Title

Unexpected predicted length variation for the coding sequence of the sleep related gene, *BHLHE41* in gorilla amidst strong purifying selection across mammals

Short title

Mammalian BHLHE41 evolution

Krishna Unadkat[1] and Justen B. Whittall[1]*

[1]Department of Biology

Santa Clara University

500 El Camino Real

Santa Clara, CA 95053

* Corresponding author: jwhittall@scu.edu (JW)

Abstract

There is a molecular basis for many sleep patterns and disorders involving circadian clock genes. In humans, “short-sleeper” behavior has been linked to specific amino acid substitutions in *BHLHE41* (DEC2), yet little is known about variation at these sites and across this gene in mammals. We compare *BHLHE41* coding sequences for 27 mammals. The coding sequence alignment length was 1794bp, of which 55.0% of base pairs were invariant among the sampled mammals. The mean pairwise nucleotide identity was 92.2%. Of the 598 residue amino acid alignment for mammals, 71.7% of amino acids were identical. The pairwise percent identity for amino acids was 94.8%. No other mammals had the same “short-sleeper” amino acid substitutions previously described from humans. Phylogenetic analyses based on the nucleotides of the coding sequence alignment are consistent with established mammalian relationships. Significant purifying selection was detected in 66.2% of variable codons. No codons exhibited significant signs of positive selection. Unexpectedly, the gorilla *BHLHE41* sequence has a 318 bp insertion at the 5’ end of the coding sequence and a deletion of 195 bp near the 3’ end of the coding sequence (including the two short sleeper variable sites). Given the strong signal of purifying selection across this gene, phylogenetic congruence with expected relationships and generally conserved function among mammals investigated thus far, we suggest the unexpected indels predicted in the gorilla *BHLHE41* may represent an annotation error and warrant experimental validation.

Keywords: *BHLHE41*, DEC2, gorilla, indel, mammal, purifying selection

Introduction

Sleep plays a vital function for survival in animals [1-3], especially vertebrates and even some invertebrates [4]. It is essential in maintaining both physical and mental health, especially in humans [5]. Adequate sleep is critical for normal daily functioning for humans including thinking, learning, reacting, and memory [6]. In addition to having a negative impact on daily life, a lack of sleep can lead to an increased risk of chronic health problems such as diabetes, high blood pressure, obesity, and decreased immune function [7]. The timing and duration of sleep varies widely among mammals [8] and is regulated by a plethora of intricate mechanisms including many circadian clock genes [9].

Among the genes responsible for circadian regulation in mammals, is basic helix-loop-helix family member e41 [5, 10, 11], also known as differentially expressed in chondrocytes protein 2 (*DEC2*). It is an essential clock protein that acts as a transcription factor which plays an important role in maintaining the negative feedback loop in the circadian clock by repressing E-box-mediated transcription [5]. Specifically, by binding to the promoter region on the *prepro-orexin* gene, *BHLHE41* acts as a repressor of orexin expression in mammals. Furthermore, disabling orexin results in narcolepsy in mammals, suggesting orexin plays a vital role in sleep regulation [5].

BHLHE41 has several conserved functional domains including a bHLH region and the “orange” domain. As a member of the bHLH family, *BHLHE41* contains a ~60 amino acid bHLH conserved domain that promotes dimerization and DNA binding [10]. Specifically, the HLH domain is composed of a DNA-binding region, E-box/N-box specificity site, and a dimerization interface for polypeptide binding. The DNA-binding region is followed by two alpha-helices surrounding a variable loop region. As part of the group E bHLH subtype, this protein specifically binds to an N-box sequence (CACGCG or CACGAG) based on *BHLHE41* amino acid site 53 (glutamate) [12]. The other well studied conserved domain in *BHLHE41* is the orange domain which provides specificity as a transcriptional repressor [13]. These domains are conserved between humans and zebrafish in both their amino acid composition and function [14]. Unfortunately, there is no 3D structure described for a mammalian *BHLHE41* in Genbank's Protein Data Bank [15].

Because of its essential function in sleep regulation, anomalies in clock genes can lead to abnormal patterns of sleep that can manifest in a wide variety of ways, ranging from insomnia to oversleeping [1]. A rare point mutation in the *BHLHE41* gene of *Homo sapiens* (P384R in NM_030762, also referred to as P385R as in He et al. 2009) confers a "short-sleeper phenotype" [10]. The mutation involves a transversion from a C to G in the DNA sequence of *BHLHE41*, which results in a non-synonymous substitution from proline to arginine at amino acid position 385 of the *BHLHE41* protein. Since proline (nonpolar) and arginine (electrically charged, basic) have chemically dissimilar structures and since substituting these amino acids is relatively rare (BLOSUM62 value of -2), it is not surprising that this mutation has a substantial phenotypic effect. Subjects with this allele reported shorter daily sleep patterns than those with the wild type allele, without reporting any other adverse effects [10]. This might be attributable to higher orexin expression in mutants due to altered repressor function. The function of *BHLHE41* in controlling sleep and circadian clocks is conserved between humans and mice, but untested in most other mammals [10]. In zebrafish, the *BHLHE41* has similar structure (five exons separated by four introns) and high sequence similarity to human homologue [14], but no variation at this residue. In *Drosophila melanogaster*, the most similar gene to *BHLHE41* is CG17100 (clockwork orange), but is only weakly similar (<11% amino acid identity; [16]). However, transgenically introducing the short-sleeper allele P385R into *Drosophila* still resulted in the short-sleeper phenotype [10] suggesting the existence of a similar regulatory network. Another nonsynonymous substitution in *BHLHE41* that correlates with altered sleep behavior in humans is Y362H [17]. This mutation reduced the ability of *BHLHE41* to suppress CLOCK/BMAL1 and NPAS2/BMAL1 transactivation in vitro [17].

These short-sleeper variants could provide adaptive functions in other mammals. In such case, we may detect the signature of positive selection on those codons. However, genes such as *BHLHE41* are essential for survival and reproduction (e.g., housekeeping genes) are often highly conserved and are more likely to show patterns of purifying selection. Purifying selection can be manifested as higher rates of synonymous substitutions compared to rates of non-synonymous substitutions (dN-dS) [18]. Negative overall dN-dS values indicate purifying selection and are often evidence that a gene is involved in some essential function (like the circadian clock), yet a codon-by-codon dN/dS analysis can detect signs of positive selection (e.g. adaptation at the molecular level) on specific codons. To date, no one has compared dN and dS in *BHLHE41*.

In fact, neither nucleotide, nor amino acid sequence comparisons have been made for *BHLHE41* among mammals beyond human-mouse comparisons. With the rapid accumulation of mammalian genome sequences, a plethora of homologous sequences likely exist (see Ledent et al. 2002 for phylogenetic analysis of all bHLH, but only includes two mammals - human and mouse; see Abe et al. 2006 for a comparison of zebrafish and human that calls for further sampling of mammals). Furthermore, the well-resolved mammalian phylogeny [19, 20] provides a robust foundation for which to detect unusual patterns of evolution (incongruent relationships and/or unusual branchlengths). For most non-model mammalian species with whole-genome sequences, genes are predicted using algorithms that locate open reading frames (e.g., [21]), yet rarely are the predicted genes validated experimentally [22, 23]. Some algorithms compare putative open reading frames with model-species to confirm length and expected sequence variation. Accounting for any differences in the length of coding sequences can be a challenge, due to both the existence of alternative mRNA isoforms and an increasing time of divergence [24]. A comparative approach across a diversity of lineages can help elucidate any unusual patterns of sequence variation.

In order to further explore the function of the *BHLHE41* gene, we analyzed the evolutionary relationships among the *BHLHE41* coding sequence in humans and other mammals. The aim of this study was to examine how the effects of selection at the molecular scale manifest themselves across mammals. Specifically, the study utilized pre-existing data in Genbank to determine whether any mammals other than humans have the “short-sleeper” allele or exhibit variation at amino acid sites P385R and Y362H. Additionally, through this study, we examined many additional amino acid substitutions, assessed the degree of biochemical changes and searched for the footprints of selection (dN/dS). We compared *BHLHE41* sequences from 27 species of mammals and a reptilian outgroup that came from sequenced cDNA and full genome sequencing projects. After creating a multiple sequence alignment, we used Bayesian and maximum likelihood analyses to investigate the evolutionary relationships underlying this gene among mammals. Finally, we used the multiple sequence alignment to test for purifying and positive selection across codons.

Materials and Methods

Query Sequence Search

In order to find the complete mRNA coding sequence for *BHLHE41* from *H. sapiens*, we searched ENTREZ using the “RefSeq” filter with the following query to the “Gene” database: “DEC2 AND *H. sapiens* [organism]”. We confirmed that the same sequence was obtained when searching for *BHLHE41* AND *H. sapiens*.

Locating Homologous Sequences with BLAST

After locating the accession number for our sequence of interest from *H. sapiens*, we used NCBI’s nucleotide BLAST [25] to find other mammalian homologues to the *H. sapiens* *BHLHE41* mRNA. We searched the nucleotide collection (nr/nt) using Megablast with default parameters (Word Size: 28, Match/Mismatch: 1, -2, Gap Costs: Linear). We downloaded sequences with E-values < 10^{-3}, local percent identity > 70%, and query coverages ~100% as Genbank complete flatfiles.

In order to find an outgroup sequence, we performed another BLAST search using Discontiguous Megablast with default parameters (Word size: 1, Match/Mismatch: 2, -3, Gap Existence/Extension Costs: 5, 2) except excluding mammals from the search results. We included the reptile, *Pelodiscus sinensis* or Chinese Softshell Turtle, as our outgroup based on the aforementioned E-value, identity, and query coverage cut-offs. GenBank flatfiles for each species coding sequence was downloaded and imported into Geneious Prime (Biomatters, New Zealand).

Multiple Sequence Alignment

In order to create an alignment with sequences that represent homology to the *H. sapiens* BHLHE4 mRNA, we used the Geneious Aligner within Geneious Prime. To prevent single nucleotide gaps and ensure all remaining nucleotide gaps were in multiples of three, since this is coding sequence, we applied a cost matrix of 70% similarity (match/mismatch of 5.0/-4.5), a gap open penalty of 90, a gap extension penalty of one, and two refinement iterations.

Phylogenetic Analyses

We compared maximum likelihood and Bayesian phylogenetic analyses to examine the evolutionary relationships of the *BHLHE41* coding sequence among mammals. In order to construct a maximum likelihood tree, we used the RAxML v.4.0 [26] plugin in Geneious Prime. We applied the GTR+CAT+I model of sequence evolution, with the Rapid Bootstrapping algorithm, 1,000 bootstrap replicates, and a Parsimony Random Seed of one. This is the most complex model of sequence evolution available for the RAxML plugin in Geneious Prime. It accounts for six rates of nucleotide substitution with categories for rate variation instead of a gamma distribution for efficiency, while simultaneously estimating the proportion of invariant sites [26].

To compare our maximum likelihood results with another method, we constructed a phylogenetic tree using the MrBayes v.2.2.4 [27] plugin from within Geneious Prime. For this analysis, we used the GTR (General Time Reversible) model of sequence evolution with “gamma” rate variation. The search ran for 2,000,000 generations, subsampling every 1,000 generations after 1,000,000 generations of burnin. Two parallel runs were conducted using four chains each with a heated chain temp of 0.2. In order to confirm sufficient number of generations were sampled in the Bayesian analysis, we recorded the standard deviation of split frequencies comparing the two runs. Furthermore, we examined the trace depicting the maximum likelihood value at each generation to ensure there was no slope (S1 Fig). After both maximum likelihood and Bayesian trees were generated, we rooted them with the reptilian outgroup, *P. sinensis* (Chinese Softshell Turtle).

Molecular Evolution

By comparing the rates of nonsynonymous (dN) and synonymous (dS) substitutions, we tested for selection at the molecular scale. In MEGA7 [28], we used the codon-based Z-test of selection to test for pairwise dN-dS values, using “In Sequence Pairs” as the scope, “Positive Selection” as the test hypothesis, the “Nei-Gojobori method (Proportional)” as the model [29], and “Pairwise Deletion” to account for gaps without removing sites entirely. We then repeated this process using “Purifying Selection” as the test hypothesis. Purifying selection was represented by negative dN-dS values, positive selection was represented by positive dN-dS values. dN-dS values of zero represent neutrality. For the codon-based Z-test of selection, p-values under 0.05 were considered significant.

In order to determine if there was directional selection on any specific codons, we used HyPhy [30] from within MEGA7. We used a “Neighbor-Joining tree”, “Maximum Likelihood” statistical method, “Syn-Nonsynonymous” substitution, and the “General Time Reversible” model of sequence evolution to analyze the alignment codon-by-codon. We applied the partial deletion option if <70% of the sequences had a gap. After running HyPhy, we removed invariant codons where dN and dS could not be calculated and examined the remaining codons with significant P-values. Values greater than 0.95 were considered significant evidence of purifying selection. We estimated the average dN-dS values for both conserved domains compared to the remaining codons outside the conserved domains.

Results

Query Sequences

We found a single hit for the *Homo sapiens BHLHE41* gene with the RefSeq accession number NM_030762 [31]. The coding sequence for this gene is 1449 base pairs long. According to EMBL (ENSGGOT00000015550.3), there are five introns (yet see Abe et al. 2006 where they report only four introns). All subsequent analyses are based solely on the coding sequence as determined by EMBL.

Locating Homologous Sequences with BLAST

From the results of the BLAST search using *BHLHE41* from *H. sapiens*, we downloaded 27 mammalian sequences (including the query) and one reptile sequence as an outgroup for a total alignment of 28 species.The E-values for all sequences were 0.0 and the local identity scores from the BLAST report ranged from 87.50% to 100% (Table 1). The coding sequences ranged in length from 1368 (*P. sinensis*) to 1569 (*Gorilla gorilla gorilla*) base pairs. The query coverage from the BLAST report ranged from 38% to 100% (Table 1).

Table 1. Sequences used in creating the multiple sequence alignment and their BLAST scores using the mRNA from the human basic helix-loop-helix family member e41 as the query.

Latin name	Accession number	Query coverage (%)	Identity (%)
Bos indicus x Bos taurus	XM_027541573	64	91.97
Callorhinus ursinus	XM_025879601	82	92.94
Cebus capucinus	XM_017507035	94	96.60
Cercocebus atys	XM_012093655	46	98.21
Chlorocebus sabaeus	XM_007967990	99	97.96
Delphinapterus leucas	XM_022577811	95	87.54
Homo sapiens[1]	NM_030762	100	100

Gorilla gorilla gorilla	XM_019037881	89	98.92
Lagenorhynchus obliquidens	XM_027129408	89	87.27
Lipotes vexillifer	XM_007446307	38	93.63
Macaca fascicularis	XM_005570417	100	98.25
Macaca mulatta	XM_015151321	49	98.13
Macaca nemestrina	XM_011759130	100	98.08
Marmota flaviventris	XM_027934162	52	92.03
Microcebus murinus	XM_012739537	93	93.71
Orcinus orca	XM_004270956	51	89.93
Ovis aries	XM_015093964	67	92.73
Panthera pardus	XM_019452268	60	92.65
Pan troglodytes	XM_520805	49	99.15
Pelodiscus sinesis [2]	XM_006127674	49	88.53
Physeter catodon	XM_024128992	85	87.50
Piliocolobus tephrosceles	XM_023209042	100	97.17
Pongo abelii	XM_002823045	48	98.98
Rousettus aegyptiacus	XM_016119294	92	91.98
Sus scrofa	XM_003355541	79	92.49
Theropithecus gelada	XM_025402281	94	97.23
Tursiops truncatus	XM_019936346	95	87.54
Zalophus californianus	XM_027593397	87	92.89

[1] Query sequence
[2] Reptile outgroup

Multiple Sequence Alignment

All sequences in the multiple sequence alignment are complete from start codon (AUG) to stop codon (all use TGA) (S3 Table). Indels ranged from three base pairs to 318 base pairs - always in multiples of three. Of the 1,794 bp nucleotide alignment for mammals, 986 bp were identical (55.0%). The average nucleotide pairwise identity among the mammalian sequences was 92.2%. At the amino acid level, of the 598 residues for mammals, 71.7% were identical. The pairwise percent identity in amino acids was 94.8% (S4 Table).

There are two large indels in the gorilla sequence (Fig 1; S3-4 Table). The first 318 base pairs are only present in accession XM_019037881 - a predicted protein from the *G. gorilla gorilla* genome sequence [32]. Additionally, the sequence for *G. gorilla gorilla* has a 195 base pair deletion starting at nucleotide alignment site 1,360 and ending at 1,555bp. Both of these indels are multiples of three and therefore maintain the reading frame throughout the coding sequence yielding a predicted *G. gorilla gorilla BHLHE41* amino acid sequence 522 residues. The average non-gorilla mammalian amino acid sequence is 482aa long.

Fig 1. Multiple sequence alignment of the *BHLHE41* mRNA for 27 mammals and one reptile outgroup. The alignment shows a 318 base pair insertion in the gorilla sequence on the 5' end. Additionally, the sequence for gorilla has a 195 base pair gap starting at bp 1360, which includes the two amino acid variants known to affect sleep behavior in humans (P385R and Y362H indicated with arrows). Sequence identity is shown immediately below the consensus (green = 100% identical; gold = 25-99% identical; red < 25% identical).

There are no amino acid substitutions in our alignment at either residue previously described to confer alternative sleep behaviors in humans (Y362H and P385R, site numbers refer to human sequence). In our multiple sequence alignment, Y362H is at amino acid alignment position 476 (S4 Table) and nucleotide alignment positions 1423-1425bp (S3 Table). There is also no nucleotide variation for this codon. Alternatively, P385R is at amino acid alignment position 498 (S4 Table) and nucleotide alignment positions 1489-1491bp (S3 Table). Although there are no amino acid substitutions at this residue, there is synonymous variation. All but four sequences have the codon CCG, which codes for proline. The exceptions are synonymous substitutions in *Sus scrofa* (CCA), *Rousettus aegyptiacus* (CCC)*,* and *P. sinensis* (CCC) - all of which still code for proline. However, in the *G. gorilla gorilla* sequence, both residues 362 and 385 fall within the 195 base pair deletion described above.

Phylogenetic Analyses

Both maximum likelihood and Bayesian phylogenetic analyses were highly congruent. There were 20 significantly supported branches in both the maximum likelihood phylogenetic analysis (Fig 2) and in the Bayesian phylogenetic analysis (S2 Fig). In both trees, *H. sapiens* and *Pan troglodytes* are strongly supported sister species (bootstrap = 97%, posterior probability = 0.99). Additionally, the Great Apes are monophyletic in both phylogenetic analyses. While the two trees support the same evolutionary relationships, they have one minor differences in terms of support values. In the tree generated using Bayesian analysis, two of the Old World monkeys (*Cercocebus atys* and *Theropithecus gelada*) are sister species with a strong posterior probability value of 0.99, while in the tree generated using maximum likelihood, these species have bootstrap values of 68%, which is just below the frequently used cut-off for reliability of 70% [34].

Fig 2. Maximum likelihood phylogenetic analysis of mammalian *BHLHE41* coding sequence. We used the GTR+CAT+I parameter settings with 100 bootstrap replicates which are indicated next to the branches. The tree is rooted with the reptilian outgroup, *Pelodiscus sinensis*.

Molecular Evolution and Variation around Conserved Domains

Among the species, there were no significant pairwise dN-dS values in the test for positive selection (all comparisons had $p = 1.0$). On the other hand, the Z-test for purifying selection revealed 96.8% of the pairwise species comparisons had dN-dS values significantly less than zero (S1 Table). The mean dN-dS value was -6.13 suggesting strong purifying selection.

After removing invariant codons and those with a gap in >70% of the sequences, we found 227 of 343 codons had significantly higher dS than dN values (66.2%) indicating strong purifying selection (S2 Table; Fig 3). The dN-dS value for the “short-sleeper” allele (P385R [10]) had a dN-dS value of -3.02 ($p < 0.01$), consistent with strong purifying selection. When compared to all 343 codons, P385R had the 45th most negative dN-dS value. Another variant known to affect sleep behavior in humans, Y362H [17], exhibited no variation in the codon and therefore no p-value could be calculated (S2 Table).

Fig 3. Codon by codon comparison of dN-dS across the mammalian alignment of *BHLHE41*. Positive dN-dS values represent positive selection, negative dN-dS values represent purifying selection, and zero dN-dS values represent neutrality. Codon # comes from the HyPhy output and does not include codons removed because >70% of sequences in the alignment had gaps (e.g. the first 106 amino acids in gorilla). All the codons with significant p-values (red) have dN < dS. Blue points have dN-dS that are not significantly different from zero. There are no codons with significant dN > dS. Conserved domains in the *Homo sapiens BHLHE41* protein are indicated with black bars above the graph representing the codon positions for bHLH and the orange domain. Invariant codons are not shown because a p-value could not be calculated.

Although 3D structures are an integral part of determining a protein's function, there was no known 3D structure for *H. sapiens BHLHE41* protein. Additionally, a BLAST search revealed there were no sequences that had known protein structures in NCBI's protein data bank with E-values below 0.042, which is above the commonly used threshold for homology ($<10^{-3}$; [33]).

Instead, we compared variation in the two conserved domains from the *H. sapiens BHLHE41* protein GenBank flat file - the bHLH domain and the orange domain. There is no variation in the amino acid alignment across the 59 amino acids in the bHLH domain (S4 Table). The average dN-dS value for these codons is -0.891 suggesting purifying selection. All the p-values of variable codons in this region are <0.01 (S4 Table; Fig 3). The orange domain spans amino acids in the human coding sequence 129-175 (amino acids 235-281 in our alignment, S4 Table). There are two variable sites at human amino acid sites S147A (variants appear in pig, whales, dolphins, sheep and cow; dN-dS = 0.50) and P157Q (variants appear in leopard, northern fur seal and California sea lion; dN-dS = -0.141). The average dN-dS for the 47 codons in the orange domain is -0.811 (Fig 3). Of the 27 p-values that could be calculated in the orange domain, all but five have p-values <0.05. In general, there are large stretches of invariant amino acids among the mammalian samples (e.g., residues 148 to 252 of our alignment). Furthermore, there are seven poly-alanine residues ranging from four to 16 amino acids in length between amino acid alignment positions 407 and 547.

Discussion

Through this study, we explored patterns of molecular evolution in the sleep-related, circadian clock gene *BHLHE41* in mammals. Overall, this gene is highly conserved among mammals consistent with its essential function. For example, the HLH conserved domain shows no amino acid variation among mammals (and even the reptilian outgroup!) (amino acid alignment positions 152-210 in S4 Table). Furthermore, the evolutionary history of this gene among mammals is consistent with well-established species-level phylogenetic relationships [19, 20].

Yet, unexpectedly the gorilla homologue indicates there are two large, suspicious indels. The 318 base pair insertion at the 5' end of the coding sequence suggests a start codon 106 amino acids upstream from the remaining mammalian start codons. It is noteworthy that the gorilla sequence still contains AUG at the site where the remaining sequences start translation. Additionally, the gorilla sequence contains a 195 base pair deletion near the 3' end of the coding sequence. This predicted deletion includes both short-sleeper variants previously described (Y362H and P385R) - essential amino acids for proper circadian clock function [10, 14]. Although these indels may reflect novel function of *BHLHE41* in gorilla, these animals are not known to have particularly unusual sleep patterns, nor disrupted circadian clocks as would be expected from the addition of 106 amino acids on the 5' end and the deletion of 65 amino acids from near the 3' end.

The existence of these indels seems especially unlikely given the widespread pattern of purifying selection on this gene across mammals (>97% of pairwise species comparisons) and across codons (~50% of codons). The 5' insertion is especially suspicious because it is unique among the 27 mammal sequences investigated and without it, the sequence aligns perfectly with the rest of the mammalian start codons. Although this insertion does not immediately affect the HLH conserved domain, such a large insertion within 50 residues seems very likely to disrupt protein folding in this region. Without a known 3D structure, confident determination of the effects of these indels on the 3D structure and therefore function remain unknown.

The gorilla *BHLHE41* sequence was produced during whole-genome sequencing and was predicted using an annotation pipeline [32]. There is no literature discussing this unusual gorilla *BHLHE41* sequence. Unfortunately, there is no cDNA sequence for this gene from gorilla in Genbank release 233.0 (April 2019). Therefore, we suggest that there may have been an error in the identification of the start codon by the open reading frame search algorithm [35]. We searched the *Gorilla gorilla gorilla* chromosome 12 whole genome shotgun sequence (NC_018436) between bp 58,885,949 and 58,889,015 and found that although the unusual 318bp upstream from the mammalian start codon exists, the gorilla annotation actually identified the correct start codon (no 318bp insertion on the 5' end). Yet, regarding the 195bp deletion near the 3' end, we found 224 N's between exon 5 and exon 6 which likely includes both intron 5 and the missing 195bps of exon 6.

An error in the open reading frame detection algorithm may account for the incorrectly identified start codon. It is noteworthy that He et al. (2009) suggested only 4 introns, yet EMBL identified 5 introns. Furthermore, EMBL indicates this gorilla amino acid sequence is only 419aa long compared to the Genbank accession which measures 522 residues (S3 Fig). Experimental determination of the length of the gorilla *BHLHE41* protein by sequencing cDNA or RNA-Seq will be necessary in order to determine the true start codon in gorilla (or start codons if there are multiple isoforms of this gene) and the validity of the 195bp deletion near the 3' end of the coding sequence.

There is no evidence of alternative splice variants for *BHLHE41* in Gorilla according to EMBL (ENSGGOG00000015498; accessed June 13, 2019). Furthermore, although there are 11 paralogues in EMBL, all are less than 37% identical to *BHLHE41* indicating significant sequence divergence and unlikely to be mistaken for *BHLHE41*. Furthermore, paralogous sequences would most likely show incongruent relationships with the well-established mammalian phylogeny. The status in the UNIPROTKB database indicates it is still only a predicted protein with an Annotation Score of 2/5 (G3RHJ7_GORGO). It is noteworthy that the EMBL transcript protein sequence contains neither the early start codon, nor the 195bp deletion in the coding sequence (ENSGGOT00000015550.3). However, the Genbank Annotation Release 101 of the

Gorilla gorilla gorilla genome (Nov 4 2016) still contains these two large indels. A very recent, new genome sequence of a different gorilla individual (Kamilah, CM017859, Aug. 28, 2019) no longer exhibits the 195bp deletion near the 3' end of the coding sequence. No annotations were available for this genome, but hopefully it eventually includes a start codon that matches the rest of mammals.

The annotation of protein-coding genes is currently based on gene prediction algorithms [36]. Gene prediction algorithms have been through several revolutions since their initial application. Majoros et al. (2003) evaluated the quality of gene prediction algorithms. An evaluation of gene finders based on hidden markov models (HMMs) was done by Knapp & Chen (2007); the authors reported that no significant improvement in the quality of de novo gene prediction methods occurred during the previous 5 years. Bakke et al. (2009) evaluated three second-generation gene annotation systems on the genome of the archaeon *Halorhabdus utahensis* from the performance of the gene-prediction models to the functional assignments of genes and pathways. Comparison of gene-calling methods showed that 90% of all three annotations share exact stop sites with the other annotations, but only 48% of identified genes share both start and stop sites [40]. Palleja et al. (2008) performed an interesting investigation of overlapping CDS in prokaryotic genomes. They compared overlapping genes with their corresponding orthologues and found that more than 900 reported overlaps larger than 60 bp were not real overlaps, but annotation errors. Given that *BHLHE41* is just one of the 46,653 coding sequences predicted in gorilla, we are cautious about making any widespread conclusions about the remaining loci.

To avoid annotation mistakes, Armengaud (2009) recommends using proteomics in association with translations in all six reading frames. Prasad et al. (2017) provide a method combining transcriptome and proteomics to aid in genome annotation. However, genes that are expressed only under special conditions or in rarely sampled tissues, or whose expression is below the detection level, pose a challenge even for proteomic and cDNA validation.

Conclusions

We sought to determine if there was a footprint of positive selection on *BHLHE41* in mammals in light of its effect on sleep behaviors. We found that the majority of the *BHLHE41* coding sequence exhibits a history of purifying selection (especially the conserved domains), indicating the gene has an essential function for survival and reproduction. In particular, if adaptive sleep behaviors are conferred by *BHLHE41*, we predicted residues 362 and/or 385 to show a history of positive selection. Both sites were invariant across mammals consistent with strong purifying selection on the underlying codons. The evolutionary history of *BHLHE41* is largely congruent with the well-established mammalian phylogeny. From the available mammalian sequences, it appears that the “short-sleeper” variant is only present in humans [10], and is otherwise undergoing purifying selection in all other mammalian species. However, during our investigation, we discovered an unusually annotated sequence for *G. gorilla gorilla*. We suggest that the early start codon and deletion near the 3’ end are annotation errors that warrant experimental verification.

Acknowledgements

The authors thank Santa Clara University's Department of Biology for supporting the computer lab and access to Geneious software, especially Daryn Baker and Steve Hines. We acknowledge the students in BIOL178 who provided valuable feedback during the early stages of this research. Aleezah Salmaan provided valuable editorial advice on an earlier draft of this study.

References

1. Allada, R., & Siegel, J. M. (2008). Unearthing the phylogenetic roots of sleep. Current biology, 18(15), R670-R679.

2. Joiner, W. J. (2016). Unraveling the evolutionary determinants of sleep. Current Biology, 26(20), R1073-R1087.

3. Miyazaki, S., Liu, C. Y., & Hayashi, Y. (2017). Sleep in vertebrate and invertebrate animals, and insights into the function and evolution of sleep. Neuroscience research, 118, 3-12.

4. Tosches, M. A., Bucher, D., Vopalensky, P., & Arendt, D. (2014). Melatonin signaling controls circadian swimming behavior in marine zooplankton. Cell, 159(1), 46-57.

5. Hirano, A., Hsu, P. K., Zhang, L., Xing, L., McMahon, T., Yamazaki, M., ... & Fu, Y. H. (2018). DEC2 modulates orexin expression and regulates sleep. Proceedings of the National Academy of Sciences, 115(13), 3434-3439.

6. American Academy of Sleep Medicine. "Sleep Deprivation ." *American Academy of Sleep Medicine – Association for Sleep Clinicians and Researchers*, AASM, 2008, aasm.org/.

7. Medic, Goran, Micheline Wille, and Michiel EH Hemels. "Short-and long-term health consequences of sleep disruption." *Nature and science of sleep* 9 (2017): 151.

8. Capellini, I., Barton, R. A., McNamara, P., Preston, B. T., & Nunn, C. L. (2008). Phylogenetic analysis of the ecology and evolution of mammalian sleep. Evolution: International Journal of Organic Evolution, 62(7), 1764-1776.

9. Albrecht, U. (2002). Invited review: regulation of mammalian circadian clock genes. Journal of Applied Physiology, 92(3), 1348-1355.

10. He, Ying et al. “The transcriptional repressor DEC2 regulates sleep length in mammals” *Science (New York, N.Y.)* vol. 325,5942 (2009): 866-70.

11. Kato, Yukio, et al. "DEC1/STRA13/SHARP2 and DEC2/SHARP1 coordinate physiological processes, including circadian rhythms in response to environmental stimuli." Current topics in developmental biology. Vol. 110. Academic Press, 2014. 339-372.

12. Ledent, Valérie, Odier Paquet, and Michel Vervoort. "Phylogenetic analysis of the human basic helix-loop-helix proteins." Genome biology 3.6 (2002): research0030-1.

13. Dawson, S. R., Turner, D. L., Weintraub, H., & Parkhurst, S. M. (1995). Specificity for the hairy/enhancer of split basic helix-loop-helix (bHLH) proteins maps outside the bHLH domain and suggests two separable modes of transcriptional repression. Molecular and cellular biology, 15(12), 6923-6931.

14. Abe, Tomotaka, Tomoko Ishikawa, Tomohiro Masuda, Kanta Mizusawa, Toshiro Tsukamoto, Hiroshi Mitani, Tadashi Yanagisawa, Takeshi Todo, and Masayuki Iigo. "Molecular analysis of Dec1 and Dec2 in the peripheral circadian clock of zebrafish photosensitive cells." Biochemical and biophysical research communications 351, no. 4 (2006): 1072-1077.

15. Parasuraman, S. (2012). Protein data bank. Journal of pharmacology & pharmacotherapeutics, 3(4), 351.

16. Kadener, S., Stoleru, D., McDonald, M., Nawathean, P., & Rosbash, M. (2007). Clockwork Orange is a transcriptional repressor and a new *Drosophila* circadian pacemaker component. Genes & development, 21(13), 1675-1686.

17. Pellegrino, Renata, et al. "A novel *BHLHE41* variant is associated with short sleep and resistance to sleep deprivation in humans." *Sleep* 37.8 (2014): 1327-1336.

18. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24(8), 1586-1591.

19. Kemp, T. S. 2005. The Origin and Evolution of Mammals. Oxford University Press.

20. Tarver, James E., Mario Dos Reis, Siavash Mirarab, Raymond J. Moran, Sean Parker, Joseph E. O'Reilly, Benjamin L. King et al. "The interrelationships of placental mammals and the limits of phylogenetic inference." Genome Biology and Evolution 8, no. 2 (2016): 330-344.

21. Swiss Institute for Bioinformatics (SIB). ExPASY Translate tool. website: https://web.expasy.org/translate/ [accessed August 30, 2019].

22. Parker, J. Encyclopedia of Genetics, 2001, Page 1375.

23. Sieber, P., Platzer, M., & Schuster, S. (2018). The definition of open reading frame revisited. Trends in Genetics, 34(3), 167-170.

24. Wang, Y., Liu, J., Huang, B.O., Xu, Y.M., Li, J., Huang, L.F., Lin, J., Zhang, J., Min, Q.H., Yang, W.M. and Wang, X.Z., 2015. Mechanism of alternative splicing and its regulation. Biomedical reports, 3(2), pp.152-158.

25. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

26. Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21), 2688-2690.

27. Huelsenbeck, J.P., Ronquist, F., Nielsen, R. and Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. science, 294(5550), pp.2310-2314.

28. Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution, 33(7), 1870-1874.

29. Nei M. and Gojobori T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3:418-426.

30. Kosakovsky Pond, S. L., Simon D. W. Frost, Spencer V. Muse, HyPhy: hypothesis testing using phylogenies, Bioinformatics, Volume 21, Issue 5, 1 March 2005, Pages 676–679,

31. Grottke C, Mantwill K, Dietel M, Schadendorf D and Lage H. Identification of differentially expressed genes in human melanoma cells with acquired resistance to various antineoplastic drugs. Int. J. Cancer 88 (4), 535-546 (2000).

32. Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T. and McCarthy, S., 2012. Insights into hominid evolution from the gorilla genome sequence. Nature, 483(7388), p.169.

33. Butler, T., Dick, C., Carlson, M. L., & Whittall, J. B. (2014). Transcriptome analysis of a petal anthocyanin polymorphism in the arctic mustard, Parrya nudicaulis. PloS one, 9(7), e101338.

34. Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic biology, 42(2), 182-192.

35. Trimble, W. L., Keegan, K. P., D'Souza, M., Wilke, A., Wilkening, J., Gilbert, J., & Meyer, F. (2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC bioinformatics, 13(1), 183.

36. Prasad, T.S. Keshava, Ajeet Kumar Mohanty, Manish Kumar, Sreelakshmi K. Sreenivasamurthy, Gourav Dey, Raja Sekhar Nirujogi, Sneha M. Pinto, Anil K. Madugundu, Arun H. Patil, Jayshree Advani, Srikanth S. Manda, Manoj Kumar Gupta, Sutopa B. Dwivedi, Dhanashree S. Kelkar, Brantley Hall, Xiaofang Jiang, Ashley Peery, Pavithra Rajagopalan, Soujanya D. Yelamanchi, Hitendra S. Solanki, Remya Raja, Gajanan J. Sathe, Sandip Chavan, Renu Verma, Krishna M. Patel, Ankit P. Jain, Nazia Syed, Keshava K. Datta, Aafaque Ahmed Khan, Manjunath Dammalli, Savita Jayaram, Aneesha Radhakrishnan, Christopher J. Mitchell, Chan-Hyun Na, Nirbhay Kumar, Photini Sinnis, Igor V. Sharakhov, Charles Wang, Harsha Gowda, Zhijian Tu, Ashwani Kumar, and Akhilesh Pandey. Method: Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res. January 2017 27: 133-144.

37. Majoros, W. H., Pertea, M., Antonescu, C., & Salzberg, S. L. (2003). GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders. Nucleic acids research, 31(13), 3601-3604.

38. Bakke, P., Carney, N., DeLoache, W., Gearing, M., Ingvorsen, K., Lotz, M., McNair, J., Penumetcha, P., Simpson, S., Voss, L. and Win, M., 2009. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. PloS one, 4(7), p.e6291.

39. Knapp, K. & Chen, Y. P. (2007). An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy. Nucleic Acids Res 35, 317–324.

40. Poptsova, M. S., & Gogarten, J. P. (2010). Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiology, 156(7), 1909-1917.

41. Pallejà, A., Harrington, E. D., & Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC genomics, 9(1), 335.

42. Armengaud, J. (2009). A perfect genome annotation is within reach with the proteomics and genomics alliance. Current opinion in microbiology, 12(3), 292-300.

Supporting information captions

S1 Fig. Verification that the Bayesian MCMC phylogenetic search reached stationarity.

S2 Fig. Phylogenetic Tree of Euteleostomi *BHLHE41* mRNA using Bayesian analysis.

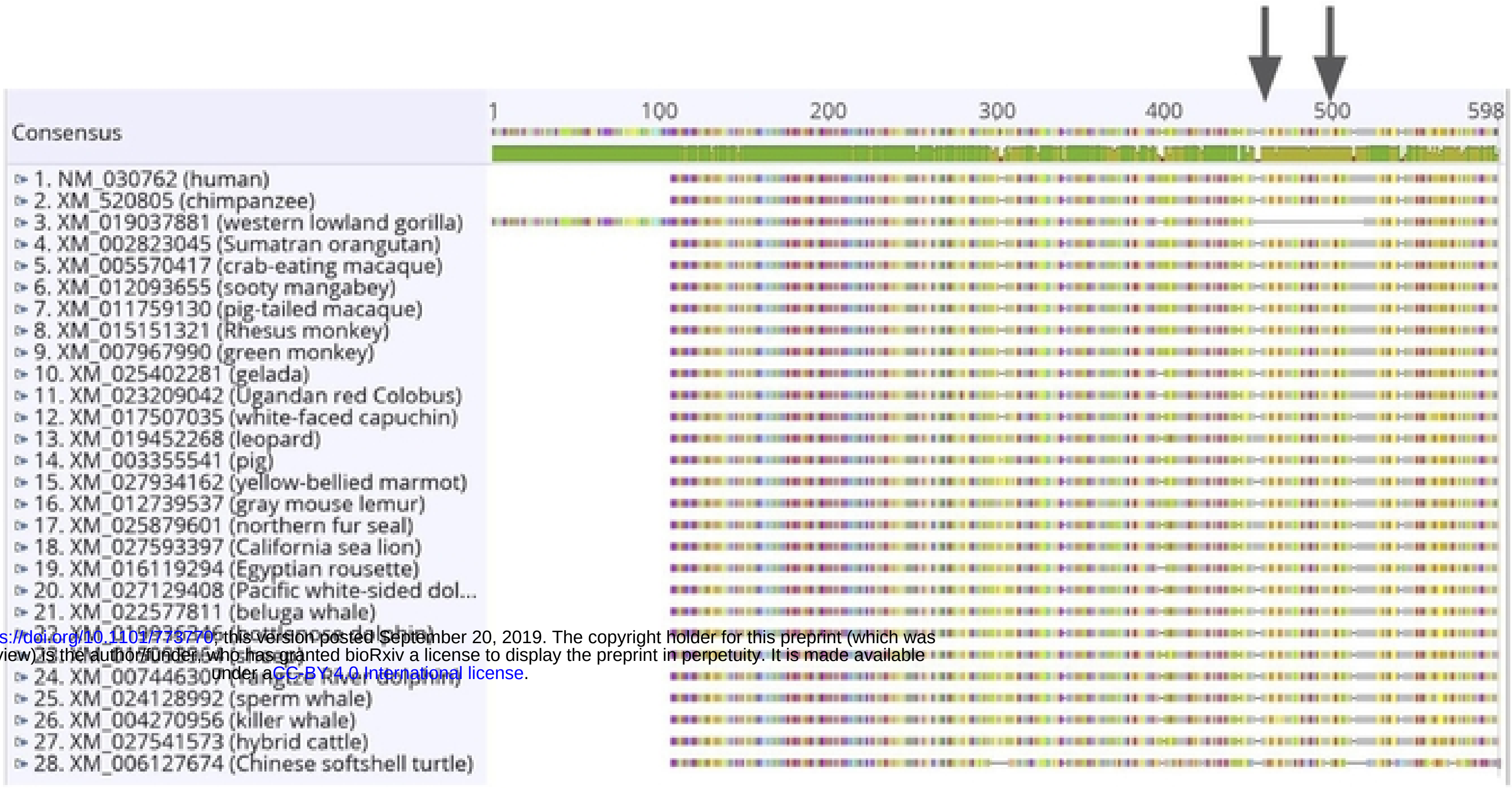
S3 Fig. EMBL structure of the transcript for *BHLHE41* from *Gorilla gorilla gorilla* with conserved domains indicated.

S1 Table. Pairwise codon-based test of purifying selection for mammalian *BHLHE41*.

S2 Table. Codon-by-codon test for selection.

S3 Table. *BHLHE41* mammalian nucleotide alignment with reptile outgroup.

S4 Table. *BHLHE41* mammalian amino acid alignment with reptile outgroup.



Consensus
1
100
200
300
400
500
598
1. NM_030762 (human)
2. XM_520805 (chimpanzee)
3. XM_019037881 (western lowland gorilla)
4. XM_002823045 (Sumatran orangutan)
5. XM_005570417 (crab-eating macaque)
6. XM_012093655 (sooty mangabey)
7. XM_011759130 (pig-tailed macaque)
8. XM_015151321 (Rhesus monkey)
9. XM_007967990 (green monkey)
10. XM_025402281 (gelada)
11. XM_023209042 (Ugandan red Colobus)
12. XM_017507035 (white-faced capuchin)
13. XM_019452268 (leopard)
14. XM_003355541 (pig)
15. XM_027934162 (yellow-bellied marmot)
16. XM_012739537 (gray mouse lemur)
17. XM_025879601 (northern fur seal)
18. XM_027593397 (California sea lion)
19. XM_016119294 (Egyptian rousette)
20. XM_027129408 (Pacific white-sided dol...
21. XM_022577811 (beluga whale)
22. (bottlenose dolphin)
24. XM_007446307 (Yangtze River dolphin)
25. XM_024128992 (sperm whale)
26. XM_004270956 (killer whale)
27. XM_027541573 (hybrid cattle)
28. XM_006127674 (Chinese softshell turtle)

Figure 1. Multiple sequence alignment of the *BHLHE41* mRNA for 27 mammals and one reptile outgroup. The alignment shows a 318 base pair insertion in the gorilla sequence on the 5' end. Additionally, the sequence for gorilla has a 195 base pair gap starting at bp 1360, which includes the two amino acid variants known to affect sleep behavior in humans (P385R and Y362H indicated with arrows). Sequence identity is shown immediately below the consensus (green = 100% identical; gold = 25-99% identical; red < 25% identical).

Fig1

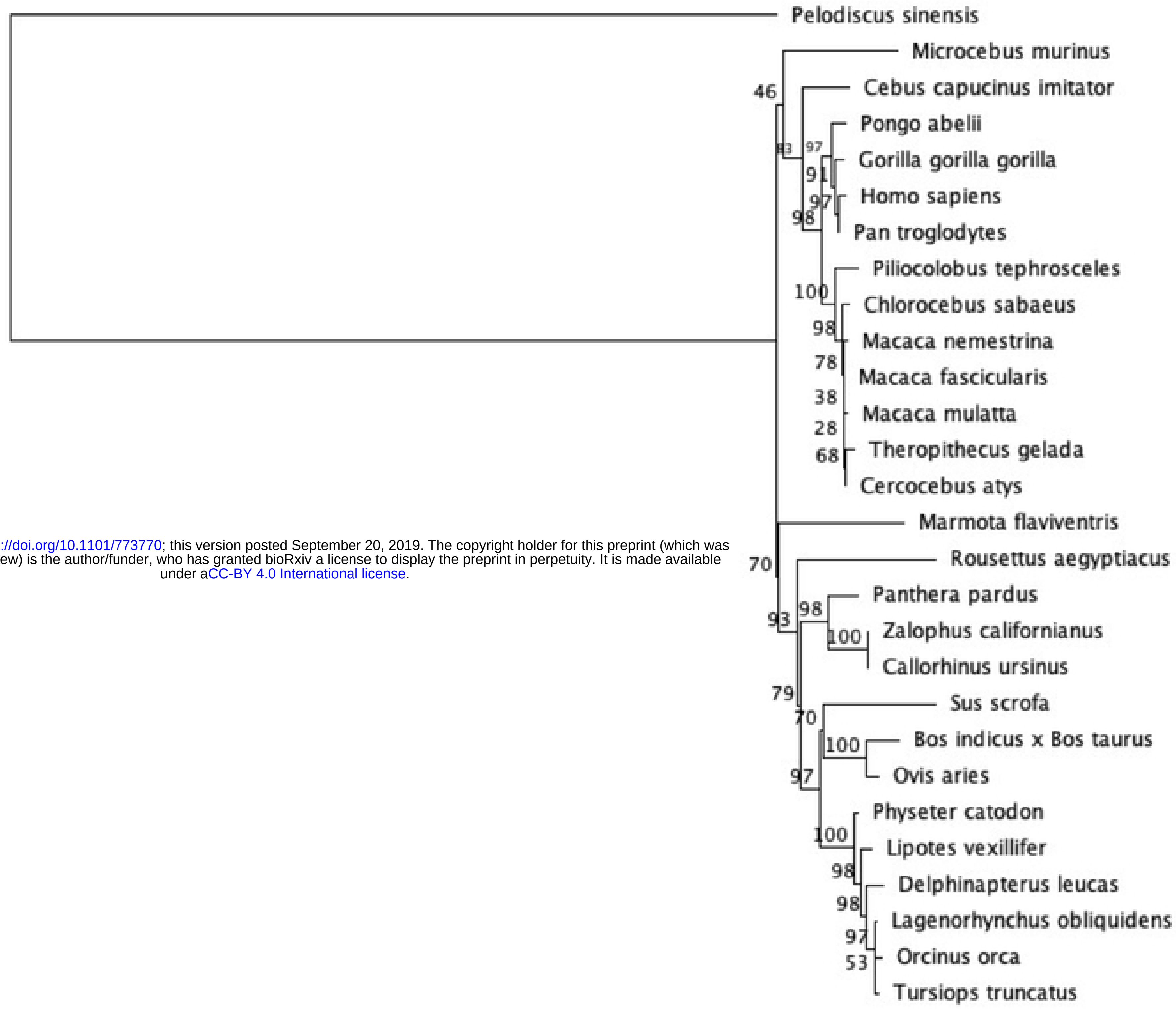


Figure 2. Maximum likelihood phylogenetic analysis of mammalian *BHLHE41* coding sequence. We used the GTR+CAT+I parameter settings with 100 bootstrap replicates which are indicated next to the branches. The tree is rooted with the reptilian outgroup, *Pelodiscus sinensis*.

Fig2

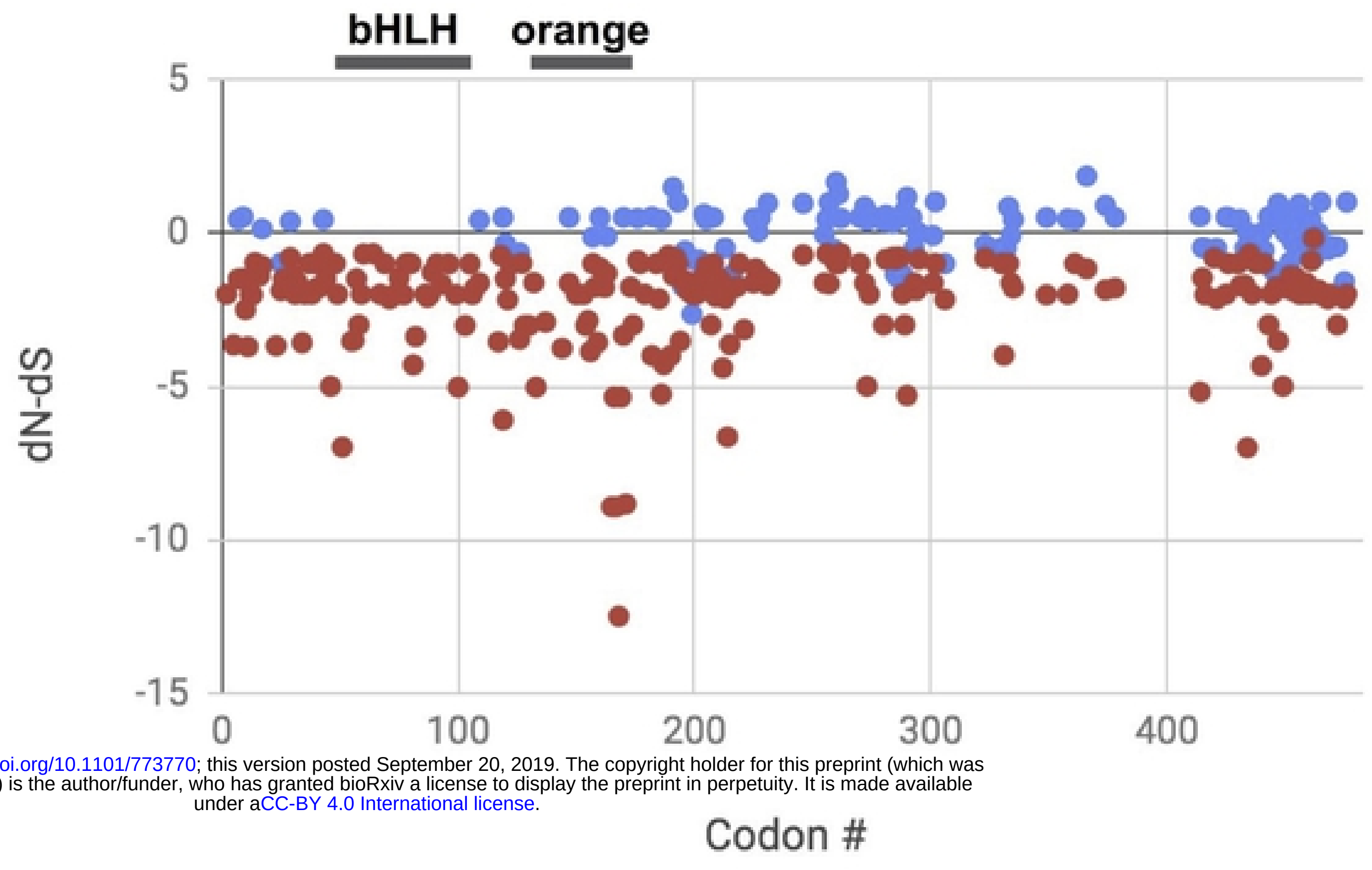


Figure 3. Codon by codon comparison of dN-dS across the mammalian alignment of *BHLHE41*. Positive dN-dS values represent positive selection, negative dN-dS values represent purifying selection, and zero dN-dS values represent neutrality. Codon # comes from the HyPhy output and does not include codons removed because >70% of sequences in the alignment had gaps (e.g. the first 106 amino acids in gorilla). All the codons with significant p-values (red) have dN < dS. Blue points have dN-dS that are not significantly different from zero. There are no codons with significant dN > dS. Conserved domains in the *Homo sapiens BHLHE41* protein are indicated with black bars above the graph representing the codon positions for bHLH and the orange domain. Invariant codons are not shown because a p-value could not be calculated.

Fig3