

# Molecular Evolution of Teleost Neural Isozymes

Ryan R. Auld · Joseph M. Quattro ·  
Thomas J. S. Merritt

Received: 12 July 2012 / Accepted: 9 November 2012 / Published online: 25 November 2012  
© Springer Science+Business Media New York 2012

**Abstract** Isozymes, homologous enzymes coded by separate loci within a genome, present interesting systems for examining molecular and functional divergence through natural selection. Isozyme pairs for a number of metabolic enzymes, including *Triosephosphate isomerase* (*Tpi*), *Malate dehydrogenase* (*Mdh*), *Phosphoglucose isomerase* (*Pgi*), and *Guanylate kinase* (*Guk*), appear to all result from a single, large duplication event early in teleost evolution. These small gene families include two forms, a generally expressed form with no apparent charge and a neurally expressed form with a pronounced negative charge although the canalization of expression of the second form varies across families. Using ancestral sequence reconstructions and standard comparisons of rates of nonsynonymous and synonymous change, combined with the examination of the specific amino acid changes observed and predicted we examined the evolution of the *Tpi* and *Guk* families using all available vertebrate sequences and all four families using a smaller, common, dataset. We find that post-duplication, the neural *Tpi* and *Guk* isozymes evolved through similar periods of positive selection as evidenced by elevated rates of nonsynonymous change and accumulation of negative amino acids. Over the same evolutionary period our analysis suggests that

*Mdh* and *Pgi* isozymes appear to have evolved under a less divergent pattern of selection. These distinct results likely reflect functional differences between the isozymes, possibly a result of differences in expression patterns.

**Keywords** Guanylate kinase · Triosephosphate isomerase · Positive selection · Duplication · Neural isozyme

## Introduction

Isozymes, closely related paralogous enzymes, share enzymatic function, but often have distinct expression patterns and amino acid sequences, suggesting some degree of functional specialization. These observed differences are generally thought to result from natural selection, but direct evidence of selection has been difficult to demonstrate (Ohno 1970; Fisher and Whitt 1978; Whitt 1983; Merritt and Quattro 2001, 2003). Such duplicated enzymes and other products of gene and genome duplication are however major contributors to gene and genome complexity. Considerable effort has been made to study and model their evolution, but the mechanism by which this complexity and novel gene function evolve is still not well understood (Ohno et al. 1968; Ohno 1970; Soltis and Soltis 2000; Wendel 2000). Positive selection is often proposed to play a substantial role in the evolution of isozymes and other duplicated genes (e.g., Ohno 1970; Goodman et al. 1975; Li and Gojobori 1983; Fisher et al. 1980; Ohta 1991), but ruling out simple neutral evolution following relaxation of functional constraint is often challenging (Dykhuizen and Hartl 1980; Kimura 1983; King and Jukes 1969; Yang 2002). One approach that has had some success is the use of ancestral sequence reconstruction to focus phylogenetic analysis on the evolutionary period directly following a

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-012-9532-1) contains supplementary material, which is available to authorized users.

R. R. Auld · T. J. S. Merritt (✉)  
Department of Chemistry and Biochemistry,  
Laurentian University, Sudbury, ON P3E 2C6, Canada  
e-mail: tmerritt@laurentian.ca

J. M. Quattro  
Department of Biological Sciences, University of South  
Carolina, Columbia, SC, USA

duplication event (Yang et al. 1995; Messier and Stewart 1997; Zhang and Nei 1997; Merritt and Quattro 2001).

Modern proteins are generally the result of millions of years of adaptation to function, and as such are generally under purifying selection to maintain function (Charlesworth et al. 1993). In at least some cases, protein diversity appears to result from brief periods of positive selection following a gene or genome duplication, followed by a return to purifying selection (Hughes 1994; Messier and Stewart 1997; Kosiol et al. 2008). Purifying selection is characterized by a predominance of synonymous substitutions and a dearth of nonsynonymous substitutions; while the hallmark of diversifying selection is an excess of nonsynonymous substitutions. Comparison of only modern, extant, sequences might encompass multiple modes of selection and may therefore fail to identify evidence of a relatively short period of positive selection (Messier and Stewart 1997). For this reason, methods have been developed to identify selective pressures, purifying to diversifying, by gene phylogenies and a branch-by-branch examination in an attempt to isolate distinct periods of evolutionary interest. Ancestral sequences are often reconstructed allowing comparison of all nodes (modern and inferred ancestral genes and proteins) in a pairwise fashion (Zhang et al. 1998) to quantify selective pressures. If sequences are evolving without selective constraint, i.e., neutrally, the amount of observed nonsynonymous/synonymous ( $n/s$ ) change would be expected to equal the number of potential nonsynonymous and synonymous sites ( $N/S$ ) sites (Kimura 1983). Functional genes are expected to be under purifying selection with fewer nonsynonymous changes than expected, given the number of nonsynonymous sites (Zhang et al. 1998; Nei and Kumar 2000). Genes evolving new functions are expected to be under positive or directional selection with more observed nonsynonymous change than expected (Zhang et al. 1997, 1998; Yang and Bielawski 2000). Testing for these differences is not trivial and various tests have come under fire at different times (e.g., Zhang et al. 1998; Yokoyama et al. 2008; Hughes and Friedman 2008). In the work presented here, we attempt to address this issue by means of multiple tests and by directly addressing the biological context of changes (e.g., protein isoelectric point and the location of changes in the three-dimensional (3D) space of the protein).

Complete genome duplications are thought to have had a role in creating genomic diversity in a wide variety of organisms (Pebusque et al. 1998; Kellis et al. 2004). In fish, a genome duplication event appears to have occurred early in teleost evolution, approximately 320–350 mya (Vandepoele et al. 2004; Meyer and Van De Peer 2005; Taylor et al. 2001), after the divergence of teleosts and more basal non-teleost fish species (e.g., sturgeon or gar, Jaillon et al. 2004; Meyer and Van De Peer 2005; Amores et al. 1998). This teleost-specific duplication event was followed by extensive gene

loss and a concomitant return to functional diploidy, as well as apparent divergence and specialization within the remaining gene pairs. The event is likely the duplication that produced multiple isozyme pairs observed in teleost fish (Fisher et al. 1980). Many of these small isozyme gene families include a broadly, or ubiquitously, expressed isozyme and an isozyme with a more narrow expression pattern, often largely restricted to “neural” tissues such as the eye and brain identified in the early protein electrophoresis studies (e.g., Fisher et al. 1980; Gaida 1995). Many of these isozymes differ in charge, isoelectric point, as well as pattern of expression, with the neural isozyme having a pronounced net negative charge (Shaklee et al. 1973; Champion and Whitt 1976; Fisher and Whitt 1978; Fisher et al. 1980) and the non-neural, or generally expressed, isozyme having a neutral charge that is also a characteristic of homologous enzymes in single-gene species. These negative neural isozymes appear to reflect a general trend observed across neural molecules toward negative charge (Moore and McGregor 1965; Moore 1973; Margolis and Margolis 1993; Novak and Kaye 2000; Elkin et al. 2010). Pairs of negatively charged neural, and neutrally charged general, isozymes have been reported in *Triosephosphate isomerase* (TPI: EC. 5.3.1.1, Pontier and Hart 1981; Merritt and Quattro 2001), *Aldolase* (ALD: EC. 4.1.2.13, Merritt and Quattro 2002), *Malate dehydrogenase* (MDH: EC. 1.1.1.37, Merritt and Quattro 2003), *Phosphoglucose isomerase* (PGI: EC. 5.3.1.9, Fisher et al. 1980), and *Lactate dehydrogenase* (LDH: EC. 1.1.1.27, Whitt 1970; Markert et al. 1975; Fisher et al. 1980). The exact functional repercussions of the charge differences are unknown although the production of a negative intracellular environment for impulse transmission in neural tissues has been suggested (Merritt and Quattro 2001).

The distinct charges and expression patterns between these teleost isozymes suggest a functional difference and provide a good model system for investigating the role of selection in protein functional divergence and molecular evolution. Merritt and Quattro (2001) reported that the negative charge of the neural *Tpi* isozyme resulted from the specific accumulation of negatively charged amino acids during a period of positive (directional) selection directly following the duplication event, although Yang (2002) cautioned that the evidence was somewhat equivocal. Similarly *Mdh* also includes a negative isozyme, which was also found to have evolved through a period of positive selection following the duplication although the strong bias in amino acid change was not observed in this gene family (Merritt and Quattro 2003). Neural and general *Pgi* isozymes have also been described in teleost fish although no evidence for positive selection between the two loci was found (Sato and Nishida 2007). Several other gene families with tissue-specific isozymes have also been described,

including *Guanylate kinase* (GUK: EC. 2.7.4.8, Gaida 1995; Kettler and Whitt 1986). The *Guk* gene family has not previously been examined for selective pressure in teleost fish, but may provide another example of positive selection after duplication with specialization to a neural environment. GUK is a phosphotransferase that catalyzes the phosphorylation of GMP to GDP (Agarwal et al. 1978). Teleost fish express two *Guk* isozymes, a negatively charged neural form and a neutral, generally expressed, form (Fisher et al. 1980; Gaida 1995) similar to *Tpi*, *Mdh*, and *Pgi*, suggesting that *Guk* may have evolved under similar selective constraints as these other gene families.

Here, we compare the patterns of DNA and amino acid change across four gene families—*Tpi*, *Mdh*, *Pgi* and *Guk*—all of which contain neural and general isozymes in teleost fish. The *Tpi*, *Mdh*, and *Pgi* gene families have all been examined previously and tested for evidence of diversifying selection along their neural branches; this is the first examination of the fish *Guk* family. Based on the charge difference between the *Guk* isozymes, we suspect that the GUK proteins may have evolved under similar selective pressures following the genome duplication in teleost fish. We examine the evolution of the *Guk* and *Tpi* gene families in greater depth with “large” datasets that include all fish sequences currently in public genomic databases. We use phylogenetic analysis and ancestral sequence reconstruction to identify possible differences in selective pressure along the branch leading to the ancestral neural isozyme. We also compiled and analyzed “comparison” datasets containing identical species sequences for *Tpi*, *Mdh*, *Pgi* and *Guk* and reanalyzed and examined the evolution of these genes in a comparative framework to yield a better understanding of the selective pressures and patterns of evolution observed within isozymes after duplication.

## Methods

### Phylogenetic Analysis

Coding sequences for *Guk*, *Tpi*, *Mdh*, and *Pgi* genes were gathered from public databases by means of the Molecular Evolutionary Genetic Analysis (MEGA 4) software (Tamura et al. 2007) and BLAST search tools (Altschul et al. 1990). Six main datasets were compiled for molecular analyses: two “large” datasets, one for *Tpi* and one for *Guk* containing all teleost sequences, and four smaller “comparison” sets containing only sequences from species with orthologs available for all four loci (i.e., available orthologs that the gene families have in common). As an example, both isozyme sequences are available for all four gene families from *Danio rerio* (Cypriniformes) and

*Oryzias latipes* (Beloniformes) so sequences from these two species are present in both the complete and comparison datasets. Both *Guk* isozyme sequences are available for *Gasterosteus aculeatus* (Gasterosteiformes), *Takifugu rubripes* (Tetraodontiformes), and *Tetraodon nigroviridis* (Tetraodontiformes), but not the sequences from the other gene families, so these genes are included in the large *Guk* set, but not common datasets. The “large” datasets are a comprehensive survey of available coding sequences for *Guk* and *Tpi*, while the “comparison” datasets, containing sequences available for the four gene families, allow direct comparison of analyses across the four families. The species and corresponding accession numbers, included in the large *Tpi* and large *Guk* datasets are found in the Online Resource Tables 1 and 2, respectively. Similarly, the species and accession numbers for the sequences used in the comparison datasets for all four genes are found in the Online Resource Table 3. Sequences were aligned by means of ClustalW (Thompson et al. 1994) as implemented in MEGA 4 (Tamura et al. 2007). Neighbor joining (NJ) phylogenetic analysis was also performed by means of MEGA 4 with the maximum composite likelihood model and complete deletion. Third codon positions were excluded because of saturation at these sites. Two thousand bootstrap replicates were performed in all analyses. Phylogenetic analysis under the maximum likelihood [ML] (PAML) criterion was conducted by means of PHYLIP (Felsenstein 1989) with the F84 nucleotide substitution following Yang et al. (1994) and 100 bootstrap replicates. Bayesian phylogenetic analysis was conducted by means of MrBayes (Ronquist and Huelsenbeck 2003) using the general GTR evolutionary model found to previously perform well in a test model (Bollback 2002), gamma distributed rate variation, and 100 bootstrap replicates.

### Ancestral Sequence Reconstruction and Tests of Positive Selection

Ancestral sequence reconstructions and estimates of nucleotide substitutions were performed by the ML methods implemented in PAML, version 4.3, with nucleotide and amino acid sequences inferred for all nodes using the free ratio model and F3X4 codon frequency (Yang 2007). The large *Tpi* dataset included 20 sequences, the large *Guk* included 25, and the comparison datasets all contained eight sequences each. Only complete coding sequences were included in all analyses. Alignment gaps were excluded in all phylogenetic and PAML analyses. The numbers of nonsynonymous and synonymous changes and sites were calculated along each branch, using the free ratio model and a F3X4 codon frequency. The ratio of *n/s* (observed change) along the neural branch was compared

to the  $N/S$  (potential change) ratio of the entire tree for evidence of positive selection. Patterns of selection were also inferred and tested across tree topologies using various models of rates of nonsynonymous and synonymous change ( $\omega$ ) implemented in PAML.

### Protein Analysis

The predicted ancestral sequences (from PAML, above) for the ancestral “single” node (basal to both isozymes—AncestralS), the predicted ancestral neural isozyme node (AncestralN), and the predicted ancestral general isozyme node (AncestralG) were similar and trivial to align by eye. The isoelectric points of reconstructed and existing protein sequences were calculated by means of Sequence Manipulation Suite (Stothard 2000). Within each gene family, the average isoelectric points—the pH at which the protein carries no net charge—of the general isozyme was calculated from all modern and inferred general sequences and compared to the single inferred ancestral neural isozyme by a one sample Z test (Daly and Bourke 2008). All reconstructed and modern sequences were also compared along branches, i.e., between two aligned sequences, to determine the number and type of amino acid substitutions within each family. Observed amino acid substitutions were scored as either maintaining (conservative) or changing (radical) physiochemical properties (i.e., within or crossing amino acid groupings). Amino acids were grouped by either polarity (polar and non-polar), charge (positive, negative, neutral), or size (five categories from small to large; changes in size categories were defined as crossing at least two IMGT size classes, as classes defined in Pommie et al. (2004)). The pattern of conservative and radical change across the S–N (see above) neural branch was then compared to the pattern across the rest of the tree for each gene family by Fisher’s exact test.

The general 3D location of amino acids, i.e., surface versus internal or buried, was determined by comparison of inferred amino acid sequences with known protein structures. Crystal structures for the teleost fish proteins are not available so inferred sequences were compared to their homologs in the mouse (GUK and PGI), human (TPI), and wild boar (MDH). The protein databank was used to find crystal structures for all four proteins, GUK, TPI, MDH, and PGI (protein databank IDs: 1LVG, 1WYI, 4MDH, and 1UOE). Amino acid location, calculated as solvent accessibility, was determined by means of Accelrys DS Visualizer version 2.0 (<http://accelrys.com/products/discovery-studio/visualization-download.php>), where exposed amino acids had a solvent accessibility surface (SAS) greater than 25 % and buried amino acids had a SAS of less than 10 %. The accumulated negative amino acids were then compared to their positioning to the native protein.

## Results

### Triosephosphate Isomerase

The large *Tpi* dataset (Online Resource Table 1) includes 20 *Tpi* coding sequences from a variety of vertebrates including tetrapods, basal vertebrates, and pre- and post-genome duplication fish. NJ and ML phylogenetic methods both produce the same tree, Online Resource Fig. 1, which is consistent with the literature on vertebrate and fish evolution and earlier work on the *Tpi* gene family in vertebrates (Merritt and Quattro 2001). All major nodes are identical in the Bayesian tree although the Zebrafish (*D. rerio*) and Salmon (*Salmo salar*) neural sequences are interchanged producing a tree that is inconsistent with the current consensus on teleost evolution (e.g., Chiu et al. 2004; Steinke et al. 2006; Volff 2005). All data presented here are from analyses using the “correct” ML tree, but running our analyses by means of the Bayesian-derived topology did not significantly alter any results or conclusions (data not shown).

To test different evolutionary hypotheses, log likelihood values were produced and compared for the large *Tpi* dataset under specific  $\omega$  ratio models (rates of  $dn/ds$  change) by means of PAML 4.3 (following Yang 1998, 2002). The log likelihood score under the free ratio model, which allows every branch of the phylogeny to have its own  $\omega$  ratio, was significantly higher, when compared to a  $\chi^2$  distribution, than under all other models used. This result indicates that a model in which  $\omega$  can vary across all branches, best fits all the datasets, i.e., that there is significant variation in  $\omega$  across branches across all genes examined here. The free ratio model was therefore used to infer ancestral sequences, and to calculate the number of synonymous and nonsynonymous substitutions and amino acid changes along each branch of the *Tpi* tree. The numbers of nucleotide and amino acid changes (see below) were used in tests of positive selection. In fact, the free ratio model was found to be a significantly better fit than any other  $\omega$  ratio model for all datasets used in this study and this model was used to infer ancestral sequences in all cases.

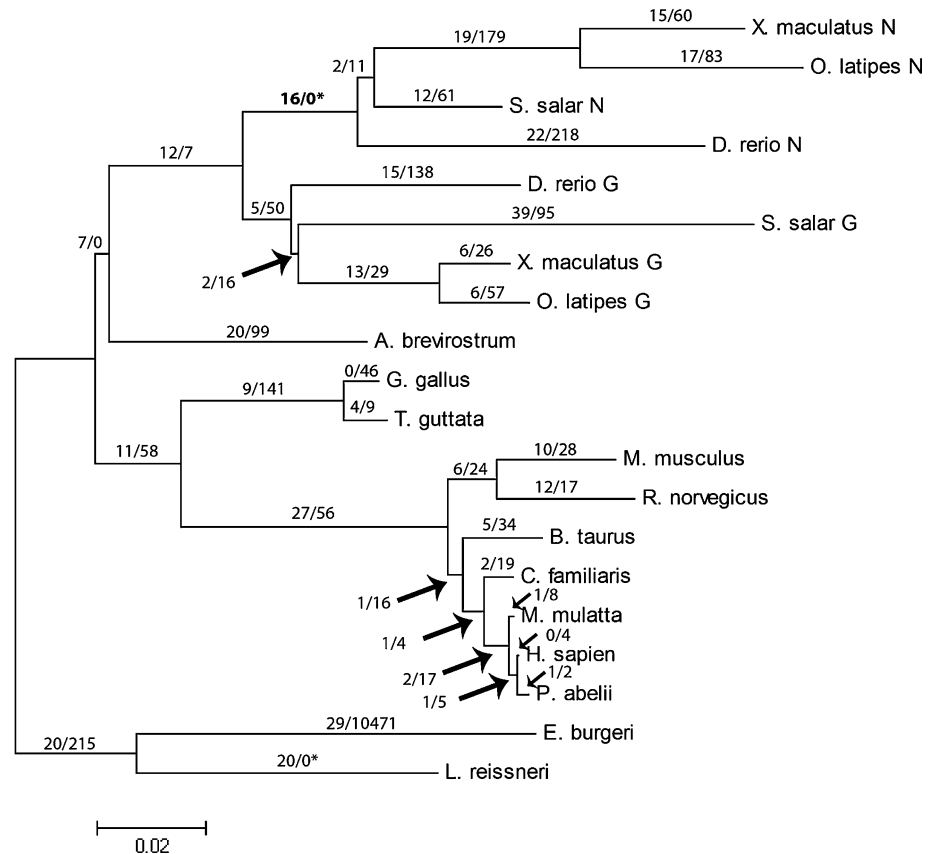
We estimated the number of nucleotide substitutions along each branch by ML methods and tested for a significant difference between  $n/s$  and  $N/S$  by Fisher’s exact tests (following Zhang et al. 1997). Figure 1 shows the number of nonsynonymous and synonymous changes ( $n/s$ ) along each branch of the large *Tpi* tree. We were particularly interested in the pattern of evolution along the branch leading to the neural isozyme, a branch that earlier work suggests evolved through a period of positive selection (Merritt and Quattro 2001). Our analysis of the large *Tpi* dataset identified 16 nonsynonymous and zero

synonymous changes along the neural branch (Fig. 1). This pattern of observed change is significantly different from the pattern of potential  $N/S$  sites ( $P < 5\%$ ), suggesting that *Tpi* likely evolved through a period of positive selection along this neural branch, in agreement with earlier analyses (Merritt and Quattro 2001). It is worth noting that in this dataset, and in fact in all datasets, after the neural branch (i.e., the terminal branches of the trees) the substitution pattern returns to a preponderance of synonymous change, i.e., following this brief period the sequences return to a period of purifying selection.

Log likelihood ratio tests were also used to specifically test for differences in selective pressures across different branches of the *Tpi* tree, an alternative method for identifying periods of positive selection. To do this, we constructed 11 models, A–K (Table 1, left most columns) that varied in the number of  $\omega$  ratios allowed, the branches along which  $\omega$  was allowed to vary, and the  $\omega$  values stipulated (following Yang 1998). To test if a particular branch has a significantly different  $\omega$  value than the rest of the tree, e.g., to test if the pattern of sequence evolution along the neural branch is different than the rest of the tree, we compared the one ratio model (which allows only one  $\omega$  ratio for all branches) with the two ratio model. When we compare the one ratio model with the two ratio model, in which the  $\omega$  ratio of the neural branch was free to differ

(compared models A–C, Table 1, right most columns), we find a significant difference between the models, indicating a difference between the  $\omega$  value of the neural branch and  $\omega$  across the rest of the tree. Similarly, when we compare the two ratio model and the three ratio model, in which the  $\omega$  ratios of both the neural and general branches were free to differ (compared models B–E and A–B respectively, Table 1, right most columns), we find a significant difference between the  $\omega$  ratio along the neural, but not general branch, indicating that the neural  $\omega$  is different from that of the background  $\omega$ . To possibly discriminate between positive selection and selective neutrality, likelihood values were also compared using the two ratio models and the three ratio models with and without forcing  $\omega$  to equal one, allowing us to directly test for positive selection (Yang 1998, 2002). Significantly different  $\omega$  values along a branch can indicate neutrality or positive selection and values significantly greater than one would indicate positive selection. The log likelihood values for the two ratio tests, with and without setting  $\omega = 1$  (compared model C–H), are not statistically different, providing no support for  $\omega$  being greater than one. Results of the three ratio tests (compared model E–J) similarly do not support  $\omega$  being greater than one. These two tests, then, do not provide any support for *Tpi* having evolved through a period of positive selection following the duplication event. Taken as a

**Fig. 1** ML tree depicting nucleotide relationship among large *Tpi* dataset sequences. The number of nonsynonymous/synonymous substitutions is shown above each branch. Boldface asterisk values represent significantly different nonsynonymous/synonymous change in comparison to nonsynonymous/synonymous sites by Fisher's exact test ( $P < 5\%$ )





**Table 1** ML tests and estimates for the large and comparison *Tpi* datasets

	Models <sup>a</sup>	Sub-models <sup>a</sup>	Log likelihood values	$\omega n^b$	$n/s^b$	Compared models <sup>c</sup>	Test	Likelihood ratio test
A	One ratio	$\omega b = \omega n = \omega g$	-3436.15	0.0631 (0.065)	19.6/105.3 (9.3/46.7)	A–C	$\omega b = \omega n$	0.04 (12.38**)
B	Two ratio	$\omega b = \omega n, \omega g$	-3435.82	0.0614 (0.065)	20/110.3 (9.3/46.7)	B–E	$\omega b = \omega n$	0.02 (12.9**)
C	Two ratio	$\omega b = \omega g, \omega n$	-3436.13	0.0710 ( $\infty$ )	20/95.3 (15.9/0)	A–D	$\omega g$ and $\omega n = \omega b$	0.18 (5.68*)
D	Two ratio	$\omega b, \omega g = \omega n$	-3436.06	0.0784 (0.196)	20.3/87.6 (13.8/23)	A–B	$\omega b = \omega g$	0.66 (0.0)
E	Three ratio	$\omega b, \omega g, \omega n$	-3435.81	0.058 ( $\infty$ )	19.8/115.4 (16.1/0)	C–E	$\omega b = \omega g$	0.64 (0.52)
F	Free ratio		-3419.53	0.0549 ( $\infty$ )	20.1/123.8 (16.2/0)			
G	Two ratio	$\omega b = \omega n, \omega g = 1$	-3435.82	0.0615 (0.064)	20/109.7 (10.3/52.7)	D–I	$\omega g$ and $\omega n > 1$	5.82* (4.06*)
H	Two ratio	$\omega b = \omega g, \omega n = 1$	-3438.24	1 (1)	20.8/7 (15.5/5.1)	C–H	$\omega n > 1$	4.22* (0.66)
I	Two ratio	$\omega b, \omega g = \omega n = 1$	-3438.97	1 (1)	20.9/7 (16.4/5.3)	E–J	$\omega n > 1$	4.82* (0.8)
J	Three ratio	$\omega b, \omega g, \omega n = 1$	-3438.22	1 (1)	20.8/7 (15.6/5.1)	B–G	$\omega g > 1$	0 (4.62*)
K	Three ratio	$\omega b, \omega g = 1, \omega n$	-3435.82	0.0590 (0.388)	19.8/113.5 (16/13.5)	E–K	$\omega g > 1$	0.02 (10.84**)

<sup>a</sup> Models and submodels indicate branches which are allowed to vary in PAML analyses

<sup>b</sup> Refer to methods for explanation on  $n/s$  and  $\omega$

<sup>c</sup> Models compared indicate  $\omega$  ratios compared by likelihood ratio tests as indicated. Parameters  $\omega b$ ,  $\omega n$ , and  $\omega g$  represent dN/ds for the background branches (all branches excluding the neural and general branch) and the neural and general branch, respectively

Bracketed values indicate results from the large *Tpi* dataset, while unbracketed values correspond to the comparison *Tpi* dataset

\* Significant  $P < 5\%$ , \*\* significant at  $P < 1\%$

whole, the results from our PAML analysis suggest that following the duplication, the neural *Tpi* isozyme experienced a change in selective pressure, but cannot eliminate the possibility that this period was one of neutral evolution, not positive selection. Comparisons of the patterns of amino acid changes and the “comparison” *Tpi* dataset, below, and the comparison of observed and potential sites (above) do, however, suggest that this was in fact a period of positive selection.

Pairwise comparisons of all reconstructed (ancestral) and modern (extant) sequences were used to infer amino acid substitutions across the large dataset *Tpi* tree. Amino acids were grouped by charge, size, or polarity, and the number of conservative (within a group) and radical (between groups) changes were calculated along individual branches across the entire tree. Comparison of changes along a specific branch with those across the rest of the tree allowed us to test for significantly different patterns of amino acid substitution during different periods of

evolutionary time (branches) following Merritt and Quattro (2001). A significant difference in the pattern of radical amino acid changes between a branch and the whole tree would suggest that, across this branch, sequences evolved under different selective pressure than across the rest of the tree. The pattern of charge change along the large *Tpi* neural branch, the evolutionary period directly following the duplication event and leading to evolution of the neural form, was significantly different from the rest of the gene tree (Fisher’s exact test, Table 2), consistent with previous results (Merritt and Quattro 2001). Eight radical amino acid charge changes occurred along the neural branch, six of which are to negative amino acids while two are from positive to neutral amino acids; an overall substantial increase in negative charge of the protein. When amino acids were classed by size or polarity, no significant difference was found between the neural *Tpi* branch and the whole tree (Table 2, also consistent with previous results, Merritt and Quattro 2001).

**Table 2** Amino acid alterations that alter or conserve given amino acid properties along the neural branch in comparison to the remaining tree for *Tpi*, *Guk*, *Mdh*, and *Pgi*

	Charge		Polarity		Size	
	Conserve	Alter	Conserve	Alter	Conserve	Alter
<i>Tpi</i>						
Neural branch	(7) 8	(8) 9	(10) 13	(3) 4	(10) 12	(3) 5
Tree	(236) 102	(95) 35	(225) 93	(106) 44	(258) 104	(73) 33
<i>P</i> value		(0.044*) 0.022*		(0.37) 0.34		(0.58) 0.41
<i>Guk</i>						
Neural branch	(12) 8	(6) 7	(15) 12	(3) 3	(18) 14	(0) 1
Tree	(296) 92	(187) 72	(333) 114	(150) 50	(398) 127	(85) 37
<i>P</i> value		(0.42) 0.52		(0.15) 0.30		(0.033*) 0.13
<i>Mdh</i>						
Neural branch	9	3	8	4	10	2
Tree	136	35	115	56	142	29
<i>P</i> value		0.47		0.60		0.67
<i>Pgi</i>						
Neural branch	18	5	20	3	20	3
Tree	230	131	263	98	294	67
<i>P</i> value		0.12		0.10		0.37

Amino acid changes along the neural branch and rest of the gene tree based on ML reconstructions under the free ratio model. Amino acid change are based on size, polarity, or charge as classified using IMGT system (Pommie et al. 2004). Bracketed values represent the corresponding large datasets alterations, while the unbracketed values indicate changes from the comparison datasets

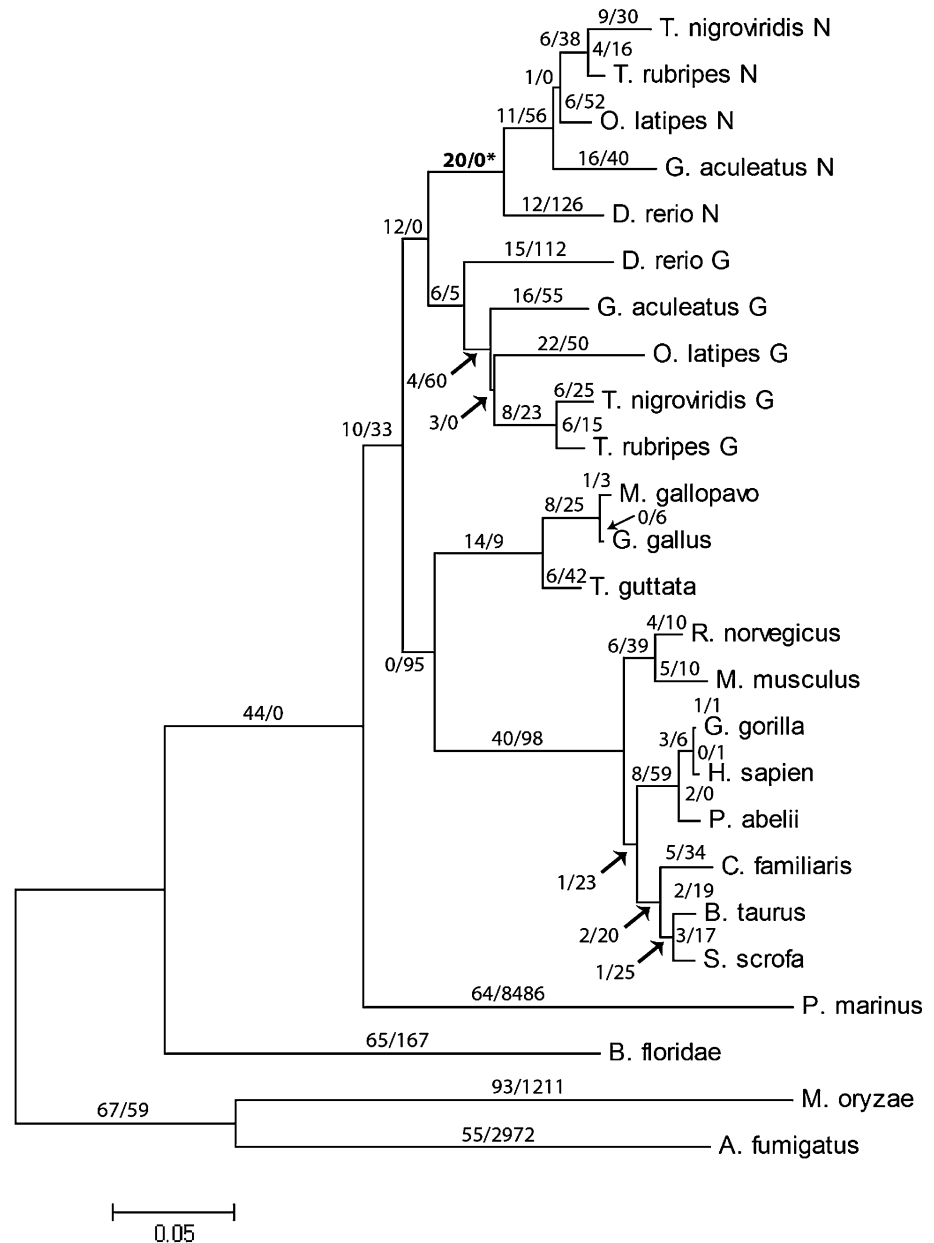
\* Significant at  $P < 5\%$ , \*\* significant at  $P < 1\%$  (Fisher's exact test)

### Guanylate Kinase

The large *Guk* dataset includes 25 sequences from a variety of vertebrates including 10 teleost fish (post-genome duplication fish; Online Resource Table 2). No pre-duplication, basal fish *Guk* sequences are currently publically available. Early enzyme electrophoresis studies (Morizot et al. 1991; Gaida 1995) suggest that similar to the *Tpi* family, the *Guk* enzyme family in fish has a negatively charged, neural, isozyme and a neutrally charged, generally expressed, isozyme. In agreement with these papers, our phylogenetic analysis (Online Resource Fig. 2) indicates two well-supported subfamilies within the teleost *Guk* sequences and a single *Guk* in all non-fish taxa. Predicted isoelectric points of the GUK proteins indicate that all the ancestral locus proteins (pre-duplication) and one of the fish subfamilies are neutrally charged, while one of the fish subfamilies has a pronounced negative charge (Fig. 4). Based on this charge differential, we have identified the negatively charged subfamily as “neural” and the neutral subfamily as “general.” NJ analysis of the large *Guk* dataset produces a tree with identical large-scale topology as that found for *Tpi* (Figs. 1, 2; Merritt and Quattro 2001), *Mdh* (Merritt and Quattro 2003), and *Pgi* (Sato and Nishida 2007), consistent with a single, similarly timed, duplication event producing the neural and general orthologs in teleost

fish (Online Resource Fig. 2). ML and Bayesian trees both support separate neural and general subfamilies, but place the neural subfamily at the base of the vertebrate tree although with weak bootstrap support (Online Resource Fig. 2). Placement of the neural branch at the root of the tree requires two independent gene duplication events, followed by two subsequent losses. Conversely, the NJ tree only requires a single duplication event and no losses. This simpler pattern, and its agreement with studies of a series of other genes suggesting a genome duplication early in teleost evolution (Jaillon et al. 2004; Meyer and Van De Peer 2005), strongly indicates that the NJ tree correctly reflects the evolution of this gene family and this tree was used in all analyses. The basal placement of the negative, neural *Guk* subfamily, may be the result of long-branch attraction (Felsenstein 1978) although if this is the case it is surprising that our distance-based method, but not ML or Bayesian methods, recovered the correct topology (Huel- senbeck and Hillis 1993; Huelsenbeck 1995; Philippe 2000). Placement of neural *Guk* at the base of the tree by ML and Bayesian analysis may instead be caused by the lack of basal (pre-duplication) fish species in this dataset. The addition of basal fish species, such as gar, bowfin, or sturgeon, may resolve this issue and future work will investigate the evolution of these genes in relatively ancient fish species. There are a few minor differences in

**Fig. 2** ML tree depicting nucleotide relationship among large *Guk* dataset sequences. The number of nonsynonymous/synonymous substitutions is shown above each branch. Boldface asterisk values represent significantly different nonsynonymous/synonymous change in comparison to nonsynonymous/synonymous sites by Fisher's exact test ( $P < 5\%$ )



placement of the tetrapod sequences between ML, Bayesian, and NJ methods (e.g., Bayesian and ML method swap *O. latipes* and *T. rubripes*), but these small changes were not found to affect amino acid or nucleotide results when tested (data not shown).

Both comparisons of the number of observed and potential substitutions and  $\omega$  ratio tests suggest that the neural *Guk* gene evolved through a period of evolution similar to that of the *Tpi* isozyme. Our analysis of the observed ( $n/s$ ) and potential ( $N/S$ ) changes along the neural branch of the large *Guk* tree estimated by ML methods indicates 20 nonsynonymous and no synonymous changes along this branch (Fig. 2), which is significantly different from potential sites (399/114). This pronounced excess of

nonsynonymous change is almost identical to the pattern along the neural branch of the large *Tpi* tree (Fig. 1), suggesting that the neural *Guk* locus, like the neural *Tpi* locus, evolved through a period of positive selection following the duplication event. Results from our log likelihood ratio tests of the large *Guk* dataset are similar to those from the large *Tpi* dataset. The neural  $\omega$  ratio is significantly different from that of the rest of the tree (rightmost column—compared models A–C, and B–E, Table 3). However, as in the *Tpi* dataset, the direct likelihood ratio tests for  $\omega > 1$  (rightmost column models C–H and E–J) do not indicate evidence of positive selection acting along the neural branch of the large *Guk* tree (Table 3), again leaving open the possibility that the neural isozyme evolved



through a period of neutrality, not positive selection. Interestingly, likelihood estimates of  $\omega n$  (neural  $dn/ds$ ) for all models which allow the  $\omega n$  to vary do, however, produce  $\omega$  values greater than one (free ratio, three ratio, and two ratio— $\omega o = \omega g$ ,  $\omega n$  models, Table 3).

As in the large TPI dataset analysis, we used inferred ancestral protein sequences, and modern sequences to reconstruct all amino acid changes across each branch of the large GUK tree. We find a large number of radical amino acid charge changes along the neural branch although charge change along this branch is not statistically different to that of the entire large GUK tree (Table 2). This lack of significant difference appears to reflect the relatively large amount of charge change across the entire GUK tree, not a lack of charge change across the neural branch. We identified six radical charge changes, lowering the isoelectric point of the protein: five changes to negative amino acids and one from positive to a neutral amino acid. Across the rest of the tree, we identified 483 changes, 187 that involved charge, i.e., 39 % of changes involved charge compared with 29 % in TPI. These numbers are significantly different between *Guk* and *Tpi* (Fisher's exact test,  $P = 0.0019$ ). The pattern of amino acid change along the neural branch was also not different from that across the

rest of the tree when amino acids were classified by size or polarity. We lack a basal, single gene, fish GUK sequence in our analysis and it is possible that the absence of this sequence results in the incorrect assignment of charge changes across our tree, but the wide distribution of the observed changes (i.e., changes are not restricted to a few branches, but instead spread across the tree) makes this unlikely. Future research will include such basal species sequences in our analysis.

### Comparison Datasets

Comparison datasets for *Guk*, *Tpi*, *Mdh*, and *Pgi* include eight sequences: four tetrapods, and four sequences from two post-duplication teleosts (Online Resource Table 3). Phylogenetic analysis was performed by NJ, Bayesian, and ML methods, identical to that of the large *Tpi* and *Guk* datasets. The comparison *Guk* and *Tpi* datasets produce the same trees by all three phylogenetic methods (Online Resource Fig. 3) with bootstrap support generally greater than 80 % for all nodes. Both the *Mdh* and *Pgi* comparison datasets produce the same tree as compared to the *Guk* and *Tpi* datasets with a single duplication immediately before the teleost radiation. Two of the three phylogenetic

**Table 3** ML tests and estimates for the large and comparison *Guk* datasets

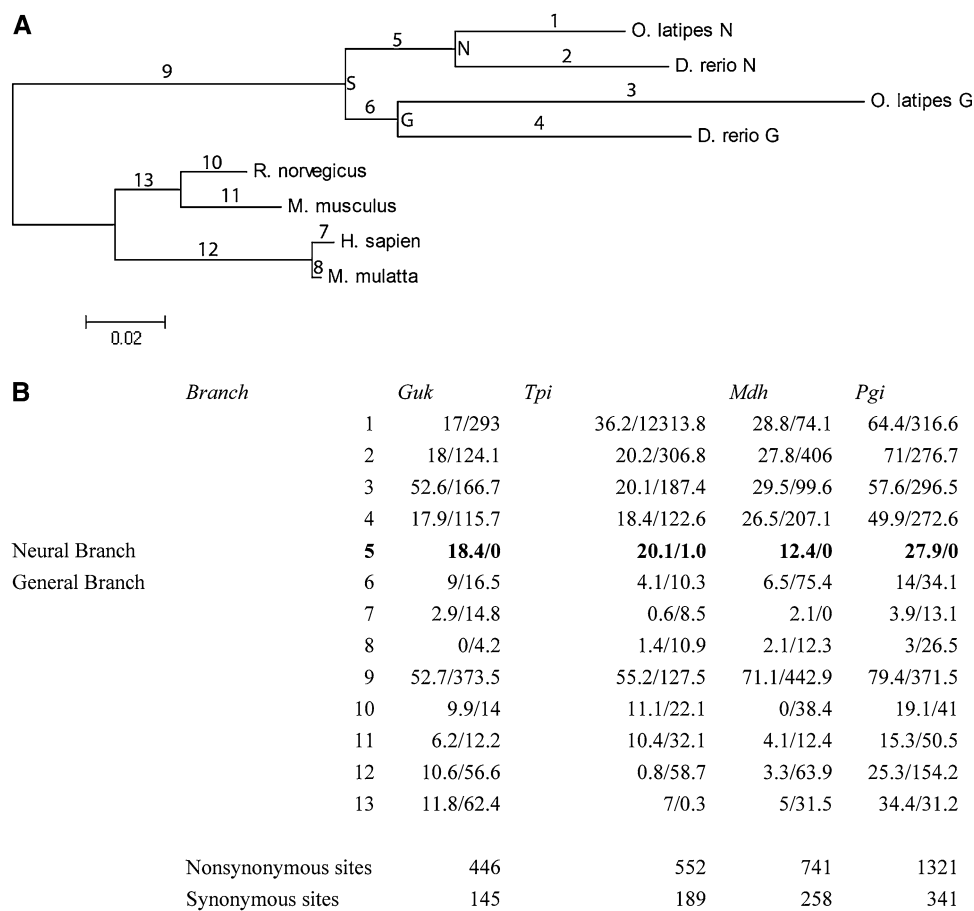
	Models	Sub-models	Likelihood values	$\omega n$	$n/s$	Compared models	Test	Likelihood ratio test
A	One ratio	$\omega b = \omega n = \omega g$	−2910.12	0.0666 (0.064)	14.3/69.9 (10.9/48)	A–C	$\omega b = \omega n$	3.02 (12.14**)
B	Two ratio	$\omega b = \omega n, \omega g$	−2909.36	0.0642 (0.63)	14.6/74.6 (11/50.1)	B–E	$\omega b = \omega n$	1.34 (10.14**)
C	Two ratio	$\omega b = \omega g, \omega n$	−2908.61	$\infty$	17.5/0 (19.2/0)	A–D	$\omega g$ and $\omega n = \omega b$	2.98 (11.82**)
D	Two ratio	$\omega b, \omega g = \omega n$	−2908.63	0.2366 (0.78)	17.4/24 (18.5/6.8)	A–B	$\omega b = \omega g$	1.52 (2.66)
E	Three ratio	$\omega b, \omega g, \omega n$	−2908.69	0.2136 ( $\infty$ )	17.5/26.8 (18.9/0)	C–E	$\omega b = \omega g$	0.16 (0.66)
F	Free ratio		−2897.04	$\infty$ (47.1)	18.4/0 (19.6/0.1)			
G	Two ratio	$\omega b = \omega n, \omega g = 1$	−2909.40	0.0643 (0.063)	14.5/74 (10.9/49.6)	D–I	$\omega g$ and $\omega n > 1$	0.58 (0.02)
H	Two ratio	$\omega b = \omega g, \omega n = 1$	−2908.63	1 (1)	17.5/5.7 (18.8/5.3)	C–H	$\omega n > 1$	0.04 (0.54)
I	Two ratio	$\omega b, \omega g = \omega n = 1$	−2908.92	1 (1)	17.7/5.8 (18.7/5.3)	E–J	$\omega n > 1$	0.22 (0.36)
J	Three ratio	$\omega b, \omega g, \omega n = 1$	−2908.58	1 (1)	17.5/5.7 (18.6/5.3)	B–G	$\omega g > 1$	0.08 (0.08)
K	Three ratio	$\omega b, \omega g = 1, \omega n$	−2908.69	0.192 (3.01)	17.5/29.8 (18.8/1.8)	E–K	$\omega g > 1$	0 (0.94)

Refer to Table 1 for explanation of all models and tests

Bracketed values indicate results from the large *Guk* dataset, while unbracketed values correspond to the comparison *Guk* dataset

\* Significant  $P < 5$  %, \*\* significant at  $P < 1$  %

**Fig. 3** Common *Guk*, *Tpi*, *Mdh*, and *Pgi* datasets produce the same tree topology shown in (a). Labelled branches along this common tree correspond to the number of *n/s* changes shown in (b) for *Guk*, *Tpi*, *Mdh*, and *Pgi*. Boldface values represent the neural branch *n/s*, all of which contain significantly elevated levels of nonsynonymous change by Fisher's exact test. The numbers of nonsynonymous and synonymous sites for each tree is shown at the bottom of the table



methods produced the “correct” tree for both comparison *Mdh* and *Pgi* with generally over 70 % bootstrap support (Online Resource Fig. 3). Surprisingly, the ML tree topology for *Mdh* and Bayesian tree topology for *Pgi* do not resolve the duplication although previous analysis of larger datasets did (*Mdh*, Merritt and Quattro 2003; *Pgi*, Sato and Nishida 2007). All subsequent analyses were conducted using “correct” trees resolving the duplication (Online Resource Fig. 3).

As in the analyses of the larger *Tpi* and *Guk* datasets, above, we compared observed (*n/s*) to potential (*N/S*) nonsynonymous and synonymous change across the four comparison datasets. We found significantly higher amounts of observed nonsynonymous change along the neural branches of all four comparison datasets (Fig. 3). In this analysis, the human *Tpi* sequence was excluded from the comparison *Tpi* data because inclusion of the human sequence resulted in an unexpectedly large amount of synonymous change along the neural branch, inconsistent with both the large *Tpi* dataset and earlier findings by Merritt and Quattro (2001). Exclusion of the human *Tpi* sequence results in a pattern consistent with both the large *Tpi* dataset and previous work done by Merritt and Quattro (2001). Amino acid change and likelihood ratio tests both

produce identical results with or without the human *Tpi* sequence in the comparison dataset (not shown). The human *Tpi* nucleotide sequence is somewhat diverged from the general teleost *Tpi* sequences which may cause the inconsistent results when the human *Tpi* sequence is included in the smaller “comparison” dataset.

We also compared log likelihood values to test for differences in  $\omega$  ratios, and evidence of  $\omega > 1$ , along the neural branches of the comparison *Tpi*, *Guk*, *Mdh*, and *Pgi* datasets. In general, the tests suggest differences between the neural branches and their corresponding trees, but the patterns are complicated and in no case is there unambiguous evidence of positive selection. In one test, we find support for  $\omega$  being greater than one along the neural *Tpi* branch (rightmost models C–H and E–J, Table 1), but in another test we find no support that this  $\omega$  value is different from that across the rest of the tree (A–C and B–E, Table 1). The  $\omega$  ratios along the neural branch of the *Mdh*, and *Pgi* trees are all significantly different from those across the rest of their corresponding trees, but neither are significantly greater than one (Tables 4, 5), leaving open the possibility of a period of neutrality, not positive selection. In the analysis for the comparison *Guk* dataset, the neural value is only marginally significantly different

**Table 4** ML tests and estimates from the comparison *Mdh* dataset

	Models	Sub-models	Likelihood values	$\omega n$	$n/s$	Compared models	Test	Likelihood ratio test
A	One ratio	$\omega b = \omega n = \omega g$	-4236.91	0.055	8.0/51.3	A–C	$\omega b = \omega n$	6.18*
B	Two ratio	$\omega b = \omega n, \omega g$	-4236.72	0.0532	8.2/54.4	B–E	$\omega b = \omega n$	5.86*
C	Two ratio	$\omega b = \omega g, \omega n$	-4233.82	451.99	12.9/0	A–D	$\omega g$ and $\omega n = \omega b$	4.58*
D	Two ratio	$\omega b, \omega g = \omega n$	-4234.62	0.2304	12.4/18.9	A–B	$\omega b = \omega g$	0.38
E	Three ratio	$\omega b, \omega g, \omega n$	-4233.79	$\infty$	12.9/0	C–E	$\omega b = \omega g$	0.06
F	Free ratio		-4216.27	279.07	12.4/0			
G	Two ratio	$\omega b = \omega n, \omega g = 1$	-4236.78	0.0519	8.5/57.8	D–I	$\omega g$ and $\omega n > 1$	0.94
H	Two ratio	$\omega b = \omega g, \omega n = 1$	-4233.86	1	12.7/4.5	C–H	$\omega n > 1$	0.08
I	Two ratio	$\omega b, \omega g = \omega n = 1$	-4235.09	1	13.5/4.7	E–J	$\omega n > 1$	0.12
J	Three ratio	$\omega b, \omega g, \omega n = 1$	-4233.85	1	12.7/4.5	B–G	$\omega g > 1$	0.12
K	Three ratio	$\omega b, \omega g = 1, \omega n$	-4234.85	0.2726	12.9/16.6	E–K	$\omega g > 1$	2.12

Refer to Table 1 for explanation of all models and tests

\* Significant  $P < 5\%$ , \*\* significant at  $P < 1\%$

**Table 5** ML tests and estimates from the comparison *Pgi* dataset

	Models	Sub-models	Likelihood	$\omega n$	$n/s$	Compared models	Test	Likelihood ratio test
A	One ratio	$\omega b = \omega n = \omega g$	-7460.19	0.0664	18.8/73	A–C	$\omega b = \omega n$	11.44**
B	Two ratio	$\omega b = \omega n, \omega g$	-7460.05	0.1112	13.6/31.5	B–E	$\omega b = \omega n$	11.2**
C	Two ratio	$\omega b = \omega g, \omega n$	-7454.47	$\infty$	27.9/0	A–D	$\omega g$ and $\omega n = \omega b$	7.7*
D	Two ratio	$\omega b, \omega g = \omega n$	-7456.34	0.403	27.2/17.4	A–B	$\omega b = \omega g$	0.28
E	Three ratio	$\omega b, \omega g, \omega n$	-7454.45	$\infty$	27.9/0	C–E	$\omega b = \omega g$	0.04
F	Free ratio		-7444.46					
G	Two ratio	$\omega b = \omega n, \omega g = 1$	-7460.43	0.064	19.8/79.8	D–I	$\omega g$ and $\omega n > 1$	0.2
H	Two ratio	$\omega b = \omega g, \omega n = 1$	-7454.78	1	27.4/7.1	C–H	$\omega n > 1$	0.62
I	Two ratio	$\omega b, \omega g = \omega n = 1$	-7456.44	1	27.9/7.2	E–J	$\omega n > 1$	0.64
J	Three ratio	$\omega b, \omega g, \omega n = 1$	-7454.77	1	27.4/7.1	B–G	$\omega g > 1$	0.76
K	Three ratio	$\omega b, \omega g = 1, \omega n$	-7456.36	$\infty$	28.1/0	E–K	$\omega g > 1$	3.82

Refer to Table 1 for explanation of all models and tests

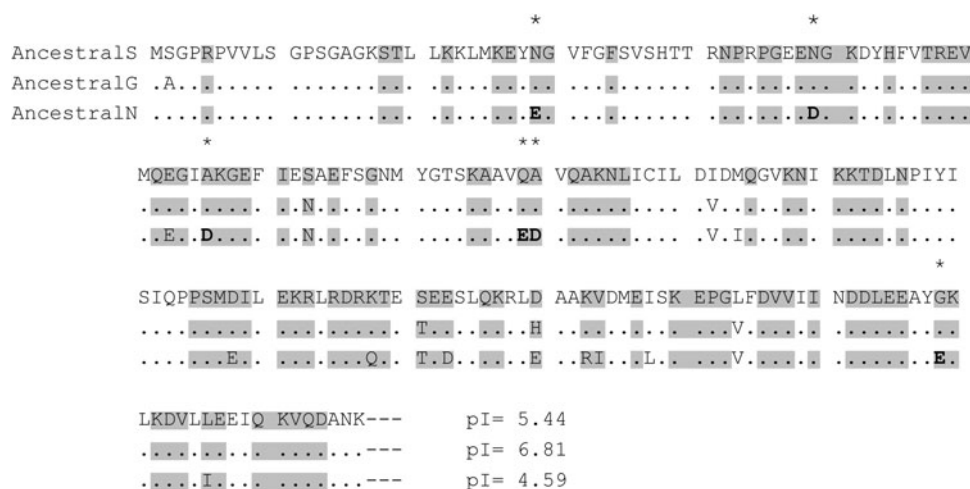
\* Significant  $P < 5\%$ , \*\* significant at  $P < 1\%$

( $0.05 < P < 0.10$ ) from the rest of the tree (rightmost models A–C and B–E, Table 3). Given the comparison using the larger *Guk* dataset (above, Table 3), we are, however, confident that the  $\omega n$  ratio is different from all other branches of the *Guk* tree like that of *Mdh* and *Pgi*. In addition, all models examining the four comparison datasets which allow the neural  $\omega n$  to vary freely, find  $\omega$  ratios to be greater than one, consistent with positive selection along these branches (Tables 1, 3, 4, 5).

We compared amino acid changes along the neural branches with changes across the rest of the trees for each of the four gene families using modern and inferred protein sequences. There is substantial charge change along the neural branches of both the comparison *Tpi* and *Guk* datasets, but, as in our analysis of the larger datasets, only the changes along the neural *Tpi* branch were significantly different from the rest of the tree ( $P = 0.022$  and Table 2).

Eight substitutions conserving charge and nine changing charge, occurred along the neural *Tpi* branch. Across the rest of the tree, we find that 102 changes conserved amino acid charge, while 35 changes altered it. If we correct for multiple hypothesis testing with a simple Bonferroni's correction (a conservative correction; Anisimova and Yang 2007), the pattern of charge change is suggestive, but not quite significant (Bonferroni's  $P = 0.0167 < 0.022$ , Table 2). In the *Guk* dataset, we observed a similar amount of charge change along the neural branch, seven substitutions change charge, while eight changes maintain charge, however, these numbers are not significantly different from those across the rest of the tree (Fisher's exact test,  $P = 0.52$ ), where 72 change charge and 92 maintain charge (Table 2). The large amount of charge change across the *Guk* tree (in both the large and comparison dataset analyses) suggests that charge change is relatively

**Fig. 4** (*Guk*) Inferred amino acid sequences for the duplication node (AncestralS), node leading to the neural (AncestralN) and general (AncestralG) isozyme branches of the common *Guk* tree. Highlighted amino acids were found to reside on the outside of the protein by solvent accessibility of >25 %. Asterisks indicate amino acid substitutions toward negative amino acids, and isoelectric points are shown as the 3' end of the proteins



common in this gene family (78 % of all substitutions are radical charge changes as compared with 34 and 40 % in common and large *Tpi*, 26 % in *Mdh*, 57 % in *Pgi*). No significant differences were found when amino acids were grouped by charge for the *Mdh* and *Pgi* datasets or when amino acids were grouped by size or polarity in any of the four gene families.

In all four gene families, the predicted neural isozymes are all significantly more negatively charged (i.e., have lower isoelectric points) than their non-neural counterparts (one sample Z test; Fig. 5). We also inferred the location within the protein structure of the amino acids responsible for this charge differential by comparison of modern and reconstructed ancestral protein sequences with modern sequences with known 3D structures. All predicted amino acid charge changes occurred on the outside of the protein structure for the comparison *Guk* (Fig. 4), *Tpi* (Online Resource Fig. 4), and *Mdh* (Online Resource Fig. 5) datasets. Similarly, all charge changes predicted for the large *Guk* and *Tpi* datasets also occurred only on the surface of the protein (data not shown). Surprisingly, our reconstructions do not predict any unique negatively charged amino acids in comparison of the neural PGI with the single gene or generally expressed form (data not shown), despite an overall reduced isoelectric point of the neural *Pgi* isozyme.

## Discussion

In summary, our analysis of the patterns of nucleotide and amino acid change suggests that the neural isozymes from the four gene families, *Guk*, *Tpi*, *Mdh*, and *Pgi*, evolved through a period of distinct selective pressure. Our goal was to look for general patterns or trends across these gene families with distinct biochemical functions in the evolution of “neural” isozymes, and use this multi-gene

comparative approach to build on the model of evolution of these isozymes proposed by Merritt and Quattro (2001). In each of the four families, we found significantly elevated rates of nonsynonymous substitution, and distinct  $\omega$  ratios, along the branch leading to the neural isozyme when compared to the branches across the rest of tree. Both *Guk* and *Tpi* seem to have substantial amino acid charge change along this branch, suggesting positive selection along the neural branch, but neutral evolution could not be ruled out in all cases. In all four gene families, the neural isozymes have significantly lower predicted isoelectric points than their generally expressed counterparts, but the charge differential is noticeably more pronounced in TPI and GUK than in MDH or PGI. This difference in charge differential across the gene families may reflect variations in selective pressures or functional constraints across the gene families, possibly reflected in different expression patterns between isozymes across the four gene families (Merritt and Quattro 2001, 2003; Sato and Nishida 2007).

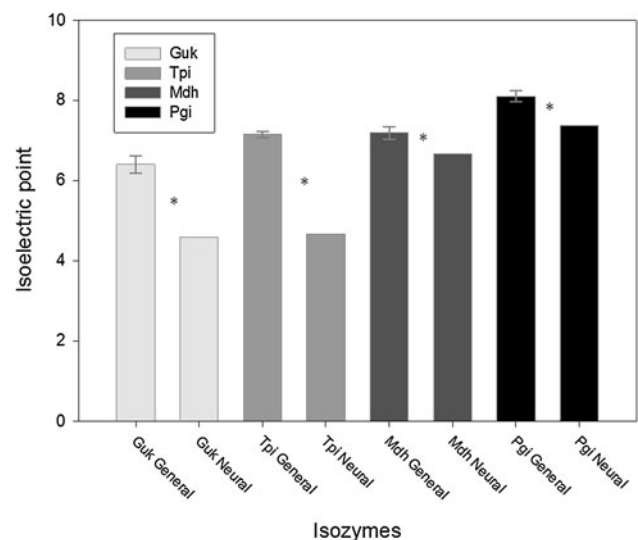
Since the initial examination of the molecular evolution of the neural TPI isozyme in fish (Merritt and Quattro 2001), considerable support has emerged for a complete genome duplication after the split of fish and tetrapods, likely early in teleost evolution (e.g., Amores et al. 1998; Postlethwait et al. 1998; Jaillon et al. 2004; Meyer and Van De Peer 2005). Our phylogenetic analysis of all four gene families, and previous work on the *Tpi*, *Mdh*, and *Pgi* gene families (Merritt and Quattro 2001, 2003; Sato and Nishida 2007, respectively) supports a common duplication event hypothesis for these isozymes, likely the whole genome duplication event implicated in teleost evolution. The neural isozymes in each family have then been evolving for the same period of time and differences and similarities observed across the gene families should represent differences and similarities in selective constraint or pressure, not timing or duration of selective pressure. As such, the patterns of DNA and protein change that we document here

suggest that the *Tpi* and *Guk* gene families likely share more similar protein function, and likely patterns of expression, than either do with *Mdh* or *Pgi*. Our analysis of the large *Guk* dataset includes teleost fish, tetrapods, and ancestral chordate sequences, but lacks a basal fish sequence as used in earlier studies of teleost neural isozymes (Merritt and Quattro 2001, 2002, 2003; Sato and Nishida 2007). Based on these earlier studies, we expect that basal fish, having diverged from more derived fish lineages before the genome duplication, will contain a single *Guk* locus. In our initial examination of the evolution of the neural isozymes, we used the ancestral single gene taxa, to help us isolate the evolutionary period (branch) of interest—that directly follows the duplication event. Given the central role of these single-gene taxa in defining our period of interest, it is impressive that even without a single-gene fish taxa, our analysis is able to find evidence for a distinct pattern of evolution, implying difference in selective pressures, along the neural branch in the *Guk* gene family. Presumably, inclusion of a single *Guk* gene (basal fish taxa) would only strengthen this result, and future research including a broader range of species will test this point.

In each of the datasets, by examining DNA-based likelihood ratio and nucleotide change tests, we find evidence suggesting that the neural isozymes evolved differently than other sequences in the trees. Overall, our results suggest that the isozymes likely evolved through a period of positive selection, but we cannot rule out a period of neutral evolution in all cases from the DNA-based tests alone. In general, inferring the focus of selective pressure from DNA-based tests is difficult or impossible (see Hughes and Friedman 2008; Yokoyama et al. 2008) so we also examined the patterns of amino acid change in each gene family. Merritt and Quattro (2001) found an accumulation of net negative charge (change to negatively charged amino acids) along the branch leading to the TPI neural isozyme, and inferred that an increase in net negative charge in the neural isozyme was selected for in this gene family. In fact, all four neural isozymes show significantly lower isoelectric points than their generally expressed, or single isozyme counterparts (Fig. 5, and Fisher et al. 1980), suggesting that selection was toward the accumulation of net negative charge in all four families. To test this possible source of the selective pressure, we quantified the pattern of amino acid change for each of the gene families. Consistent with the earlier analysis, we find significantly elevated accumulation of negative charge along the neural branch in both large and comparison *Tpi* trees. We also found a large amount of change toward negative charge along the *Guk* neural branch, but overall charge change across whole *Guk* tree is also greater than found in *Tpi* or the other gene families. Given the pattern in

the *Tpi* gene family and the similarities in the *Guk* gene family, it is likely that the neural *Guk* isozyme also evolved through a period of positive selection similar to the neural *Tpi*, but the large amount of charge change across the *Guk* tree makes it impossible to completely rule out relaxed selective constraint and neutral evolution along this branch. The *Mdh* and *Pgi* neural branches show random patterns of amino acid change when amino acids are grouped by charge (or size or polarity) even though the neural sequences are more negatively charged than their generally expressed counterparts. The charge differentials are less pronounced in these two gene families than either the *Tpi* or *Guk* gene families, and the expression patterns seem to be broader, less canalized, at least than that of *Tpi*, suggesting less pronounced functional differences between the isozymes. In sum, our analysis suggests that both the *Tpi* and *Guk* neural isozymes evolved under similar selective pressures, while selection on the *Mdh* and *Pgi* neural isozymes seems to have been distinct and, perhaps, less pronounced.

Two lines of evidence suggest that the accumulation of negative charge in the neural isozymes is from positive, or directional, selection and not simply a random accumulation of amino acid changes under neutrality. First, in all four gene families, selective pressures return to purifying selection, in which significantly lower than random amounts of nonsynonymous substitution occur, after the relatively brief period of elevated nonsynonymous substitution, suggesting that the proteins are then functional. It seems unlikely, though not impossible, that all four isozymes would have evolved



**Fig. 5** Isoelectric points of the “neural” and general *Guk*, *Tpi*, *Mdh* and *Pgi* isozymes. The neural isoelectric points are calculated from the reconstructed neural node, while the general isoelectric points are averaged from all general sequences as well as reconstructed general nodes. Statistical significance indicated as asterisks were tested by one sample Z test



through periods of neutrality, a lack of selective constraint suggesting a non-functional protein, only to return to function, and purifying selection, once a negative charge had been randomly established. In addition, in TPI, GUK, and MDH, the negatively charged amino acids are found only on the surface of the proteins (Fig. 4; Online Resource Figs. 4, 5, we found no unique charged amino acid change in the PGI sequence). Charge changes are likely to have large impact on 3D structure if located within the interior of a protein. In functional proteins, then, we expect to find changes restricted to the surface, while in non-functional proteins changes are likely to occur throughout the protein. The observed restriction, of all amino acid charge changes to the surface of the protein suggests that the changes occurred in functional proteins, and that the amino acid changes are of functional importance (i.e., the charge differential between the isozymes reflects functional divergence). Combined, the pronounced accumulation of negative charge, the restriction of charge changes to the surface of the proteins, and the return to purifying selection after a period of increased nonsynonymous change, supports our claim that the neural isozymes are functional proteins which have evolved a potentially novel function through a period of positive selection after the duplication event. The actual function of the neural isozymes is unknown so it is difficult to conclusively say if these proteins have evolved to be better adapted to a function within the original protein function (i.e., subfunctionalization) or to a completely novel function (i.e., neofunctionalization). The novel charge and distinct pattern of expression of the neural isozymes, however, suggests that these proteins are an example of neofunctionalization (Ohno 1970; Lynch and Force 2000). The continued expression of the general isozyme in the neural tissues (both isozymes are present) also suggests that the neural isozyme have not simply replaced the general form in these tissues further supporting neofunctionalization and not subfunctionalization. Evidence for a combination of both sub- and neofunctionalization has been suggested (He and Zhang 2005), and neofunctionalization has been found in yeast, fruit flies, snakes, humans, and teleost evolution to name a few (He and Zhang 2005; Lynch 2007; Beisswanger and Stephan 2008; Douard et al. 2008). The neural teleost isozymes, especially the examples with the most pronounced divergence, *Tpi* and *Guk*, appear to likely be a further example of neofunctionalization following duplication. Further examination of functional difference between the isozymes will allow explicit testing of this possibility.

Similarities and differences in the pattern of sequence change suggests that the four gene families can be placed in two groups with the *Tpi* and *Guk* families showing more pronounced patterns of nucleotide and amino acid change—and more negatively charged neural isozymes—than *Mdh* or *Pgi*. This divergence in patterns suggests that

the neural isozymes within either group likely share more similar functions than between groups, a statement supported by the limited expression data available so far. By means of rather crude presence/absence PCR assays, Merritt and Quattro (2001) demonstrated that the neural *Tpi* gene was expressed in the eye, brain, and ovary, but not in a variety of other tissues. In contrast, they found that the “neural” *Mdh* gene was expressed in muscle as well as the eye, brain, and ovary (Merritt and Quattro 2003). In both gene families, the general gene was found to be expressed in all tissues examined. By means of more sensitive and accurate qPCR assays, Sato and Nishida (2007) found that the “neural” *Pgi* gene was also more broadly expressed than the neural *Tpi*. Expression data is not available for *Guk*, but based on the results shown here we predict that the neural *Guk* gene, like the neural *Tpi* gene, will have a relatively canalized pattern of expression. Future research on the expression patterns of the *Guk* isozymes is warranted. Variation in isozyme expression is also found in other similar gene families in fish. The fish *LDH* gene family contains three isozymes, likely a result of the same duplication that created the *Tpi*, *Pgi*, *Mdh*, and *Guk* gene families and a second, more recent, duplication (Quattro et al. 1993). The LDH isozymes not only have significantly different biochemistry (Whitt 1970; Markert et al. 1975), implying functional divergence but also appear to vary in expression across taxa making the evolution of this gene family difficult to reconstruct. The biochemical specialization of the *Ldh* isozymes supports our proposal of functional divergence of the isozymes. The variety of similar gene families suggests that a broader bioinformatic search for additional gene families, including those that would have been missed by traditional isozyme and gel electrophoretic studies is warranted.

Expression studies and further examination of basal, single gene, fish taxa may help to unravel the remaining functional mystery of this system: the biological significance of the negative charge of these isozymes. The pattern of sequence change, DNA and amino acid, and especially the surface localization of the charge changes strongly suggest that this charge has a functional impact, but the nature of this impact remains elusive. Merritt and Quattro (2001) speculated that the negative charge may be involved with repolarization of neural tissues. The correlation of weaker negative charge with broader expression pattern is consistent with this speculation, but certainly not definitive. A finer-scale understanding of the expression patterns of all four “neural” isozymes and their generally expressed counterparts, and perhaps some biochemical analysis on the functionality of these isozymes will hopefully shed light on the functional differences our analyses suggest.

**Acknowledgments** We thank Teresa Rzezniczak, David Bing, Jose Knee, Brendan McConkey, Katharine Coykendall, and Efe Sezgin for constructive comments on this manuscript, as well as Jennifer Fenske and Stefan Siemann for phylogenetic and protein structure assistance. This study was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery (3414-07) and a Canada Research Chair (950-215763) to TJSM.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Agarwal KC, Miech RP, Parks RE Jr (1978) Guanylate kinases from human erythrocytes, hog brain, and rat liver. *Methods Enzymol* 51:483–490
- Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410
- Amores A, Force A, Yan Y, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang Y, Westerfield M, Ekker M, Postlethwait JH (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science* 282(27):1711–1714
- Anisimova M, Yang Z (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol* 24(5):1219–1228
- Beisswanger S, Stephan W (2008) Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. *Proc Natl Acad Sci USA* 105(14):5447–5452
- Bollback JP (2002) Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol* 19(7):1171–1180
- Champion MJ, Whitt GS (1976) Differential gene expression in multilocus isozyme systems of the developing green sunfish. *J Exp Zool* 196(3):263–281
- Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303
- Chiu C-h, Dewar K, Wagner GP, Takahashi K, Ruddle F, Ledje C, Bartsch P, Scemama J-L, Stellwag E, Fried C, Prohaska SJ, Stadler PF, Amemiya CT (2004) Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res* 14(1):11–17
- Daly LE, Bourke GJ (2008) Interpretation and uses of medical statistics, 5th edn. Blackwell Science, Oxford
- Douard V, Brunet F, Boussau B, Ahrens-Fath I, Vlaeminck-Guillem V, Haendler B, Laudet V, Guiguen Y (2008) The fate of the duplicated androgen receptor in fishes: a late neofunctionalization event? *BMC Evol Biol* 8(1):336
- Dykhuizen D, Hartl DL (1980) Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics* 96:801–817
- Elkin BS, Shaik MA, Morrison B (2010) Fixed negative charge and the donnan effect: a description of the driving forces associated with brain tissue swelling and oedema. *Philos Trans R Soc A* 368(1912):585–603
- Felsenstein J (1989) PHYLIP—phylogeny inference package. *Cladistics* 5 Version 3.2:164–166
- Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27(4):401–410
- Fisher SE, Whitt GS (1978) Evolution of isozyme loci and their differential tissue expression creatine kinase as a model system. *J Mol Evol* 12(1):25–55
- Fisher SE, Shaklee JB, Ferris SD, Whitt GS (1980) Evolution of five multilocus isozyme systems in the chordates. *Genetica* 52–53(1):73–85
- Gaida IH (1995) Evolutionary aspects of gene expression in the Pacific Angel Shark, *Squatina California* (Squatiniformes: Squatinidae). *Copeia* 3:532–554
- Goodman M, Moore GW, Matsuda G (1975) Darwinian evolution in the genealogy of haemoglobin. *Nature* 253(5493):603–608
- He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169(2):1157–1164
- Huelsenbeck JP (1995) The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol Biol Evol* 12(5):843–849
- Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the four-taxon case. *Syst Biol* 42(3):247–264
- Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proc R Soc B* 256(1346):119–124
- Hughes AL, Friedman R (2008) Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. *Immunogenetics* 60(9):495–506
- Jaillon O, Aury J, Brunet F, Petit J, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau J, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volf J, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crolius H (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431(7011):946–957
- Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983):617–624
- Kettler MK, Whitt GS (1986) An apparent progressive and recurrent evolutionary restriction in tissue expression of a gene, the lactate dehydrogenase-c gene, within a family of bony fish (Salmoniformes: Umbridae). *J Mol Evol* 23(2):95–107
- Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
- King JL, Jukes TH (1969) Non-Darwinian evolution. *Science* 164:788–798
- Kosiol C, Vinař T, Da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet* 4(8):E1000144–E1000416
- Li WH, Gojobori T (1983) Rapid evolution of goat and sheep globin genes following gene duplication. *Mol Biol Evol* 1(1):94–108
- Lynch VJ (2007). Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. *BMC Evol Biol* 7(2). doi:10.1186/1471-2148-7-2
- Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154(1):459–473
- Margolis RK, Margolis RU (1993) Nervous tissue proteoglycans. *Experientia* 49(5):429–446
- Markert CL, Shaklee JB, Whitt GS (1975) Evolution of a gene. *Science* 189:102–114
- Merritt TJS, Quattro JM (2001) Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase. *Genetics* 159:689–697
- Merritt TJS, Quattro JM (2002) Negative charge correlates with neural expression in vertebrate aldolase isozymes. *J Mol Evol* 55(6):674–683

- Merritt TJS, Quattro JM (2003) Evolution of the vertebrate cytosolic malate dehydrogenase gene family: duplication and divergence in actinopterygian fish. *J Mol Evol* 56(3):265–276
- Messier W, Stewart C (1997) Episodic adaptive evolution of primate lysozymes. *Nature* 385(6612):151–154
- Meyer A, Van De Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays* 27(9):937–945
- Moore BW (1973) Brain specific proteins. In: *Proteins of the nervous system*. Raven, New York, pp 1–12
- Moore BW, McGregor D (1965) Chromatographic and electrophoretic fractionation of soluble proteins of brain and liver. *J Biol Chem* 240:1647–1653
- Morizot DC, Slaugenhaupt SA, Kallman KD, Chakravarti A (1991) Genetic linkage map of fishes of the genus *Xiphophorus* (Teleostei: Poeciliidae). *Genetics* 127(2):399–410
- Nei M, Kumar S (2000) *Molecular evolution and phylogenetics*. Oxford University Press, Oxford
- Novak U, Kaye AH (2000) Extracellular matrix and the brain: components and function. *J Clin Neurosci* 7(4):280–290
- Ohno S (1970) *Evolution by gene duplication*. Springer, Berlin
- Ohno S, Wolf U, Atkin NB (1968) Evolution from fish to mammals by gene duplication. *Heredity* 59(1):169–187
- Ohta T (1991) Multigene families and the evolution of complexity. *J Mol Evol* 33(1):34–41
- Pebusque MJ, Coulier F, Birnbaum D, Pontarotti P (1998) Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol Biol Evol* 15(9):1145–1159
- Philippe H (2000) Opinion: long branch attraction and protist phylogeny. *Protist* 151:307–316
- Pommie C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* 17:17–32
- Pontier PJ, Hart NH (1981) Developmental expression of glucose and triose phosphate isomerase genes in teleost fishes (*Brachydanio*). *J Exp Zool* 217(1):53–71
- Postlethwait JH, Yan Y, Gates MA, Horne S, Amores A, Brownlie A, Donovan A, Egan ES, Force A, Gong Z, Goutel C, Fritz A, Kelsh R, Knapik E, Liao E, Paw B, Ransom D, Singer A, Thomson M, Abduljabbar TS, Yelick P, Beier D, Joly J-S, Larhammar D, Rosa F, Westerfield M, Zon LI, Johnson SL, Talbot WS (1998) Vertebrate genome evolution and the Zebrafish gene map. *Nat Genet* 18(4):345–349
- Quattro JM, Woods HA, Powers DA (1993) Sequence analysis of teleost retina-specific lactate dehydrogenase C: evolutionary implications for the vertebrate lactate dehydrogenase gene family. *Proc Natl Acad Sci USA* 90(1):242–246
- Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574
- Sato Y, Nishida M (2007) Post-duplication charge evolution of phosphoglucose isomerases in teleost fishes through weak selection on many amino acid sites. *BMC Evol Biol* 7(1):204
- Shaklee JB, Kepes KL, Whitt GS (1973) Specialized lactate dehydrogenase isozymes: the molecular and genetic basis for the unique eye and liver LDHS of teleost fishes. *J Exp Zool* 185(2):217–240
- Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. *Proc Natl Acad Sci USA* 97(13):7051–7057
- Steinke D, Hoegg S, Brinkmann H, Meyer A (2006) Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC Biol* 4(16):16
- Stothard P (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28:1102–1104
- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* 24(8):1596–1599
- Taylor JS, De Peer Y, Braasch I, Meyer A (2001) Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc B* 356(1414):1661–1679
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22(22):4673–4680
- Vandepoele K, De Vos W, Taylor JS, Meyer A, Van De Peer Y (2004) Major events in the genome evolution of vertebrates: paraneome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA* 101(6):1638–1643
- Vollf J-N (2005) Genome evolution and biodiversity in teleost fish. *Heredity* 94(3):280–294
- Wendel JF (2000) Genome evolution in polyploids. *Plant Mol Biol* 42(1):225–249
- Whitt GS (1970) Developmental genetics of the lactate dehydrogenase isozymes of fish. *J Exp Zool* 175(1):1–35
- Whitt GS (1983) Isozymes as probes and participants in developmental and evolutionary genetics. *Isozymes Curr Top Biol Med Res* 10:1–40
- Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15(5):568–573
- Yang Z (2002) Inference of selection from multiple species alignments. *Curr Opin Genet Dev* 12(6):688–694
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591
- Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15(12):496–503
- Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol* 11(2):316–324
- Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650
- Yokoyama S, Tada T, Zhang H, Britt L (2008) From the cover: elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. *Proc Natl Acad Sci USA* 105(36):13480–13485
- Zhang J, Nei M (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol* 44:139–146
- Zhang J, Kumar S, Nei M (1997) Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol Biol Evol* 14(12):1335–1338
- Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* 95:3708–3713