UPPSALA
UNIVERSITET

Evolution of the quaking gene

Giulia Tuveri

Degree project in biology, Master of Science (2 years), 2014
Examensarbete i biologi 45 hp till masterexamen, 2014
Biology Education Centre, Evolution and Development, Uppsala University
Supervisors: Åsa Tellgren-Roth and Elena Jazin
External opponent: Allison Perrigo

Table of Contents

ABSTRACT

The quaking gene has attracted attention for its role in vertebrate development and its possible involvement in certain mental disorders, such as schizophrenia. Zebrafish (*Danio rerio*) is a potential animal model for researching quaking functions, however it presents the problem of having three copies of the gene (*qkia*, *qkib* and *qki2*). The scope of this study was to elucidate the evolutionary relationship among the zebrafish quaking genes and thus find the true ortholog to the human quaking gene. The phylogeny of the quaking gene was investigated through bioinformatics analysis based on multiple sequence alignments. Bayesian inference was used to construct the phylogenetic trees in the program MrBayes 3.2. Due to high conservation of the sequences, the estimation of phylogenetic trees was problematic. jModelTest2 and MEGA6 were employed to test various models of nucleotide substitution. The most fitted models were used to construct the trees and were then compared to each other to test the strength of the topology. General time reversible (GTR), molecular clock and Tamura Nei (TN93) models confirmed the topology, assigning the zebrafish *qkia* to a separate clade, indicating this gene to be paralog to the tetrapods' quaking gene. *qkib* and *qki2* were found in the same clade of the tetrapods, and so were considered orthologs.

TERMINOLOGY

Common terms in phylogenetics

Phylogenetic tree graphic representation (in diagram-fashion) of the evolutionary relationship of organisms or genes based upon genetic (or morphological) differences

Node in general, the intersection between two lineages that diverge
Internal nodes hypothetical ancestors
Root the common ancestor to all species in the tree
Leaves, or tips the existing species
branch length measures the amount of evolution between nodes

Homology similarity between two or more sequences due to shared ancestry

Orthologs homologous sequences in different species that arose through speciation

Paralogs homologous sequences in two different species or in the same species that arose as consequence of duplication event before speciation

Consensus tree tree that represents the information that a set of trees has in common

Evolutionary distance the average number of substitution occurred at each nucleotide/amino acid site

Nucleotide substitution models

A substitution model is a formal (mathematical) description of the changes that occur in a sequence. Models differ from each other in mutation rate parameters.

Tamura Nei (TN93) model allows rates to vary, but only for transition (when a purine changes in another purine or a pyrimidine into another pyrimidine), which is more common than transversion (when a purine changes into a pyrimidine and vice versa)

Transition model (TIM2) accounts for variable transition rates and two transversion rates

General Time Reversible (GTR) model allows all rates to differ.

Molecular clock model assumes that the expected changes per site between two lineages are proportional to divergence time, therefore expected changes increase with time after the split from the common ancestor. Violations from this assumption are common in distantly related species, in genes created by duplication, and for sequences under strong selection. Direction is another important characteristic of the molecular clock theory. The trees are rooted, so there can only be one direction of evolutionary change, from the root, fixed from the start, towards the rest of the branches. So the closer a node to the root, the older it is. Time is indirectly expressed on branch lengths as amount of sequences divergence.

Relaxed molecular clock relaxes the assumption of the strict clock, allowing rate variation across the tree.

Methods for models selection

Likelihood ratio tests (LRTs) compare two models at the time against the sequence data and a reasonable topology, looking for the most fitted against a null model, uses p value to score likelihood of different models.

Akaike information criterion (AIC) is similar to LRTs, but it judges models according to the number of parameters. No hypothesis is used.

Bayesian Information criterion (BIC) is similar to AIC but accounts for sample size.

Bayes factors uses Bayesian inference to deduce the likelihoods, this method is replacing LRTs (ref)

Methods for phylogenetic inference

Markov chain Monte Carlo (MCMC) is a class of algorithms used to sample probability distributions

Metropolis-coupled Markov chain Monte Carlo (MCMCMC) a variant of the above with the additional feature of sampling from different chains (independent searches) and exchanging information among them. This is a resourceful technique because it solves the problem of a search being stuck on a "suboptimal" hill of probability.

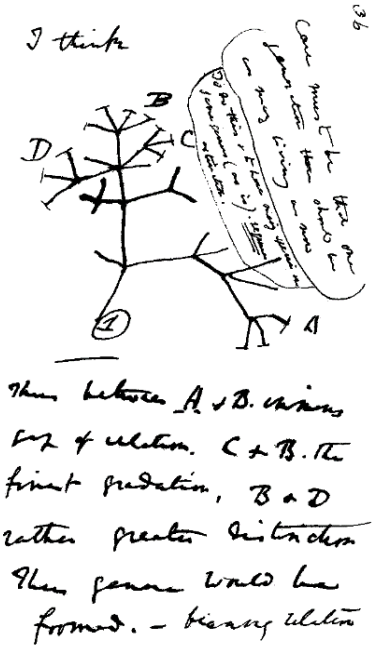
Bayesian inference (BI) is a statistical approach to phylogenetics. BI is based on the model of evolution (that one must specify), on the prior assumptions and on the information drawn from the data. It uses MCMC algorithms to search the probability space. In this manner the "posterior probabilities" are created. BI method finds the best set of trees that are consistent with both the model and the data. Posterior probabilities are an alternative to bootstrapping when estimating reliability of a topology, this is also referred to as clade or branch support.

Maximum likelihood (ML) Similar to BI, character-based as BI, it's a statistical method that seek the tree that makes the data the most likely. It produces only one tree and it generally make use of the bootstrap method to assess the topology of a phylogenetic tree. However it is possible to use ML approach coupled to other methods for estimating reliability. For example aLRT statistics, available in PhyML program.

Neighbour joining (NJ) a distance-based method for phylogenetic inference. Uses a simple algorithm based on a distance matrix. Here distance is intended as the fraction of sites that vary between two aligned sequences (pairwise differences). It does not account for multiple nucleotide changes at the same location. Another disadvantage is the lack of branch support values.

INTRODUCTION

Why a phylogenetic study



Phylogenetics is the science that unveils the evolutionary relationships among living organisms. Nowadays it's not restricted anymore only to systematics but it has become an essential tool in biology (Baum and Smith, 2013). Its paradigm is that organisms, or genes, can be tracked back in time to a common ancestor. Darwin was the first to introduce the "tree thinking" concept, the understanding and inferring from evolutionary trees, or phylogenies. Symbolic is his famous sketch of a phylogenetic tree, dated 1837. The tree represents the evolution of species, which are the tip of the branches, from a common ancestor, symbolized with the root. Common ancestry is also called homology and is a key notion in phylogenetics. It implies evolution: if two species have a common ancestor, then all the traits (genes) that diverge must have changed after the split from the common ancestor, accumulating mutations, therefore evolving.

Figure 1. The first phylogenetic tree, Charles Darwin 1837.

Human, fish and fruit flies have radiated at different points in time but going far enough backwards on the tree of life, a common ancestor can be found. Because of shared ancestry, species like the previous mentioned, which are very distant from each other, are considered together in phylogenetic studies. The aim of phylogenetics is to bring to light the path of genetic changes that form the characters defining an organism and serving specific functions. The history of a gene becomes a complementary tool to molecular studies for the understanding of a particular trait. What makes the comparative approach possible and very powerful today is: new technologies, such as genome sequencing; online databases like NCBI and Ensembl, that offer readily available sequences and analyzing tools; and state-of-the-art bioinformatics methods.

Vertebrate evolution

Vertebrates evolved from a common ancestor around 500 million years ago (mya), yet conserved genes are often found in their genomes. Since the advent of new techniques in evolutionary biology the phylogeny of vertebrates has been explored. Genomes diverge through time because of accumulation of mutations and the strategy to uncover their relationship is to identify true homology (shared ancestry). Inference of homology is complicated by genome rearrangement, gene duplication and whole genome duplication (WGD). In his *Evolution by gene duplication* (1970), Ohno was the first to argue the importance of duplication as an evolutionary factor and the first to suggest whole genome duplication events in vertebrates (Ohno, 1970). Today these facts are almost universally

accepted, as many showed that the evolution of vertebrates has been facilitated by the duplication events (Kasahara et al., 2007), (Cañestro et al., 2013). The evidence for WGD events in the history of vertebrates comes from the number of gene copies found in vertebrates versus non-vertebrates, often in the characteristic proportion of many : 1. For example, genes within the HOX family (Lemons and McGinnis, 2006) occur in a 4:1 ratio between vertebrates and the cephalochordate amphioxus (Donoghue and Purnell, 2005). Duplicate copies of genes are usually lost but occasionally can acquire new functions and so be maintained. At the beginning of the vertebrate lineage, two rounds of WGD occurred (called 1R and 2R), between 700 mya and 450 mya (Meyer and Schartl, 1999), (Panopoulou and Poustka, 2005), (Putnam et al., 2008), (Cañestro, 2012). Another key event in vertebrates evolution is the split between ray-finned fish (Actinopterygii) and the lobe-finned fish (Sarcopterigii), around 450 mya (Hedges, 2009). After this split, the ray-finned group went through another genome duplication (Braasch and Postlethwait, 2012), known as Teleost-specific whole genome duplication (TGD). Teleosts, like zebrafish (*Danio rerio*), are part of the ray-finned fish lineage, while tetrapods, such as human, chicken, frog, etc., belong to the lobe-finned lineage. Coelacanth also belongs to the lobe-finned fish lineage and is particularly interesting for phylogenetic studies because it is considered a living fossil of the lobe-finned lineage (Amemiya et al., 2013). These species are thought to be key species in vertebrate phylogeny and now their genomes are freely available at online databases.

Phylogenetic analysis of vertebrate groups requires the use of an outgroup, a species that is not included in a monophyletic clade and/or is evolutionary distant enough to carry sufficient variation to work as an external comparison. In this regard, the fruit fly *Drosophila melanogaster* has been a favorite outgroup. The tunicate *Ciona intestinalis*, less distant to vertebrates than the fruit fly, is another good example. Amphioxus (*Branchiostoma floridae*), commonly known as lancelet, is also a proper outgroup candidate, even more suitable than fruit fly and tunicates. It has been shown that amphioxus belongs to the most ancient subphylum of Chordata, with over 500 million years of independent evolution. After the split from amphioxus, vertebrates underwent two rounds of WGD, while amphioxus retained its diploid genome unduplicated. This makes amphioxus a perfect outgroup in a study where understanding gene duplication within vertebrates is critical.

The quaking gene

The quaking gene encodes a RNA-binding protein, belonging to the signal transduction and activation of RNA (STAR) family. They are characterized by containing a single K-Homology (KH) domain, the part of the protein where RNA binds and interacts (Musco et al., 1996). The 70-100 amino acids that form the domain are evolutionarily conserved in the STAR proteins, from the mammalian quaking gene (named “qkI” in mouse and “QKI” human) to the fruit fly HOW gene (Held Out Wing, named after the phenotype), homologous to QKI. The KH domain is even found in bacteria (Protein Sequence Analysis & Classification Database, assession number: IPR004087)

Alternative splicing is another characteristic of this gene. During gene expression, alternative splicing allows coding regions (exons) within the gene to be translated differently, creating multiple proteins from the same gene, the so called protein isoforms. In QKI three main protein isoforms are known: QKI5, QKI6, QKI7 and later 7b isoform was also found (Kondo et al., 1999). The names reflect the length of 5, 6 and 7 kilobases (Kb) (Ebersole et al., 1996). These splice variants are also conserved

throughout evolution, in fact corresponding transcripts are found in the fruit fly HOW and the nematode *Caenorhabditis elegans* ASD-2 (Volk, 2010), with similar structure to QKI. The protein isoforms differ from each other in length, as some exons might be included or excluded in the transcription process, and at the end of the sequence. The end (or tail) of the protein, called C-terminal, determines the location of the protein in the cell. For instance, the QKI5 C-terminus carries a nuclear specific signal, while QKI6 and 7 are found in the cytoplasm (Hardy et al., 1996). Diverse splice forms and their locations serve different functions at different times. QKI5 is the most abundant during early embryogenesis while the other isoforms show higher expression in post-natal development and adulthood (Ebersole et al., 1996), (Chénard and Richard, 2008).

The most important known functions operated by QKI proteins are stability, location and alternative splicing of mRNA. It's interesting to note that the same function of alternative splicing is observed in invertebrates such as fruit flies and nematodes, despite that they shared a last common ancestor with humans 780 mya and 930 mya, respectively (Hedges et al., 2006). Such conserved function is a clear indication of the importance of quaking functions in animals.

QKI has been investigated since 1996 when Ebersole et al. explained the "quaking viable" phenotype in mouse as a consequence of a genetic mutation in the qkI locus. This phenotype is characterized behaviorally by a distinct tremor (quake) of the hind limbs and seizures and anatomically by defective formation of myelin in the nervous system, a condition called dysmyelination (Poser, 1978). Myelin is the protective and isolating layers of dense membrane forming a sheath around the axons of neurons and constituting the white matter. A type of glial cells, the oligodendrocytes, are responsible for myelin formation (Compston et al., 1997).

Hardy *et al.* (1998) first described the expression of the quaking gene in the central nervous system of adult mouse, precisely in glial cells (astrocytes and oligodentrocytes), but not neurons (Hardy, 1998). Laroque *et al.* (2005) and later Haroutunian *et al.*, (2006) showed that QKI6 and QKI7 promote oligodendrocyte differentiation (Larocque and Richard, 2005), (Haroutunian et al., 2006). Again, from an evolutionary perspective, it's worth noticing that the same fundamental role of quaking in glial differentation is also true for HOW in fruit fly (Volk, 2010). This example underlies the validity of non-mammalian animal models for investigating this gene.

Quaking role in human diseases

The reduction of oligodendrocytes in the white matter of qkI viable mouse brain is similar to what can be observed in humans affected by schizophrenia, a complex mental disorder (Haroutunian et al., 2006). Ålberg *et al.* (2006) found a link between schizophrenia and the locus where human QKI gene is found, on chromosome 6 (Åberg et al., 2006)a. The same authors observed, in post-mortem human brains, an association between variation in QKI expression and myelin alteration in schizophrenic patients (Åberg et al., 2006)b. Another human disease linked with the quaking gene is glioma, a malignant tumor in the nervous system (Yin et al., 2009). Ataxia, a neurological condition that causes problems in controlling voluntary movements, is also reported to be connected with QKI expression in humans (Chénard and Richard, 2008).

Zebrafish as animal model

The research around quaking has been done almost exclusively in mice and has focused mainly on the post-natal neurodevelopment and myelin formation. Thanks to the viable mouse mutants, the link between the gene and myelination, seizures, reduced lifespan and Purkinje cells alternation has been established (Richard, 2010). Because most qkI mutations are lethal in mouse embryos, the neurodevelopmental processes in which qkI is involved are still unknown. However, this lethality highlights the fundamental role of quaking in development, even before the start of the myelination process.

Zebrafish is an excellent animal model and the study of the quaking functions has been initiated in this species during early development. Zebrafish presents several advantages (Howe et al., 2013). It's oviparous, which means eggs develop outside the parent's body, and the embryos are transparent, allowing for *in vivo* studies. While it is true that zebrafish is well suited to study the developmental roles of quaking, it presents the problem of having not one but three copies of the quaking gene. One gene, named *qkia*, is located on chromosome 17, the second quaking gene, *qkib*, is found on chromosome 13 and the third, *qki2*, is found on chromosome 12. The duplicated copies obfuscate direct comparison to the human QKI without first establishing the true orthologs.

Aim

The present study investigates the evolution of the quaking gene. No prior evolutionary analysis has been done on the gene, despite the attention that it has attracted for its importance in vertebrate development and its possible involvement in certain mental disorders. The main question addressed is: of the three zebrafish *qki* genes, which one is the true ortholog to the human QKI gene?

Most of the work is based on bioinformatics resources, such as search tools for sequences identification, software for alignment and statistical analysis. The logical (simplified) approach carried out in this thesis goes as follow: 1) selection of species, 2) identification of homologous sequences, 3) multiple alignment of the chosen sequences, 4) selection of statistical methods and models of nucleotide substitution, 5) construction of phylogenetic trees. However, this order was not maintained at the first stage of the project, as shown in the methods, figure 2.

MATERIALS AND METHODS

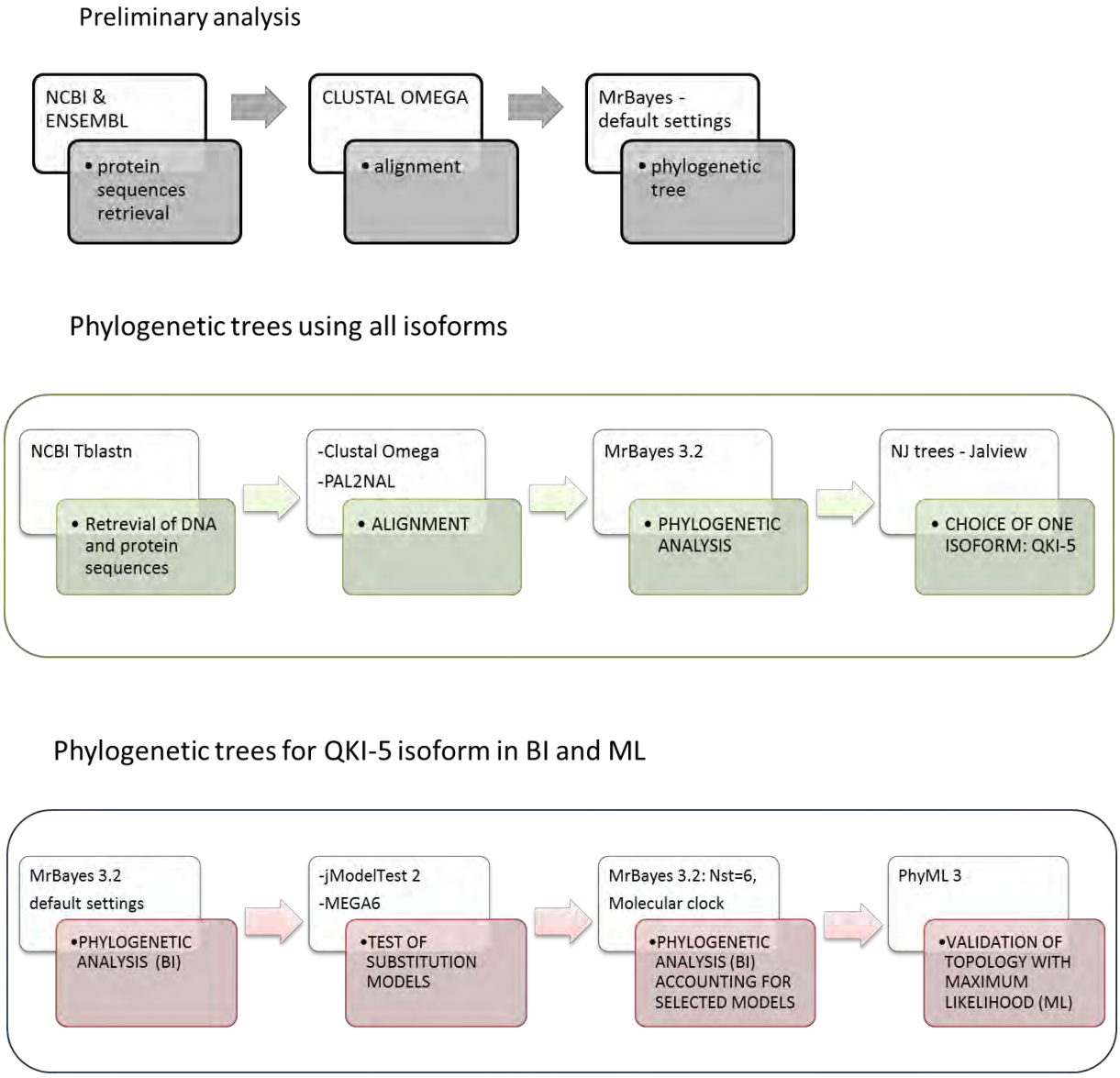


Figure 2. Flow chart of the methods. This study is divided into three parts: 1) preliminary analysis to assess the information of the sequences at the protein level, 2) phylogenetic analysis of all quaking isoforms and 3) phylogenetic analysis using only the QKI-5 isoform accounting for specific models of nucleotide substitution.

Retrieval of sequences

The zebrafish protein qkia-QKI5 and qkib-QKI5 were used as queries in the tblastn search (translated nucleotide database using a protein query) in NCBI Genomes (chromosome) and whole-genome shotgun contigs (wgs) databases (http://blast.st-va.ncbi.nlm.nih.gov/Blast.cgi). The Ensembl database (www.ensembl.org) was also used, to compare sequence identity and quality. In some occasions Ensembl sequences were preferred. Additionally, to retrieve possible further protein isoforms yet to be annotated, human QKI amino acids sequences were used as query against the genomes of the selected species. Human exons were used to annotate splice sites of new identified sequence, (Katarzyna J. Radomska, 2014) which are awaiting publication.

For each species, different isoforms were named accordingly to the similarity to the human isoforms known in the literature. For example: chicken QKI5, coelacanth QKI7b. The sequences used are shown in the alignment tables 1 and 2 in the Appendix and in the trees. No list of sequences is presented in this report. All the sequences are available upon request.

Alignment

Clustal-Omega (Sievers et al., 2011) is an efficient, powerful and accurate program designed for the alignment of multiple sequences of DNA, RNA and proteins (Sievers et al., 2011). The program creates a temporary distance matrix from the pairwise distances of the input sequences, then a "guide tree" based on the distance matrix is produced to decide the order of the sequences in the alignment. To construct the trees Clustal-Omega uses Muscle's (Edgar, 2004) UPGMA (Unweighted Pair Group Method with Arithmetic Mean), a fast method for hierarchical clustering. An iteration of the guide trees after an initial alignment allows an improvement in the final alignment, a process called "progressive alignment". Clustal-Omega is freely available at EBI-EMBL website.

All alignments in this study are performed using Clustal-Omega with default settings and output format FASTA. Manual editing of the alignment was often necessary, especially for the outgroup sequence. The outgroup supposedly being very different from the rest of the sequences in the alignment, some errors in the alignment were expected.

Visualization, editing, analysis, conversion

Jalview was the program used for visualization of the alignments. It's a free program for multiple sequence alignment visualization, editing and analysis (Waterhouse et al., 2009). It operates in Java. All alignments generated for the current investigation were visualized in Jalview version 2.8.0b1 to verify the accuracy. A few editing operations were performed, mostly addition of gaps, especially towards the end of the sequences, where more variation is found.

Jalview also provides the possibility to create phylogenetic trees using neighbour joining (NJ), which is a simple extremely fast distance algorithm that generates a single strictly bifurcating tree. The NJ method is straight forward and computationally not demanding. NJ trees are made through the Jalview function "calculate tree: Neighbour joining using percentage identity (PID)" (Waterhouse et al., 2009). It's useful for quick visualization of sequences clustering according to their similarity in tree format, but it's not reliable for complex phylogenetic analysis. This function was used to create trees of different isoforms.

PAL2NAL – amino acids & nucleotide sequences in one alignment

Pal2nal v14 (2011) was used to create the alignments. It pairs a multiple protein alignment (first input data, in CLUSTAL format) to the corresponding DNA (or mRNA) sequences (second input data, in FASTA format) and converts both into one codon alignment based on the protein alignment (Suyama et al., 2006). In other words, it builds a codon (so DNA) alignment using the nucleotide sequences but aligning them according to the corresponding amino acids alignment, maintaining the information carried at the protein level (to avoid mismatching amino acids positions in the multiple protein alignment). Settings were maintained as default, using a codon table (universal codon), gaps not removed, output format CLUSTAL. It's a web server, accessible at http://www.bork.embl.de/pal2nal/.

Phylogenetic Analysis

jModelTest 2 was the software chosen for statistical selection of nucleotide substitution models (Posada, 2008). It's built on the Phyml algorithm, based on maximum likelihood method (Guindon and Gascuel, 2003). It compares a wide variety of nucleotide substitution models and finds the most fitted for the input alignment. It also allows model averaging, in case one doesn't want to choose and rely on a single model, and model-averaged phylogeny (estimating a consensus tree for every model). This opportunity was taken in the study and used as additional support for the topology inferred in Bayesian Infernce (BI) and Maximum Likelihood (ML). Available for free download from https://code.google.com/p/jmodeltest2/.

MEGA 6 (Tamura et al., 2013) was used as a complementary tool for the model selection analysis. MEGA is a useful phylogenetic package available for download at www.megasoftware.net. It deals with alignments, multiple analysis of sequences, and construction of trees by different methods, such as NJ and Maximum Likelihood (ML). Like jModelTest2, MEGA6 can also test different models of nucleotide substitution. The results for models selection were compared to those of jModelTest2 and to confirm the tree topology.

MrBayes 3.2.1 (Ronquist et al., 2012) is the most used program (Hall, 2004) for Bayesian analysis of phylogeny. It works on a command-line interface and because of the computational weight of Bayesian inference, it requires a fairly large amount of computer memory. Input sequences are in the NEXUS format. The models of nucleotide substitution and the parameters are specified with the data in the "mr bayes block" or after loading the data in the MrBayes window. The most important two commands required to define the analysis are "lset" and "prset". "Lset" formalizes the structure of the model chosen while "prset" draws the prior distribution for the parameters. **The most important settings that were altered from default are reported below.**

Lset:

Nucmodel sets the general type of nucleotide substitution model. 4by4 (standard DNA model with only four states A,C,G,T/U) and "codon" were used.

Nst determines rate variation. It was changed from value Nst=1 constrains all rates to be equal, to Nst=6, which allows all rates to vary as in the general time reversible (GTR) model.

Rates specifies among-sites rate change. In general this is an unknown random variable.
gamma distribution of different rates across sites
invgamma same as above plus a proportion of sites with invariable rates.

Prset:

When a molecular clock was assumed, the following settings were applied:
Brlenspr sets the distribution of prior probability on branch lengths.
clock:uniform the tree is constrained to the clock model and rooted

Clockvarpr specifies the type of molecular clock model
IGR independent gamma rates, each branch is allowed to have independent gamma rate

At the end of the run MrBayes 3.2 provides the average standard deviation of split frequencies. These values are reported in the results, as indication of convergence between the two independent runs that are performed simultaneously. Values below 0.01 are considered as appropriate convergence and indicate a successful run.

Phylogenetic inference based on Maximum Likelihood method was also part of this study. PhyML 3 is a program for phylogenetic inference based on ML. Its advantage lies in its simplicity, accuracy and speed (Guindon et al., 2010). It can be run online at www. atgc-montpellier.fr/phyml/. Another virtue of this program is that it offers a good alternative to the time-consuming bootstrap method to evaluate clade support: the fast likelihood based methods, one of which is the abayes method. A disadvantage of this program is that it doesn't allow calculation of trees under molecular clock assumptions. Its results are comparable to Bayesian inference in accuracy (Anisimova et al., 2011).

PhyML online execution panel is presented in four sections. In *Input data*, one uploads the data and specifies the type (DNA/amino acid). *Substitution model* allows the user to indicate the model of choice from seven options, among which are the GTR and TN93 (Tamura and Nei, 1993), and the gamma shape parameter. *Tree searching* permits the use of a starting tree and to select a method of tree improvement. *Branch support* is where one chooses between bootstrapping and fast likelihood-based methods. All the settings in the four sections were changed according to the needs of this analysis. See the results for details.

FigTree is a great program for visualizing and editing phylogenetic trees. It's intuitive and very user friendly. It was used to choose what values to present (usually the posterior probability) and in what format, using the "node label" option. Re-rooting the tree was done when a non-clock evolutionary model was used. It's freely available at http://tree.bio.ed.ac.uk/software/figtree.

The following web resources were used to convert formats as necessary:
http://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html,
http://genome.nci.nih.gov/tools/reformat.html, http://sing.ei.uvigo.es/ALTER/ .

RESULTS

We first identified orthologous sequences to the zebrafish *qkia* gene through tblastn search. A preliminary phylogenetic analysis was carried out to assess the level of conservation and information of the quaking protein sequences. Following the results of this preliminary analysis, some species were excluded and a more comprehensive detailed analysis was performed. The species subsequently selected were considered more informative and appropriate. Finally different methods and models of phylogenetic inference were performed and compared to give robustness to the final tree.

Preliminary analysis

Identification of orthologous protein sequences and assessment of sequence information

The BLAST query was the zebrafish QKI5 protein from *qkia*, accession number NP_571299.1. The species selected with homologous sequences were: *Latimeria chalumnae* (African coelacanth), *Lepisosteus oculatus* (spotted gar), *Gasterosteus aculeatus* (three-spined stickleback), *Oryzias latipes* (Japanese medaka), *Anolis carolinensis* (green anole lizard), *Xenopus (Silurana) tropicalis* (western clawed frog), *Gallus gallus* (chicken), *Mus musculus* (mouse), *Canis familiaris* (dog), *Macaca mulatta* (Rhesus monkey), *Myotis lucifugus* (little brown bat), and *Homo sapiens* (human). For outgroups we chose: the tunicate *C. intestinali*, fruit fly, and amphioxus. The search for homologous sequences was done mostly in NCBI, in some cases Ensembl was also used.

In these species different isoforms of the quaking protein were found and included in the alignment. The alignment was made in Clustal Omega, using default settings. The resulting alignment was transformed in a NEX file and run in MrBayes 3.2.1 with the following settings: mcmc ngen=100000 printfreq=10000 samplefreq=100 nchains=4 savebrlens=yes; sumt burnin=50%. See figure 3 for the resulting tree.

The topology of the tree made some sense, having the outgroup species in the expected “root position”, suggesting enough information at the protein level between the outgroups and the vertebrates. On the other hand some cluttered branching was observed within more closely related taxa. However “imperfect” this first tree looked, it was a first hint of two main clades: zebrafish chr.17 (*qkia*), Coelacanth (named Latimeria novel gene in the tree) and “Spotted gar chr6” on one and all tetrapods plus zebrafish chr.12 (*qki2*) and chr.13 (*qkib*) and spotted gar chr16 on the other. *C. Intestinalis* was discarded because of apparent poor quality of the (predicted sequence, and because tunicates are a sister group of vertebrates (Putnam et al., 2008), thus as recent as the vertebrate group, and so more close to vertebrates than fruit fly or amphioxus.

The teleost fish three-spined stickleback and the Japanese medaka were excluded from further analysis due to their convoluted phylogenies resulting from complicated genome rearrangement and gene loss after the TGD (Braasch and Postlethwait, 2012). The rhesus monkey was substituted by *Gorilla gorilla* (western gorilla), to try another primate. The little brown bat was also excluded from further analysis as it did not add any new insight into the phylogeny.

The preliminary results are not included in the discussion.

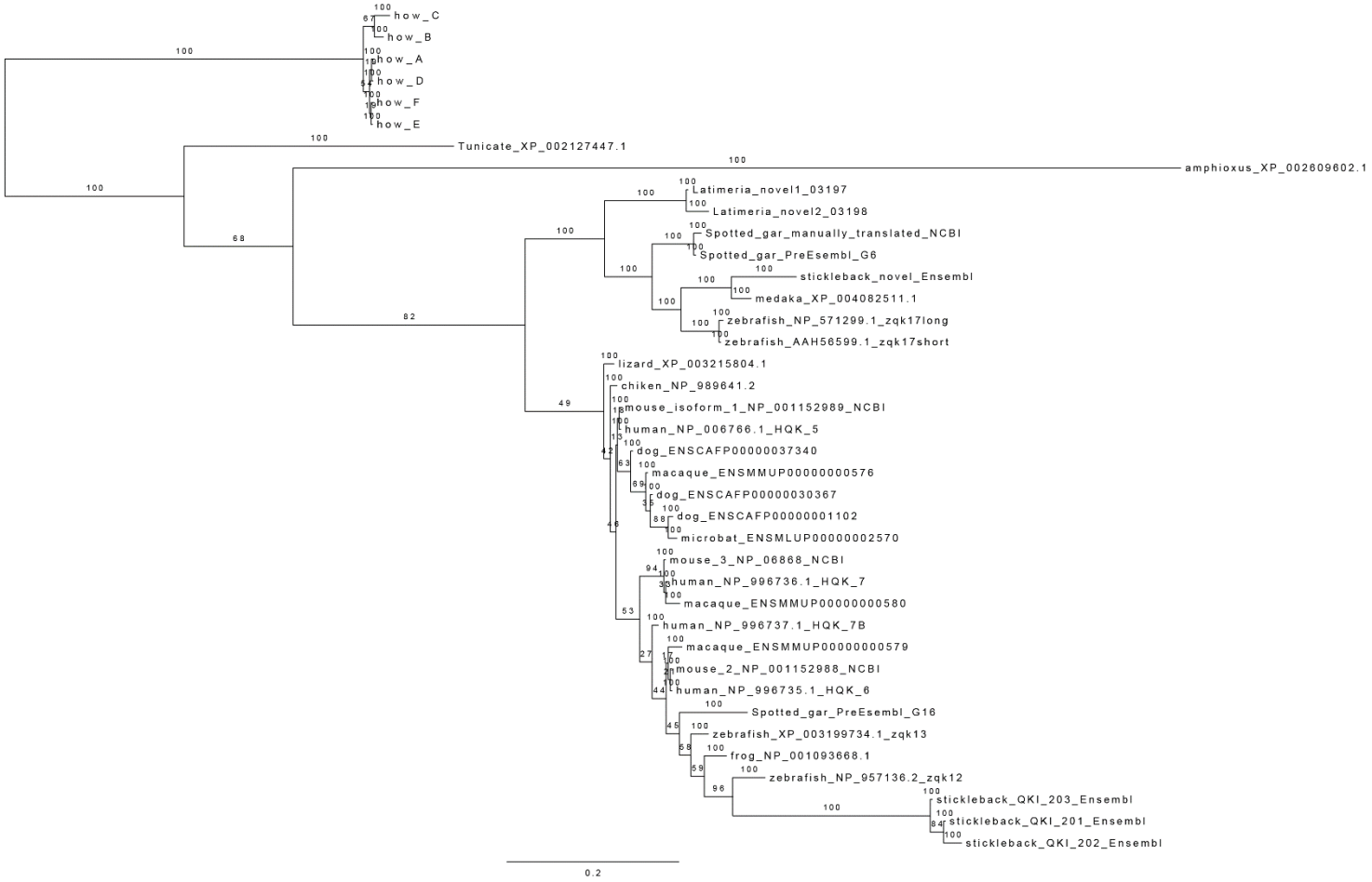


Figure 3. Phylogenetic tree of quaking proteins. Bayesian analysis were run in MrBayes 3.2.1 with default settings.

Analysis of quaking isoforms

Sequences retrieval

A new blast search in NCBI of both genomic and protein sequences was done for the following species: *Danio rerio* (zebrafish), *Latimeria chalumnae* (coelacanth), *Lepisosteus oculatus* (spotted gar), *Xenopus (Silurana) tropicalis* (western clawed frog), *Gallus gallus* (chicken), *Mus musculus* (mouse), *Monodelphis domestica* (gray short-tailed opossum), *Canis familiaris* (dog), *Sus scrofa* (pig) *Gorilla gorilla* (gorilla), *Homo sapiens* (human), *Drosophila melanogaster* (fruit fly) and *Brachiostoma floridae* (amphioxus).

Alignment

At this point, the protein alignment was coupled to the nucleotide sequences through PAL2NAL, to create a codon alignment. This alignment offers both the advantage of a protein alignment (functional sense) and of a DNA alignment (more informative, less conserved). See in the Appendix the codon alignment (Table 1) for comparison to the protein alignment (Table 2).

Phylogenetic tree of all the isoforms

A MrBayes run was performed with the following settings, using fruit fly as outgroup: datatype: DNA nucmodel:codon ngen=850000. The resulting tree is shown in figure 4 and had an average standard

deviation value of 0.04, indicating poor convergence. This was interpreted as a consequence of including all the splice isoforms in one alignment. The tree in figure 4 shows a better clustering than the protein tree (fig. 3). The different isoforms of one species cluster together, confirming the expectation that at DNA level sequences are more informative. However, the same doesn't happen for human and gorilla sequences, which are too closely related for the sequences to carry enough divergence to distinguish between them.

Ideally that is what a "real" protein tree should show, different isoforms of a protein clustering together according to isoform not to species, since the alternative splicing that creates the different transcripts and protein isoforms has likely originated in the common ancestor, and have not evolved in each species. In virtue of this considerations, the incorporation of all the available isoforms was believed to be redundant. Thus only one isoform was chosen for the rest of the analysis, as illustrated in figure 2 (methods).

Of note, the quaking gene in frog presented a difficulty in accurately diverging with known speciation events, which is a common problem with other frog genes (personal communication, Dr. Sager). The position of the frog sequence in the topology was inconsistent with the generally accepted phylogeny in all the trees derived up to this point of the study potentially due to higher rate of mutation. This problem was additional evidence for the necessity to adopt a more specific model of evolution.

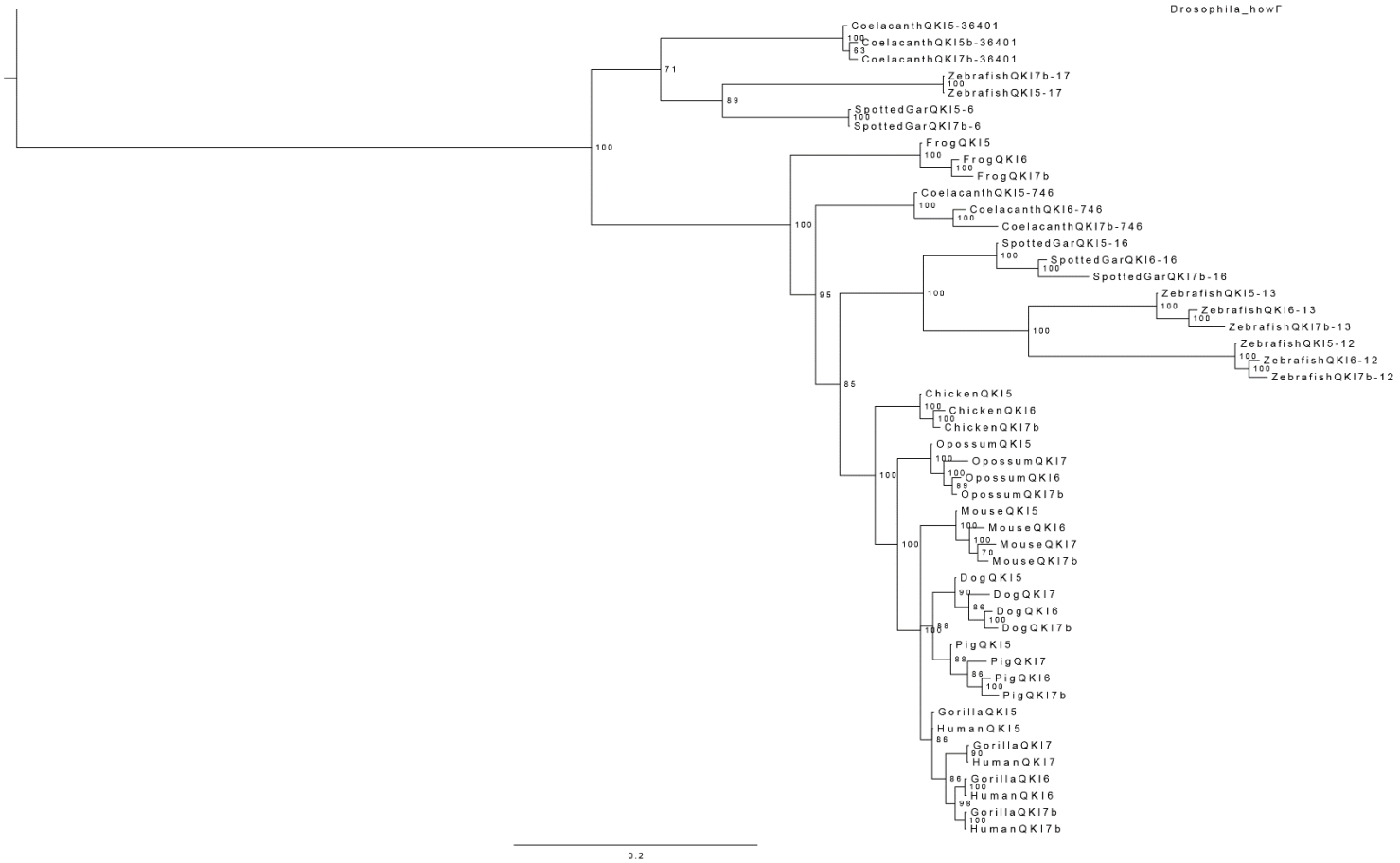


Figure 4. Phylogenetic tree based on Bayesian Inference of all the protein isoforms. Sequences (amino acids and DNA) used for the PAL2NAL alignment on which this tree is based were collected in NCBI. Phylogenetic analysis were performed in MrBayes 3.2, using "codon" nucleotide model. Although branch support values were high (> 70), the average standard deviation of split frequencies was 0.04, suggesting low convergence.

Neighbour Joining trees

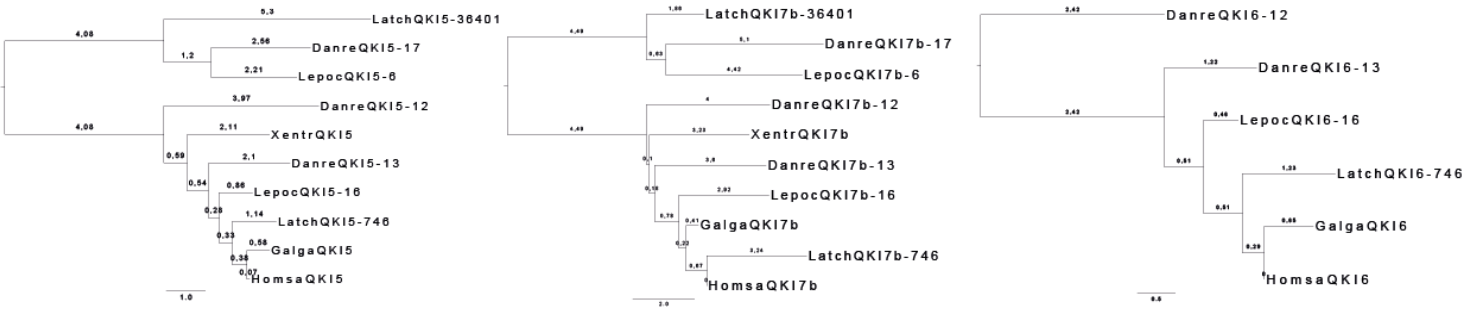


Figure 5. Neighbour Joining trees for different quaking splice variants in different species. Trees are rooted at midpoint. Values on branches represent branch lengths. In order, QKI5, QKI7b and QKI6. Trees were made in Jalview.

Neighbour Joining (NJ) trees of the different isoforms were made through Jalview to assess similarity between the same isoform (i.e. QKI5, QKI6, QKI7b) in the different species (Fig. 5). NJ tree for QKI7 is not shown because it's very similar to 7b.

QKI5 is the transcript variant found in all the species blasted and in the species with more than one quaking gene, for example in coelacanth or zebrafish, each quaking gene has its own QKI5 like variant. This is not the case for the other transcripts, which makes the topology incomplete. For example, the QKI6 transcript was not found in zebrafish *qkia*, nor was it found for the spotted gar quaking gene on chromosome 6, the resulting tree is therefore missing two key sequences from the topology. For QKI7b, it appeared that the tail of the sequences are very different due to high variation (see the long branch of Lat7b), which can create incorrect branching.

Phylogenetic tree of the QKI5 isoform

To make a phylogenetic tree based on the QKI5 isoform I used MrBayes 3.2 with default settings and the 4by4 DNA model, which is simple and fast. This tree reconstruction aimed to test the information carried by the most common isoform.

The consensus tree is only made by 14 trees. I ran it for 900000 generations and it had a standard deviation value of 0.000668.

The QKI5 variant tree below (fig. 6) gives strong nodes values and a very reasonable topology. However, to validate it, a model selection test was carried out through MEGA6 and jModelTest2.

The clustering of mammal species didn't reflect the species known phylogeny, as shown in the species tree in the Appendix. This problem was taken into account. Mouse, grey short-tailed opossum, dog, pig, gorilla, human QKI5 proteins alignment showed almost perfect conservation. Likewise, at the nucleotide level the similarity were still very high, compared to the other species (data not shown).

It was decided that, among mammals, only human and mouse would be useful to the later analysis, the first for being the other essential end of the tree we're looking for (tracking down the evolution of quaking from zebrafish to human) and the latter for being the most used mammal species in studies of quaking.

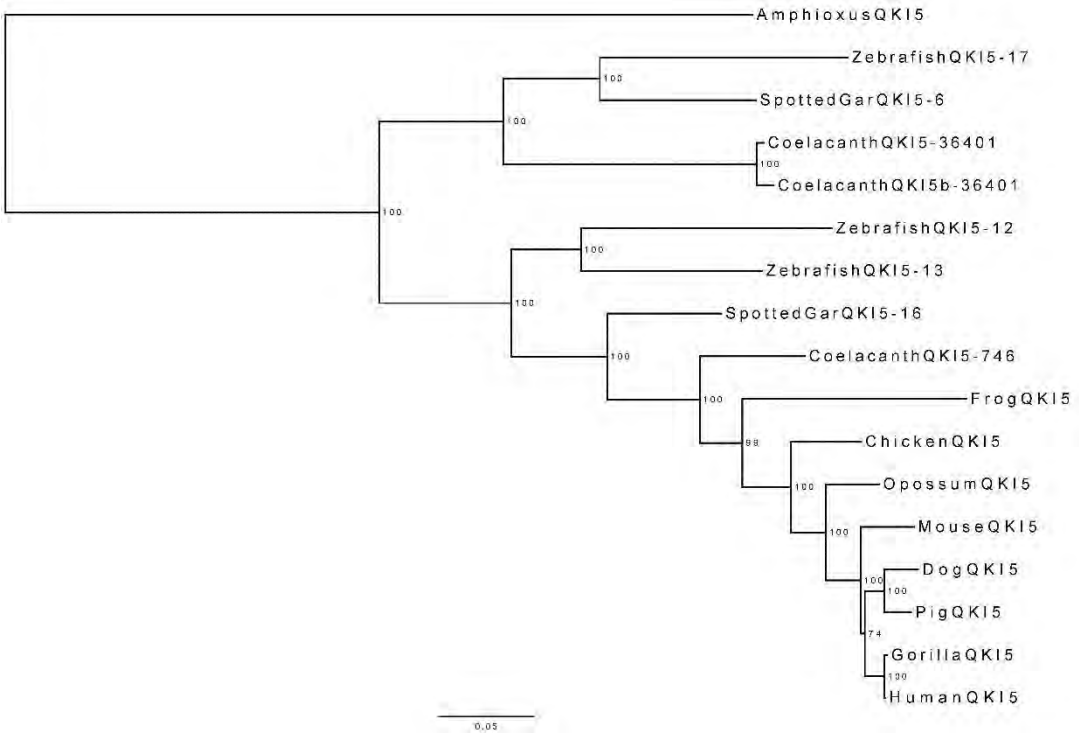


Figure 6. Phylogenetic tree in BI of QKI5 isoform. The tree is based on a codon alignment. Phylogenetic analysis was performed in MrBayes 3.2, using default settings: simple 4by4 nucleotide model, all rates equal. Branch support is very high (>99), with exception of the node between dog-pig-gorilla-human clade.

QKI5 isoform and models of evolution

Testing models of nucleotide substitution

jModelTest2 was run with the following settings: candidate models = 88, number of substitution schemes = 11, including models with equal/unequal base frequencies (+F), including models with/without a proportion of invariable sites (+I), including models with/without rate variation among sites (+G) (nCat = 4), optimized free parameters (K) = substitution parameters + 21 branch lengths + topology, base tree for likelihood calculations = ML tree, tree topology search operation = BEST.

The GTR model resulted first with LRT method, while AIC and AICc had TIM2 first and GTR second. Finally BIC method had TIM2 second and GTR sixth, however with very similar likelihood values to the models above it. Adding that the GTR model is easily implemented in MrBayes 3.2 (simply by changing Nst to 6), compared to the others, GTR was our choice. See table 3 in the Appendix.

In order to compare different model selection programs, MEGA 6 was also employed. See results for jModelTest2 and MEGA6 in table 3 and 4 respectively (Appendix). MEGA6 authors believe the BIC method to be the most appropriate. This method, in MEGA6, selected the model TN93 to be the most fitted for our data. Both GTR and TN93 were therefore used and compared.

GTR and molecular clock model

When dealing with distantly related sequences one must take into account unequal rate of evolution, therefore the GTR model and a relaxed molecular clock were assumed, following the results of the tests for model selection. The analysis were run with the following: Datatype: DNA, Nucleotide model: 4by4 (standard model of nucleotide substitution in which there are only four states (A/C/G/T), Nst: 6 (e.g. a GTR model), rates: invgamma. Prset brlens=clock:uniform (specifies a base for the relaxed clock), prset clockvarpr=igr, Mcmcp savebrlens=yes samplefreq=100 printfreq=1000 Mcmc ngen=1000000, Sumt relburnin=yes burninfrac=0.25. The result is shown in figure 7. The topology is consistent with the theory of duplication and loss of the different quaking genes in vertebrates. Spotted gar chr.16 grouped with zebrafish *qkib* and *qki2* sequences and coelacanth 746 (the *qkib*-like sequence) with the other tetrapods, with a branch support value of 99%, reflecting the expected relationship of the gene in different species. A very good value of average standard deviation of split frequencies (0.003137) was achieved. The consensus tree is built from a credible sets of trees (122 trees sampled), with 90% credible sets containing 7 trees, 95% containing 14 trees, 99% containing 38 trees.

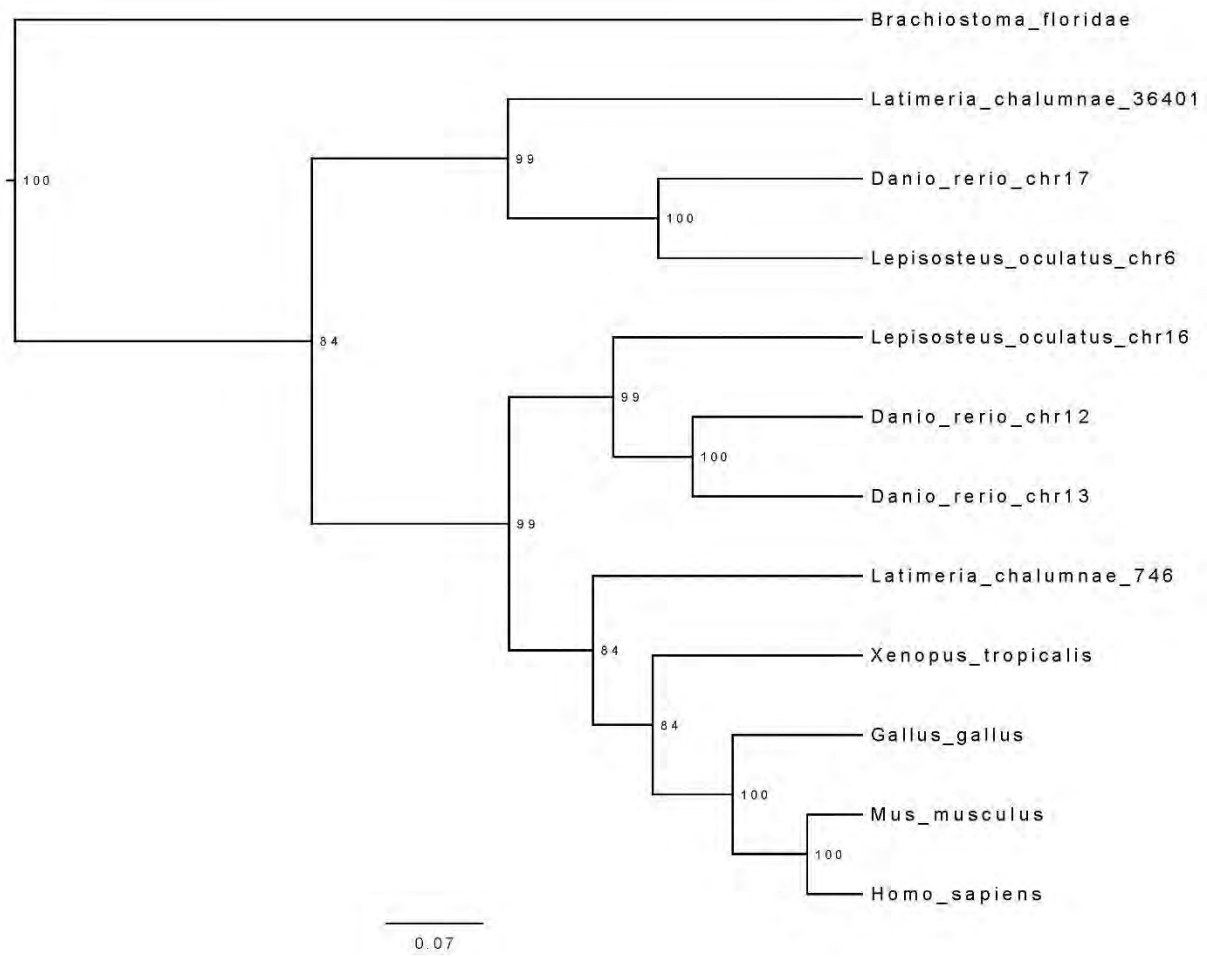


Figure 7. Phylogenetic tree bases on BI using QKI5 isoform and a molecular clock model. The tree is constructed using a codon alignment. Phylogenetic analysis was performed in MrBayes 3.2, using the simple 4by4 nucleotide model, all rates allowed to vary as predicted by GTR model (Nst=6). The IGR relaxed molecular clock model was included. Branch support shows very high values (84-100).

In order to test the strength of the topology, non-clock and strict clock models were tested. These trees are not shown. The non-clock model reported the same topology as the one shown in figure 7 while the strict clock model showed branching inconsistent with the known species phylogeny.

Validation of the topology with ML

MEGA6 under the assumptions of the BIC method selected TN93 to be the most appropriate. This model was used on the ML tree.

The Maximum Likelihood method was adopted to see how robust our data wass, and if it will give similar results given different methods. It was run in PhyML 3 with the following settings: Datatype: DNA, Sequence file: interleaved, Substitution model: TN93, Equilibrium frequencies: optimized, Proportion of invariable sites: 0.0, Number of substitution rate categories: 4, Gamma shaped parameter: 0.34, Starting tree: BIONJ, Type of tree improvement SPR&NNI, Topology and branch length: optimized. Branch support aBayes. The result is reported in figure 8. The topology was in accordance with the BI tree shown in figure 7.

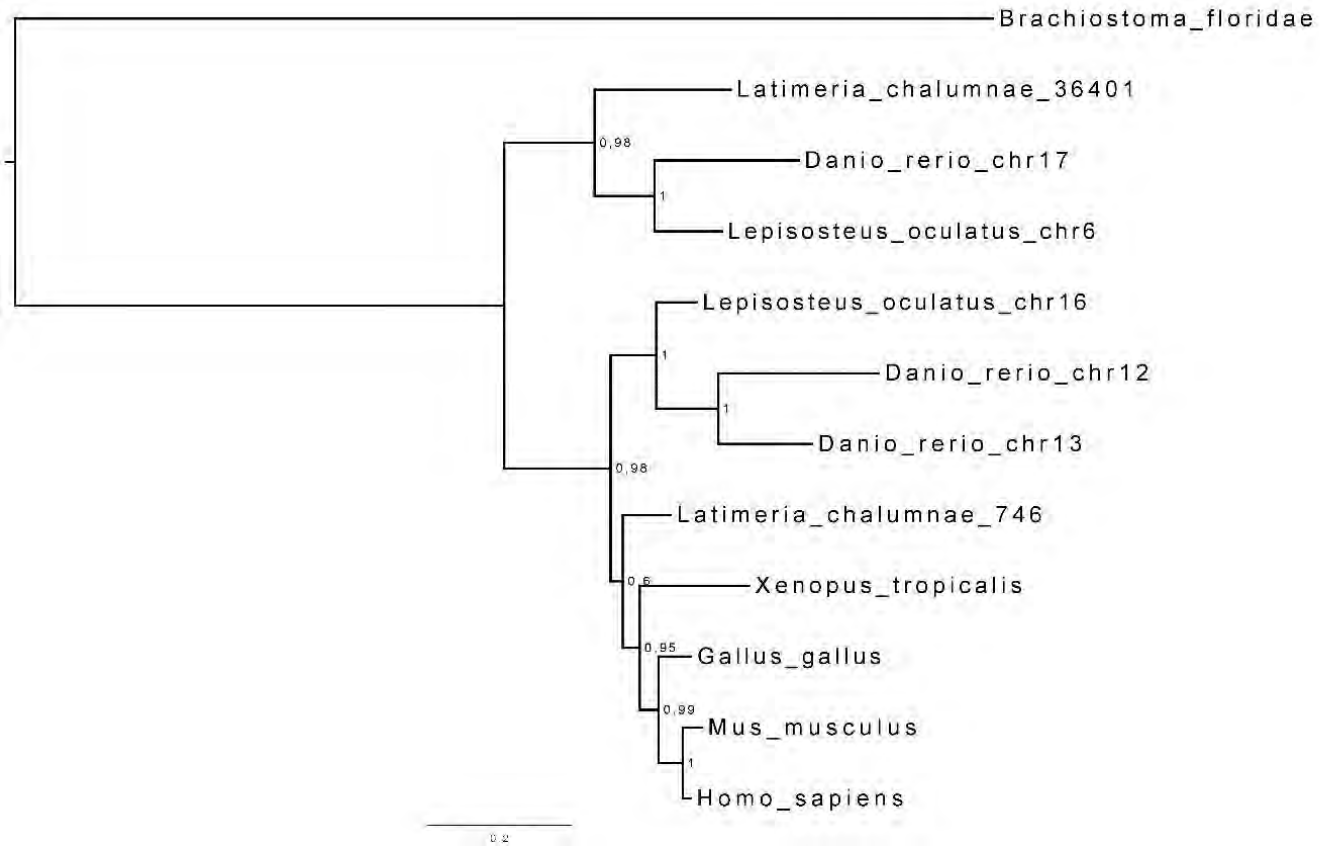


Figure 8. ML tree of the QKI5 isoform. The tree is based on a codon alignment. Phylogenetic analysis was performed in PhyML 3 under the assumptions of the TN93 model. Method for branch support abayes.

DISCUSSION

The main achievement of the study is the resolved phylogeny of the quaking gene in chordates, a finding strongly supported by different methods of phylogenetic inference. The phylogeny shows a clear split in two main clusters: one containing zebrafish *qkia*, spotted gar and coelacanth *qkia*-like genes, the other one including all tetrapods plus coelacanth and spotted gar *qkib*-like gene, along with zebrafish *qkib* and *qki2*. The analysis concludes that the true zebrafish orthologs to the mammalian quaking are *qkib* and *qki2*, while *qkia* is paralogous to all tetrapods.

Evolution of the qki genes revealed

This study was the first to assess the phylogeny of the quaking gene as most of the work on the gene has been done in mouse. Murata et al. (2005) compared QKI cDNA of pig, cow, and horse finding great similarity at nucleotide level (about 95%) and the deduced amino acid sequences were identical to the murine ones (Murata et al., 2005). Their study pointed out the link between extreme conservation of the quaking gene and its significance in biological functions within mammals. Most studies on quaking focused on the functional aspect of the gene without looking at the evolution. The present work elucidates the relationship between the three *qki* genes in zebrafish, a new animal model for quaking, and the mammalian QKI.

From the trees topology, one can infer that the quaking gene must have duplicated early in the history of chordates. For the cephalochordate amphioxus, the outgroup, only one copy of the gene was found, while key species early in jawed vertebrate evolution, such as coelacanth and spotted gar, harbor two copies. This is consistent with the theory of the two whole genome duplications at the root of vertebrates. Following the tetrapods lineage, only one gene is found in frog, chicken, mouse and human and in all the other tetrapods analyzed. This is explained by gene loss, the most common fate of a duplicated gene, in the ancestor of tetrapods. It's possible that the two quaking genes preserved in coelacanth and spotted gar acquired new functions, escaping gene loss. Zebrafish has three quaking genes. *qkia* is not orthologous to the tetrapod quaking, while *qki2* and *qkib* are. *qkib* shows higher similarity to human QKI compared to *qki2,* with similar values for the other tetrapods (sequence distance test in MEGA6, data not shown) suggesting that the *qki2* is a result of the teleost-specific whole genome duplication.

The higher rate of mutation of *qki2*, as shown in tree figures 4 and 6, is another measure of dissimilarity between *qki2* and the rest of the clade. An explanation for *qki2*'s high mutation rate might be neofunctionalization following mutation and positive selection. The function of the zebrafish *qki2* and *qkib* are under present investigation. Synteny analysis (Katarzyna J. Radomska, 2014) supports the teleost specific origin of *qki2*. The region on chromosome 13 where *qkib* is located shows more conservation between zebrafish and human than the region on chromosome 12 where *qki2* is. While *qki2* shares synteny only with other teleosts, *qkib* shares synteny with spotted gar as well.

Methods and models of evolution, the importance of making the right choice

The Bayesian method was the preferred one to perform the phylogenetic analysis in this study, accompanied by ML. They are both sophisticated methods that employ mathematical models in order to describe, with precision, the evolutionary processes occurring at nucleotide and amino acid levels. Simpler methods such as NJ and parsimony are convenient in matter of speed but they lack precision, in that they cannot discern among different kinds of genetic change (i.e. transition-transversion, codon-selection, unequal rate of variation among sites) nor they assign different weights to different mutations. Indeed, today BI and ML are the most commonly used.

Models used by BI and ML are based on character change, so called substitution models. What makes them so critical is the fact that one can never have an "assumption free" model for phylogenetic analysis. In statistical terms, the assumptions about character change become parameters. One of the most important parameters that were included in the substitution models used is the "alpha parameter". The "alpha parameter" accounts for unequal rate of change among sites, also known as the "shape parameter". It's what changes the shape of the gamma distribution. Because of the nature of the quaking genes, one can assume strong selection acting on maintaining the functional domain, while relaxed selection toward the tails, possibly to allow mutation in cases of a novel splice variant acquiring a new function. This is potentially supported by the fact that the location where the quaking proteins act is determined by the 3' end.

The results also show the need for the correct utilization of models of evolution. This can clearly be seen in the difference of topology between figure 6 (assuming equal rate of variation – Nset=1 in MrBayes) and figure 7 (assuming unequal rate of variation – Nset=6). In general, over fitting models are better than under fitting models (Kelchner and Thomas, 2007). For this reason a variety of model parameters were adopted.

Once the type of model is chosen, in our case the GTR with gamma distribution, there is still room for additional specifications. In the general GTR model, just like in other evolutionary models, there is no specification for direction of change. Theoretically, any tip of a tree could be the root, as the resulting trees are unrooted and require manual rooting. In molecular clock models, the trees are rooted. Comparing a non-clock tree against a strict clock tree and a relaxed clock tree was particularly interesting, because the results showed again how important it is to make the right assumptions. A tree constructed with a strict clock (not shown) resulted in an incorrect topology, different from both the relaxed clock tree and the non-clock tree. The non-clock tree (also not shown), constructed accounting for rate variation (gamma distribution) and a proportion of invariable site, was considered a very reasonable outcome. Nevertheless the non-clock tree is arguably biologically unrealistic, especially when ranging from amphioxus to human, two species separated by more than 700mya. Drummond *et al.* (2006) indicated the accuracy of different relaxed clock models, demonstrating the advantage of those in phylogenetic inference in comparison to the unrooted models (Drummond et al., 2006). Following this logic, I tested the relaxed molecular clock IGR. The IGR is a type of uncorrelated relaxed clock that falls in the category of relaxed clock models that were found to perform best compared to other relaxed clock models. The resulting tree confirmed the topology of the non-clock tree. This support the idea that a carefully chosen relaxed molecular clock model gives appropriate outcomes and it would be in the interest of biological realism to not dismiss the challenge of assuming the time-dependent nature of genes.

Because of the differences between programs that test models of nucleotide substitution, it seemed worth including in the analysis a couple of models of those suggested by jModelTest2 and MEGA6, again with the goal of comparing methods and confirming the topology of the trees. jModelTest2 investigation reported GTR models to be among most fitted for our sequences. MEGA6 agreed on that with regards of AICc and likelihood ratio test methods (see table 4), however the authors of MEGA6 report BIC to be the most credible method, and BIC methods selected TN93 as best fitted model for our sequences. Therefore, the TN93 model was tested with PhyML 3. The topology found with the latter confirmed once more the topology found in the GTR in BI. The results together show a very strong topology, once given the correct assumptions, which were repeated using different methods of phylogenetic inferences. Branch lengths vary with the model, however maintained similar values. For example, most importantly, zebrafish *qki2* was revealed to have a higher mutation rate irrespective of the model and method.

Quaking family and future studies

This was the first study on the quaking gene phylogeny. As mentioned in the introduction, quaking belongs to the STAR protein family. Interestingly, a review on STAR family exists, including a NJ tree of the known members of this family (Biedermann et al., 2010). Their study shows similarities among STAR members, including quaking and quaking related proteins. It appears from their resulting topology that quaking like genes are present in *Arabidopsis thaliana*, a plant, those genes being closer to SAM68, another member of the STAR family than to quaking related gene. A phylogenetic study of the STAR family can explain the origin of quaking and enlighten the evolution of the complex molecular systems in which STAR proteins are involved.

ACKNOWLEDGMENTS

I would like to thank Elena Jazin for welcoming me in her group and for giving me this very interesting project. I did enjoy the challenges that the project presented and the freedom I was given to tackle them. Thank you to my supervisor, Åsa Tellgren-Roth, for the help and for supervising even when she couldn't be there. I think somehow we made the distance work! I would also like to thank Lage Cerenius for being a very helpful and kind coordinator.

A special thank you goes to my officemates: on my left Henrik and Filipa, on my right JJ and Bryn. You were both a fantastic company and of big help for any sort of problem! I consider myself lucky to have spent these months with you and I hope to have officemates like you in the future. I will certainly miss you! I should also thank you, Bryn and JJ, for the discussions about proper English. As well as… thank you JJ for the generosity in explaining anything anytime.

Thank you Simo for showing me the way! Scandinavia is the right place indeed. Arshi, you lightened up the atmosphere and the mood. There should be more people like you! The biggest thanks to my parents, to whom I owe all my education and my life in Sweden, my gratitude will never end. Tusen tack to my dear friends and the nice people in the department, you made everything easier and more pleasant. My awesome dude Rado, I can't imagine doing this or anything without you. There is not enough space here to thank you so I'll do it privately.

I truly want to thank my opponent, Allison Perrigo, for being so cool and for giving me a lot of useful insights. It was fun meeting with you!

And thank you Meher for the speech.

REFERENCES

Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., and Rohner, N. (2013). The African coelacanth genome provides insights into tetrapod evolution. Nature *496*, 311-316.

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., and Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Systematic biology *60*, 685-699.

Baum, D.A., and Smith, S.D. (2013). Tree thinking: An introduction to phylogenetic biology (Roberts).

Biedermann, B., Hotz, H.-R., and Ciosk, R. (2010). The Quaking family of RNA-binding proteins: coordinators of the cell cycle and differentiation. Cell Cycle *9*, 1929-1933.

Braasch, I., and Postlethwait, J.H. (2012). Polyploidy in fish and the teleost genome duplication. In Polyploidy and genome evolution (Springer), pp. 341-383.

Cañestro, C. (2012). Two rounds of whole-genome duplication: evidence and impact on the evolution of vertebrate innovations. In Polyploidy and genome evolution (Springer), pp. 309-339.

Cañestro, C., Albalat, R., Irimia, M., and Garcia-Fernàndez, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. Paper presented at: Seminars in cell & developmental biology (Elsevier).

Chénard, C.A., and Richard, S. (2008). New implications for the QUAKING RNA binding protein in human disease. Journal of neuroscience research *86*, 233-242.

Compston, A., Zajicek, J., Sussman, J., Webb, A., Hall, G., Muir, D., Shaw, C., Wood, A., and Scolding, N. (1997). Review: Glial lineages and myelination in the central nervous system. Journal of anatomy *190*, 161-200.

Donoghue, P.C., and Purnell, M.A. (2005). Genome duplication, extinction and vertebrate evolution. Trends in Ecology & Evolution *20*, 312-319.

Drummond, A.J., Ho, S.Y., Phillips, M.J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. PLoS biology *4*, e88.

Ebersole, T.A., Chen, Q., Justice, M.J., and Artzt, K. (1996). The quaking gene product necessary in embryogenesis and myelination combines features of RNA binding and signal transduction proteins. Nature genetics *12*, 260-265.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research *32*, 1792-1797.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology *59*, 307-321.

Guindon, S., and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology *52*, 696-704.

Hall, B.G. (2004). Phylogenetic trees made easy: a how-to manual, Vol 547 (Sinauer Associates Sunderland).

Hardy, R.J. (1998). QKI expression is regulated during neuron-glial cell fate decisions. J Neurosci Res *54*, 46-57.

Hardy, R.J., Loushin, C.L., Friedrich, V.L., Jr., Chen, Q., Ebersole, T.A., Lazzarini, R.A., and Artzt, K. (1996). Neural cell type-specific expression of QKI proteins is altered in quakingviable mutant mice. The Journal of neuroscience : the official journal of the Society for Neuroscience *16*, 7941-7949.

Haroutunian, V., Katsel, P., Dracheva, S., and Davis, K.L. (2006). The human homolog of the QKI gene affected in the severe dysmyelination "quaking" mouse phenotype: downregulated in multiple brain regions in schizophrenia. The American journal of psychiatry *163*, 1834-1837.

Hedges, S.B. (2009). Vertebrates (Vertebrata). The timetree of life, 309-314.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics *22*, 2971-2972.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., and Matthews, L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., and Kasai, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. Nature *447*, 714-719.

Katarzyna J. Radomska, Å.T.-R., Bryn Farnsworth, Jonathan Sager, Giulia Tuveri, Elena Jazin, Petronella Kettunen, Lina Emilsson (2014). The zebrafish qkib is essential for nervous system development.

Kelchner, S.A., and Thomas, M.A. (2007). Model use in phylogenetics: nine key questions. Trends in Ecology & Evolution *22*, 87-94.

Kondo, T., Furuta, T., Mitsunaga, K., Ebersole, T.A., Shichiri, M., Wu, J., Artzt, K., Yamamura, K.-i., and Abe, K. (1999). Genomic organization and expression analysis of the mouse qkI locus. Mammalian genome *10*, 662-669.

Larocque, D., and Richard, S. (2005). Point of View QUAKING KH Domain Proteins as Regulators of Glial Cell Fate and Myelination. RNA biology *2*, 37-40.

Lemons, D., and McGinnis, W. (2006). Genomic evolution of Hox gene clusters. Science *313*, 1918-1922.

Meyer, A., and Schartl, M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Current opinion in cell biology *11*, 699-704.

Murata, T., Yamashiro, Y., Kondo, T., Nakaichi, M., Une, S., and Taura, Y. (2005). Nucleotide sequence of complementary DNA encoding for quaking protein of cow, horse and pig: Short Communication. Mitochondrial DNA *16*, 300-303.

Musco, G., Stier, G., Joseph, C., Morelli, M.A.C., Nilges, M., Gibson, T.J., and Pastore, A. (1996). Three-dimensional structure and stability of the KH domain: molecular insights into the fragile X syndrome. Cell *85*, 237-245.

Ohno, S. (1970). Evolution by gene duplication (London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.).

Panopoulou, G., and Poustka, A.J. (2005). Timing and mechanism of ancient vertebrate genome duplications–the adventure of a hypothesis. TRENDS in Genetics *21*, 559-567.

Posada, D. (2008). jModelTest: phylogenetic model averaging. Molecular biology and evolution *25*, 1253-1256.

Poser, C.M. (1978). DYsmyelination revisited. Archives of Neurology *35*, 401-408.

Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., and Yu, J.-K. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature *453*, 1064-1071.

Richard, S. (2010). Reaching for the STARs. In Post-Transcriptional Regulation by STAR Proteins (Springer), pp. 142-157.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol *61*, 539-542.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7*, 539.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic acids research *34*, W609-612.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular biology and evolution *10*, 512-526.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular biology and evolution *30*, 2725-2729.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Volk, T. (2010). Drosophila star proteins. In Post-Transcriptional Regulation by STAR Proteins (Springer), pp. 93-105.

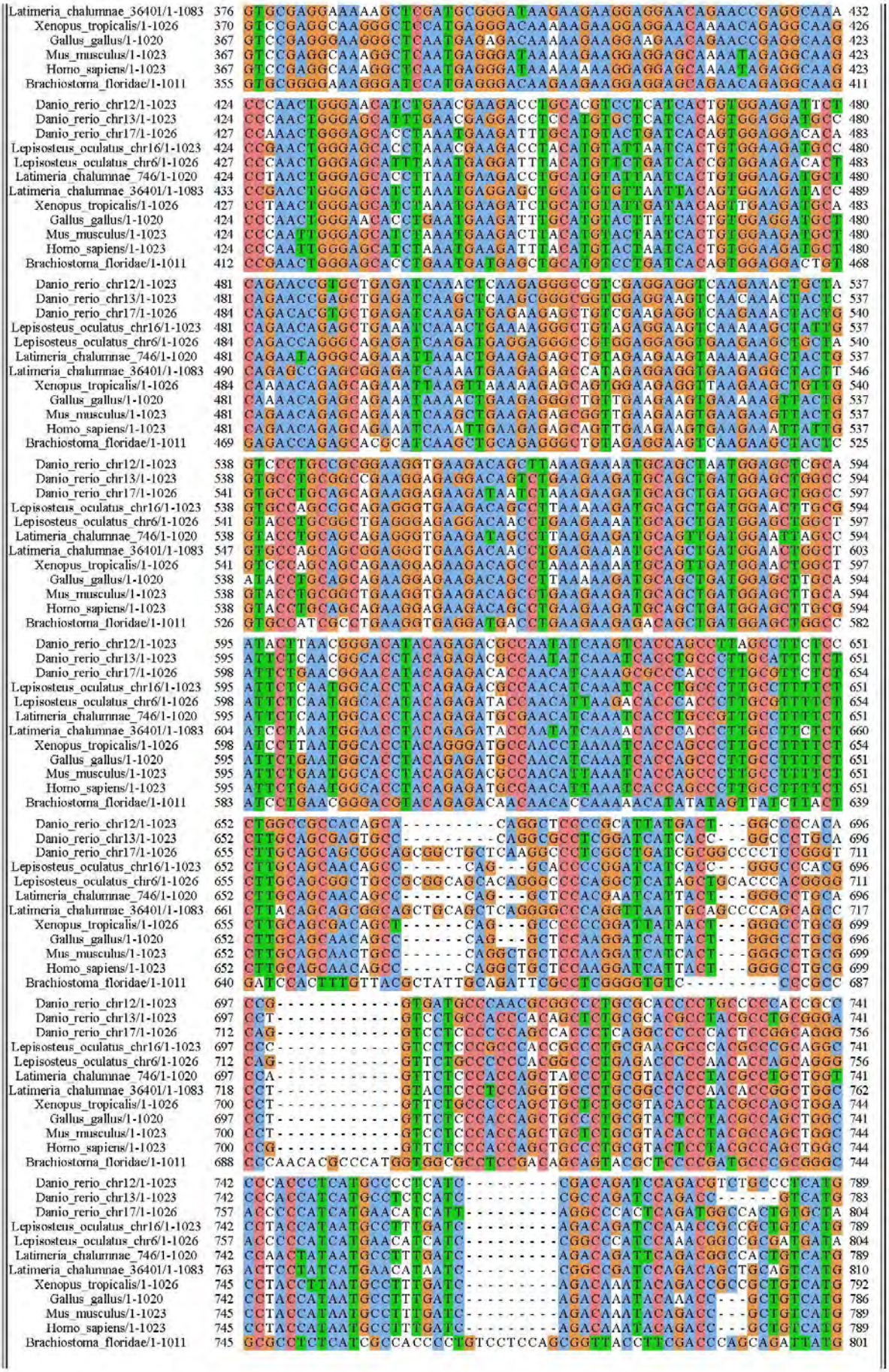
Yin, D., Ogawa, S., Kawamata, N., Tunici, P., Finocchiaro, G., Eoli, M., Ruckert, C., Huynh, T., Liu, G., Kato, M*., et al.* (2009). High-resolution genomic copy number profiling of glioblastoma multiforme by single nucleotide polymorphism DNA microarray. Molecular cancer research : MCR *7*, 665-677.

Åberg, K., Saetre, P., Lindholm, E., Ekholm, B., Pettersson, U., Adolfsson, R., and Jazin, E. (2006). Human QKI, a new candidate gene for schizophrenia involved in myelination. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics *141B*, 84-90.

Websites in the references (http://www.ebi.ac.uk/interpro/entry/IPR004087)

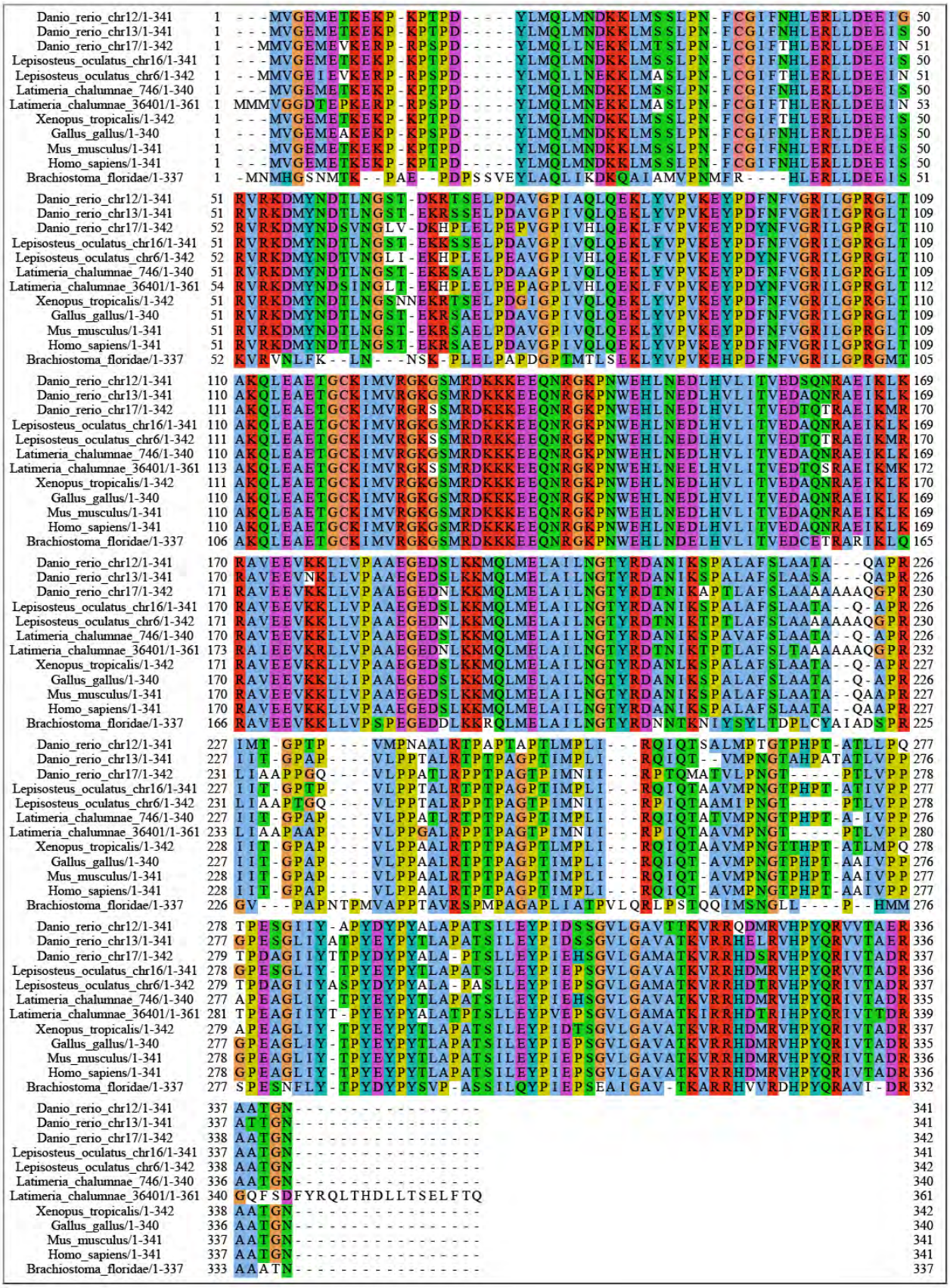
APPENDIX

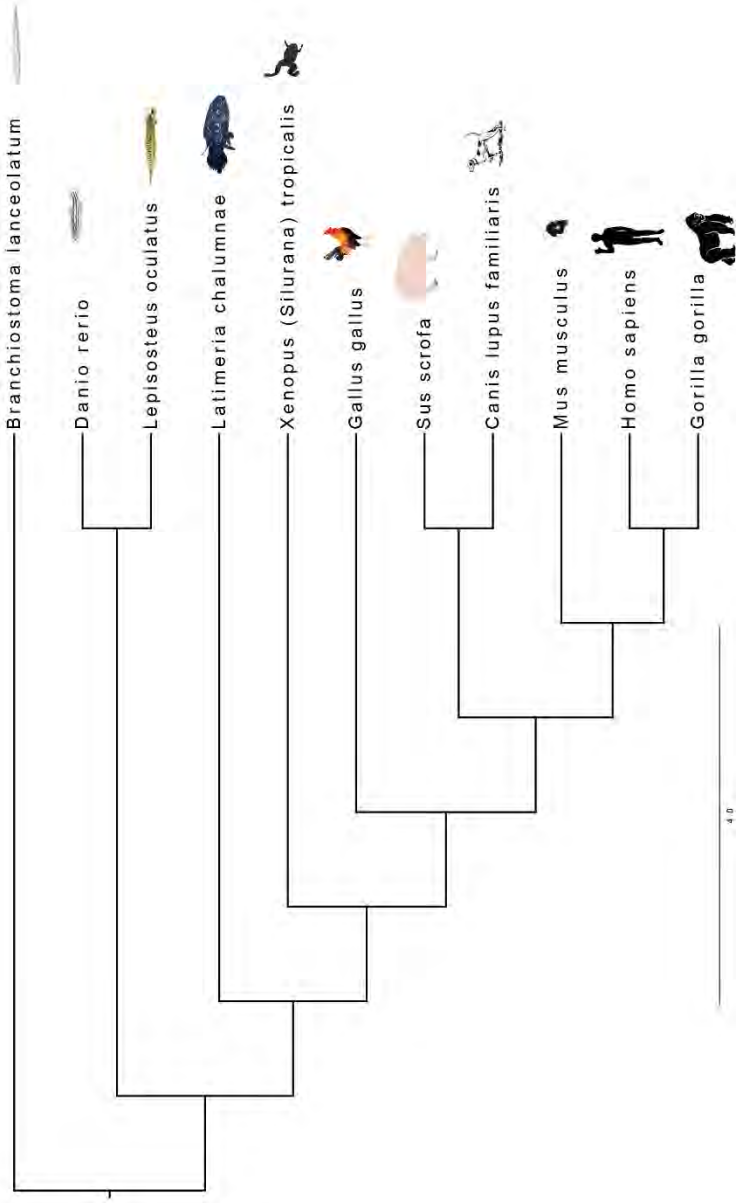
Table 1. Nucleotide alignment (PAL2NAL codon alignment)


Latimeria_chalumnae_36401/1-1083 376 GTGCGAGGAAAAAGCTCGATGCGGGATAAGAAGAAGGAGGAACAGAACCGAGGCAAA 432
Xenopus_tropicalis/1-1026 370 GTCCGAGGCAAGGGCTCCATGAGGGACAAAAAGAAGGAGGAACAAAACAGAGGCAAG 426
Gallus_gallus/1-1020 367 GTCCGAGGGAAGGGCTCAATGAGAGACAAAAAGAAGGAAGAACAGAACCGAGGCAAG 423
Mus_musculus/1-1023 367 GTCCGAGGCAAAGGCTCAATGAGGGATAAAAAGAAGGAGGAGCAAAATAGAGGCAAG 423
Homo_sapiens/1-1023 367 GTCCGAGGCAAAGGCTCAATGAGGGATAAAAAAAGGAGGAGCAAAATAGAGGCAAG 423
Brachiostoma_floridae/1-1011 355 GTGCGGGGAAAGGGATCCATGAGGGACAAGAAGAAGGAGGAGCAGAACAGAGGCAAG 411
Danio_rerio_chr12/1-1023 424 CCCAACTGGGAACATCTGAACGAAGACCTGCACGTCCTCATCACTGTGGAAGATTCT 480
Danio_rerio_chr13/1-1023 424 CCCAACTGGGAGCATTTGAACGAGGACCTCCATGTGCTCATCACAGTGGAGGATGCC 480
Danio_rerio_chr17/1-1026 427 CCAAACTGGGAGCACCTAAATGAAGATTTGCATGTACTGATCACAGTGGAGGACACA 483
Lepisosteus_oculatus_chr16/1-1023 424 CCGAACTGGGAGCACCTAAACGAAGACCTACATGTATTAATCACTGTGGAAGATGCC 480
Lepisosteus_oculatus_chr6/1-1026 427 CCCAACTGGGAGCATTTAAATGAGGATTTACATGTTCTGATCACCGTGGAAGACACT 483
Latimeria_chalumnae_746/1-1020 424 CCTAACTGGGAGCACCTTAATGAAGACCTGCATGTATTAATCACTGTGGAAGATGCT 480
Latimeria_chalumnae_36401/1-1083 433 CCGAACTGGGAGCATCTAAATGAGGAGCTGCATGTGTTAATTACAGTGGAAGATACC 489
Xenopus_tropicalis/1-1026 427 CCTAACTGGGAGCATCTAAATGAAGATCTGCATGTATTGATAACAGTTGAAGATGCA 483
Gallus_gallus/1-1020 424 CCCAACTGGGAACACCTGAATGAAGATTTGCATGTACTTATCACTGTGGAGGATGCT 480
Mus_musculus/1-1023 424 CCCAATTGGGAGCATCTAAATGAAGACTTACATGTACTAATCACTGTGGAAGATGCT 480
Homo_sapiens/1-1023 424 CCCAATTGGGAGCATCTAAATGAAGATTTACATGTACTAATCACTGTGGAAGATGCT 480
Brachiostoma_floridae/1-1011 412 CCGAACTGGGAGCACCTGAATGATGAGCTGCATGTCCTGATCACAGTGGAGGACTGT 468
Danio_rerio_chr12/1-1023 481 CAGAACCGTGCTGAGATCAAACTCAAGAGGGCCGTCGAGGAGGTCAAGAAACTGCTA 537
Danio_rerio_chr13/1-1023 481 CAGAACCGAGCTGAGATCAAGCTCAAGCGGGCGGTGGAGGAAGTCAACAAACTACTC 537
Danio_rerio_chr17/1-1026 484 CAGACACGTGCTGAGATCAAGATGAGAAGAGCTGTCGAAGAGGTCAAGAAACTACTG 540
Lepisosteus_oculatus_chr16/1-1023 481 CAGAACAGAGCTGAAATCAAACTGAAAAGGGCTGTAGAGGAAGTCAAAAAGCTATTG 537
Lepisosteus_oculatus_chr6/1-1026 484 CAGACCAGGGCAGAGATCAAGATGAGGAGGGCCGTGGAGGAGGTGAAGAAGCTGCTA 540
Latimeria_chalumnae_746/1-1020 481 CAGAATAGGGCAGAAATTAAACTGAAGAGAGCTGTAGAAGAAGTAAAAAAGCTACTG 537
Latimeria_chalumnae_36401/1-1083 490 CAGAGCCGAGCGGAGATCAAAATGAAGAGAGCCATAGAGGAGGTGAAGAGGCTACTT 546
Xenopus_tropicalis/1-1026 484 CAAAACAGAGCAGAAATTAAGTTAAAAAGAGCAGTGGAAGAGGTTAAGAAGCTGTTG 540
Gallus_gallus/1-1020 481 CAAAACAGAGCAGAAATAAAACTGAAGAGGGCTGTTGAAGAAGTGAAAAAGTTACTG 537
Mus_musculus/1-1023 481 CAGAACAGAGCAGAAATCAAGCTGAAGAGAGCGGTTGAAGAAGTGAAGAAGTTACTG 537
Homo_sapiens/1-1023 481 CAGAACAGAGCAGAAATCAAATTGAAGAGAGCAGTTGAAGAAGTGAAGAAATTATTG 537
Brachiostoma_floridae/1-1011 469 GAGACCAGAGCACGCATCAAGCTGCAGAGGGCTGTAGAGGAAGTCAAGAAGCTACTC 525
Danio_rerio_chr12/1-1023 538 GTCCCTGCCGCGGAAGGTGAAGACAGCTTAAAGAAAATGCAGCTAATGGAGCTCGCA 594
Danio_rerio_chr13/1-1023 538 GTGCCTGCGGCCGAAGGAGAGGACAGTCTGAAGAAGATGCAGCTGATGGAGCTGGCC 594
Danio_rerio_chr17/1-1026 541 GTGCCTGCAGCAGAAGGAGAAGATAATCTAAAGAAGATGCAGCTGATGGAGCTGGCC 597
Lepisosteus_oculatus_chr16/1-1023 538 GTGCCAGCCGCAGAGGGTGAAGACAGCCTTAAAAAGATGCAGCTGATGGAACTTGCG 594
Lepisosteus_oculatus_chr6/1-1026 541 GTACCTGCGGCTGAGGGAGAGGACAACCTGAAGAAAATGCAGCTGATGGAGCTGGCT 597
Latimeria_chalumnae_746/1-1020 538 GTACCTGCAGCAGAGGGTGAAGATAGCCTTAAGAAGATGCAGTTGATGGAATTAGCC 594
Latimeria_chalumnae_36401/1-1083 547 GTGCCAGCAGCGGAGGGTGAAGACAACCTGAAGAAAATGCAGCTGATGGAACTGGCT 603
Xenopus_tropicalis/1-1026 541 GTCCCAGCAGCAGAAGGAGAAGACAGCCTAAAAAAATGCAGTTGATGGAACTGGCT 597
Gallus_gallus/1-1020 538 ATACCTGCAGCAGAAGGAGAAGACAGCCTTAAAAAGATGCAGCTGATGGAGCTTGCA 594
Mus_musculus/1-1023 538 GTACCTGCGGCTGAAGGTGAAGACAGCCTGAAGAAGATGCAGCTGATGGAGCTTGCA 594
Homo_sapiens/1-1023 538 GTACCTGCAGCAGAAGGAGAAGACAGCCTGAAGAAGATGCAGCTGATGGAGCTTGCG 594
Brachiostoma_floridae/1-1011 526 GTGCCATCGCCTGAAGGTGAGGATGACCTGAAGAAGAGACAGCTGATGGAGCTGGCC 582
Danio_rerio_chr12/1-1023 595 ATACTTAACGGGACATACAGAGACGCCAATATCAAGTCACCAGCCTTAGCCTTCTCC 651
Danio_rerio_chr13/1-1023 595 ATTCTCAACGGCACCTACAGAGACGCCAATATCAAATCACCTGCCTTGCATTCTCT 651
Danio_rerio_chr17/1-1026 598 ATTCTGAACGGAACATACAGAGACACCAACATCAAAGCGCCCACCCTTGCGTTCTCT 654
Lepisosteus_oculatus_chr16/1-1023 595 ATTCTCAATGGCACCTACAGAGACGCCAACATCAAATCACCTGCCCTTGCCTTTTCT 651
Lepisosteus_oculatus_chr6/1-1026 598 ATTCTCAATGGCACATACAGAGATACCAACATTAAGACACCCACCCTTGCGTTTTCT 654
Latimeria_chalumnae_746/1-1020 595 ATTCTCAATGGCACCTACAGAGATGCGAACATCAAATCACCTGCCGTTGCCTTTTCT 651
Latimeria_chalumnae_36401/1-1083 604 ATCCTAAATGGAACCTACAGAGATACCAATATCAAAACACCCACCCTTGCCTTCTCT 660
Xenopus_tropicalis/1-1026 598 ATCCTTAATGGCACCTACAGGGATGCCAACCTAAAATCACCAGCCCTTGCCTTTTCT 654
Gallus_gallus/1-1020 595 ATTCTGAATGGCACCTACAGAGATGCCAACATCAAATCACCAGCCCTTGCCTTTTCT 651
Mus_musculus/1-1023 595 ATTCTGAATGGCACCTACAGAGACGCCAACATTAAATCACCAGCCCTTGCCTTTTCT 651
Homo_sapiens/1-1023 595 ATTCTGAATGGCACCTACAGAGATGCCAACATTAAATCACCAGCCCTTGCCTTTTCT 651
Brachiostoma_floridae/1-1011 583 ATCCTGAACGGGACGTACAGAGACAACAACACCAAAAACATATATAGTTATCTTACT 639
Danio_rerio_chr12/1-1023 652 CTGGCCGCCACAGCA---------CAGGCTCCCCGCATTATGACT---GGCCCCACA 696
Danio_rerio_chr13/1-1023 652 CTTGCAGCGAGTGCC---------CAGGCGCCTCGGATCATCACC---GGCCCTGCA 696
Danio_rerio_chr17/1-1026 655 CTTGCAGCAGCGGCAGCGGCTGCTCAAGGCCCTCGGCTGATCGCGGCCCCTCCGGGT 711
Lepisosteus_oculatus_chr16/1-1023 652 CTTGCAGCAACAGCC-------CAG---GCACCCCGGATCATCACC---GGGCCCACG 696
Lepisosteus_oculatus_chr6/1-1026 655 CTTGCAGCGGCTGCCGCGGCAGCACAGGGCCCCAGGCTCATAGCTGCACCCACGGGG 711
Latimeria_chalumnae_746/1-1020 652 CTTGCAGCAACAGCC-------CAG---GCTCCACGAATCATTACT---GGGCCTGCA 696
Latimeria_chalumnae_36401/1-1083 661 CTTACAGCAGCGGCAGCTGCAGCTCAGGGGCCCAGGTTAATTGCAGCCCCAGCAGCC 717
Xenopus_tropicalis/1-1026 655 CTTGCAGCGACAGCT-------CAG---GCCCCCCGGATTATAACT---GGGCCTGCG 699
Gallus_gallus/1-1020 652 CTTGCAGCAACAGCC-------CAG---GCTCCAAGGATCATTACT---GGGCCTGCG 696
Mus_musculus/1-1023 652 CTTGCAGCAACTGCC-------CAGGCTGCTCCAAGGATCATCACT---GGGCCTGCG 699
Homo_sapiens/1-1023 652 CTTGCAGCAACAGCC-------CAGGCTGCTCCAAGGATCATTACT---GGGCCTGCG 699
Brachiostoma_floridae/1-1011 640 GATCCACTTTGTTACGCTATTGCAGATTCGCCTCGGGGTGTC----------CCCGCC 687
Danio_rerio_chr12/1-1023 697 CCG-----------GTGATGCCCAACGCGGCCCTGCGCACCCCTGCCCCCACCGCC 741
Danio_rerio_chr13/1-1023 697 CCT-----------GTCCTGCCACCCACAGCTCTGCGCACGCCTACGCCTGCGGGA 741
Danio_rerio_chr17/1-1026 712 CAG-----------GTCCTCCCCCCAGCCACCCTCAGGCCCCCCACTCCGGCAGGG 756
Lepisosteus_oculatus_chr16/1-1023 697 CCC-----------GTCCTCCCGCCCACCGCCCTGCGAACGCCCACGCCCGCAGGC 741
Lepisosteus_oculatus_chr6/1-1026 712 CAG-----------GTTCTGCCCCCCACGGCCCTGAGACCCCCAACACCAGCAGGG 756
Latimeria_chalumnae_746/1-1020 697 CCA-----------GTTCTCCCACCAGCTACCCTGCGTACACCTACGCCTGCTGGT 741
Latimeria_chalumnae_36401/1-1083 718 CCT-----------GTACTCCCTCCAGGTGCCCTGCGGCCCCCAACACCGGCTGGC 762
Xenopus_tropicalis/1-1026 700 CCT-----------GTTCTGCCCCCAGCTGCTCTGCGTACACCTACGCCAGCTGGA 744
Gallus_gallus/1-1020 697 CCT-----------GTTCTCCCACCAGCTGCCCTGCGTACTCCTACGCCAGCTGGC 741
Mus_musculus/1-1023 700 CCT-----------GTCCTCCCACCAGCTGCTCTGCGTACACCTACGCCAGCTGGC 744
Homo_sapiens/1-1023 700 CCG-----------GTTCTCCCACCAGCTGCCCTGCGTACTCCTACGCCAGCTGGC 744
Brachiostoma_floridae/1-1011 688 CCCAACACGCCCATGGTGGCGCCTCCGACAGCAGTACGCTCCCCGATGCCCGCGGGC 744
Danio_rerio_chr12/1-1023 742 CCCACCCTCATGCCCCTCATC---------CGACAGATCCAGACGTCTGCCCTCATG 789
Danio_rerio_chr13/1-1023 742 CGCACCATCATGCCTCTCATC---------CGCCAGATCCAGACC------GTCATG 783
Danio_rerio_chr17/1-1026 757 ACCCCCATCATGAACATCATT---------AGGCCCACTCAGATGGCCACTGTGCTA 804
Lepisosteus_oculatus_chr16/1-1023 742 CCTACCATAATGCCTTTGATC---------AGACAGATCCAAACCGCCGCTGTCATG 789
Lepisosteus_oculatus_chr6/1-1026 757 ACCCCCATCATGAACATCATC---------CGGCCCATCCAAACGGCCGCGATGATA 804
Latimeria_chalumnae_746/1-1020 742 CCAACTATAATGCCTTTGATC---------AGACAGATTCAGACGGCCACTGTCATG 789
Latimeria_chalumnae_36401/1-1083 763 ACTCCTATCATGAACATAATC---------CGGCCGATCCAGACAGCTGCAGTCATG 810
Xenopus_tropicalis/1-1026 745 CCTACCTTAATGCCTTTGATC---------AGACAAATACAGACCGCCGCTGTCATG 792
Gallus_gallus/1-1020 742 CCTACCATAATGCCTTTGATC---------AGACAAATACAAACC---GCTGTCATG 786
Mus_musculus/1-1023 745 CCTACCATAATGCCTTTGATC---------AGACAAATACAGACC---GCTGTCATG 789
Homo_sapiens/1-1023 745 CCTACCATAATGCCTTTGATC---------AGACAAATACAGACC---GCTGTCATG 789
Brachiostoma_floridae/1-1011 745 GCGCCTCTCATCGCCACCCCTGTCCTCCAGCGGTTACCTTCGACCCAGCAGATTATG 801


Danio_rerio_chr12/1-1023 790 CCTACGGGCACACCTCACCCCACC---GCCACCCTTCTGCCCCAGACCCCAGAGTCC 843
Danio_rerio_chr13/1-1023 784 CCCAACGGCACGGCCCACCGTGCCACCGCCACCCTAGTGCCGCCCGGTCCAGAGTCG 840
Danio_rerio_chr17/1-1026 805 CCCAATGGCACC--------------CCTACCCTGGTGCCCCCGACGCCTGATGCA 846
Lepisosteus_oculatus_chr16/1-1023 790 CCAAACGGAACCCCTCACCCTACA---GCTACCATTGTCCCACCTGGGCCTGAGTCC 843
Lepisosteus_oculatus_chr6/1-1026 805 CCTAATGGGACG--------------CCAACCTTAGTACCCCCCACGCCAGATGCT 846
Latimeria_chalumnae_746/1-1020 790 CCAAATGGAACGCCTCATCCGACT---GCA---ATAGTTCCTCCTGCACCTGAAGCT 840
Latimeria_chalumnae_36401/1-1083 811 CCGAATGGAACT--------------CCTACCTTAGTGCCTCCAACACCTGAGGCT 852
Xenopus_tropicalis/1-1026 793 CCAAATGGGACCACTCATCCAACT---GCTACACTCATGCCCCAAGCACCTGAAGCT 846
Gallus_gallus/1-1020 787 CCAAACGGAACTCCTCATCCCACC---GCTGCAATAGTGCCACCAGGACCCGAAGCT 840
Mus_musculus/1-1023 790 CCAAACGGAACTCCTCACCCAACT---GCTGCAATAGTCCCTCCAGGGCCTGAAGCT 843
Homo_sapiens/1-1023 790 CCAAACGGAACTCCTCACCCAACT---GCTGCAATAGTTCCTCCAGGGCCCGAAGCT 843
Brachiostoma_floridae/1-1011 802 TCCAATGGGCTTCTT------------CCG------CACATGATGTCTCCCGAGTCC 840
Danio_rerio_chr12/1-1023 844 GGCATCATCTAT---GCACCCTATGATTACCCCTACGCCCTCGCCCCTGCAACCTCC 897
Danio_rerio_chr13/1-1023 841 GGCTTGATTTACGCCACGCCCTACGAGTACCCCTACACGCTGGCACCCGCCACCTCC 897
Danio_rerio_chr17/1-1026 847 GGCATTATCTACACCACGCCCTACGACTACCCTTATGCCTTGGCA---CCCACGTCT 900
Lepisosteus_oculatus_chr16/1-1023 844 GGACTCATCTAC---ACACCCTATGAGTATCCTTACACACTGGCACCCGCGACATCC 897
Lepisosteus_oculatus_chr6/1-1026 847 GGCATCATCTACGCATCGCCCTATGATTACCCCTATGCCCTGGCC---CCCGCATCT 900
Latimeria_chalumnae_746/1-1020 841 GGTCTAATCTAC---ACACCCTATGAGTACCCATACACACTGGCACCAGCTACTTCT 894
Latimeria_chalumnae_36401/1-1083 853 GGAATTATCTACACA---CCCTATGAGTATCCGTATGCACTGGCAACTCCCACCTCT 906
Xenopus_tropicalis/1-1026 847 GGTTTGATCTAC---ACACCTTATGAGTACCCTTATACACTGGCCCCAGCTACTTCA 900
Gallus_gallus/1-1020 841 GGTTTAATTTAC---ACCCCGTACGAGTACCCCTACACGCTGGCACCAGCTACATCA 894
Mus_musculus/1-1023 844 GGGTTAATCTAC---ACACCCTATGAATACCCCTACACATTGGCACCAGCTACATCA 897
Homo_sapiens/1-1023 844 GGTTTAATCTAT---ACACCCTATGAGTACCCCTACACATTGGCACCAGCTACATCA 897
Brachiostoma_floridae/1-1011 841 AACTTTCTGTAC---ACGCCTTACGACTACCCCTACTCCGTGCCT---GCCTCATCC 891
Danio_rerio_chr12/1-1023 898 ATTCTCGAATATCCTATTGACTCCAGTGGTGTTTTAGGTGCAGTGACTACTAAAGTA 954
Danio_rerio_chr13/1-1023 898 ATCCTGGAGTATCCCATCGACTCCAGCGGGGTGTTAGGTGCTGTGGCAACTAAAGTT 954
Danio_rerio_chr17/1-1026 901 CTGCTGGAATACCCCATCGAGCACAGCGGAGTCCTAGGTGCAATGGCTACTAAAGTG 957
Lepisosteus_oculatus_chr16/1-1023 898 ATCCTGGAGTACCCCATCGAGCCCAGCGGTGTATTAGGTGCGGTGGCTACTAAAGTT 954
Lepisosteus_oculatus_chr6/1-1026 901 CTGCTGGAGTACCCCATCGAGCCCAGCGGGGTTCTAGGAGCAATGGCAACTAAAGTG 957
Latimeria_chalumnae_746/1-1020 895 ATCCTGGAGTACCCAATTGAACATAGTGGAGTATTAGGTGCGGTGGCTACTAAAGTT 951
Latimeria_chalumnae_36401/1-1083 907 CTTCTGGAATACCCTGTTGAGCCCAGTGGAGTTCTGGGTGCCATGGCTACCAAAATC 963
Xenopus_tropicalis/1-1026 901 ATCCTTGAGTACCCCATTGATACAAGTGGTGTATTAGGTGCGGTGGCTACTAAAGTA 957
Gallus_gallus/1-1020 895 ATCCTTGAGTATCCCATTGAACCTAGTGGTGTATTAGGTGCGGTGGCTACTAAAGTT 951
Mus_musculus/1-1023 898 ATCCTTGAGTACCCTATTGAACCCAGTGGTGTGTTAGGTGCGGTGGCTACTAAAGTT 954
Homo_sapiens/1-1023 898 ATCCTTGAGTATCCTATTGAACCTAGTGGTGTATTAGGTGCGGTGGCTACTAAAGTT 954
Brachiostoma_floridae/1-1011 892 ATTCTGCAGTACCCGATTGAGCCCAGCGAAGCGATAGGTGCAGTG---ACAAAAGCG 945
Danio_rerio_chr12/1-1023 955 CGAAGGCAGGATATGCGCGTCCATCCTTACCAAAGGGTAGTGACCGCAGAACGAGCC 1011
Danio_rerio_chr13/1-1023 955 CGAAGGCACGAACTGCGCGTCCATCCTTACCAAAGGGTAGTGACCGCAGAAAGAGCC 1011
Danio_rerio_chr17/1-1026 958 CGGAGACATGATTCACGGGTCCATCCTTACCAAAGGATAGTGACCGCAGACAGAGCC 1014
Lepisosteus_oculatus_chr16/1-1023 955 CGAAGGCACGATATGCGGGTCCATCCTTACCAAAGGGTTGTGACCGCAGACCGAGCC 1011
Lepisosteus_oculatus_chr6/1-1026 958 CGGCGACATGACACACGGGTCCATCCTTACCAAAGGATTGTGACCGCAGACAGAGCC 1014
Latimeria_chalumnae_746/1-1020 952 CGAAGGCACGATATGCGTGTCCATCCTTACCAAAGGATTGTGACCGCAGACAGAGCC 1008
Latimeria_chalumnae_36401/1-1083 964 CGGCGACATGACACTCGTATCCATCCTTACCAAAGGATTGTGACCACAGACAGAGGT 1020
Xenopus_tropicalis/1-1026 958 AGAAGGCATGATATGCGGGTCCATCCTTACCAAAGGATTGTGACCGCAGACAGAGCC 1014
Gallus_gallus/1-1020 952 CGAAGGCACGATATGCGTGTCCATCCTTACCAAAGGATTGTGACCGCAGACCGAGCC 1008
Mus_musculus/1-1023 955 CGAAGGCACGATATGCGTGTCCATCCTTACCAAAGGATTGTGACCGCAGACCGAGCC 1011
Homo_sapiens/1-1023 955 CGAAGGCACGATATGCGTGTCCATCCTTACCAAAGGATTGTGACCGCAGACCGAGCC 1011
Brachiostoma_floridae/1-1011 946 CGGAGGCATGTGGTTAGGGATCACCCCTATCAAAGGGCTGTCATA---GACCGAGCC 999
Danio_rerio_chr12/1-1023 1012 GCCACCGGCAAC 1023
Danio_rerio_chr13/1-1023 1012 ACCACTGGCAAC 1023
Danio_rerio_chr17/1-1026 1015 GCCACCGGCAAC 1026
Lepisosteus_oculatus_chr16/1-1023 1012 GCCACCGGCAAC 1023
Lepisosteus_oculatus_chr6/1-1026 1015 GCCACCGGCAAC 1026
Latimeria_chalumnae_746/1-1020 1009 GCCACCGGCAAC 1020
Latimeria_chalumnae_36401/1-1083 1021 CAGTTCAGTGACTTCTACCGGCAACTAACCCATGACCTTCTGACCTCTGAACTCTTC 1077
Xenopus_tropicalis/1-1026 1015 GCCACCGGCAAC 1026
Gallus_gallus/1-1020 1009 GCCACCGGCAAC 1020
Mus_musculus/1-1023 1012 GCCACCGGCAAC 1023
Homo_sapiens/1-1023 1012 GCCACCGGCAAC 1023
Brachiostoma_floridae/1-1011 1000 GCCGCCACCAAC 1011
Danio_rerio_chr12/1-1023 ------
Danio_rerio_chr13/1-1023 ------
Danio_rerio_chr17/1-1026 ------
Lepisosteus_oculatus_chr16/1-1023 ------
Lepisosteus_oculatus_chr6/1-1026 ------
Latimeria_chalumnae_746/1-1020 ------
Latimeria_chalumnae_36401/1-1083 1078 ACCCAA 1083
Xenopus_tropicalis/1-1026 ------
Gallus_gallus/1-1020 ------
Mus_musculus/1-1023 ------
Homo_sapiens/1-1023 ------
Brachiostoma_floridae/1-1011 ------

Table 2. Multiple protein alignment for the isoform QKI 5





Species tree. Made in http://www.ncbi.nlm.nih.gov/guide/taxonomy/

Table **3**. jModelTest2 statistical selection of nucleotide substitution models

Likelihood scores

					----------base frequencies----------					----------Substitution model: rate matrix----------						G
ID	Name	-lnL	Partition	parameters	fA	fC	fG	fT	ti/tv	R(a)	R(b)	R(c)	R(d)	R(e)	R(f)	shape
88	GTR+I+G	7170.8016	12345	32	0.2604	0.2949	0.2653	0.1794	-	2.08318	5.15447	2.65998	1.30327	7.28638	1.0	0.3570
87	GTR+G	7170.8019	12345	31	0.2604	0.2949	0.2653	0.1794	-	2.08129	5.14542	2.65477	1.30117	7.27012	1.0	0.3570
63	TIM2+G	7172.0176	10232	29	0.2604	0.2933	0.2655	0.1809	-	1.94237	4.33483	1.94237	1.0	6.1165	1.0	0.3570
64	TIM2+I+G	7172.0177	10232	30	0.2605	0.2932	0.2655	0.1808	-	1.94198	4.33363	1.94198	1.0	6.11834	1.0	0.3570
80	TVM+I+G	7173.6159	12314	31	0.2527	0.3036	0.2569	0.1867	-	2.1962	6.27176	2.75619	1.33463	6.27176	1.0	0.3540
79	TVM+G	7173.6161	12314	30	0.2527	0.3036	0.257	0.1867	-	2.19389	6.25957	2.75022	1.33182	6.25957	1.0	0.3540
39	TPM2uf+G	7174.8281	10212	28	0.2527	0.3022	0.2571	0.1879	-	2.00171	5.19156	2.00171	1.0	5.19156	1.0	0.3540
40	TPM2uf+I+G	7174.8281	10212	29	0.2527	0.3022	0.2571	0.1879	-	2.00082	5.19073	2.00082	1.0	5.19073	1.0	0.3540
72	TIM3+I+G	7182.0594	12032	30	0.2755	0.2944	0.2501	0.18	-	0.93612	2.83314	1.0	0.93612	4.01243	1.0	0.3610
71	TIM3+G	7182.0595	12032	29	0.2755	0.2944	0.2501	0.18	-	0.93616	2.83326	1.0	0.93616	4.01092	1.0	0.3610
55	TIM1+G	7182.0602	12230	29	0.276	0.2931	0.2497	0.1812	-	1.0	3.04398	1.06606	1.06606	4.29362	1.0	0.3610
56	TIM1+I+G	7182.0602	12230	30	0.276	0.2931	0.2497	0.1812	-	1.0	3.04434	1.06637	1.06637	4.29632	1.0	0.3610
24	TrN+I+G	7182.1467	10020	29	0.2757	0.293	0.2501	0.1812	-	1.0	2.94638	1.0	1.0	4.16483	1.0	0.3610
23	TrN+G	7182.1468	10020	28	0.2756	0.293	0.2501	0.1812	-	1.0	2.94642	1.0	1.0	4.16178	1.0	0.3610
32	TPM1uf+I+G	7184.8435	12210	29	0.2678	0.3022	0.2417	0.1882	-	1.0	3.59792	1.0758	1.0758	3.59792	1.0	0.3580
31	TPM1uf+G	7184.8437	12210	28	0.2678	0.3022	0.2417	0.1882	-	1.0	3.59676	1.07537	1.07537	3.59676	1.0	0.3580
48	TPM3uf+I+G	7184.898	12012	29	0.2673	0.3033	0.2421	0.1873	-	0.95075	3.36852	1.0	0.95075	3.36852	1.0	0.3580
47	TPM3uf+G	7184.8981	12012	28	0.2673	0.3033	0.2421	0.1873	-	0.9502	3.36658	1.0	0.9502	3.36658	1.0	0.3580
15	HKY+G	7184.9565	10010	27	0.2675	0.3021	0.2421	0.1883	1.6879	-	-	-	-	-	-	0.3580
16	HKY+I+G	7184.9566	10010	28	0.2675	0.3021	0.2421	0.1883	1.6884	-	-	-	-	-	-	0.3580

jModelTest2 statistical selection of nucleotide substitution models based on "Phyml" algoritm [1] [2]. ID and name (in orange) of the models are shown, followed by the likelihood value (in blue). The model with the lowest lnL value has the highest likelihood (best fitting for the data of this study), in this analysis *GTR+I+G*. Immediately after is the model *GTR+G* with almost identical likelihood score. The partition column gives the numerical notation of the rates of substitution (=substitution code), the numers representing rates index in the order: AC, AG, AT, CG, CT, GT. For each model number of parameters, relative base frequencies and transition/transversion ratio (ti/tv) values are also presented. The rate matrix defines the substitution model, values for each set of base substitution are given in this sense: R(a) being AC, R(b) being AG and so on. Last column gives the values for gamma shape parameter. The analysis involved 12 nucleotide sequences and 1146 sites.
The settings for the analysis were as follows: candidate models = 88, number of substitution schemes = 11, including models with equal/unequal base frequencies (+F), including models with/without a proportion of invariable sites (+I), including models with/without rate variation among sites (+G) (nCat = 4), optimized free parameters (K) = substitution parameters + 21 branch lengths + topology, base tree for likelihood calculations = ML tree, tree topology search operation = BEST.

1. Darriba D., Taboada G.L., Doallo R., Posada D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8), 772.
2. Guindon S. and Gascuel O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood". Systematic Biology 52: 696-704.

For each model, AIC and AICc (Akaike Information Criterion, corrected), and BIC values are also shown.

ID	Name	AIC	deltaAIC	weight	cumWeight
63	TIM2+G	14402.0352	0.0	0.4314	0.4314
87	GTR+G	14403.6037	1.5685	0.1969	0.6283
64	TIM2+I+G	14404.0354	2.0001	0.1587	0.787
88	GTR+I+G	14405.6031	3.5679	0.0725	0.8594
39	TPM2uf+G	14405.6563	3.621	0.0706	0.93
79	TVM+G	14407.2322	5.197	0.0321	0.9621
40	TPM2uf+I+G	14407.6562	5.6209	0.026	0.9881
80	TVM+I+G	14409.2317	7.1965	0.0118	0.9999
23	TrN+G	14420.2937	18.2584	0.0	0.9999
71	TIM3+G	14422.1189	20.0837	0.0	0.9999
55	TIM1+G	14422.1204	20.0851	0.0	1.0
24	TrN+I+G	14422.2934	20.2582	0.0	1.0
15	HKY+G	14423.913	21.8778	0.0	1.0
72	TIM3+I+G	14424.1189	22.0836	0.0	1.0
56	TIM1+I+G	14424.1203	22.0851	0.0	1.0
31	TPM1uf+G	14425.6873	23.6521	0.0	1.0
47	TPM3uf+G	14425.7962	23.7609	0.0	1.0
16	HKY+I+G	14425.9131	23.8779	0.0	1.0
32	TPM1uf+I+G	14427.687	25.6518	0.0	1.0
48	TPM3uf+I+G	14427.7959	25.7607	0.0	1.0

ID	Name	AICc	deltaAICc	weight	cumWeight
63	TIM2+G	14403.5944	0.0	0.4491	0.4491
87	GTR+G	14405.3847	1.7903	0.1835	0.6326
64	TIM2+I+G	14405.7035	2.1092	0.1564	0.7891
39	TPM2uf+G	14407.1102	3.5158	0.0774	0.8665
88	GTR+I+G	14407.5007	3.9063	0.0637	0.9302
79	TVM+G	14408.9004	5.306	0.0316	0.9618
40	TPM2uf+I+G	14409.2153	5.6209	0.027	0.9889
80	TVM+I+G	14411.0127	7.4183	0.011	0.9999
23	TrN+G	14421.7476	18.1532	1.0E-4	0.9999
71	TIM3+G	14423.6781	20.0837	0.0	0.9999
55	TIM1+G	14423.6795	20.0851	0.0	0.9999
24	TrN+I+G	14423.8526	20.2582	0.0	1.0
15	HKY+G	14425.2654	21.6711	0.0	1.0
72	TIM3+I+G	14425.787	22.1927	0.0	1.0
56	TIM1+I+G	14425.7885	22.1941	0.0	1.0
31	TPM1uf+G	14427.1412	23.5469	0.0	1.0
47	TPM3uf+G	14427.2501	23.6557	0.0	1.0
16	HKY+I+G	14427.367	23.7726	0.0	1.0
32	TPM1uf+I+G	14429.2462	25.6518	0.0	1.0
48	TPM3uf+I+G	14429.3551	25.7607	0.0	1.0

ID	Name	BIC	deltaBIC	weight	cumWeight
39	TPM2uf+G	14546.8892	0.0	0.6486	0.6486
63	TIM2+G	14548.3122	1.423	0.3184	0.9671
40	TPM2uf+I+G	14553.9331	7.0439	0.0192	0.9862
64	TIM2+I+G	14555.3563	8.4671	0.0094	0.9956
79	TVM+G	14558.5532	11.664	0.0019	0.9975
87	GTR+G	14559.9687	13.0795	9.0E-4	0.9985
15	HKY+G	14560.1019	13.2127	9.0E-4	0.9994
23	TrN+G	14561.5266	14.6374	4.0E-4	0.9998
80	TVM+I+G	14565.5967	18.7075	1.0E-4	0.9998
31	TPM1uf+G	14566.9199	20.0307	0.0	0.9999
88	GTR+I+G	14567.0122	20.123	0.0	0.9999
47	TPM3uf+G	14567.0288	20.1396	0.0	0.9999
16	HKY+I+G	14567.146	20.2568	0.0	1.0
71	TIM3+G	14568.3959	21.5067	0.0	1.0
55	TIM1+G	14568.3973	21.5081	0.0	1.0
24	TrN+I+G	14568.5703	21.6811	0.0	1.0
32	TPM1uf+I+G	14573.964	27.0748	0.0	1.0
48	TPM3uf+I+G	14574.0729	27.1837	0.0	1.0
72	TIM3+I+G	14575.4399	28.5507	0.0	1.0
56	TIM1+I+G	14575.4413	28.5521	0.0	1.0

Table **4.** Maximum Likelihood fits of 24 different nucleotide substitution models

Model	#Param	BIC	AICc	lnL	Invariant	Gamma	R	Freq A	Freq T	Freq C	Freq G	A=>T	A=>C	A=>G	T=>A	T=>C	T=>G	C=>A	C=>T	C=>G	G=>A	G=>T	G=>C
TN93+G	27	*12991,138*	12793,457	-6369,661	n/a	0,34681	1,794268	0,296563	0,192842	0,256410	0,254184	0,03	0,05	0,13	0,05	0,21	0,04	0,05	0,16	0,04	0,15	0,03	0,05
HKY+G	26	12997,994	12807,630	-6377,752	n/a	0,34440	1,807956	0,296563	0,192842	0,256410	0,254184	0,03	0,05	0,16	0,05	0,16	0,05	0,05	0,12	0,05	0,19	0,03	0,05
TN93+G+I	28	13000,465	12795,467	-6369,661	0	0,34681	1,794268	0,296563	0,192842	0,256410	0,254184	0,03	0,05	0,13	0,05	0,21	0,04	0,05	0,16	0,04	0,15	0,03	0,05
GTR+G	30	13004,509	*12784,879*	*-6362,357*	n/a	0,34393	1,779738	0,296563	0,192842	0,256410	0,254184	0,04	0,06	0,13	0,06	0,21	0,02	0,06	0,16	0,04	0,15	0,02	0,04
HKY+G+I	27	13007,321	12809,640	-6377,752	0	0,34440	1,807956	0,296563	0,192842	0,256410	0,254184	0,03	0,05	0,16	0,05	0,16	0,05	0,05	0,12	0,05	0,19	0,03	0,05
K2+G	23	13012,406	12843,995	-6398,948	n/a	0,35057	1,737779	0,25	0,25	0,25	0,25	0,05	0,05	0,16	0,05	0,16	0,05	0,05	0,16	0,05	0,16	0,05	0,05
GTR+G+I	31	13013,835	12786,890	-6362,357	0	0,34393	1,779738	0,296563	0,192842	0,256410	0,254184	0,04	0,06	0,13	0,06	0,21	0,02	0,06	0,16	0,04	0,15	0,02	0,04
T92+G	24	13015,490	12839,761	-6395,827	n/a	0,34827	1,748814	0,244703	0,244703	0,255297	0,255297	0,04	0,05	0,16	0,04	0,16	0,05	0,04	0,16	0,05	0,16	0,04	0,05
K2+G+I	24	13021,733	12846,003	-6398,948	0	0,35057	1,737779	0,25	0,25	0,25	0,25	0,05	0,05	0,16	0,05	0,16	0,05	0,05	0,16	0,05	0,16	0,05	0,05
T92+G+I	25	13024,817	12841,769	-6395,827	0	0,34827	1,748814	0,244703	0,244703	0,255297	0,255297	0,04	0,05	0,16	0,04	0,16	0,05	0,04	0,16	0,05	0,16	0,04	0,05
TN93+I	27	13238,581	13040,900	-6493,383	0,41473	n/a	1,465560	0,296563	0,192842	0,256410	0,254184	0,04	0,05	0,13	0,06	0,18	0,05	0,06	0,13	0,05	0,15	0,04	0,05
HKY+I	26	13240,033	13049,669	-6498,772	0,41702	n/a	1,487680	0,296563	0,192842	0,256410	0,254184	0,04	0,05	0,15	0,06	0,15	0,05	0,06	0,11	0,05	0,18	0,04	0,05
GTR+I	30	13253,305	13033,675	-6486,755	0,41591	n/a	1,459773	0,296563	0,192842	0,256410	0,254184	0,05	0,05	0,13	0,08	0,18	0,03	0,06	0,13	0,05	0,15	0,03	0,05
K2+I	23	13254,970	13086,558	-6520,230	0,41608	n/a	1,436519	0,25	0,25	0,25	0,25	0,05	0,05	0,15	0,05	0,15	0,05	0,05	0,15	0,05	0,15	0,05	0,05
T92+I	24	13258,742	13083,012	-6517,452	0,41659	n/a	1,445452	0,244703	0,244703	0,255297	0,255297	0,05	0,05	0,15	0,05	0,15	0,05	0,05	0,14	0,05	0,14	0,05	0,05
JC+G	22	13277,300	13116,206	-6536,058	n/a	0,38114	0,5	0,25	0,25	0,25	0,25	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08
JC+G+I	23	13286,626	13118,215	-6536,058	0	0,38114	0,5	0,25	0,25	0,25	0,25	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08
JC+I	22	13492,639	13331,546	-6643,728	0,41384	n/a	0,5	0,25	0,25	0,25	0,25	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08
TN93	26	13795,662	13605,298	-6776,586	n/a	n/a	1,255371	0,296563	0,192842	0,256410	0,254184	0,04	0,06	0,12	0,07	0,18	0,06	0,07	0,13	0,06	0,13	0,04	0,06
GTR	29	13810,987	13598,674	-6770,259	n/a	n/a	1,257581	0,296563	0,192842	0,256410	0,254184	0,05	0,06	0,12	0,07	0,18	0,04	0,07	0,13	0,06	0,14	0,03	0,06
HKY	25	13812,699	13629,652	-6789,768	n/a	n/a	1,254703	0,296563	0,192842	0,256410	0,254184	0,04	0,06	0,14	0,07	0,14	0,06	0,07	0,11	0,06	0,16	0,04	0,06
K2	22	13813,599	13652,506	-6804,208	n/a	n/a	1,254920	0,25	0,25	0,25	0,25	0,06	0,06	0,14	0,06	0,14	0,06	0,06	0,14	0,06	0,14	0,06	0,06
T92	23	13821,967	13653,555	-6803,728	n/a	n/a	1,256517	0,244703	0,244703	0,255297	0,255297	0,05	0,06	0,14	0,05	0,14	0,06	0,05	0,14	0,06	0,14	0,05	0,06
JC	21	14026,601	13872,826	-6915,372	n/a	n/a	0,5	0,25	0,25	0,25	0,25	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08

NOTE. In MEGA6 models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best [1], in this case *TN93+G* (Tamura and Nei 1993), in green. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (lnL), and the number of parameters are also presented [2]. AICc (yellow) and lnL (in blue) best values correspond to the *GTR+G* model (General Time Reversible). Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+G) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+I). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown. Assumed or estimated values of transition/transversion bias (R) are shown for each model. They are followed by nucleotide frequencies and rates of base substitutions (r) for each nucleotide pair. For simplicity, sum of r values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. The analysis involved 12 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 936 positions in the final dataset. Analyses were conducted in MEGA6.

1. Tamura K., Stecher G., Peterson D., Filipski A., and Kumar S. (2013). *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Molecular Biology and Evolution30: 2725-2729.

2. Nei M. and Kumar S. (2000). *Molecular Evolution and Phylogenetics.* Oxford University Press, New York.

The end.