Extreme Positive Selection on a New Highly-Expressed Larval Glycoprotein (LGP) Gene in *Galaxias* Fishes (Osmeriformes: Galaxiidae)

Lise J. Wallis[1] and Graham P. Wallis*,[1]

[1]Department of Zoology, University of Otago, Dunedin, New Zealand

***Corresponding author:** E-mail: g.wallis@otago.ac.nz.

Associate editor: Willie Swanson

Abstract

We describe the intron–exon structure and DNA/protein sequences of a new larval glycoprotein (LGP) gene from nine species of galaxiid fish. The gene has a distant similarity to *Danio* THP (Tamm–Horsfall urinary glycoprotein; uromodulin) and cichlid SPP120 (seminal plasma glycoprotein) due to conserved features of its zona pellucida (ZP) domain, including eight highly conserved cysteines and a consensus furin cleavage site. Using a combination of 454 sequencing of cDNA and exon-primed intron-spanning sequencing of genomic DNA, we obtained full sequences of the coding region (996 bp) and its intervening sequences (1,459 bp). *LGP* shows an exceptionally strong signal of positive selection over the entire coding region, as evidenced by d_N/d_S values >1. Across nine species of *Galaxias*, 87/332 (26%) amino acid residues are variable, compared with 9/386 (2%) for mitochondrial cytochrome *b* (cyt*b*) in the same group of species. Across 36 interspecific pairwise comparisons, genetic distances are in all cases larger for coding region than for introns, by a factor of 2.4-fold on average. Reading frame, gene structure, splice sites, and many ZP motifs are conserved across all species. Together with the fact that the gene is expressed in all species, these results argue clearly against the possibility of a pseudogene. We show by 454 sequencing and quantitative polymerase chain reaction that the transcript is abundant (ca. 0.5%) in newly hatched larvae and appears to be almost absent from a range of adult tissues. We postulate that the strong Darwinian evolution exhibited by this protein may reflect some type of immunoprotection at this vulnerable larval stage.

Key words: Darwinian selection, *Galaxias*, SPP120, Tamm–Horsfall, zona pellucida.

Introduction

"What is the nature of genetic variation for fitness in a population?" (Lewontin 1974) remains a central question in evolutionary genetics (Ellegren and Sheldon 2008). Because of the difficulties in measuring selection directly, many indirect methods have been employed (Barton et al. 2007). Fruitful among these have been those that involve interspecific comparison of orthologous sequences, as they integrate a large number of DNA substitutions over time (Kimura 1983). These include d_N/d_S (or k_A/k_S) ratio (Li and Gojobori 1983), k_R/k_C ratio (Hughes et al. 1990; Hanada et al. 2007), and the McDonald–Kreitman tests (McDonald and Kreitman 1991). These tests facilitate quick screens for selection across entire transcriptomes in order to identify genes that have experienced various types of selection (Fay et al. 2002; Plotkin et al. 2004; Bustamante et al. 2005; Mikkelsen et al. 2005).

The d_N/d_S ratio test tests for a significant deviation from unity (Nei and Kumar 2000; Kryazhimskiy and Plotkin 2008). Purifying selection is evidenced by a value of <1 and is the most common result (Li 1997). Conversely, a value of >1 is evidence of positive selection to the extent that it causes amino acid replacement substitutions at nonsilent sites to exceed synonymous substitution at silent sites. It is thus a particularly stringent test for positive selection, particularly when applied over the length of an exon or entire coding region of a gene. Despite the demanding nature of the test when used to detect positive selection, a number of genes have provided evidence for Darwinian selection using this approach (Endo et al. 1996; Swanson and Vacquier 2002). These include especially genes associated with reproduction (Swanson et al. 2001; Swanson and Vacquier 2002; Civetta 2003; Berlin et al. 2008; Clark et al. 2009; Dorus et al. 2010; Morgan et al. 2010), immunity (Hughes et al. 1990; Endo et al. 1996), neurotoxins (Duda and Palumbi 1999; Weinberger et al. 2010), vision (Spady et al. 2005; Larmuseau et al. 2010), apoptosis (Bustamante et al. 2005), transcription factors (Bustamante et al. 2005), and genes whose function has changed (Stewart et al. 1988). In marine invertebrates, fast evolution of egg–sperm binding proteins has been used as evidence for reproductive character displacement (reinforcement) (Geyer and Palumbi 2003). This is an intriguing case because we would normally expect strong positive frequency–dependent selection against new variants in proteins involved in fundamental sperm–egg recognition. Recent work suggests that under conditions of high gametic density, one might expect an arms race between egg and sperm recognition proteins; egg receptor proteins may show negative frequency dependence to minimize polyspermy, whereas sperm-binding proteins are subject to purifying selection to maximize egg recognition (Levitan and Stapper 2010). This does not, however, explain the

Table 1. Sources of *Galaxias* Material for LGP cDNA Sequence Determination.

Species	Source	*N* Parents	*N* Hatchlings
Galaxias brevipinnis	Extruded eggs and sperm from wild adults	3 f, 2 m	100
Galaxias vulgaris	Captive eggs	1 f, 1 m	60
Galaxias depressiceps	Wild egg mass	1 f, 1 m?	60
Galaxias "sp D"	Wild egg mass	1 f, 1 m?	60
Galaxias eldoni	Wild egg mass	1 f, 1 m?	60
Galaxias gollumoides	Captive eggs	1 f, 1 m	60
Galaxias anomalus	Wild eggs	??	60
Galaxias macronasus	Wild eggs	??	4
Galaxias argenteus	Captive eggs	1 f, 1 m	4

NOTE.—?, indicates number of parent(s) is unknown for wild eggs. f, female; m, male; N, number.

prevalence of adaptive evolution in bird gamete recognition proteins on the egg envelope, where polyspermy is the norm (Berlin et al. 2008), so other explanations of positive selection are required.

Salmoniform and osmeriform fishes are renowned for their diadromous life history, whereby growth and reproductive phases alternate between marine and freshwater environments (McDowall 1988; Crespi and Teo 2002). The ancestral state of the Galaxiidae is amphidromy: migration between sea and freshwater that is not related to breeding (cf. anadromy and catadromy) (McDowall 1992, 1997). These species do most of their growing in freshwater and reproduce there. On hatching, larvae flow out to sea, where they spend the first 4–6 months of their lives, before returning to freshwater permanently. Galaxiids have repeatedly lost this marine larval phase (Waters et al. 2000), producing freshwater-limited lineages (Allibone et al. 1996; Waters and Wallis 2001a, 2001b). We are interested in the underlying molecular genetic causes and consequences of this change in life history. Without candidate genes, in a nonmodel organism, we employed Roche 454 pyrosequencing technology to search for sequence or expression differences in transcripts that might mediate this change in life history (Vera et al. 2008). During the course of this work, we have found a larval-expressed zona pellucida (ZP)-domain gene that shows unusually strong evidence for adaptive evolution.

Materials and Methods

Fish

We have restricted our transcriptome analysis to newly emerged larvae, as the life history difference among species involves a difference in larval behavior. In order to control for environmental effect on gene expression, we targeted egg masses so that we could raise them under controlled conditions. Eggs were collected from seven closely related species of the *Galaxias brevipinnis* group (Waters and Wallis 2001a, 2001b; Waters et al. 2010), as well as the more divergent *Galaxias macronasus* (McDowall and Waters 2003) and *Galaxias argenteus* (table 1). *Galaxias* "sp D" refers to a genetically distinct, but as yet undescribed, member of the *G. brevipinnis* group (Allibone et al. 1996). Of these nine species, only *G. brevipinnis* and *G. argenteus* are diadromous; the other seven stream-resident species probably arose from a *G. brevipinnis*–like ancestor. Eggs were obtained in different ways: by extrusion from wild-caught fish by hand, from tanks containing captive fish, or as egg masses from the wild. In the last case, species identity of eggs was unambiguous because the location was known to have only one fish species breeding in that way at that time.

Eggs from *G. brevipinnis* and *G. argenteus* were incubated on gauze in high humidity chambers at 10–12 °C; eggs from all other species were incubated fully immersed in aerated fresh water and held at 10–12 °C. All eggs were examined on a daily basis—any eggs that appeared to have fungal infections were removed and discarded. After 4–5 weeks, eggs from *G. brevipinnis* and *G. argenteus* were repeatedly immersed in spring water (10–12 °C) over a period of days. Hatchlings were harvested upon emergence and placed immediately in RNA Later (Ambion). Eggs incubating in water were left to hatch naturally (4–6 weeks). Daily examination of eggs meant that hatchlings could be removed and placed in RNA Later within, at most, 2 h of hatching. Newly emerged larvae were 7–9 mm in length, depending on species. DNA was isolated from a sample of larvae to confirm species identity by cytochrome *b* (cyt*b*) sequence (Waters and Wallis 2001a). Some fresh adult tissues were also used for expression profiling, and these were similarly placed in RNA Later.

cDNA Preparation

Total RNA was extracted using TriReagent and a Ribopure Kit according to conditions specified by manufacturers (Ambion). Approximately, 40 mg of larvae (between 40 and 100 depending on the species) and between 20 and 40 mg of tissue were used for the total RNA extraction. Double-stranded cDNA was prepared from total RNA using a SMART PCR cDNA synthesis kit (Clontech). Total RNA from *G. brevipinnis* larvae was also used to prepare SMART RACE cDNA (Clontech) in order to perform 5′ or 3′ rapid amplification of cDNA ends.

454 Sequencing

Samples of cDNA from *G. brevipinnis* and *Galaxias depressiceps* larvae were sequenced by GS FLX pyrosequencing (Roche). DNA reads were assembled into contigs using GS De Novo Assembler GS FLX Data Processing Software (Roche). Identities of assembled contigs were determined

Table 2. Primers Used to Amplify Regions of LGP in *Galaxias*.

Name	cDNA Position	Primer Sequence
SPG120F1	1–23	5′-CTGAACTGCAAATTCAAGGGAGG-3′
SPG120F2	51–73	5′-CTGCCCATAAGTCAACTGAAATC-3′
SPG120R2a	131–151	5′-CTGACCTGACAATAGCACCAC-3′
SPG120F6	193–213	5′-ACGGGTGAATTTTGTAATAC-3′
SPG120R6	230–249	5′-GATGCATTTGGGTCGCACTG-3′
SPG120F3	336–360	5′-GCATGAACATAGACACTCTGCACCT-3′
SPG120Rex4	336–360	5′-AGGTGCAGAGTGTCTATGTTCATGC-3′
SPG120F5	389–413	5′-CAAGGTCACGTCGTGACTTTCAGCT-3′
SPG120R3	457–481	5′-GCCGTTGTTAAACAGAATGTGAGTG-3′
SPG120Fex5	499–523	5′-CTCCTCCAACGTCATCACTCGCGAG-3′
SPG120R5	499–523	5′-CTCGCGAGTGATGACGTTGGAGGAG-3′
SPG120F4	828–852	5′-ATAGTGTTCGCTGGGACCTGATAAA-3′
SPG120Fex7	968–989	5′-GTGTTCGTGCACTGCAAAGTGA-3′
SPG120R1	1080–1103	5′-TCTTCCGATTGAATGAGAAGGAGA-3′
SPG120Ra	1108–1127	5′-CTCTGGAGTTGGAGCCAACC-3′
SPG120Fin5.2	Intron 6	5′-ATGTGTGGTTGGAAATACCAG-3′

by BlastN and Blastx searches of the NCBI database through PLAN (http://bioinfo.noble.org/plan/).

Larval Glycoprotein Amplification

Primers for larval glycoprotein (*LGP*) were designed (table 2) on the basis of 454 sequencing results for these two species, including introns by exon-primed intron-crossing polymerase chain reaction (PCR). Sequences were amplified from larval double-stranded cDNA with iTaq DNA polymerase (iNTRON Biotechnology) following the manufacturer's specifications. Genomic DNA samples from 2 to 5 specimens of each species were amplified with Advantage 2 DNA polymerase (Clontech) following the manufacturer's specifications.

Sequencing

PCR DNA products were prepared for sequencing using an Exonuclease1 and Shrimp Alkaline Phosphatase Presequencing Kit (USB Corporation). DNA was sequenced using Big Dye Terminator Version 3.1 (Applied Biosystems) and fragments were separated on an ABI3730 DNA Analyzer.

Quantitative Polymerase Chain Reaction

Total RNA samples from tissues and larvae were used in quantitative polymerase chain reactions (qPCRs). RNA was reverse transcribed using random nonomer and oligo-dT primers with Superscript III Reverse Transcriptase (Invitrogen) following the manufacturer's instructions. A region spanning an intron/exon boundary was then amplified with SensiMix*Plus*SYBR Kit (Quantace) and the products were analysed using a Stratagene Mx3000P Real Time PCR System.

Sequence Searches/Alignment

Sequences were aligned with CLC Sequencer Viewer 6 (www.clcbio.com) and Se-Al v2.0a11 (http://evolve.zoo.ox.ac.uk). All analyses of aligned sequences were performed using MEGA 4 (www.megasoftware.net) (Tamura et al. 2007). All nine complete gene sequences have been deposited in GenBank (accession numbers HM629948–HM629956).

Results

Our initial 454 screen of a cDNA library from *G. brevipinnis* and *G. depressiceps* larvae returned sequences that showed little or no interspecific amino acid sequence differentiation, as would be expected for closely related species showing *d* <0.08 across 5 kb of mtDNA (Waters and Wallis 2001a, 2001b). One highly expressed transcript, however, showed 27 nonsynonymous and 4 synonymous differences between the two taxa. This result was a gross outlier in terms of both the absolute number of amino acid differences and the d_N/d_S ratio, so we designed primers to look at larval cDNA and genomic DNA samples of these two species and the seven others.

Identity of eggs collected to represent nine *Galaxias* was unambiguous because of habitat, timing, and geographic location, but identity was confirmed by cyt*b* sequences and BLAST searches. All nine species closely matched predictions.

Gene Structure and Homologies

BLASTN/BLASTX searches for this cDNA returned five genes with approximately equal amino acid similarities (34–39% over at least 300 amino acid residues; $P < 10^{-47}$): uromodulin precursor (Tamm–Horsfall urinary glycoprotein; THP), zymogen granule membrane glycoprotein 2, alpha tectorin, Fc fragment of IgG binding protein, and seminal plasma glycoprotein (SPP) 120. The first four were from zebrafish (*Danio rerio*); SPP120 matches were with cichlids (*Astatotilapia*, *Haplochromis*, *Melanochromis*, *Oreochromis*, *Pseudocrenilabrus*, *Pseudotropheus*, and *Pundamilia*) (Gerrard and Meyer 2007). No salmoniform (sister Order), medaka (*Oryzias*), or *Fugu* hits were returned. These homologies were evident over most of the length of the gene and included 17 invariable cysteine residues, with no other cysteines found elsewhere in the protein. The cDNA is dominated by a ZP domain, found in many extracellular proteins of diverse function (Bork and Sander 1992;

Table 3. Maximum Composite Likelihood Estimate of the Pattern of Nucleotide Substitution (coding exons only).

	A	T	C	G
A	—	6.24	6.59	15.47
T	6.07	—	9.73	5.81
C	6.07	9.21	—	5.81
G	16.16	6.24	6.59	—

Jovine et al. 2005). ZP domains consist of approximately 260 amino acids close to the C-terminus, characterized by 8–10 conserved cysteine residues that form disulphide bridges to give higher order structures (Callebaut et al. 2007). We specifically identify all ten cysteines and a consensus furin cleavage site (CFCS) of four consecutive arginine residues near the C-terminus (Jovine et al. 2005). Our gene includes ten exons and nine introns (fig. 1) and does not possess anything homologous to the 3′-terminal sperm-combining region of ZP3. The 5′ exon 1 and part of exon 2 are noncoding. Introns 2–5 and intron 7 all have microsatellite regions within them. There is a repeat of 54 bp in intron 7 in several individuals of *Galaxias eldoni*, with one having six repeats. A "strong" Kozak sequence (Kozak 1981) was conserved across all species (gtaATGg).

LGP Sequence Variation

Entire gene sequences (with the exception of 3′ mRNA UTR) were compiled for all nine taxa. These include 996 coding nucleotide sites (332 amino acids) and 1,459 intron nucleotide sites (excluding microsatellite and repeat regions).

Exon Variation

All of our exon analysis includes only the coding region (part of exon 2 to stop codon at the end of exon 9). Bases are represented approximately equally (24.6% A, 25.2% T, 26.7% C, and 23.5% G). Transition/transversion rate ratios were 2.661 (purines) and 1.478 (pyrimidines), with an overall bias of $R = 1.027$. Maximum composite likelihood estimation (pairwise deletion option) of the pattern of nucleotide substitution (Tamura et al. 2004) suggested similar rates for the six classes of transversion, with two different (higher) transition rates (table 3), conforming to the Tamura–Nei model (Tamura and Nei 1993).

Across all nine species, 87 of 332 amino acid residues were variable (including a single deletion in *G. macronasus*; table 4 and fig. 2). Many of these amino acid substitutions were profound/nonconservative. To give this finding perspective, this proportion is over ten times as large as the proportion variable for mitochondrial cyt*b* (9/386) across the same nine taxa (Waters et al. 2000; 2002), despite the fast-evolving nature of mtDNA.

Several variants existed as polymorphisms within species. These were conservatively coded as ambiguities because linkage phase was unknown in heterozygotes. Polymorphisms within species thus make no contribution to the differences we report. Only two indels were observed in exons. One was a 3-bp in-frame deletion toward the 3′ end of exon 5 in *G. macronasus*; the other was a 1-bp insertion in exon 1 of *G. macronasus*. This latter indel, and an in-frame UGA in *G. argenteus* just 5′ of the posited Kozak sequence, argues for exon 1 and the first 34 bp of exon 2 being noncoding.

Interspecific pairwise comparisons across 996 bp of coding exon showed much higher counts of nonsynonymous than synonymous differences (ranges 5–73 and 0–20, respectively; table 5). The modified Nei–Gojobori method (Nei and Gojobori 1986; Nei and Kumar 2000) was used to test whether $d_N > d_S$. We used a one-tailed *Z*-test because the excess of amino acid substitution was so strikingly high compared with all other proteins sequenced. Over all coding exons across all species, there was highly significant evidence for positive selection ($Z = 2.5303$; $P = 0.0063$). When broken down by exon, there was some loss of power to detect selection, but exons 6–8 showed significant excess d_N/d_S, and exon 2 was close to significance (table 6). Exon 5 had the lowest (negative) value (i.e., consistent with purifying selection), but nonetheless has 19/61 variable amino acids, and over twice as much nonsynonymous variation as synonymous. To test whether the effect was limited to a subset of species, we performed the *Z*-test of selection over the length of the coding region using pairwise species comparisons (table 7). Most significant values involved the two more divergent species: *G. macronasus* and *G. argenteus*. Seven of the eight comparisons involving *G. macronasus* were significant. Comparisons involving *Galaxias anomalus* or *G. eldoni* were not significant (with the exception of their respective comparisons with *G. macronasus*). All other species were involved in at least two significant comparisons. Thus, the evidence is, overall, strongly indicative of positive selection over the entire coding region across these species.

Transcript Localization by qPCR

To determine which adult tissues expressed *LGP* transcript, double-stranded cDNAs from muscle, heart, brain, gill, eye,

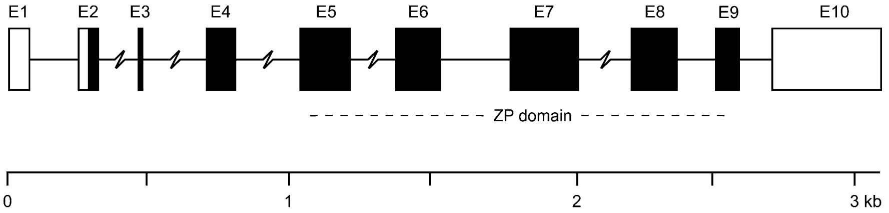

Fig. 1 Exon structure of *Galaxias LGP* from genomic DNA. Blocked boxes represent coding exons, empty boxes non-coding. Concertinas represent microsatellite regions. The ZP domain starts 9 bp into exon 5.

Table 4. Variable Amino Acid Residues by Exon for LGP Across Nine Species of *Galaxias* (codons spanning exons assigned by second base location).

Region	Nucleotides	Codons Substituted	Codon Indels	% Codons Variable
Exon 1	75	—	—	—
Intron 1	169	—	—	—
Exon 2	71	2	0	16
Intron 2	133	—	—	—
Exon 3	15	0	0	0
Intron 3	222	—	—	—
Exon 4	105	10	0	29
Intron 4	224	—	—	—
Exon 5	182	19	1	33
Intron 5	155	—	—	—
Exon 6	161	10	0	19
Intron 6	243	—	—	—
Exon 7	249	19	0	23
Intron 7	183	—	—	—
Exon 8	166	17	0	31
Intron 8	130	—	—	—
Exon 9	84	9	0	33
Intron 9	109	—	—	—
All coding	996	86	1	26

liver, ovary, and testes from *G. brevipinnis* and *G. depressiceps* were used in standard PCR amplifications. cDNA from testes was the only sample to produce any visible PCR product after 40 cycles. When sequenced, this product was identical to that found in larval cDNA with no other splice variants identified. Further analysis by qPCR indicated that the level of expression in the testes was approximately 0.2% of that observed in the newly hatched larvae but may well be dependent on ripeness. Expression level of the larval transcript is high, similar to that of the control gene, elongation factor 1α. This level of expression was maintained for at least the first week after hatching. This is in keeping with our 454 sequencing results, in which this LGP transcript made up 926/103,293 (0.9%) and 576/133,436 (0.4%) of all sequences for *G. brevipinnis* and *G. depressiceps*, respectively.

Interspecific Genetic Distance

Direct interspecific pairwise comparisons across 1,459 bp of intron showed lower overall number of differences (range 7–59) than occurred in the shorter 996 bp of coding exon (range 6–92; supplementary table S1, Supplementary Material online).

MDRSLLLLVVLFSALVSCAFAQTCTPTCVTGEFCNTSISTCQCDPNASNSS

DFGSQVVCSGSSATMYLSYCLLSSAGMNIDTLHLNDGNCTGQVQGHVVTF

SFNSTHTCGSTITTNATHILYNNGILVDVNSSNVITREEAVQLDMSCALE

ISAPELTIPLSLKISDSGAVLINLTSGAWTYTLAIAAFLDSKFTQAINST

TDLLLNEDIYVDLVATGLDEAGIVVVVDSCWATPSNDSGNSVRWDLIKHG

CPNKKDKSVVIQKNGNSTESSFSFKMFEFTGLPQKLVFVHCKVNLCVTAN

HTCTPSCGSRRRRRSLQNIDNNVISFSFNRKI

Fig. 2 Amino acid sequence for *Galaxias brevipinnis* LGP, in blocks of 50 (332 residues). The zona pellucida domain (261 residues) is gray shaded, with its diagnostic ten cysteines and CFCS underlined. All invariable amino acids are emboldened.

Using the Tamura–Nei model of sequence evolution, *d*-values were calculated independently across coding exons and intron sequences, excluding the microsatellite and repeat regions found in five introns (supplementary table S2, Supplementary Material online). For the *G. brevipinnis* group, *d* ranges were 0.006–0.034 (coding) and 0.005–0.012 (intron). Across all species, *d* ranges were 0.006–0.099 (coding) and 0.005–0.044 (intron). The *d*-values for coding regions are on average 2.4-fold greater than those for introns. In the 55 bp of non-coding 3′ UTR sequence that we have available for all nine species, there are only three variants, contrasting again with the high exon variation. From cDNA sequence, *G. brevipinnis* and *G. depressiceps* are identical over 198 bp of this 3′ UTR.

Discussion

Whereas the ZP domain is indisputable (fig. 2), the identity of *LGP* is obscure. The exon–intron structure is similar to the eight-exon 1.85-kb *zp2a-c* genes that make up the *zp2* cluster encoding the major egg envelope protein of

Table 5. Interspecific Pairwise Differences Among Nine *Galaxias* Species Across Eight Coding Exons (996 bp; upper values nonsynonymous, lower values synonymous) (see table 1 for full species names).

	spD	*dep*	*vul*	*ano*	*gol*	*eld*	*bre*	*mac*	*arg*
spD	—	24	21	14	6	24	16	64	49
dep	7	—	12	6	26	16	27	73	60
vul	7	0	—	5	20	15	19	67	53
ano	4	1	1	—	13	7	13	61	42
gol	2	5	5	4	—	22	16	63	46
eld	10	6	5	6	8	—	21	71	52
bre	4	4	4	3	2	5	—	60	43
mac	17	19	19	15	17	20	15	—	59
arg	12	14	13	12	10	15	9	19	—

Table 6. *Z*-Test of Selection (one tail) Across Nine *Galaxias* Species by Exon (codons spanning exons assigned by second base location).

Exon	2	4	5	6	7	8	9
Z-value	1.429	0.497	−0.416	1.698	1.871	1.727	0.180
P value	0.078	0.310	1.000	0.046	0.032	0.043	0.429
Codons	12	35	61	54	83	55	27

zebrafish (Mold et al. 2001), but our gene is 50% longer. Additionally, this *zp2* cluster showed extreme sequence conservation of exons 1–6 among tandem duplicates, whereas our gene is fast evolving along its length. In contrast to this ovary/liver-expressed family (Modig et al. 2006), ours is highly expressed in larvae, and undetectable in adult tissue, except gonad at much lower levels. Although sequence similarity with *SPP120* is clear (Gerrard and Meyer 2007), our gene is much shorter, corresponding only to the ZP domain, as is the case for *Gasterosteus* and *Danio*, that is, it does not include the upstream von Willebrand factor (VWF) D domain or N-terminal uncharacterized domain (Mochida et al. 2002). More tellingly, the region giving evidence for positive selection in *SPP120* is almost exclusively (14/17 codon sites) limited to the N-terminal region upstream of both VWF and ZP (Gerrard and Meyer 2007), completely absent from LGP. Similarly, LGP has just two amino acid residues between the final cysteine and the CFCS box (fig. 2), in contrast to the fast-evolving 48-residue region of *Peromyscus* ZP3 containing the sperm-combining region (Turner and Hoekstra 2006). A number of studies of ZP2 and ZP3 from a variety of mammals and birds show a pattern of positive selection similar to ours (Swanson et al. 2001; Calkins et al. 2007; Berlin et al. 2008). The vast majority of sites under positive selection in these genes are in two clusters: one just following the signal sequence and the other in the sperm-combining site. Ironically, then, the nature of the molecular adaptation that we characterize here is completely unrelated to that described for SPP120 in cichlids or ZP in birds and mammals, excepting the fact that these genes all contain a ZP domain. LGP is special because the ZP domain shows evidence of positive selection over its full length.

We show evidence for strong positive selection over the entire coding region for *LGP* in this group of congeneric osmeriform fishes. Because our gene bears approximately equal and low similarity to *SPP120*, uromodulin, and zymogen glycoprotein, it must be something different, with only deep homology evidenced by possession of a ZP domain. So, we have uncovered a new example of positive selection on a new ZP-domain gene.

Table 7. Interspecific Pairwise *Z*-Tests of Selection (one tail) Across All Exons (*Z*-values upper; *P*-values lower [$P < 0.05$ emboldened]) (see table 1 for full species names).

	spD	*dep*	*vul*	*ano*	*gol*	*eld*	*bre*	*mac*	*arg*
spD	—	0.899	0.513	0.692	0.301	0.066	1.075	1.978	2.020
dep	0.185	—	3.627	1.016	1.869	0.353	2.626	2.085	2.417
vul	0.305	0.000	—	0.802	1.119	0.466	1.597	1.727	2.152
ano	0.245	0.156	0.212	—	0.546	−1.234	1.150	2.037	1.403
gol	0.382	0.032	0.133	0.293	—	0.464	2.239	1.916	2.456
eld	0.474	0.362	0.321	1.000	0.322	—	1.261	1.722	1.449
bre	0.142	0.005	0.056	0.126	0.014	0.105	—	2.035	2.247
mac	0.025	0.020	0.043	0.022	0.029	0.044	0.022	—	1.070
arg	0.023	0.009	0.017	0.082	0.008	0.075	0.013	0.143	—

Such a mode of evolution may be indicative of a Red Queen arms race involving negative frequency–dependent selection, such as that seen in genes involved in protection against pathogens (Swanson and Vacquier 2002). The cysteine-rich ZP domain leads to bridges that facilitate both folding within the protein and binding of the entire protein to other structures, as in membranes and cuticles (Jovine et al. 2005; Monné et al. 2008). LGP also has six additional invariable cysteines in the 53-residue region 3′ to the start of the ZP domain (fig. 2). Mouse ZP3, for example, is implicated as the oocyte sperm receptor and initiates the acrosome reaction (Bork and Sander 1992; Wassarman and Litscher 2009). THP, produced in mammalian kidney, has strong immunoregulatory activity affording protection against urinary tract infection by binding to bacterial type 1 fimbriae (Säemann et al. 2005). THP is the most abundant protein in human urine, and ablation of the gene leads to severe bladder and kidney infections in knockout mice (El-Achkar et al. 2008; Kemter et al. 2009). Such soluble mediators of immunoregulatory activity are postulated to be necessary in the urogenital tract as this tissue does not have mucus or ciliated epithelium for protection (Säemann et al. 2005). The high expression (some 500-fold higher than in gonad) of *LGP* in larvae without reproductive tissue leads us to speculate that it may be affording protection at this vulnerable developmental stage. Interestingly, galaxiids are scaleless, with adults possessing a thick leathery skin and mucus layer for protection (McDowall 1990). In the absence of scales, this group of fish may have developed some novel immune defense.

Rapid evolutionary rates are often evidence of lack of function (Kimura 1983), but several lines of evidence clearly argue against this gene being a pseudogene. First, we have shown through both 454 sequencing and qPCR that this gene is expressed in all species. Second, we see only two examples of indel variation in exons: The first is a single base pair insertion in exon 1 of *G. macronasus*, 5′ to our postulated start codon; the second is an in-frame asparagine codon deletion in exon 5 of *G. macronasus*. Third, diagnostic features of a ZP domain, the strong Kozak sequence and CFCS, and 17 cysteines are conserved across all species. Fourth, all introns are present in all genomic gene sequences, and there is complete conservation of all GT/AG intron–exon splice junctions. Fifth, exons are evolving faster than introns, a nonneutral pattern. Finally, and most convincingly, the significant excess of amino acid substitutions is a nonneutral pattern that can only be explained by Darwinian selection, that is, at the level of a protein product. These findings also cannot be explained by high mutation rate alone.

The fact that three female reproductive proteins and several male reproductive proteins show positive selection has been used as molecular evidence for sperm competition and sexual conflict in species with internal fertilization

(Swanson et al. 2001). Although positive evolution of egg coat ZP-domain genes is well established, our *LGP* is highly active in larvae with no reproductive tissue. If strong positive selection continues to be shown in a variety of other nonreproductive ZP-domain genes, Darwinian evolution might be a feature of ZP-domain genes per se. This mode of evolution would be consistent with recognition of the ZP-N fold as diagnostic of a new immunoglobulin superfamily subtype (Monné et al. 2008). Possibly, these genes may have a general immunological role in life history stages that have extremely high surface area to volume ratios, that is, gametes (Swanson and Vacquier 2002; Aagaard et al. 2010) and larvae. Exons 6–8 show the strongest evidence for positive selection in galaxiids (table 6), although exon 8 also contains the extremely conserved CWAT motif, diagnostic of ZP domains. Further work will use more species comparisons to enable identification, by maximum likelihood, of specific amino acid residues that are targets of selection and elucidation of the structure, function, and evolution of this interesting protein.

Gerrard and Meyer (2007) mention paralogs and gene conversion for *SPP120*. Only three of our species gave evidence for heterogeneous sequences, which are almost certainly polymorphisms, though we conservatively treat them as ambiguities. One cannot explain these data by gene duplication because one has to explain why within-individual sequence variation is very low compared with among-species differences; that is, one would expect similar paralogous differences within and among taxa.

Supplementary Material

Supplementary tables S1–S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org).

Acknowledgments

Andy Hicks supplied eggs from *Galaxias brevipinnis*, and Nicholas Dunn gave us eggs from *Galaxias gollumoides* and *Galaxias vulgaris*. Charles Mitchell provided eggs from *Galaxias argenteus*. Peter Ravenscroft (Department of Conservation, Otago) supplied eggs for *Galaxias depressiceps*, *Galaxias* "sp D," *Galaxias eldoni*, and *Galaxias macronasus*. Mark Lokman provided advice about qPCR, and David Winter helped with bioinformatics. Craig Marshall and Sigurd Wilbanks provided helpful discussion on protein biochemistry. This research was funded by contract 06-UOO-143 from the Marsden Foundation and by the University of Otago.

References

Aagaard JE, Vacquier VD, MacCoss MJ, Swanson WJ. 2010. ZP domain proteins in the abalone egg coat include a paralog of VERL under positive selection that binds lysin and 18-kDa sperm proteins. *Mol Biol Evol.* 27:193–203.

Allibone RM, Crowl TA, Holmes JM, King TM, McDowall RM, Townsend CR, Wallis GP. 1996. Isozyme analysis of *Galaxias* species (Teleostei: Galaxiidae) from the Taieri River, South Island, New Zealand: a species complex revealed. *Biol J Linn Soc.* 57:107–127.

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH. 2007. Evolution. Cold Spring Harbor, (NY): Cold Spring Harbor Laboratory Press.

Berlin S, Qu L, Ellegren H. 2008. Adaptive evolution of gamete-recognition proteins in birds. *J Mol Evol.* 67:488–496.

Bork P, Sander C. 1992. A large domain common to sperm receptors (Zp2 and Zp3) and TGF-β type III receptor. *FEBS Lett.* 300:237–240.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Calkins JD, El-Hinn D, Swanson WJ. 2007. Adaptive evolution in an avian reproductive protein: ZP3. *J Mol Evol.* 65:555–563.

Callebaut I, Mornon J-P, Monget P. 2007. Isolated ZP-N domains constitute the N-terminal extensions of zona pellucida proteins. *Bioinformatics* 23:1871–1874.

Civetta A. 2003. Positive selection within sperm-egg adhesion domains of fertilin: an ADAM gene with a potential role in fertilization. *Mol Biol Evol.* 20:21–29.

Clark NL, Gasper J, Sekino M, Springer SA, Aquadro CF, Swanson WJ. 2009. Coevolution of interacting fertilization proteins. *PLoS Genet.* 5:293.

Crespi BJ, Teo R. 2002. Comparative phylogenetic analysis of the evolution of semelparity and life history in salmonid fishes. *Evolution* 56:1008–1020.

Dorus S, Wasborough ER, Busby J, Wilkin EC, Karr TL. 2010. Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Mol Biol Evol.* 27:1235–1246.

Duda TF, Palumbi SR. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc Natl Acad Sci U S A.* 96:6820–6823.

El-Achkar TM, Wu XR, Rauchman M, McCracken R, Kiefer S, Dagher PC. 2008. Tamm-Horsfall protein protects the kidney from aschemic injury by decreasing inflammation and altering TLR4 expression. *Am J Physiol Renal Physiol.* 295:F534–F544.

Ellegren H, Sheldon BC. 2008. Genetic basis of fitness differences in natural populations. *Nature* 452:169–175.

Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol.* 13: 685–690.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* 415:1024–1026.

Gerrard DT, Meyer A. 2007. Positive selection and gene conversion in SPP120, a fertilization-related gene, during the East African cichlid fish radiation. *Mol Biol Evol.* 24:2286–2297.

Geyer LB, Palumbi SR. 2003. Reproductive character displacement and the genetics of gamete recognition in tropical sea urchins. *Evolution* 57:1049–1060.

Hanada K, Shiu S-H, Li W-H. 2007. The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. *Mol Biol Evol.* 24:2235–2241.

Hughes AL, Ota T, Nei M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol Biol Evol.* 7:515–524.

Jovine L, Darie CC, Litscher ES, Wassarman PM. 2005. Zona pellucida domain proteins. *Annu Rev Biochem.* 74:83–114.

Kemter E, Rathkolb B, Rozman J, et al. (15 co-authors). 2009. Novel missense mutation of uromodulin in mice causes renal dysfunction with alterations in urea handling, energy, and bone metabolism. *Am J Physiol Renal Physiol.* 297:F1391–F1398.

Kimura M. 1983. The neutral theory of evolution. Cambridge: Cambridge Univeristy Press.

Kozak M. 1981. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* 9:5233–5252.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet.* 4:e1000304.

Larmuseau MHD, Huyse T, Vancampenhout K, van Houdt JKJ, Volckaert FAM. 2010. High molecular diversity in the rhodopsin gene in closely related goby fishes: a role for visual pigments in adaptive speciation? *Mol Phylogenet Evol.* 55:689–698.

Levitan DR, Stapper AP. 2010. Simultaneous positive and negative frequency-dependent selection on sperm binding, a gamete recognition protein in the sea urchin Strongylocentrotus purpuratus. *Evolution* 64:785–797.

Lewontin RC. 1974. The genetic basis of evolutionary change. New York: Columbia University Press.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates Inc.

Li W-H, Gojobori T. 1983. Rapid evolution of goat and sheep globin genes following gene duplication. *Mol Biol Evol.* 1:94–108.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654.

McDowall RM. 1988. Diadromy in fishes: migrations between freshwater and marine environments. London: Croom Helm.

McDowall RM. 1990. New Zealand freshwater fishes: a natural history and guide. Auckland, (New Zealand): Heinemann Reed.

McDowall RM. 1992. Diadromy: origins and definitions of terminology. *Copeia* 1992:248–251.

McDowall RM. 1997. Is there such a thing as amphidromy? *Micronesica* 30:3–14.

McDowall RM, Waters JM. 2003. A new species of *Galaxias* (Teleostei: Galaxiidae) from the Mackenzie Basin, New Zealand. *J R Soc N Z.* 33:675–691.

Mikkelsen TS, Hillier LW, Eichler EE, et al. (67 co-authors). 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Mochida K, Matsubara T, Andoh T, Ura K, Adachi S, Yamauchi K. 2002. A novel seminal plasma glycoprotein of a teleost, the Nile tilapia (*Oreochromis niloticus*), contains a partial von Willebrand factor type D domain and a zona pellucida-like domain. *Mol Reprod Dev.* 62:57–68.

Modig C, Modesto T, Canario A, Cerdà J, von Hofsten J, Olsson P-E. 2006. Molecular characterization and expression pattern of zona pellucida proteins in gilthead seabream (*Sparus aurata*). *Biol Reprod.* 75:717–725.

Mold DE, Kim IF, Tsai C-M, Lee D, Chang C-Y, Huang RCC. 2001. Cluster of genes encoding the major egg envelope protein of zebrafish. *Mol Reprod Dev.* 58:4–14.

Monné M, Han L, Schwend T, Burendahl S, Jovine L. 2008. Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. *Nature* 456:653–659.

Morgan CC, Loughran NB, Walsh TA, Harrison AJ, O'Connell MJ. 2010. Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins. *BMC Evol Biol.* 10:39.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3:418–426.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. Oxford: Oxford University Press.

Plotkin JB, Dushoof J, Fraser HB. 2004. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*.. *Nature* 428:942–945.

Säemann MD, Weichhart T, Hörl WH, Zlabinger GJ. 2005. Tamm-Horsfall protein: a multilayered defence molecule against urinary tract infection. *Eur J Clin Investig.* 35:227–235.

Spady TC, Seehausen O, Loew ER, Jordan CJ, Kocher TD, Carleton KL. 2005. Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid species. *Mol Biol Evol.* 22:1412–1422.

Stewart C-B, Schilling JW, Wilson AC. 1988. Convergent evolution of lysozyme sequences? *Nature* 332:787–788.

Swanson WJ, Vacquier VD. 2002. Reproductive protein evolution. *Annu Rev Ecol Syst.* 33:161–179.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc Natl Acad Sci U S A.* 98:2509–2514.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4. *Mol Biol Evol.* 24:1596–1599.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol.* 10:512–526.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A.* 101:11030–11035.

Turner LM, Hoekstra HE. 2006. Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (*Peromyscus*). *Mol Biol Evol.* 23:1656–1669.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol.* 17:1636–1647.

Wassarman PM, Litscher ES. 2009. The multifunctional zona pellucida and mammalian fertilization. *J Reprod Immunol.* 83: 45–49.

Waters JM, López JA, Wallis GP. 2000. Molecular phylogenetics and biogeography of galaxiid fishes (Osteichthyes: Galaxiidae): dispersal, vicariance, and the position of *Lepidogalaxias salamandroides*. *Syst Biol.* 49:777–795.

Waters JM, Rowe DL, Burridge CP, Wallis GP. 2010. Gene trees versus species trees: reassessing life-history evolution in a freshwater fish radiation. *Syst Biol.* 59: doi:10.1093/sysbiol/syq031.

Waters JM, Saruwatari T, Kobayashi T, Oohara I, McDowall RM, Wallis GP. 2002. Phylogenetic placement of retropinnid fishes: data set incongruence can be reduced by using asymmetric character state transformation costs. *Syst Biol.* 51: 432–449.

Waters JM, Wallis GP. 2001a. Cladogenesis and loss of the marine life history phase in freshwater galaxiid fishes (Osmeriformes: Galaxiidae). *Evolution* 55:587–597.

Waters JM, Wallis GP. 2001b. Mitochondrial DNA phylogenetics of the *Galaxias vulgaris* complex from South Island, New Zealand: rapid radiation of a species flock. *J Fish Biol.* 58:1166–1180.

Weinberger H, Moran Y, Gordon D, Turkov M, Kahn R, Gurevitz M. 2010. Positions under positive selection—key for selectivity and potency of scorpion α-toxins. *Mol Biol Evol.* 27:1025–1034.