[1]UFR des Sciences Agronomiques, de l'Aquaculture et des Technologies Alimentaires (S2ATA), Université Gaston Berger (UGB), Saint-Louis, Senegal

[2]Genome Centre at the Max-Planck Institute for Plant Breeding Research, Köln, Germany

[3]Department of Ecophysiology and Aquaculture, Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany

[4]Department of Zoology, Centre for Ecological Genomics and Wildlife Conservation, University of Johannesburg, Johannesburg, South Africa

Correspondence
Mbaye Tine, UFR des Sciences Agronomiques de l'Aquaculture et des Technologies Alimentaire (UFR S2ATA), Université Gaston Berger (UGB), Route de Ngallele BP 234, Saint-Louis, Senegal.
Email: tine@mpipz.mpg.de

Funding information
Max Planck Institute for Molecular Genetics; Genome Centre at MPI Cologne, Germany; Gaston Berger University

Abstract

Several studies have attempted to understand the origin and evolution of single-exon genes (SEGs) in eukaryotic organisms, including fishes, but few have examined the functional and evolutionary relationships between SEGs and multiple-exon gene (MEG) paralogs, in particular the conservation of promoter regions. Given that SEGs originate via the reverse transcription of mRNA from a "parental" MEGs, such comparisons may enable identifying evolutionarily-related SEG/MEG paralogs, which might fulfill equivalent physiological functions. Here, the relationship of SEG proportion with MEG count, gene density, intron count, and chromosome size was assessed for the genome of the European sea bass, *Dicentrarchus labrax*. Then, SEGs with an MEG parent were identified, and promoter sequences of SEG/MEG paralogs were compared, to identify highly conserved functional motifs. The results revealed a total count of 1,585 (8.3% of total genes) SEGs in the European sea bass genome, which was correlated with MEG count but not with gene density. The significant correlation of SEG content with the number of MEGs suggests that SEGs were continuously and independently generated over evolutionary time following species divergence through retrotranscription events, followed by tandem duplications. Functional annotation showed that the majority of SEGs are functional, as is evident from their expression in RNA-seq data used to support homology-based genome annotation. Differences in 5′UTR and 3′UTR lengths between SEG/MEG paralogs observed in this study may contribute to gene expression divergence between them and therefore lead to the emergence of new SEG functions. The comparison of nonsynonymous to synonymous changes (Ka/Ks) between SEG/MEG parents showed that 74 of them are under positive selection (Ka/Ks > 1; $p = .0447$). An additional fifteen SEGs with an MEG parent have a common promoter, which implies that they are under the influence of common regulatory networks.

KEYWORDS

comparative genomics, *Dicentrarchus labrax*, European sea bass, evolution, promoter, single-exon gene

1 | INTRODUCTION

Early comparative genomic studies on eukaryotes showed that the majority of their genes consist of multiple exons, including coding sequences and untranslated regions, or UTRs. These are the precursors of mRNA and are interrupted by noncoding sequences called introns (Long, 2002; Rogozin et al., 2005; Smith, 1988). These introns are spliced out and exons are concatenated to form the mature mRNA, which is then expressed as a protein product (Rogozin et al., 2005). Recent studies using high-throughput sequencing technologies have revealed that the genomes of both unicellular and multicellular organisms are composed of a significant proportion of single-exon genes (SEGs), also called intronless genes (Tay et al., 2009; Tine et al., 2011; Venter, 2001; Yang et al., 2009; Zou et al., 2011). This discovery has raised serious questions about the origin, evolution, and function of this type of genes (Fablet et al., 2009; Savisaar & Hurst, 2016; Shabalina et al., 2010; Zou et al., 2011). Most studies on the origin of SEGs and evolution suggest that they are the result of the reverse transcription of mRNA from a "parental" gene into cDNA and its insertion elsewhere in the genome (Ostertag & Kazazian, 2001). This process, known as retrotransposition, is mediated by long interspersed nuclear element 1 (LINE 1)-derived enzymes, which encode a reverse transcriptase enzyme that can produce a DNA copy from any RNA molecule in the cell (Cordaux & Batzer, 2008; Doenecke & Albig, 2005; Kaessmann et al., 2009; Sakharkar et al., 2006). There are two possible outcomes concerning the fate of the retro-transcribed DNA copy. It can either be integrated into a silent location (the most frequent case) where there are no regulatory elements that can promote its transcription (Cooper, 2005; Kaessmann et al., 2009). These sequences, often called retropseudogenes, are under relaxed selection and remain dormant because they lack a regulatory region, and they will most likely eventually be deleted (Cooper, 2005). Alternatively, the retro-transcribed DNA copy can be integrated near a resident functional promoter that can promote its activation (Xing et al., 2006). This latter case results in active retrogenes that, during the course of evolution, may undergo subfunctionalization and then either share function with the parents, develop a new function through a neofunctionalization process, or completely replace the parental multiple-exon gene. These retrogenes may then expand in number by duplication and/or recombination events (Altschul et al., 1997; Gentles, 1999).

Most comparative genomic studies on SEGs have focused on their inventory and relative proportions in the genome (Jorquera et al., 2016; Navarro & Galante, 2013; Sakharkar & Kangueane, 2004). Few studies have explored the evolution and functional divergence of SEGs (Grzybowska, 2012; Sakharkar et al., 2006; Shabalina et al., 2010; Zou et al., 2011), and to our knowledge, no study has investigated the adaptive roles that this type of gene may play in living organisms. We have previously demonstrated that a relatively small proportion of claudin SEGs, the largest SEG family in vertebrates, may originally have coexisted with claudin MEGs in the common ancestor of all vertebrate species (Tine et al., 2011). The claudin SEGs were likely inherited from the common ancestor of fishes and other vertebrates. Further tandem duplications may have occurred in teleost fishes, resulting in multiple copies, which may explain the greater number of claudin SEG paralogs in this lineage compared with mammals (Loh et al., 2004). We have also demonstrated that many of newly emerged teleost SEGs may have evolved new functions (Tine et al., 2011), through adaptive functional divergence of encoded proteins, as previously demonstrated in both vertebrates and invertebrates (Emerson et al., 2004; Rosso et al., 2008), where retrogenes have evolved into novel protein-coding genes with new functions (Marques et al., 2005; Rosso et al., 2008).

The SEGs encoding functional proteins are involved in various biological processes, from early developmental to mature stages and resistance to diverse stressors that organisms must deal with during their life cycle (Kaessmann et al., 2009). The regulation of the expression of genes is crucial for the accomplishment of their biological functions and is under the strict control of various regulatory elements including promoters, enhancers, and repressors (Gordon et al., 2009; Lettice et al., 2003; Vavouri et al., 2006). Promoters contain short motifs where transcription factors bind to regulate the transcription, and the sequences of many of them have been characterized in eukaryotic organisms, including fishes (Molina et al., 2001; Streelman & Kocher, 2002; Tchoudakova et al., 2001; Velan et al., 2015). Although the length and motif content of promoters, as well as their position relative to the 5′UTR, may vary considerably (up to several Kbp upstream of a gene's transcription start site, or TSS) (Placido et al., 2006), some core promoters can occur in short sequences from 100 to 300 bp upstream or downstream of the TSS (Smale & Kadonaga, 2003). Given that genes are organized into common pathways to accomplish their activity (Segal et al., 2003), those that are under the influence of common regulatory networks (i.e., genes that share the same regulatory elements, including promoter motifs) may have similar expression patterns, which implies that they may be involved in the same biological or physiological processes. Genes that arise from retroduplication need to recruit regulatory elements to be transcribed and are therefore more likely to have evolved new functions compared with genes that resulted from segmental DNA duplications (Kaessmann et al., 2009).

In a previous study, 78 SEGs (5.30% of the total gene count) were identified on three different sea bass chromosomes (Tine et al., 2011). Comparative analyses revealed that the fraction of SEGs predicted on these chromosomes is slightly higher than that found in the whole genome of other teleosts (*Takifugu rubripes*, *Tetraodon nigroviridis*, *Oryzias latipes*, *Gasterosteus aculeatus*, and *Danio rerio*). The comparison with stickleback *G. aculeatus* revealed that the count, composition, and order of SEGs varied significantly among corresponding chromosomes. Accordingly, Tine et al. (2011) proposed that these genes have continuously and independently evolved through retrotranscription followed by tandem duplications.

The main objective of the present study was to assess the proportion of SEGs found in the complete European sea bass genome, and to identify features conserved or divergent between SEG and MEG paralogs, which might give new insights into the evolution of

SEGs in teleost fishes. Identifying conserved features that may confer specific functions may facilitate a better understanding of the biological functions of SEGs. Using annotations of the sequenced genome of the European sea bass, *Dicentrarchus labrax*, we first identified all SEGs present in the genome and then described their occurrence across chromosomes. We then investigated the relationship between SEGs and other components of the genome such as chromosome size, MEG counts, gene density, and intron counts, to infer their origin. The ratio of nonsynonymous to synonymous changes was compared between SEG and MEG paralogs. More importantly, the promoter sequences of SEG and MEG paralogs were compared to identify SEGs that share the same promoter with their parental MEG. The results revealed a significant correlation between SEG and MEG counts over the genome and allowed identifying SEG/MEG paralogs that share the same promoter sequence, suggesting that they are under the influence of common regulatory networks.

2 | MATERIALS AND METHODS

2.1 | SEG inventory

The SEGs present in the *D. labrax* genome were extracted from the sea bass UCSC Genome Browser (http://seabass.mpipz.de/index.html) using the Table browser option. A single-exon filtering step (ExonCount ≤1 exon) was performed to select genes with only one exon. The nucleotide and encoded protein sequences of these genes were downloaded from the browser, together with the table containing information on exon/intron counts for each gene. Each SEG sequence was blasted against a local database comprising all genes identified as SEGs. The results were filtered to identify duplicates (SEGs with the same nucleotide sequence) and singleton genes (unique genes).

2.2 | Identification of SEGs and MEG parents

The SEG nucleotide sequences were queried against the European sea bass genome to retrieve paralog MEGs using the BLAT algorithm (Kent, 2002). The BLAT results were carefully checked and manually edited, if necessary. Two SEG loci were considered to be duplicates if the two corresponding sequences matched aligned blocks with an average length of at least 100 bp of nucleotide sequence with ≥40% identity. The same approach was used to identify the number of SEGs with an MEG parent located on the same or a different chromosome. The search for reciprocal best hits (RBH) was performed against the European sea bass genome using both nucleotide and amino acid sequence queries. An MEG was assumed to be a parent of an SEG if they matched on an average length of ≥70% with at least 40% identity. The use of these strict criteria allowed identification of SEGs and their MEG parent loci with high confidence, and to ascertain which SEGs are represented multiple times in the European sea bass genome.

2.3 | Correlations between SEGs and the other gene entities

The relationships between SEG and MEG counts in the genome were estimated using a Pearson correlation test (Pearson, 1895). The correlation between SEG content and other entities of the genome, such as chromosome size, gene density, and intron count, was also evaluated using the same correlation test. The tests were performed with the R software v.64.3.1.1 (Ihaka & Gentleman, 1996). A probability of less than 5% ($p < .05$) was considered as fiducial level of significance.

2.4 | Test for natural selection

The ratio of nonsynonymous to synonymous (d_N/d_S) substitutions, also called Ka/Ks, which is indicative of the type of natural selection acting on protein sequences, was estimated for all pairs of SEG duplicates using the Nei and Gojobori (NG) method implemented in Ka/Ks_Calculator software (Zhang et al., 2006). The tests were conducted using nucleotide sequences to determine whether singleton genes are under comparable evolutionary constraints, or whether they are subjected to different rates of evolution. Several methods have been incorporated into the Ka/Ks software calculator for the estimation of Ka/Ks ratios, which include NG (Nei & Gojobori, 1986), LPB (Molina et al., 2001; Pamilo & Bianchi, 1993), MLPB (modified LPB) (Tzeng et al., 2004), MLWL (modified LWL) (Tzeng et al., 2004), MYN (modified YN) (Zhang et al., 2006), and YN (Yang et al., 2009). The Ka/Ks_Calculator software uses a maximum-likelihood framework extended from the method of Zhang et al. (2006), which takes into account transition/transversion rate ratio and nucleotide frequencies, and incorporates these evolutionary features into a codon-based model. All these methods were tested, and the results were not significantly different between methods. Hence, only the results of the NG method were reported, and Fisher's exact test (Fisher, 1922) was used to test the significance of differences in Ka, Ks, and Ka/Ks ratios between SEG/MEG parents and between SEG/SEG paralogs.

2.5 | Identification of 3′UTR and 5′UTR, and promoter region

Sequences of the three-prime untranslated region (3′UTR) and the five-prime untranslated region (5′UTR) of SEGs and MEG parents were extracted from the BLAT results. Although the length of promoter sequences and the size of motifs can vary considerably, it has been demonstrated that core promoters can be contained in short sequences from ~100 bp upstream and downstream of the start site of the transcription. As we expected that some promoters might be present in the region 1,000 bp upstream of the transcription start site (TSS) of these genes, we extracted 1,000 bp sequences upstream of the potential promoter of SEGs and MEGs

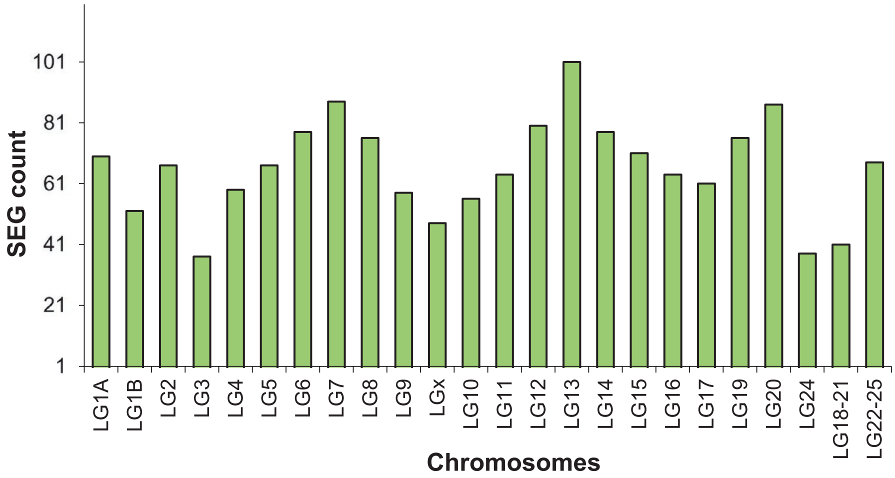

FIGURE 1 Single-exon gene counts shown for each chromosome in the genome of European sea bass

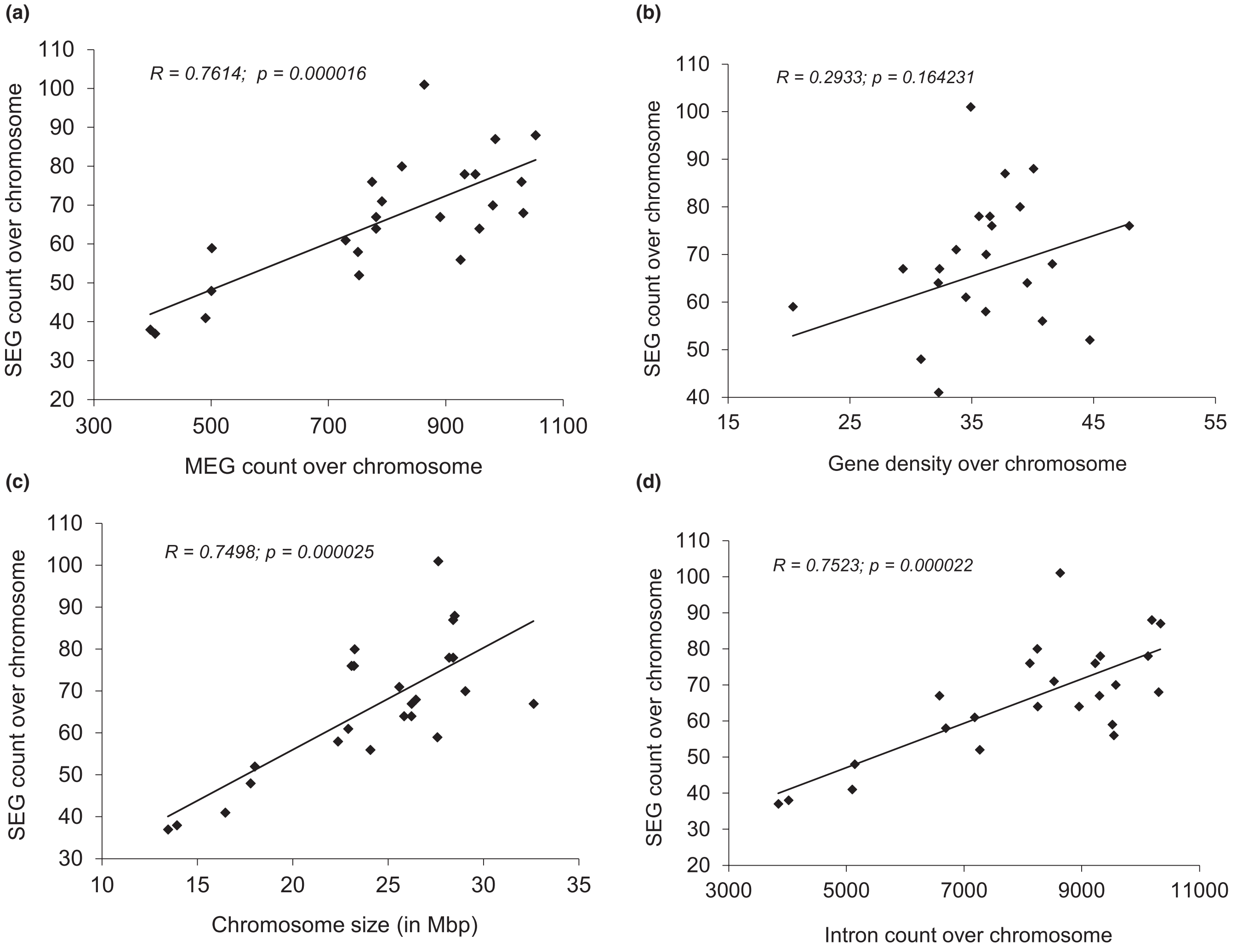

FIGURE 2 Correlation of single-exon gene (SEG) count with multiple-exon gene (MEG) count (a), gene density (b), chromosome size (c), and intron counts over chromosome (d)

from the European sea bass UCSC Genome Browser. The promoter motif characteristics of promoter regions were identified using the PROMOTER PREDICTOR software, version 2.2 (March 1999), implemented on the Berkeley Drosophila Genome Project website (http://www.fruitfly.org/seq_tools/promoter.html). This software allows promoter predictions for eukaryotic sequences and can be used for promoter prediction in teleost fishes. A score cutoff of 0.80 was used as a threshold. When motif characteristics of a promoter region were identified, they were aligned against the existing motifs in public sequences databases, and the corresponding sequences were manually compared between SEG and MEG paralogs to confirm that they are shared.

Given that the aim of these analyses was to specifically identify shared promoters between SEG and MEG paralogs, the promoter region analysis was performed only with SEG and MEG sequences where it was possible to extract the true promoter region with a high level of confidence. A conserved sequence filtering approach using an in-house script was applied to eliminate SEG and MEG pairs with erroneously identified common regulatory elements. This filtering procedure allowed to select a high-confidence subset of SEG and MEG parents with conserved promoter regions and to avoid contamination of the dataset with other types of conserved sequence. Then, an all-against-all promoter sequence comparison was performed on the first set of SEG/MEG sequences with identified promoter pairs using PROMOTER PREDICTOR, and it was determined how effectively the expected pairs were recovered. The results from the all-against-all comparison have the advantage to indicate whether the conserved regions are specific to particular promoters or whether they reflect more general signals that appear in many promoter regions.

2.6 | Functional annotation

Gene ontology terms assigned to all SEGs identified were downloaded from the European sea bass UCSC genome browser (http://seabass.mpipz.mpg.de/). This functional annotation was performed for all genes that were characterized and annotated in the European sea bass genome, including SEGs. The annotation was done using Blast2GO (https://www.blast2go.com/). Here, only the second level of GO terms (based on cellular component, biological process, and molecular function) is presented.

3 | RESULTS

3.1 | SEG counts and correlation analyses

The total number of SEGs recorded from the 24 chromosomes of the European sea bass genome was 1,585. The highest counts of SEGs were recorded on chromosomes LG13 (101), LG7 (88), and LG20 (87), followed by LG6 and LG14 (78 each), and followed by LG19 (76), LG1A (70), and LG5 (67), and the lowest numbers were observed on LG3 (37), LG24 (38), LG18-21 (41), LGx (48), and LG1B (52) (Figure 1).

The name and chromosomal location of all MEGs and SEGs used in the correlation analyses are indicated in Appendix S1. The number of SEGs on chromosomes was significantly positively correlated with the number of MEGs over chromosome ($R = 0.7614$; $p = .000016$) (Figure 2a), but no significant correlation was found between the proportion of SEGs over chromosome and gene density ($R = 0.2933$; $p = .164231$) (Figure 2b). The SEG count was significantly correlated with chromosome size ($R = 0.7498$; $p = .000025$) (Figure 2c). Likewise, there was a statistically significant positive correlation between the number of SEGs and the proportion of introns over chromosomes ($R = 0.7523$; $p = .000022$) (Figure 2d), which was strongly positively correlated with the number of genes ($R = 0.8665$; $p < .00001$).

3.2 | SEGs with parental MEGs

Of the 1,585 SEGs identified in the European sea bass genome, 312 have MEG paralogs. Based on their similarity, these MEGs can be considered to be parents of their SEG paralogs. The number of SEGs with an MEG parent located on different chromosomes was higher than the number of SEGs with an MEG parent located on the same chromosome. Chromosome LG7 has more SEGs with parental MEGs (20), whereas chromosome LG1B has fewer (5) SEGs with parental MEGs. The number of SEGs with parent MEGs was not correlated with the number of MEGs or SEGs over chromosomes. Likewise, there was no significant correlation between the number of SEGs with a parental MEG with chromosome size, intron count, or gene density.

3.3 | UTR conservation between SEGs and MEGs

Most of the SEGs identified in the European sea bass genome are without 3′UTR or 5′UTR, or even lacked both ends, which is likely an artifact of using protein sequences from other fish species that did not include UTR sequences for annotation. For most of the chromosomes, the comparison of the average length of the 3′ and 5′ ends between SEG and MEG showed that the latter have the largest 3′UTR (Figure 3a). The median length of the MEGs was also larger than that of the SEGs (Figure 3a). Given that both SEGs and MEGs were annotated using the same procedure, that is, using protein sequences from other teleosts, it is improbable that these results reflect an annotation bias indicative of a more reliable MEG annotation. By contrast, for most of the chromosomes, the average length of 5′UTR of SEGs was larger than that of MEGs, but the median of these latter was larger (Figure 3b). Among all SEGs with an MEG parent analyzed, most had a 3′UTR that was larger than the corresponding 5′UTR (Figure 4a). Likewise, for most of the chromosomes, the average length of the 3′UTR end was greater than that of the 5′UTR end (Figure 4b). However, the median length of 5′UTRs of both SEGs and MEGs was longer than that of 3′UTRs (Figure 4a,b).

3.4 | Functional categories

The Blast2GO annotation allowed classifying 797 SEGs in functional categories of cellular components. These SEGs are distributed in the following categories as follows: *integral to membrane* (43.22%), *membrane* (16.22%), *extracellular region* (5.87%), *nucleus* (5.03%), *mitochondrion* (23.64%), *intracellular part* (2.52%), *troponin complex* (2.24%),

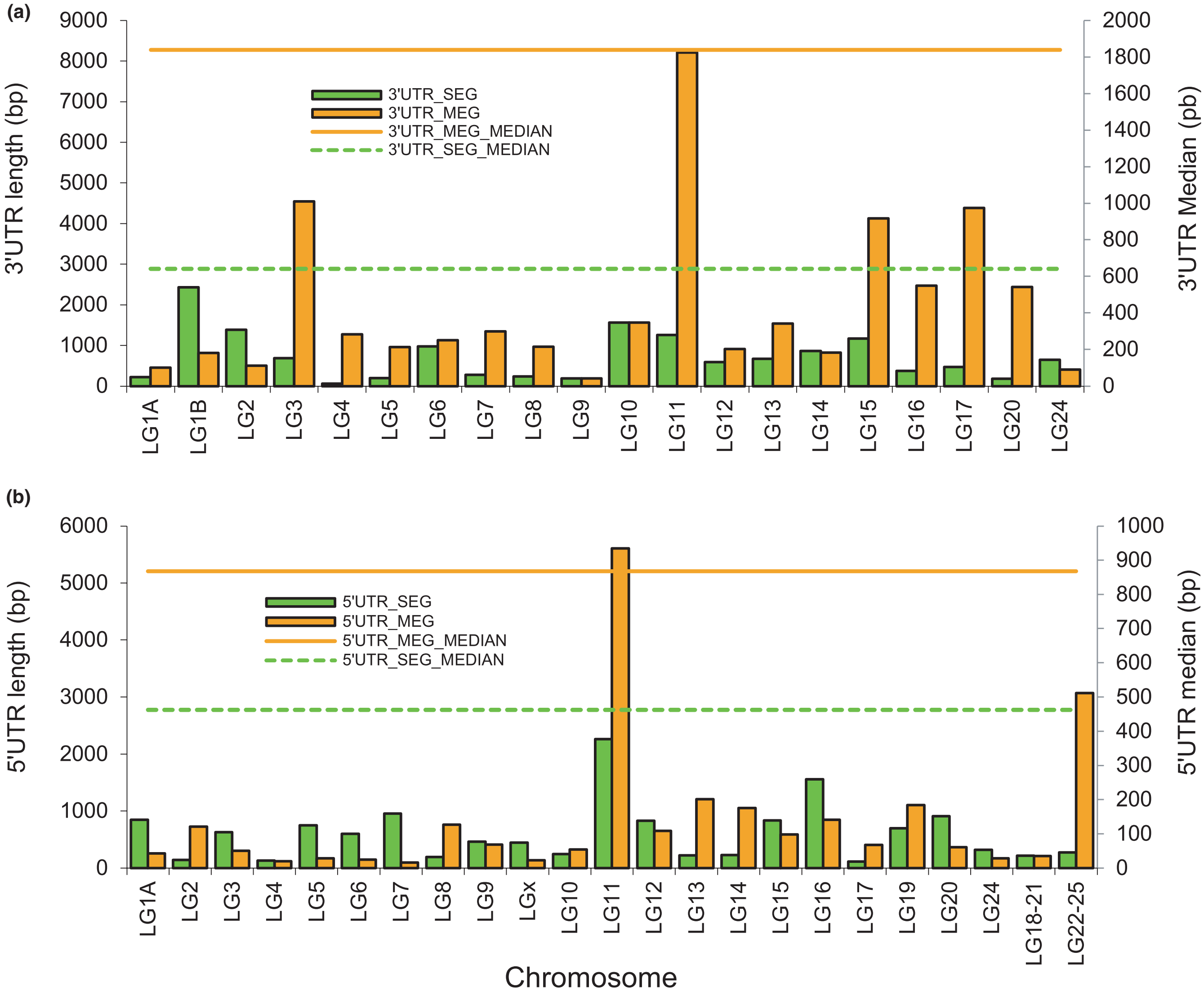

FIGURE 3 Comparison of the average length of (a) 3′UTR length between single-exon genes (SEGs) and multiple-exon genes (MEGs) and (b) 5′UTR length between SEG and MEG paralogs for each chromosome

endoplasmic reticulum (1.68%), *cell part* (1.68%), *cytoskeleton* (1.12%), and *others* (12.73%). The category *others* comprises 42 different subfunctional categories with one to two SEGs each. The GO annotation also allowed the classification of 295 SEGs in different functional categories of biological processes (FCBPs). These SEGs were assigned to 30 FCBPs, of which the most frequently represented are *regulation of apoptotic process*, *response to chemical stimulus*, *multicellular organismal development*, *termination of G-protein-coupled receptor signaling pathway*, *cell–cell signaling*, and *defense response to bacterium* (2%–2.80% each), and *cell redox homeostasis*, *response to stress*, *single-organism cellular process* (3.15% each), *cell process* (4.90%), *single-organism process* (4.90%), and *signal transduction* (5.59%). Other, less frequently represented (1%–1.75% each) FCBPs include *autophagy*, *hemopoiesis*, *nervous system development*, *spermatogenesis*, *cell wall macromolecular catabolic process*, *intracellular signal transduction*, *lipid metabolic process*, *phosphorylation*, and *others*. The group *others* includes 82 categories comprising only one SEG each (25.52%). Genes annotated with the GO biological process term for *cell process*, *single-organism process*, and signal transduction are found predominantly in a subset of 295 SEGs. Finally, 1,524 SEGs could be classified into different functional categories of molecular function. The most important of these are as follows: *protein binding* (60.76%), *GTP binding* (4.82%), *carbohydrate binding* (4.55%), *zinc ion binding* (4.27%), *nucleic acid binding* (2.93%), *ATP binding* (2.93%), *calcium ion binding* (2.58%), *DNA binding* (2.31%), *hydrolase activity* (1.55%), and *others* (51 categories) (11.26%). Genes taking part in protein binding activities are found predominantly in this set of 1,524 SEGs that could be classified into molecular function categories.

The GO annotation of SEGs with an MEG parent revealed that 63.02% of them have a protein name, whereas 31.16% were not annotated (referred as NA in Appendix S1). The remaining SEGs with an MEG parent (5.82%) are annotated as uncharacterized or unnamed proteins. The 63.02% of SEGs with a protein name belong to several gene families, including leucine-rich repeat protein (3.15%), claudin (2.64%), G-protein-coupled receptor (2.54%), e3 ubiquitin–protein

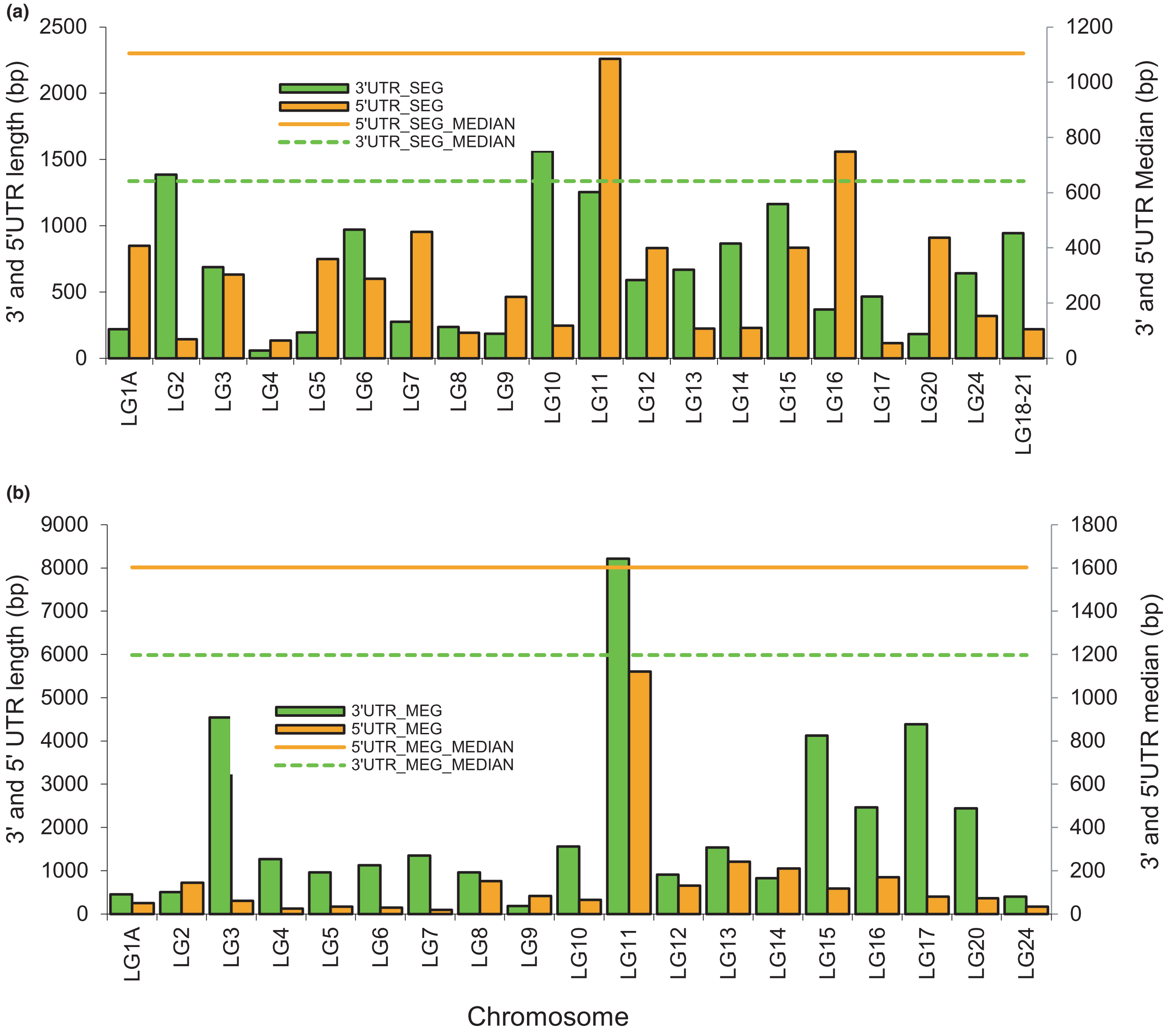

FIGURE 4 Comparison of the average length of (a) 3′UTR and 5′UTRs between single-exon genes (SEGs) with multiple-exon genes (MEG) and (b) 3′UTR and 5′UTR length of SEGs and MEGs over chromosome paralogs

ligase (2.13%), forkhead box protein (2.13%), and reverse transcriptase protein (2.13%). Besides these well-represented gene families, there were other, less common gene families, which include odorant receptor (1.93%), transcription factor (1.83%), transposase (1.83%), transmembrane protein (1.73%), transposable element tc1 and tcb2 transposase (1.73,%), gap junction alpha, beta and gamma (1.62%), tripartite motif-containing protein (1.62%), and zinc finger protein (1.62%). We also identified other well-known gene families including histone protein (0.91%), nuclear factor ovary-like (0.91%), ion channels (0.91%), c-c chemokine receptor type 11-like (0.81%), interferon-induced very large GTPase 1 (0.51%), and fibroblast growth factor-binding protein 1 (0.41%). Heat shock proteins including heat shock protein 30 and heat shock protein 70 as well as transcriptional activator protein pur-alpha and pur-beta are also represented (0.30% each).

3.5 | Comparison of potential promoters between SEGs and MEGs

Of the 312 SEG and MEG pairs analyzed, 34 have very similar sequence 1,000 bases upstream of the TSS, which might contain the promoter, whereas 278 others did not show similarity in the upstream sequence. The similarity search for sequences characteristic of promoter regions allowed the identification of 15 promoter motifs with a cutoff score of 0.80 (Appendix S2). The average size of the motifs found was 50 bp, including a ~10 bp upper bound of the transcription start binding site (Appendix S2).

3.6 | Natural selection

The ratios of nonsynonymous to synonymous (Ka/Ks) substitutions could be estimated for 110 SEG/MEG pairs of the 312 SEGs with an MEG parent. The Ka/Ks could not be estimated for the 201 remaining SEG/MEG parents because one of the paralogs lacked the translation start codon required by the Ka/Ks_Calculator program, which might be due to an annotation bias. The average ratio of nonsynonymous to synonymous changes (Ka/Ks) for all SEG/MEG pairs was 1.40 ($p = .0480$), which is indicative of positive selection. The comparison of Ka/Ks between 110 SEG/MEG parents showed that 74 of them are under positive selection (Ka/Ks > 1; $p = .0447$) (Appendix S1), whereas the remaining 36 were not (Ka/Ks < 1). SEG/MEG pairs with a Ka/Ks > 1 include transposase, reverse transcriptase-like protein, nuclear factor ovary-like, immunoglobulin light chain precursor, c-c chemokine receptor type 4-like, and transposable element/transposase and sry (sex-determining region y)-box 4 (Appendix S1). Few of them are not annotated genes or were annotated as uncharacterized protein, which means that the corresponding protein is not present in Gene Ontology repertory.

The average Ka/Ks ratio estimated for 33 SEG/SEG paralogs was 1.01. Of these 33 SEG duplicates for which the Ka/Ks could be estimated, 16 are under positive selection (Ka/Ks > 1; $p = .0307$), whereas the 17 remaining paralog pairs were not (all Ka/Ks < 1).

4 | DISCUSSION

The main objective of the current study was to identify features conserved or divergent between SEG and MEG paralogs, which may confer a specific function. This may improve our understanding of the biological roles of this type of gene, which has long been considered to be marginal and dysfunctional. The proportion of SEGs in the European sea bass genome was accessed at both chromosomal and genome levels. The results showed significant correlations of SEG count with the proportion of MEGs, chromosome size, and intron count, but no significant correlation with gene density was found. All these correlations indicate that SEGs are evenly spread over the genome. The results also showed that SEG order and composition varied among corresponding chromosomes. The SEG fraction on a particular chromosome is also correlated with the chromosome's total gene content, which suggests that SEGs are distributed across the genome. Functional annotation by gene ontology indicated that SEGs code for a variety of protein families, including leucine-rich repeat protein, claudins, forkhead box protein, olfactory receptors, histones, and ion channels, all of which are essential for various biological functions. In addition, several ion channels were identified, including potassium voltage-gated channel and potassium sodium channels, which play important roles in hydromineral balance. Likewise, heat shock proteins, including heat shock protein 30 and heat shock protein 70, which are involved in responses to environmental constraints (Currie et al., 2000; Tine et al., 2010; Oksala et al., 2014), were also identified. The GO functional annotation indicated that a significant number of SEGs belong to different functional categories of cellular component, biological process, and molecular function. The GO results indicated that 295 SEGs are involved in important biological pathways. This finding is supported by the transcriptomic results from the RNA-seq data used to support the European sea bass genome annotation, which indicates that of the 1,587 SEGs identified, 1,234 are expressed, suggesting that they are functional. However, given that SEGs result from the reverse transcription of mRNA from a "parental" gene into cDNA and its insertion elsewhere in the genome, the transcriptomic data were unsuitable to differentiate between SEGs and their MEG parents. The RNA-seq data were not specially produced to compare the expression profiles. For that reason, they could not be used to distinguish any SEG that is highly expressed from its parental MEG. Such information could provide strong evidence of SEG functionality compared with their MEG parents. Further transcriptomic analyses using real-time PCR or RNA-seq, specially designed to compare expression profiles, are required to identify differentially expressed SEG/MEG parents.

Many features that contribute to the stability of mRNA, as well as its translation and regulation, are located within untranslated regions (UTRs). The 3′UTR that is located downstream from the coding region is not translated into protein but contains several regulatory elements, including polyA adenylation signals and binding sites for micro-RNAs (Tuller et al., 2009). The 5′UTR at the upstream region also harbors regulatory elements such as sequences functioning as binding sites for regulatory proteins that may affect mRNA regulation and its stability (Lin & Li, 2012; Tuller et al., 2009). Also, the presence of secondary structures, upstream start codon (AUG), and open reading frames (ORFs) in the 5′UTR region affect the overall gene transcription (Mignone et al., 2002; Tuller et al., 2009). The results of this study show that for most of the chromosomes, the average length of the 3′UTR end was overall longer than that of the 5′UTR end for both MEGs and SEGs, in agreement with previous observations that 3′UTRs in metazoans are much longer than 5′UTRs (200–800 nucleotides and 100–200 nucleotides, respectively (Mignone et al., 2002). These results may reflect an important role of 3′UTRs in the regulation of gene expression, especially at the translational level. It has been found that 3′UTR on average is longer and has evolved faster in cichlids compared with other teleosts, which might be due to their meta-regulation and regulation roles in post-transcriptional regulation mechanisms (Xiong et al., 2018). The present study also shows that the average length of the 5′UTR end of the SEGs was much longer than that of the MEGs, whereas the latter have longer 3′UTRs. These differences in 5′UTR and 3′UTR lengths might contribute to gene expression divergence between SEGs and their parental MEGs, and therefore lead to sub- or neo-functionalization of new SEGs. They may reflect the involvement of 5′UTR in mechanisms governing transcriptional regulation. Indeed, the longer 5′UTRs of SEGs may reflect lower translation rates compared with MEGs, in agreement with previous findings that mRNAs with high translation rates often contain short 5′UTRs (Larsen & Michael, 2014).

The promoter sequences are less conserved than the coding regions (Chiba et al., 2008; Hemberg et al., 2012), which implies that high similarity of promoter regions between SEGs and their MEG parents could be indicative of their involvement in common regulatory networks. In this study, the comparison of 1,000 bp upstream sequence, potentially containing the promoter motifs, indicated that 34 SEG/MEG parents share high similarity. The search for conserved motifs in this region indicated that 15 SEG promoter sequences have an equivalent with strongly conserved motif signals in the same genome. By contrast, the comparison of regions potentially harboring promoter sequences of the 312 MEGs with SEG paralogs failed to identify MEGs that share common promoters with them. These results suggest that promoter sequences evolve in a manner that is closely linked to the genes they control. It also implies that there is little or no identifiable promoter similarity between more distantly related genes. Overall, these results indicate that gene retrotransposition, which is presumably followed by an insertion of a retrotranscript elsewhere in the genome, is likely accompanied by substantial changes in the promoters. This is interesting but not entirely surprising since following retrotransposition, it is more likely that an active retrogene is inserted next to the parental gene with which it can share a common promoter (Fablet et al., 2009). This is consistent with the observation that most SEGs with a common MEG promoter identified in this study are nested genes.

The SEG/MEG pairs that share motif characteristics of functional promoter regions are probably under the influence of common regulatory networks. It has been demonstrated that paralog genes that display high similarities in their promoter regions are likely to be involved in the same physiological pathways (Huning & Kunkel, 2020). The results of the current study thus support previous findings that SEGs in eukaryotic organisms are genes that are just as functional as are MEGs, thus providing strong evidence that they may play crucial roles in the genome. Indeed, depending on their occurrence and frequency, retrogenes may contribute considerably to the diversification of genomes and may therefore be responsible for the emergence of species-specific features (Kubiak & Makałowska 2017). It can be, therefore, speculated that retrogenes might be involved in specific adaptive processes in many organisms, including teleosts. The alignments produced by the method used, which combined two different approaches, reflect *bona fide* functional sequence rather than background synteny. However, the low number of SEG/MEG parents with common promoter motifs found in this study indicates that few SEGs in the genome are under the influence of the same regulatory networks. This can also be explained by a failure of the approach used to identify promoter motifs. It is possible that the promoters of some SEG/MEG parents are by chance not contained in the 1,000 bp upstream of the coding sequence. It has been demonstrated that some promoters occur several kb upstream of the transcription factor-binding sites (Khambata-Ford et al., 2003). By extending the search several kb upstream of the coding sequences, it may be possible to find additional SEG and MEG parents that share common promoter motifs. Although this study allowed identifying SEG and MEG parents that share common promoter motifs, further experimental evidence is needed to confirm that they are under the regulatory control of the same promoter.

A significant number of SEGs identified in the European sea bass genome have MEG paralogs either on the same or on different chromosomes, suggesting that these might originate from LINE1 retrotransposon-mediated reverse transcription of mRNA from “parental” source genes (Brosius, 1991; Long, 2002). Given that the probability for an inserted retrocopy to meet a functional regulatory element that could promote its transcription is higher in genomes with higher gene density, it can be expected that genomic regions with more coding sequences harbor more potentially functional SEGs. However, the results of this study revealed that the fraction of SEGs is not significantly correlated with gene density at the chromosomal level, in disagreement with the above assumption. This may be explained by the fact that the SEG content in the genome is also dependent on L1 element activity (Seleme et al., 2006), which is related to the chromatin status in which these elements are located in the genome.

The fraction of SEGs found in the European sea bass genome in this study showed significant differences between chromosomes. If retrotransposition is the main mechanism generating SEGs, these results may indicate that differences in the frequency of gene retrotransposition activity may have occurred between chromosomes. The proportion of SEGs in the European sea bass genome was slightly higher than previously reported in other teleost fishes (Tine et al., 2011), which may be explained by the fact that most of SEGs found on different chromosomes in this study were not considered to be duplicates, but of different origin and/or location. The evolutionary analyses based on the estimation of nonsynonymous to synonymous ratios showed that 33 SEG/SEG paralogs have an average Ka/Ks of 1.01 (p = .0307), suggesting that they are under positive selection. These SEG/SEG paralogs under positive selection may have different functions, despite not being sufficiently divergent at the nucleotide sequence level. A large proportion of these SEGs of European sea bass are nested genes (i.e., they overlap with the MEG parent), which may explain why they are under positive selection. These nested SEGs have the same gene name as the host MEG paralog (Appendix S2) and may have originated from retroduplication from their host MEGs. It has been demonstrated that duplicated paralogs evolve faster than paralogs with similar levels of divergence and similar function (Kondrashov et al., 2002). The proportion of SEGs (8.3%) found in the European sea bass genome is lower than the fraction reported in other vertebrates, including human (12.3%) and mouse (15.8%) (Sakharkar et al., 2006) and in plants (rice: 19.9%; *Arabidopsis*: 21.7%) (Jain et al., 2008), which may reflect less retrotransposition activity in fish genomes compared with other vertebrates such as human, chimp, dog, cow, rat, and mouse, as previously demonstrated for *Tetraodon nigroviridis* (Yu et al., 2007). About 1% of the genome of this species consists of retrotransposons (Roest et al., 2000), implying a lower frequency of retrotransposition events in its genome. This might be a feature common to teleost fishes, as is evident from the lower fraction of SEGs found in this lineage compared with those previously reported in vertebrates and plants.

5 | CONCLUSION

This study showed that a large proportion (8.3%) of the European sea bass genes are SEGs. The proportion of SEGs in the European sea bass genome is highly correlated with the number of MEGs, suggesting that SEGs are continually being created by retrotransposition events. This is supported by the significant number of SEGs with a parental MEG in the genome. A significant number of SEGs showed high similarity in the promoter region with their MEG paralogs, which implies that both have the same biological function. The present study is the first to illustrate that some SEGs have conserved features in their promoter regions that are shared with their MEG parents, while others did not conserve motif characteristics of regulatory regions with their MEG parents. The results suggest that certain SEGs have evolved new functions after their genesis by natural selection that has acted on their promoter regions, especially for those with promoter sequences that are dissimilar from the promoter sequences of their potential MEG parents.

ACKNOWLEDGMENTS

This project has been financially supported by the Max Planck Institute for Molecular Genetics (Berlin, Germany) and the Genome Centre at MPI Cologne, Germany. Financial support was provided by Gaston Berger University. We thank Dr. Bruno Louro for his comments on the manuscript. We also thank the Max Planck Institute for Molecular Genetics (Berlin, Germany) and the Genome Centre at MPI Cologne (Germany) for allowing us to use the European sea bass genome resources.

CONFLICT OF INTEREST

Eventual conflicts of interest (including personal communications or additional permissions, related manuscripts), sources of financial support, corporate involvement, and patent holdings are disclosed.

AUTHOR CONTRIBUTIONS

Mbaye Tine: Investigation (lead); writing-original draft (lead). **Heiner Kuhl:** Formal analysis (supporting); methodology (supporting); supervision (supporting); writing-review & editing (supporting). **Peter R. Teske:** Formal analysis (supporting); writing-review & editing (supporting). **Richard Reinhardt:** Funding acquisition (lead); project administration (lead); supervision (equal).

DATA AVAILABILITY STATEMENT

All datasets supporting the results of this article are included within the article and its Additional Files. This study was based on European sea bass genome resource data available at the following link: http://seabass.mpipz.de/index.html?org=European+seabass&db=dicLab1&hgsid=1895.

ORCID

Mbaye Tine https://orcid.org/0000-0003-2427-1027
Heiner Kuhl https://orcid.org/0000-0001-7623-9227
Peter R. Teske https://orcid.org/0000-0002-2838-7804
Richard Reinhardt https://orcid.org/0000-0001-9376-2132

REFERENCES

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 Research*, *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Brosius, J. (1991). Retroposons – Seeds of evolution. *Science*, *251*, 753. https://doi.org/10.1126/science.1990437

Chiba, H., Yamashita, R., Kinoshita, K., & Nakai, K. (2008). Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs. *BMC Genomics*, *9*, 152. https://doi.org/10.1186/1471-2164-9-152

Cooper, D. N. (2005). *Pseudogenes and their evolution. Encyclopedia of life sciences*. John Wiley & Sons, Ltd. http://www.els.net

Cordaux, R., & Batzer, M. A. (2008). *Evolutionary emergence of genes through retrotransposition. Encyclopedia of life sciences*. John Wiley & Sons, Ltd. http://www.els.net

Currie, S., Moyes, C. D., & Tufts, B. L. (2000). The effects of heat shock and acclimation temperature on hsp70 and hsp30 mRNA expression in rainbow trout: In vivo and in vitro comparisons. *Journal of Fish Biology*, *56*, 398–408.

Doenecke, D., & Albig, W. (2005). *Intronless genes. Encyclopedia of life sciences*. John Wiley & Sons, Ltd. http://www.els.net

Emerson, J. J., Kaessmann, H., Betrán, E., & Long, M. (2004). Extensive gene traffic on the mammalian X chromosome. *Science*, *303*, 537–540. https://doi.org/10.1126/science.1090042

Fablet, M., Bueno, M., Potrzebowski, L., & Kaessmann, H. (2009). Evolutionary origin and functions of retrogene introns. *Molecular Biology and Evolution*, *26*, 2147–2156. https://doi.org/10.1093/molbev/msp125

Fisher, R. A. (1922). On the interpretation of X2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*, 87–94.

Gentles, A. J., & Karlin, S. (1999). Why are human G-protein-coupled receptors predominantly intronless? *Trends in Genetics*, *15*, 47–49. https://doi.org/10.1016/S0168-9525(98)01648-5

Gordon, C. T., Tan, T. Y., Benko, S., Fitzpatrick, D., Lyonnet, S., & Farlie, P. G. (2009). Long-range regulation at the SOX9 locus in development and disease. *Journal of Medical Genetics*, *46*, 649–656. https://doi.org/10.1136/jmg.2009.068361

Grzybowska, E. A. (2012). Human intronless genes: Functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochemical and Biophysical Research Communications*, *424*, 1–6.

Hemberg, M., Gray, J. M., Cloonan, N., Kuersten, S., Grimmond, S., Greenberg, M. E., & Kreiman, G. (2012). Integrated genome analysis suggests that most conserved non-coding sequences are regulatory factor binding sites. *Nucleic Acids Research*, *40*, 7858–7869. https://doi.org/10.1093/nar/gks477

Huning, L., & Kunkel, G. R. (2020). Two paralogous *znf143* genes in zebrafish encode transcriptional activator proteins with similar functions but expressed at different levels during early development. *BMC Molecular and Cell Biology*, *21*, 3. https://doi.org/10.1186/s12860-020-0247-7

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299–314.

Jain, M., Khurana, P., Tyagi, A. K., & Khurana, J. P. (2008). Genome-wide analysis of intronless genes in rice and *Arabidopsis*. *Functional and Integrative Genomics*, *8*, 69–78. https://doi.org/10.1007/s10142-007-0052-9

Jorquera, R., Ortiz, R., Ossandon, F., Cardenas, J. P., Sepulveda, R., Gonzalez, C., & Holmes, D. S. (2016). SinEx DB: A database for single exon coding sequences in mammalian genomes. *Database*, *2016*, 1–8. https://doi.org/10.1093/database/baw095

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: Mechanistic and evolutionary insights. *Nature Reviews Genetics*, *10*, 19–31. https://doi.org/10.1038/nrg2487

Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research*, *12*, 656–664. https://doi.org/10.1101/gr.229202

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, *3*, research0008.1. https://doi.org/10.1186/gb-2002-3-2-research0008

Khambata-Ford, S., Liu, Y., Gleason, C., Dickson, M., Altman, R. B., Batzoglou, S., & Myers, R. M. (2003). Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Research*, *13*, 1765–1774. https://doi.org/10.1101/gr.529803

Kubiak, M. R., & Makałowska, I. (2017). Protein-Coding genes' retrocopies and their functions. *Viruses*, *9*, 80. https://doi.org/10.3390/v9040080

Larsen, C. A., & Michael, T. (2014). Howard Conserved regions of the DMD 3′ UTR regulate translation and mRNA abundance in cultured myotubes. *Neuromuscular Disorders*, *24*, 693–706. https://doi.org/10.1016/j.nmd.2014.05.006

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., & de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*, 1725–1735. https://doi.org/10.1093/hmg/ddg180

Lin, Z., & Li, W.-H. (2012). Evolution of 5′ untranslated region length and gene expression reprogramming in yeasts. *Molecular Biology and Evolution*, *29*, 81–89. https://doi.org/10.1093/molbev/msr143

Loh, Y. H., Christoffels, A., Brenner, S., Hunziker, W., & Venkatesh, B. (2004). Extensive expansion of the claudin gene family in the teleost fish, *Fugu rubripes*. *Genome Research*, *14*, 1248–1257. https://doi.org/10.1101/gr.2400004

Long, M. (2002). *Protein coding segments: Evolution of exon-intron gene structure. Encyclopaedia of life science* (pp. 1–6). Macmillan Reference Ltd.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A., & Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. *PLoS Biology*, *3*, 1970–1979. https://doi.org/10.1371/journal.pbio.0030357

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology*, *3*, REVIEWS0004.

Molina, A., Iyengar, A., Marins, L. F., Biemar, F., Hanley, S., Maclean, N., Smith, T. J., Martial, J. A., & Muller, M. (2001). Gene structure and promoter function of a teleost ribosomal protein: A tilapia (*Oreochromis mossambicus*) L18 gene. *Biochimica et Biophysica Acta*, *1520*, 195–202. https://doi.org/10.1016/S0167-4781(01)00272-X

Navarro, F. C., & Galante, P. A. (2013). RC Pedia: A database of retrocopied genes. *Bioinformatics*, *29*, 1235–1237. https://doi.org/10.1093/bioinformatics/btt104

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*, 418–426.

Oksala, N. K., Ekmekçi, F. G., Özsoy, E., Kirankaya, Ş., Kokkola, T., Emecen, G., Lappalainen, J., … Atalay, M. (2014). Natural thermal adaptation increases heat shock protein levels and decreases oxidative stress. *Redox Biology*, *3*, 25–28. https://doi.org/10.1016/j.redox.2014.10.003

Ostertag, E. M., & Kazazian, H. H. J. (2001). Biology of mammalian L1 retrotransposons. *Annual Review of Genetics*, *35*, 501–538. https://doi.org/10.1146/annurev.genet.35.102401.091032

Pamilo, P., & Bianchi, N. O. (1993). Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes. *Molecular Biology and Evolution*, *10*, 271–281.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Placido, A., Damiano, F., Sciancalepore, M., De Benedetto, C., Rainaldi, G., & Gallerani, R. (2006). Comparison of promoters controlling on the sunflower mitochondrial genome the transcription of two copies of the same native trnK gene reveals some differences in their structure. *Biochimica et Biophysica Acta*, *1757*, 1207–1216. https://doi.org/10.1016/j.bbabio.2006.05.014

Roest, C. H., Jaillon, O., Dasilva, C., Ozouf-Costaz, C., Fizames, C., Fischer, C., Bouneau, L., Billault, A., Quetier, F., Saurin, W., Bernot, A., & Weissenbach, J. (2000). Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Research*, *10*, 939–949. https://doi.org/10.1101/gr.10.7.939

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., & Koonin, E. V. (2005). Analysis of evolution of exon–intron structure of eukaryotic genes. *Henry Stewart Publications*, *Briefings in Bioinformatics*, *6*, 118–134. https://doi.org/10.1093/bib/6.2.118

Rosso, L., Marques, A. C., Weier, M., Lambert, N., Lambot, M.-A., Vanderhaeghen, P., & Kaessmann, H. (2008). Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. *PLoS Biology*, *6*, 1281–1291. https://doi.org/10.1371/journal.pbio.0060140

Sakharkar, K. R., Sakharkar, M. K., Culiat, C. T., Chowd, V. T., & Pervaiz, S. (2006). Functional and evolutionary analyses on expressed intronless genes in the mouse genome. *FEBS Letter*, *580*, 1472–1478. https://doi.org/10.1016/j.febslet.2006.01.070

Sakharkar, M. K., & Kangueane, P. (2004). Genome SEGE: A database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics*, *5*, 67.

Savisaar, R., & Hurst, L. D. (2016). Purifying selection on exonic splice enhancers in intronless genes. *Molecular Biology and Evolution*, *33*, 1396–1418. https://doi.org/10.1093/molbev/msw018

Segal, E., Wang, H., & Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, *19*, i264–i272. https://doi.org/10.1093/bioinformatics/btg1037

Seleme, M. D. C., Vetter, M. R., Cordaux, R., Bastone, L., Batzer, M. A., & Kazazian, H. H. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *PNAS*, *103*, 6611–6616. https://doi.org/10.1073/pnas.0601324103

Shabalina, S. A., Ogurtsov, A. Y., Spiridonov, A. N., Novichkov, P. S., Spiridonov, N. A., & Koonin, E. V. (2010). Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. *Molecular Biology and Evolution*, *27*, 1745–1749. https://doi.org/10.1093/molbev/msq086

Smale, T., & Kadonaga, T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, *72*, 449–479. https://doi.org/10.1146/annurev.biochem.72.121801.161520

Smith, M. W. (1988). Structure of vertebrate genes: A statistical analysis implicating selection. *Journal of Molecular Evolution*, *27*, 45–55. https://doi.org/10.1007/BF02099729

Streelman, J. T., & Kocher, T. D. (2002). Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiological Genomics*, *9*, 1–4. https://doi.org/10.1152/physiolgenomics.00105.2001

Tay, S. K., Blythe, J., & Lipovich, L. (2009). Global discovery of primate-specific genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 12019–12024. https://doi.org/10.1073/pnas.0904569106

Tchoudakova, A., Kishida, M., Wood, E., & Callard, G. V. (2001). Promoter characteristics of two cyp19 genes differentially expressed in the brain and ovary of teleost fish. *The Journal of Steroid Biochemistry and Molecular Biology*, *78*, 427–439. https://doi.org/10.1016/S0960-0760(01)00120-0

Tine, M., Bonhomme, F., McKenzie, D. J., & Durand, J. D. (2010). Differential expression of the heat shock protein Hsp70 in natural populations of the tilapia, *Sarotherodon melanotheron*, acclimatised to a range of environmental salinities. *BMC Ecology*, *10*, 11. https://doi.org/10.1186/1472-6785-10-11

Tine, M., Heiner, K. H., Beck, A., Bargelloni, L., & Reinhardt, R. (2011). Comparative analysis of intronless genes in teleost fish genomes: Insights into their evolution and molecular function. *Marine Genomics*, *4*, 109–119. https://doi.org/10.1016/j.margen.2011.03.004

Tuller, T., Ruppin, E., & Kupiec, M. (2009). Properties of untranslated regions of the *S. cerevisiae* genome. *BMC Genomics*, *10*, 391. https://doi.org/10.1186/1471-2164-10-391

Tzeng, Y.-H., Pan, R., & Li, W.-H. (2004). Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *21*, 2290–2298. https://doi.org/10.1093/molbev/msh242

Vavouri, T., McEwen, G. K., Woolfe, A., Gilks, W. R., & Elgar, G. (2006). Defining a genomicra dius for long-range enhancer action: Duplicated conserved non-coding elements hold the key. *Trends in Genetics*, *22*, 5–10. https://doi.org/10.1016/j.tig.2005.10.005

Velan, A., Hulata, G., Ron, M., Slosman, T., Shirak, A., & Cnaani, A. (2015). Association between polymorphism in the Prolactin I promoter and growth of tilapia in saline-water. *Aquaculture Reports*, *1*, 5–9. https://doi.org/10.1016/j.aqrep.2015.03.001

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The sequence of the human genome. *Science*, *292*, 1304–1351. https://doi.org/10.1126/science.1058040

Xing, J., Wang, H., Belancio, V. P., Cordaux, R., Deininger, P. L., & Batzer, M. A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. *PNAS*, *103*, 17608–17613. https://doi.org/10.1073/pnas.0603224103

Xiong, P., Hulsey, C. D., Meyer, A., & Franchini, P. (2018). Evolutionary divergence of 3' UTRs in cichlid fishes. *BMC Genomics*, *19*, 433. https://doi.org/10.1186/s12864-018-4821-8

Yang, X., Jawdy, S., Tschaplinski, T. J., & Tuskan, G. A. (2009). Genome- wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*. *Genomics*, *93*, 473–480. https://doi.org/10.1016/j.ygeno.2009.01.002

Yu, Z., Morais, D., Ivanga, M., & Harrison, P. M. (2007). Analysis of the role of retrotransposition in gene evolution in vertebrates. *BMC Bioinformatics*, *8*, 308. https://doi.org/10.1186/1471-2105-8-308

Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Wong, G. K., & Yu, J. (2006). KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*, *4*, 259–263.

Zou, M., Guo, B., & He, S. (2011). The roles and evolutionary patterns of intronless genes in deuterostomes. *Comparative and Functional Genomics*, ID 680673, 1–8. https://doi.org/10.1155/2011/680673

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Tine M, Kuhl H, Teske PR, Reinhardt R. Genome-wide analysis of European sea bass provides insights into the evolution and functions of single-exon genes. *Ecol Evol*. 2021;11:6546–6557. https://doi.org/10.1002/ece3.7507