Adaptation and convergence in genes of the circadian system in Iberian *Squalius* freshwater species

João M Moreno[1], Vitor C Sousa[1], Tiago F Jesus[1] and Maria M Coelho[1,*]

[1]cE3c – Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisboa, Portugal

*Corresponding author: e-mail: mmcoelho@fc.ul.pt

ABSTRACT

The circadian clock is a biological timing system that improves the inherent ability of organisms to deal with environmental fluctuations. It is regulated by daily alternations of light but can also be affected by temperature. Fish, as ectothermic, have an increased dependence on the environmental temperature and thus are good models to study the integration of temperature within the circadian system. Here, we studied four species of freshwater fish of *Squalius* genus, distributed across a latitudinal gradient in Portugal with variable conditions of light and temperature. We identified and characterised the expected sixteen genes belonging to four main gene families (Cryptochromes, Period, CLOCK and BMAL) involved in the circadian system. We conducted phylogenetic analysis and tested these genes for the presence of signatures of selection based on dN/dS that revealed the presence of mutations under positive selection in nine genes. Furthermore, we analysed functional features of the proteins and found that these mutations alter protein isoelectric point and aliphatic index. Additionally, in 50% of the genes we found a pattern of convergence uncovered by the phylogenetic clustering of *S. aradensis* and *S. pyrenaicus* from Almargem, which inhabit basins with similar environmental conditions. Although this clustering can be explained by incomplete lineage sorting, evidence of positive selection in one gene in the two species and similar functional features in seven proteins are consistent with convergence at protein level. Our results highlight the importance of integrating sequence-based functional protein characterisation with dN/dS-based methods when studying the molecular adaptation of species.

Key words: adaptive molecular evolution; circadian rhythm; evolutionary convergence; Iberian freshwater fish; protein characterisation

INTRODUCTION

Organisms are exposed to daily environmental fluctuations in their natural habitats. To overcome them, organisms developed biological timing systems to optimize their physiological and biochemical processes in space and time (Foulkes et al. 2016). These systems work as internal clocks and require a proper synchronization with environmental signals that specify the actual time of the day. Thus, understanding the molecular evolution of the genes involved in the circadian system provide important clues to elucidate how species adapt to their environments.

The circadian system is a universal biological timing system found virtually in all organisms (Paranjpe and Sharma 2005). Circadian system is synchronised by light–dark cycle of a day's period and present oscillations with a period of ~24h called circadian rhythms. Oscillations are generated and regulated at molecular level, but the outcomes have been shown to influence several aspects of physiology, behaviour and ecology of organisms (Paranjpe and Sharma 2005; Vaze and Sharma 2013). In fact, circadian rhythms, when properly entrained by light, have been proved to improve the inherent aptitude of several organisms to survive under changing environments by aiding them to efficiently anticipate periodic events, specifically light changes and climate seasons (Paranjpe and Sharma 2005; Vaze and Sharma 2013).

The molecular circadian system consists of a network of signalling transduction pathways regulated mainly by interconnected transcription-translation feedback loops (supplementary fig. S1) (Dunlap 1999; Pando and Sassone-Corsi 2002). The regulatory loops are sustained by the so-called core circadian genes and proteins and require about 24h to complete a cycle (Pando and Sassone-Corsi 2002; Foulkes et al. 2016). In vertebrates, several genes have been reported to be responsible for the maintenance and regulation of the circadian system (Foulkes et al. 2016). The core circadian-genes belong to four main gene families: Cryptochromes (CRY), Period (PER), CLOCK, and BMAL (Pando and Sassone-Corsi 2002). These gene families encompass several characterized genes (*cry, per, bmal* and *clock*) in vertebrates, and in fish several families possess a larger number of circadian paralogs as compared to other vertebrates (Toloza-Villalobos et al. 2015). For instance, in zebrafish (*Danio rerio*), several genes have been identified: six *cry* (*cry1aa, cry1ab, cry1ba, cry1bb, cry2, cry3*), four *per* (*per1a, per1b, per2, per3*), three *bmal* (*bmal1a, bmal1b, bmal2*), and three *clock* (*clocka, clockb, clock2/npas2*) (Wang 2008a; Wang 2008b; Wang 2009; C. Liu et al. 2015). Cryptochrome genes encode for a class of flavoproteins that are sensitive to blue light (Lin and Todo 2005), whereas *period* genes encode for proteins that also display a strong but differential light responsiveness (Pando et al. 2001; Vallone et al. 2004; Vatine et al. 2009). They were found to be key agents in the entrainment of the

circadian system, as they are photoreceptors responsible for transducing light signals into the core circadian machinery, and constitute the negative elements of the system (i.e. repressors of transcription) (Tamai et al. 2007). BMAL (Brain and muscle ARNT like) and CLOCK (Circadian locomotor output cycle kaput) families encode for canonical circadian proteins, a highly conserved bHLH (basic-Helix-Loop-Helix)-PAS (Period-Aryl hydrocarbon receptor nuclear translocator- Single minded) transcriptional factors and are the positive elements of the circadian system (i.e. activators of transcription) (Pando and Sassone-Corsi 2002; Foulkes et al. 2016). Importantly, in vertebrates the heterodimer CLOCK:BMAL binds to E-box enhancer elements in DNA to trigger transcription of target genes (Pando and Sassone-Corsi 2002; Foulkes et al. 2016), including the negative elements of the network, *per* and *cry* genes (Nakahata et al. 2008). In turn, the heterodimer PER:CRY interacts with the CLOCK:BMAL heterodimers to inhibit their function (Ishikawa et al. 2002; Pando and Sassone-Corsi 2002). To complete the circadian regulatory network, further layers of regulation are provided by post-transcriptional and post-translational modifications (McClung 2011; Lim and Allada 2013).

By sequencing the transcriptome of zebrafish exposed to light, two recent studies identified several genes whose expression depends on light, revealing a multi-level regulation of circadian rhythms by light-cycles (Weger et al. 2011; Ben-Moshe et al. 2014). This gene expression light-dependence was also recently shown for the Atlantic cod *Gadus morhua* (Lazado et al. 2014). The authors conclude that the molecular mechanism involves light-induced *cry1aa* and *per2* that encode for negative core elements of the circadian transcription-translation loop, with CRY1AA acting as a direct inhibitor of CLOCK:BMAL activation via preventing heterodimer formation (Ishikawa et al. 2002; Tamai et al. 2007), while PER2 plays a more complex role as either transcriptional coactivator or corepressor depending on its transcriptional regulatory targets (Wang et al. 2015). Photoreception is particularly interesting in fish as, contrary to most vertebrates that only perceive light through the eyes, fish also possess a photosensitive pineal gland, dermal melanophores, and brain photoreceptors (Foulkes et al. 2016). In addition, fish possess independent peripheral photoreceptors and self-sustaining circadian oscillators, essentially in every tissue (Whitmore et al. 1998).

Circadian rhythms can also be entrained by temperature (Liu et al. 1998; Tsuchiya et al. 2003; Lahiri et al. 2005; Chappuis et al. 2013; Jerônimo et al. 2017). In mammals it was demonstrated that peripheral cells *in vitro* could sense the change of room temperature as a cue for entrainment of circadian system (Tsuchiya et al. 2003). In zebrafish, temperature has also an important role in circadian clock (Lahiri et al. 2005; Jerônimo et al. 2017), and it was proposed that temperature could entrain the phase of the system by driving expression levels

of *per3*, and other clock genes (namely *cry2* and *cry1ba*) via an alternative hypothetical enhancer upstream the E-box (Lahiri et al. 2005). In this model, *per1b* (formerly known as *per4*) promoter integrates temperature and light regulatory input from the E- boxes together with regulation by other temperature-driven elements (Lahiri et al. 2005). In agreement with this hypothesis, in a study comparing transcriptome profiling of two freshwater fish species (*Squalius carolitertii* and *S. torgalensis*), Jesus et al. (2016) found two differentially expressed genes (*cry1aa* and *per1a*) between a control a thermal stress condition. Subsequent studies with those genes reinforce the dependence on temperature for proper entrainment of the circadian system in these species by showing a response in expression of *cry1aa* and *per1a* genes in *S. carolitertii* when subjected to a simulated scenario of moderate climate change (Jesus et al. 2017). Other studies connected the circadian system with stress response through a network involving heat-shock proteins (HSP) (Jerônimo et al. 2017). When exposed to thermal stress organisms increase up to 10% the expression of HSPs (Söti et al. 2005). This activation is related to heat-shock factors (HSFs) that bind regions in the DNA called heat-shock elements (HSE). In normal situations, HSP90 binds HSF1 and prevents its migration for the nucleus, but under thermal stress, these two proteins dissociate and HSF1 is free to migrate to the nucleus and bind to HSE sequence. This pathway can affect the expression of *per2* in mammals as this gene possesses in its promoter a heat-shock element (HSE) sequence (Kornmann et al. 2007; Tamaru et al. 2011; Chappuis et al. 2013). Other authors studying this same relation in zebrafish hypothesize that HSF1 activation is related to circadian regulation in at least two ways: (1) HSP activation improves the ability of the organism to deal with thermal stress, but HSP also interacts with BMAL1 to maintain the optimal function of the circadian system and (2) HSF1 activates other circadian-related genes as *per2* that may result in synchronization of the circadian machinery and maintain the outputs of the circadian system fully functional (Jerônimo et al. 2017). In the same work, Jerônimo et al. (2017) identified the channel TRPV1 (Transient receptor potential cation channel subfamily V member 1), a thermo-sensitive receptor, as an integrant part of zebrafish circadian rhythm by playing a role in heat-mediated *hsp90* and *per2* responses. By blocking the channel with a TRPV1 inhibitor, the authors demonstrated that even though *hsp90* expression was unaffected, *per2* expression was partially downregulated (Jerônimo et al. 2017).

Several studies have been conducted to understand the circadian system at its different levels of organization. In fish, these studies are particularly interesting once it is a very diverse group of animals adapted to nearly all aquatic environments and possess a larger number of circadian paralogs as compared to the other vertebrates (Toloza-Villalobos et al. 2015). Some of these studies cover the evolutionary relationships of the core-clock gene

families and the mechanisms driving their molecular evolution (Wang 2008a; Wang 2008b; Wang 2009; C. Liu et al. 2015), but several key questions still remain open, namely the reason for the higher number of paralogs in fish when compared to other vertebrates (Toloza-Villalobos et al. 2015). Toloza-Villalobos et al. (2015) described the possible evolutionary history of circadian-related gene families and pointed to a possible event of subfunctionalization among *per* genes. Even though neofunctionalization has been a widely assumed fate for these duplicated circadian-related genes during evolution, functional analysis indicate that paralogs retain ancestral functions creating a redundant pool of circadian genes. Thus, it remains unclear what are the mechanisms driving molecular evolution in these gene families.

In the Iberian Peninsula, the genus *Squalius* Bonaparte, 1837 is represented in Portuguese rivers by four known species (*S. carolitertii, S. pyrenaicus, S. torgalensis* and *S. aradensis*) distributed across a latitudinal cline (supplementary fig. S2). *S. carolitertii* (Doadrio, 1988) inhabits the northern rivers of Portugal, *S. pyrenaicus* (Günther, 1868) occurs in the Central and Southern drainages (e.g. Tagus, Guadiana and Almargem), while sister species *S. torgalensis* and *S. aradensis* (Coelho et al., 1998) are confined to small basins in the Southwest of Portugal (e.g. Mira and Arade, respectively). This distribution gains special interest from an evolutionary perspective since Portugal is at the frontier between two contrasting climate types: the Atlantic in the Northern region that is characterized by mild temperatures, and the Mediterranean in the Southern region typified by high temperatures and droughts during summer periods (Pires et al. 1999; Mesquita and Coelho 2002; Mesquita et al. 2005; Henriques et al. 2010). The photoperiod is also variable along this latitudinal cline, and typically daytime is longer in southern region of Portugal for about 15 minutes when compared to the northern region (supplementary table S1). Moreover, these basins present differences in water pH, as southern basins are typically more alkaline than northern ones (supplementary table S1). These environmental differences associated with the distribution of *Squalius* species in Portugal make them an excellent model to study evolutionary adaptation to different environmental settings, and for studying adaptation of the circadian system to different environmental conditions of light and temperature.

Here, we present an integrative study on molecular evolution of circadian system in Portuguese *Squalius* species. To identify the genes involved in the core circadian system in these species, we compared previously published transcriptomes of *S. carolitertii* and *S. torgalensis* (Jesus et al. 2015) with transcriptomes of the reference organism *Danio rerio* (Weger et al. 2011; Ben-Moshe et al. 2014). To assess the evolutionary history of the genes within each gene family identified in the *Squalius* species, we conducted phylogenetic

analysis using predicted protein sequences. We also predicted patterns of protein-protein interactions to further investigate the function of duplicated genes and to provide insights on the fate of these duplications, specifically to detect signals of post-duplication diversification. To detect signatures of molecular adaptation, we combined the use of sequence-based statistical tests to detect natural selection with analysis of functional features of the proteins that may be relevant in response to environmental differences in light, temperature and pH. This allowed us to describe the mechanisms of adaptation of the circadian system in species from the genus *Squalius* to the environmental cline. This may help understanding the mechanisms of adaptation to environmental variability in other freshwater fish species.

RESULTS

Identification of core circadian gene families in *Squalius* species

To identify the genes involved in the central circadian system in *Squalius* species, we analysed previously published transcriptome data of *S. torgalensis* and *S. carolitertii* (Jesus et al. 2015) and compared it with light-induced zebrafish transcriptomes (Weger et al. 2011; Ben-Moshe et al. 2014). From this comparison we identified four main gene families (Cryptochromes, Period, CLOCK and BMAL), involved in the main core of circadian system in the transcriptomes of *S. torgalens*is and *S. carolitertii* (supplementary table S2). Identified gene families correspond to those already described for other fish species (Wang 2008a; Wang 2008b; Wang 2009; C. Liu et al. 2015). Identified genes were re-sequenced for *S. torgalensis* and *S. carolitertii* and sequenced *de novo* for *S. pyrenaicus* and *S. aradensis.* Non-redundant sequences were deposited in the European Nucleotide Archive (ENA) under the accession numbers presented supplementary table S2.

To test if the identified genes corresponded to proteins involved in the circadian system, we used the predicted protein sequences to conduct an analysis of protein-protein interactions (PPI). This revealed that all proteins interact with other core circadian proteins (supplementary table S3). We also found proteins with interactions indirectly related with the core circadian system through circadian-dependent functions. All interactions found are summarized in supplementary table S3.

Evolutionary history of circadian-related gene families

To assess evolutionary history of gene families we conducted phylogenetic analysis independently for each family. To account for paralogous, we used *Drosophila melanogaster* protein as outgroup rather than the zebrafish (Danio rerio) but included zebrafish proteins (see supplementary table S4 for accession numbers). As detailed for each family below, all reconstructed phylogenies recovered the evolutionary relationships between paralogous as previously described in other fish species (Wang 2008a; Wang 2008b; Wang 2009; C. Liu et al. 2015). For the CRY family, six *cry* genes encoding for six proteins were identified and clustered in three major groups (CRY1, CRY2 and CRY3) (Figure 1A). Among CRY1 group, four protein-coding genes were identified (*cry1aa*, *cry1ab, cry1ba* and *cry1bb*) based on the fact that they form monophyletic groups, clustering with the orthologous proteins from zebrafish*.* Using the same approach, for CRY2 and CRY3 groups we identified one protein-coding gene for each: *cry2* and *cry3,* respectively. For the PER family, we identified four protein-coding genes, which encode proteins belonging to three different groups (Figure 1B). PER1 group comprises two proteins (PER1A and PER1B) corresponding to genes *per1a* and *per1b*, while PER2 and PER3 groups comprise a single protein each, encoded by the genes

per2 and *per3*, respectively. For the CLOCK family, we identified three protein-coding genes belonging to two groups (Figure 1C). CLOCK1group encompasses two protein-coding genes (*clocka* and *clockb*) and CLOCK2 encompasses a single protein-coding gene, *clock2.* For the BMAL family we found three genes belonging to 2 different groups according to phylogenetic analysis (Figure 1D). BMAL1 group encompasses two protein-coding genes (*bmal1a* and *bmal1b*) and BMAL2 group is composed by a single protein-coding gene, *bmal2*. In sum, we found that sequences cluster with orthologs from zebrafish for all gene families, and that for each gene there is a monophyletic group including the four *Squalius* species.

Signatures of selection in circadian-related genes

To detect signatures of selection for each gene identified, we conducted several statistical tests based on dN/dS ratio. Out of sixteen genes, we detected positive selection in five using the gene-wide level test implemented in BUSTED (supplementary table S5), and in eight using the site-level analysis implemented in MEME to detect individual sites under positive selection (Table 1). Among the genes with signatures of positive selection for at least one of the mentioned methods, we detected significant positive selection in at least one branch of the phylogeny in five of them, using a branch-based test implemented in aBSREL (supplementary table S6).

Overall, we did not find evidence of positive selection in *cry* genes with the gene-wide test (supplementary table S5), but at the site-level analysis we found sites with significant evidence of positive selection in *cry1aa, cry2* and *cry3* genes (Table 1). For *cry1aa*, we inferred a positively selected site corresponding to a non-synonymous non-conservative mutation changing a lysine to a glutamine (K275Q) in the CRY1AA protein of *S. torgalensis*. However, we did not find any evidence of positive selection in *cry* genes using the branch test (supplementary table S6). Among the PER genes, *per1b, per2* and *per3* displayed signals of gene-wide positive selection (supplementary table S5). Moreover, the site-level test indicated positive selection in all *per* genes at sites with non-conservative mutations (Table 1). For *per1b*, the branch test indicated positive selection in the branch of *S. torgalensis* (supplementary table S6), which is likely correlated with the site T1256V, a mutation in *S. torgalensis* that was significant at the site-level test. For *per2*, the branch test indicated positive selection on the branches of *S. torgalensis* and *S. aradensis*, while for *per3* only the branch of *S. pyrenaicus* from Almargem was significant (supplementary table S6). Among *clock* genes, only *clocka* and *clockb* showed patterns consistent with gene-wide positive selection (supplementary table S5), but we did not find any sites under positive selection when applying the site-level tests (Table 1). The branch test revealed significant positive selection in the branches of *S. aradensis* and *S. pyrenaicus* (Almargem) for *clocka*, while for *clockb* this was only found for the branch of *S. carolitertii* (supplementary table S6).

Among BMAL genes, *bmal2* showed significant signals of gene-wide positive seletion (supplementary table S5). Only one individual site, also in *bmal2*, was significant in the site-level test of episodic positive selection (Table 1). This site corresponds to a non-synonymous and non-conservative mutation from an arginine to a proline (R20P). However, the branch-specific test of positive selection did not detect any branch under selection for this gene (supplementary table S6).

We also tested for purifying selection by using the FEL statistical test to detect sites under negative selection. We found a considerable amount of conserved positions under negative selection, ranging from 2 sites in *cry1ab* to 59 sites in *per2* (supplementary table S7).

Functional and structural analysis of circadian-related proteins

To assess whether signatures of selection have functional impacts relevant for these populations, we characterised proteins at different structural levels. Based on the primary structure, we inferred the aliphatic index (AI) and the theoretical isoelectric point (pI). These parameters were chosen due to their relationship with protein thermostability and environmental pH, respectively. Both factors have high biological relevance for the studied species, given that they live in habitats with different water pH and temperatures. While AI is defined as the relative volume occupied by aliphatic side chains (Ala, Val, Ile, Leu) and it is positively correlated with increased thermostability of globular proteins (Ikai 1980; Gasteiger et al. 2005); pI reflects the pH at which a protein has neutral net charge, and hence it is important for its solubility, subcellular localization, and potential interactions (Gasteiger et al. 2005; Kiraga et al. 2007; Khaldi and Shields 2011).

Regarding the aliphatic index (AI), we found differences among *Squalius* species for all proteins except for BMAL1B. For CRY1AA, CRY1AB and PER1A we found that northern populations of *S. carolitertii* and *S. pyrenaicus* from Tagus presented higher values of AI, whereas for CRY3, PER2, PER3, CLOCKB and BMAL1A it was the southern populations that showed higher AI (supplementary table S8).

We found differences in the pI across *Squalius* species for most proteins, except for PER1B, CLOCKA, BMAL1A, BMAL1B (supplementary table S8). The pI values tended to be lower in southern populations of *S. torgalensis, S. aradensis* and *S. pyrenaicus* from Almargem (supplementary table S8), even though this was not the case for CRY1BA, CRY1BB, CRY3, PER3, CLOCKB, CLOCK2 and BMAL2.

We conducted a sequence-based analysis to assess for domain and/or motif location using HMM predictions. Th allowed us to map important sites of the proteins detected to be under positive and negative selection and assess whether they could structurally impact the proteins. As expected, sites detected to be under negative selection are located mainly in

domain regions, highlighting the importance of these regions to the proper function of the protein; whereas positively selected sites are more widespread (Fig. 2).

For CRY proteins, two common domains to all CRY proteins were found: the DNA photolyase domain and the FAD-binding domain. Both domains are responsible for binding chromophores and are of extreme importance for the activity of photoreception displayed by these proteins (Lin and Todo 2005). Interestingly, the mutation Q96H in CRY3, which is on a site inferred to be under positive selection, was found to be located inside the DNA Photolyase domain (Fig. 2). Except for CRY1AB, more than 60% of all negatively selected sites were found to be located inside functional domains of CRY proteins (supplementary table S7). We found two domains common to all PER proteins: Period-Arnt-Sim (PAS_3/PAS_11) domain and the Period protein 2/3C-terminal region. Additionally, PER1A, PER2 and PER3 present an additional PAS domain relatively conserved. PAS domains are of extreme importance for PER function as they serve for protein dimerization, but also display activity of photoreception (Ponting and Aravind 1997; Möglich et al. 2009). For both PER1A and PER1B, there were two mutations inferred to be under positive selection located inside functional domains, PAS and Period C-terminal, respectively. For PER2, except for mutations T514L and R697A, all sites detected to be under positive selection are located inside the functional domains. Mutation K429Y in PER3 was also found to be located inside the PAS domain of the protein. Approximately 20% of the negatively selected sites in PER proteins are located inside functional domains (supplementary table S7). For CLOCK and BMAL proteins, we found three domains conserved in all two proteins: two Period-Arnt-Sim (PAS fold and PAS_11) domain and the basic helix-loop-helix (bHLH). The bHLH is a protein structural motif that characterizes one of the largest families of dimerizing transcription factors and consists in a DNA-binding region. For CLOCK proteins, we also found a high proportion of negatively selected sites located in functional domains, specifically in CLOCKB (60%). For BMAL2 most sites under negative selection (69%) are in PAS domains.

DISCUSSION

Identification and *in silico* characterisation of circadian-related genes in *Squalius* species

We were able to identify sixteen genes belonging to the four core-circadian gene families, which are orthologous to genes described in *D. rerio* and other fish species (Wang 2008a; Wang 2008b; Wang 2009; C. Liu et al. 2015; Toloza-Villalobos et al. 2015). We characterised the network of protein-protein interactions (PPI) that indicated all studied proteins have functions related to the vertebrate circadian system, as they interact with core circadian proteins (supplementary table S3). Interestingly, we found that most PER and CRY proteins of *Squalius* interact with BHLHe41 (basic helix-loop-helix family, member e41), a protein associated with biochemical pathways activated in response to cold temperatures, as previously observed in cold shock responses in zebrafish larvae (Hung et al. 2016). Moreover, studies using adult wild-type zebrafish organs in microarray characterizations have shown that expression of BHLHe41 presents circadian-dependant oscillations compatible with those observed for *per* genes (Li et al. 2013). These findings support the interaction between PER proteins and BHLHe41 due to their co-expression patterns. We found that CRY2 interacts with HSF2 (heat-shock factor 2), a transcription factor involved in activation of heat-shock proteins (HSP) expression in conditions of thermal stress. This interaction supports other findings that connect the circadian system with the thermal-stress response via heat-shock proteins (Tamaru et al. 2011; Chappuis et al. 2013; Jerônimo et al. 2017). Altogether, our results reinforce that genes in circadian system are also involved in temperature response, in agreement with previous results obtained from gene expression measurements in *D. rerio* under controlled experimental conditions (e.g. Lahiri et al. 2005; Chappuis et al. 2013; Jerônimo et al. 2017). In addition, we found that BMAL1A, BMAL2 and CRY3 interact with NFIL3-5 (nuclear factor, interleukin 3-regulates, member 5), a protein involved in immune response, specifically in the activation of transcription from the interleukin-3 promoter in T-cells (Zhang et al. 1995). Previous experiments in zebrafish larvae show that *nfil3-5* gene is rhythmically expressed in a light-dependent way, an indication that this gene may be coordinated with the circadian rhythmicity (Li et al. 2013; Li et al. 2015). This is consistent with the previously described link between the immune and circadian systems (Li et al. 2015). We found interactions between CRY5 (cryptochrome 5) and CRY-DASH (cryptochrome DASH), two proteins putatively belonging to the cryptochrome family, but whose functions are related with UV-induced DNA damage response (Daiyasu et al. 2004; Selby and Sancar 2006). Therefore, even though circadian-related proteins retain their ancestral functions in Iberian *Squalius* fish species, the proteins we identified, as well as their interactions, suggest that they consist of a redundant pool of

regulatory proteins, possibly as a consequence of selective pressures to maintain a robust circadian system. Still, we also detected diversification in functions that may have arisen to optimize important biological processes in space and time, synchronised with the circadian oscillation. For instance, the set of interactions in CRY, PER and BMAL proteins (supplementary table S3) involve circadian related proteins but also proteins with other biological functions. This could be due to events of subfunctionalization or gene dosage followed by neofunctionalization, i.e. resulting from natural selection to maintain the ancestral functions plus new function due to new mutations (Zhang 2003; He and Zhang 2005).

Molecular evolution and positive selection within *Squalius* circadian-related genes

Phylogenetic relationships within each gene family revealed that the history of these duplications is conserved in *Squalius* species. History of the gene families associated with the circadian system reflects the whole genome duplication events during evolution of teleost fish, including posterior gene losses described for the various gene families as well (Toloza-Villalobos et al. 2015). Phylogenies were investigated individually for each protein to assess whether relationships between the studied species recovered the previously described species tree topology based on a multi-locus phylogenetic analysis with other nuclear markers (Waap et al. 2011; Sousa-Santos et al. 2019). The previously described species tree showed two main clusters: i) *S. aradensis* and *S. torgalensis*; ii) *S. pyrenaicus* from Tagus clustering with *S. carolitertii* (Waap et al. 2011; Sousa-Santos et al. 2019). Southern populations of *S. pyrenaicus* (Almargem and Guadiana) rendered paraphyletic to *S. carolitertii* (Waap et al. 2011; Sousa-Santos et al. 2019).

In the present study, eight of the sixteen studied proteins presented a resolved phylogeny, but only for PER1A, PER2 and BMAL2 we recovered the topology consistent with the species tree mentioned above. Interestingly, for CRY1BA, our phylogeny recovered a topology with both populations of *S. pyrenaicus* as a monophyletic group, recovering the topology described for these species based on mitochondrial marker *cytochrome b* (Mesquita et al. 2007; Sousa-Santos et al. 2019). Overall, the fact that genes show a phylogeny compatible with the species tree suggest that evolution of the mentioned protein-coding genes seems to have been at least partially influenced by the speciation events which is likely associated with the formation and geomorphological history of the basins inhabited by these species (e.g. Brito et al. 1997; Filipe et al. 2009; Sousa-Santos et al. 2019).

An example of a gene with a phylogeny recovering the species tree and evidence of positive selection is the PER2 protein, where the phylogeny recovered the species tree and we detected signatures of selection in both southwestern sister species *S. aradensis* and *S. torgalensis*. This protein was reported to have a role in response to UV radiation

(supplementary table S2). Presently southwestern regions experience higher solar irradiance when compared to the northern region, and even slightly higher than south-eastern region (IPMA [ipma.pt (Portuguese Institute for Sea and Atmosphere); accessed in June 2018]; data not shown). In addition, PPI for PER2 (supplementary table S3) revealed an interaction with CRY-DASH protein, a cryptochrome-related protein strongly implied in UV-induced DNA damage repair (Daiyasu et al. 2004; Selby and Sancar 2006). Thus, PER2 is very likely a key agent for dealing with stress introduced by intense radiation (e.g. Corrà et al. 2017). Our results suggest that the changes observed in PER2 for southwestern *Squalius* species could be an adaptation to cope with increased UV radiation during the evolutionary history of these species. Additionally, PER2 could be used in the future as a potential molecular marker to assess radiation-induced stress in other species in order to understand responses to increasing UV radiation associated with climate change.

Interestingly, for CRY1AA, CRY1BB, CRY3, PER3 and CLOCKB, our inferred phylogenetic gene trees had two main clustering groups: i) *S. carolitertii* and *S. pyrenaicus* from Tagus; and ii) S*. torgalensis, S. aradensis* and *S. pyrenaicus* from Almargem. This topology is incongruent with the species tree but is consistent with a North-South distribution of environmental conditions of light and temperature where these species live (supplementary table S1). These incongruent gene trees can be due to incomplete lineage sorting of neutral mutations. However, as for some genes we found signatures of positive selection acting on these genes and branches, mostly in southern populations (Table 1, supplementary table S6), such neutral scenario is less likely. As previously mentioned, southern populations inhabit basins with a Mediterranean climate type, facing periods of overflowing during the winter, followed by severe periods of droughts in the summer, when habitable areas persist only in isolated deeper areas (Pires et al. 1999; Mesquita and Coelho 2002; Mesquita et al. 2005; Henriques et al. 2010).

The signals of molecular adaptation that we found in these southern species may thus be related to selective pressures imposed by these environmental conditions, which likely shaped the evolution of these genes. Nevertheless, convergent adaptive evolution can produce gene trees patterns that are incongruent with the species tree (Dávalos et al. 2012; Harms and Thornton 2013). In fact, we found a possible pattern of convergence between *S. aradensis* and *S. pyrenaicus* from Almargem based on phylogenetic analysis (see below).

Functional and structural assessment of circadian-related proteins in *Squalius*

Our results of the protein functions indicate that there are significant differences between *Squalius* species, which seem related to differences in environmental variables along their species distribution.

Temperature is an environmental factor expected to impose strong selective pressures for protein thermostability, especially in ectothermic organisms. Thus, protein thermostability can be of extreme importance for species dealing either with extreme warm or cold environments. Higher protein thermostability can be achieved either by (i) increasing the aliphatic index AI of a protein (Ikai 1980; Gasteiger et al. 2005), (ii) by formation of stronger quaternary structures (Fraser et al. 2016), (iii) by increasing the strength of ionic interactions (Kumar et al. 2000), or a combination of these mechanisms. Our results indicate that these three aspects might affect protein thermostability in *Squalius*. Proteins with an aliphatic index (AI) larger than 70 are considered highly thermostable, and we found such values for CRY and BMAL proteins. Interestingly, this higher AI was correlated with a high proportion of aliphatic amino acids that were conserved across species, which also showed signs of negative selection (supplementary table S7). The role of quaternary structure is exemplified by our results showing that for dimers CRY:PER and BMAL:CLOCK only one of the proteins needs to increase intrinsic thermostability to increase dimer stability (i.e. thermostability of PER and CLOCK proteins increases through the formation of quaternary structure rather than by increasing aliphatic index). Evidence for the role of ionic interactions comes from the pattern we found for PER3, CLOCKB, CLOCK2 and BMAL2, where southern species have a lower pI mostly due to the presence of extra charged amino acids, associated with formation of ionic bonding. Our results also indicate variation in protein thermostability across *Squalius* species related with differences in aliphatic index. For instance, we found a higher AI in northern populations for CRY1AA, CRY1AB, CRY1BA, PER1A, PER1B, CLOCKA, CLOCK2 and BMAL2 proteins and in the southern populations for CRY3, PER2, PER3, CLOCKB and BMAL1A. Interestingly, in PER2, this higher thermostability is associated with mutations under positive selection that lead to aliphatic amino acids (Table 1). Although we cannot always reject neutrality based on dN/dS, these differences in protein thermostability across species could reflect differential selective pressures across the environmental gradient, either due to current warmer environments experienced in southern basins and/or past colder environments in the northern basins, which were more affected by Pleistocene glaciations (Schmitt 2007). A clear example where we could reject neutrality is the PER1B. For this protein, we found signatures of positive selection for this gene in *S. torgalensis*, and physicochemical parameters revealed higher AI in this species, indicating that it has the most thermostable PER1B among *Squalius* species. In addition, several sites dispersed across *per1b* codons coding for aliphatic amino acids show evidence of negative selection, consistent with purifying selection for thermostability. This protein has shown to be important in integration of temperature and light cues within the circadian system (Lahiri et al. 2005). Thus, it is likely that these signatures of selection reflect adaptation to the specific conditions of the Mira basin, the only one inhabited by *S. torgalensis* (Henriques et al. 2010). Even

though the number of light hours per day and water temperature are similar to those experienced in other southern basis, Mira has a lower annual temperature variation when compared to Arade or Almargem (supplementary table S1). This may result from a higher overshadowing of rivers in Mira basin, when compared to other southern basins, caused by the presence of riparian and overhanging vegetation (Magalhães et al. 2002; Martelo et al. 2014). Density of riparian vegetation was shown to influence several aspects in the dynamics of the slow-flowing rivers, specifically water temperature, but also the irradiance reaching the water (Garner et al. 2017; Trimmel et al. 2018). Thus, given that seasonal variation in water temperature is lower in Mira than other southern margins, we speculate that patterns of PER1B in *S. torgalensis* might reflect selection for fine tuning the temperature integration within the circadian rhythm.

Water pH is another factor that has been shown to influence several physiological aspects of fish, either at gene expression and other biochemical levels (e.g. Jesus et al. 2017; Xu et al. 2017; Jesus et al. 2018). Thus, water pH my act as a selective pressure, specifically an increased alkalinity (i.e. higher pH), and consequent increased rate in protein precipitation, should favour mutations leading to decrease in isoelectric point (pI) values. In agreement with this hypothesis, in the species inhabiting southern basins with higher pH (supplementary table S1), we found a trend of lower pI values, associated with positively selected mutations (e.g. mutation T402D in PER2 – Table 1). An interesting example is the CRY1AA in *S. torgalensis* where we found that the lower values of pI can be correlated with the mutation K275Q, a site detected to be under positive selection. This mutation consists of a substitution of an uncharged for a positively charged amino acid with acid properties, leading to a decrease in the overall charge of the protein, and consequently a lower pI. Moreover, previous studies with *S. torgalensis* and *S. carolitertii* revealed that expression of *cry1aa* is affected either by water temperature and pH in these species (Jesus et al. 2017). Previous studies have been conducted in fish species to understand the effects of extreme water pH on the internal homeostasis of the organisms (e.g. Xu et al. 2013; Kavembe et al. 2015; Jesus et al. 2017; Xu et al. 2017; Jesus et al. 2018), but to our knowledge few studies related protein properties with changes in water pH (but see Jesus et al. 2017).

It is well known several mutations can impact protein structure as certain amino acids present a propensity for specific structural arrangements (Chou and Fasman 1974). We found mutations under positive selection inside and outside functional domains (Fig. 2). For instance, mutation K275Q in CRY1AA from *S. torgalensis* is in a region between two functional domains (Fig. 2), which could affect protein flexibility and inter-domain interaction. We also found several examples of mutations under positive selection that are in functional domains and could lead to differences in protein structure (e.g. CRY3, PER1A, PER2,

PER3). Another important positively selected mutation is the R20P in BMAL2 since proline behaves as a helix breaker, and the replacement of an arginine (R) by a proline (P) can potentially disrupt helix structures in this region (Chou and Fasman 1974; Barlow and Thornton 1988).

Adaptive convergence in *Squalius aradensis* and *Squalius pyrenaicus* from similar environments

The phylogenies we inferred indicate that for 8 out of 16 proteins the best topology clusters together *S. aradensis* and *S. pyrenaicus* from Almargem. According to the species tree these two species belong to two highly divergent lineages (Sousa-Santos et al. 2019), and hence this clustering is highly unlikely due to neutral incomplete lineage sorting. Instead, this suggests a scenario of evolutionary convergence of populations inhabiting similar environments. In fact, Almargem and Arade are two basins from the south of Portugal influenced by Mediterranean climate, facing very similar environmental conditions (e.g. average water temperature, water pH and photoperiod; supplementary table S1). This pattern of sequence convergence in these genes and proteins may thus be a consequence of convergent adaptation. The exception was the *cry1bb* we do not have support for convergence at any level except the phylogenetic, and therefore it would be explained by a scenario of incomplete lineage sorting.

The convergence in the remaining genes and proteins matches the criteria for detecting evolutionary convergence proposed by Dávalos et al. (2012), namely: (1) evidence from sequences of functional parts of genes; (2) clear link between gene function and ecological conditions; and (3) evidence that selection is acting on target genes at different rates from other lineages. We could confirm all the criteria for some genes, but for CRY1BB, BMAL1A and CLOCK2 we did not find evidence for signatures of selection based on dN/dS on the respective coding genes (criterium 1). Nevertheless, for BMAL1A and CLOCK2 proteins, the functional characterisation revealed similar physicochemical patterns in *S. aradensis* and *S. pyrenaicus* from Almargem, pointing to functional convergence. Clear evidence of adaptive convergence came from *clocka*, for which we found significant gene-wide positive selection and similar thermostability in *S. aradensis* and *S. torgalensis* (criterium 1), which can be associated with similar temperature, pH and light exposure (criterium 2). There were some cases where this was not so clear. For instance, for *per3* we only found positive selection in *S. pyrenaicus* from Almargem, but protein characterisation revealed similar functions in *S. aradensis* and *S. pyrenaicus* from Almargem. For *clockb* we did not find significant signatures of positive selection in both species, but we found in *S. carolitertii,* and for *per1b* we found positive selection in *S. torgalensis.* However, protein characterisation showed possible functional modifications in both *S. aradensis* and *S. pyrenaicus* (Almargem). This

might indicate that tests used to detect positive selection based on dN/dS did not have enough statistical power to find positively selected sites. In sum, combining tests of selection based on dN/dS and functional protein characterisation allowed us to detect patterns consistent with convergence between *S. aradensis* and *S. pyrenaicus* from Almargem. Scenarios of convergence have been described for other fish species at the morphological level (e.g. Muschick et al. 2012; Alter et al. 2015; Passow et al. 2017), and at the molecular level (e.g. Protas et al. 2006; Nath et al. 2013; Natarajan et al. 2016; Zhu et al. 2018; Graham and McCracken 2019). Moreover, light was shown to be an important determinant in two studies that detected molecular convergence in fish: (1) the evolution of albinism linked to the *Oca2* gene in two independent populations of the cavefish *Astyanax mexicanus* (Protas et al. 2006)*;* and (2) functional evolution of Rhodopsin proteins in several fish species (Yokoyama et al. 2008). This last study also elucidated the need to assess functional features, as statistical tests of positive selection based on dN/dS can fail to detect convergence at molecular level. Here, we detected significant convergent evolution in freshwater fish in genes and proteins related to integration of visual and thermal stimuli within the circadian system, which are part of gene families with duplications.

CONCLUSION

This is the first study to extensively characterise the evolution of circadian-related gene families in non-model freshwater organisms. Moreover, it is the first study to provide clear insights on how the environment can shape the evolution of the circadian system, and how it can contribute to the process of adaptation to different environments.

We studied the four gene families involved in the core mechanism of the circadian system in four Portuguese *Squalius* freshwater fish. For this we integrated results from the application of statistical methods based on dN/dS ratio to detect signatures of selection acting on these genes with sequence-based protein analysis at different structural levels of organisation. Our results show that tests based on dN/dS might not be powerful enough to detect signatures of positive selection and convergence, even for cases where the characterization of protein function are compatible with adaptation to different environmental conditions and/or convergent evolution. This reinforces the idea that dN/dS tests should be complemented with other analyses, following an integrative approach including a sequence-based protein characterisation at their different structural levels of organisation.

For the first time, our findings support that not only historical factors, but also selection, drove the molecular evolution of the four studied species that live in different environmental conditions of light and temperature related with a gradient shaped by Atlantic and Mediterranean climate types. Moreover, we find evidence for adaptive convergence between the two southern populations of *S. aradensis* and *S. pyrenaicus* from Almargem, a pattern described for the first time. Our approach and findings can help exploring patterns found in other species, namely concerning the role of convergence, a process that can be more frequent than expected.

MATERIAL AND METHODS

Sampling

Muscle tissue from five wild adult fish of *S. carolitertii* and *S. torgalensis* species was available from previous work (Jesus et al. 2017) and stored at -80ºC in RNAlater® (Ambion, Austin, TX, USA). The adult fish were sampled in Portuguese basins of Mondego (40˚8’5.22"N; 8˚8’35.06"W) and Mira (37˚38'1.31"N; 8˚37’22.37"W), respectively (Jesus et al. 2017). Samples of *S. pyrenaicus* were also stored at -80ºC from previous projects. This work includes two sampling sites: Almargem (37°09'50.7"N; 7°37'13.2"W) (Machado et al. 2016) and Tagus (39°43'48.2"N; 7°45'38.1"W) (Matos et al. 2016).

Squalius aradensis individuals were captured from Portuguese basin Arade (37º17'0.53"N; 8º29'7.31"W) under the license 421/2017/CAPT issued by Portuguese authority for Conservation of endangered species [ICNF (Instituto da Conservação da Natureza e das Florestas)]. After capturing, the specimens were transported alive to the laboratory in aerated containers to minimize fish discomfort and euthanized immediately upon arrival with an overdose of tricaine mesylate (400 ppm of MS-222; Sigma-Aldrich, St. Louis, MO, USA) with sodium bicarbonate (1:2) following the recommended ethical guidelines (ASAB/ABS, 2012) and European Union regulations. Organs were stored in RNAlater® at -80ºC until further use. Distribution of the species and sampling sites are illustrated in Figure S1.

We characterized several environmental variables for each sampling site, summarized in supplementary table S1. The average water temperature and the pH between 2001 and 2016 were calculated with data retrieved from SNIRH [snirh.pt (National Information System of Water Resources); accessed in June 2018]. The number of daylight hours was retrieved from Time and Date AS (timeanddate.com; accessed in June 2018).

Identification of circadian system related genes in Iberian freshwater fish

Previously published transcriptome assemblies of *S. torgalensis* and *S. carolitertii* (Jesus et al. 2015), were used to identify the genes related to the circadian system in the study species. To identify potential genes related to the circadian system we performed BLAST searches of both transcriptomes against two *Danio rerio* light-induced transcriptomes (Weger et al. 2011; Ben-Moshe et al. 2014). To avoid retrieving paralogous or splicing isoforms we used a stringent e-value threshold of 1×10^{-7} for the BLAST searches and kept only sequences with identity higher than 85%. As expected, the light-responsive genes that we identified using this procedure included several genes already characterised as components of the circadian system, but there were also several uncharacterised genes or genes unrelated to the circadian system. To further test that the identified genes were involved in

the circadian system we performed a functional enrichment analysis using the list of top blast *Danio rerio* ENA accession numbers and the method implemented in DAVID functional annotation tool (Huang et al. 2007). For all functional analyses performed in DAVID, we considered that an EASE score <0.05 indicated statistical significance, since this score provides a statistical test to examine the significance of gene–term enrichment with a modified Fisher's exact test. Through this methodology we were able to find enriched GO terms among the genes retrieved. The most significant enriched GO terms for Biological Process and Molecular Function were filtered, by removing all the genes whose function was unrelated to the circadian system, and among the circadian related genes a threshold for adjusted p-values (Benjamini) of 0.05 was used to remove false positives. From the final list of genes only protein-coding genes whose function was related to the core circadian mechanism were maintained for further analysis (supplementary table S2).

Gene sequencing and protein sequence prediction

Based on the sequences retrieved from the transcriptomes that matched core genes of the circadian system, we designed specific primers for Polymerase chain reactions (PCRs) using PerlPrimer software v.1.1.19 (Marshall 2004) (supplementary table S9) with the purpose of amplifying those same genes for all the studied species.

Total RNA was extracted from muscle samples of 25 individuals, 5 from each population. RNA was used to facilitate gene sequencing since introns are avoided. 1 mL TRI Reagent (Ambion, Austin, TX, USA) was added to 50–100 mg of muscle samples and, after homogenization with Tissue Ruptor (Qiagen, Valencia, CA, USA), RNA was extracted according to the TRI Reagent manufacturers protocol. TURBO DNase (Ambion, Austin, TX, USA) was employed to degrade any remaining genomic DNA contaminants, followed by phenol/chloroform purification and LiCl precipitation (Cathala et al. 1983). Sample quality was checked using a NanoDrop™-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) based on the 260nm/280nm and 260nm/230nm absorbance ratios. Samples concentration was determined with Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) to ensure enough quantity of homogeneous RNA for cDNA synthesis. Synthesis of cDNA was performed according to manufacturer’s protocol using the RevertAid H Minus First Strand cDNA synthesis kit (Thermo Fisher Scientific, Waltham, MA, USA), and it was stored subsequently at -20˚C until further use. PCRs were performed in 25 µL reactions containing 10–100 ng of cDNA, 2 mM $MgCl_2$, 2 mM each dNTP, 10 µM each primer, Taq Polymerase (5 U/µL), and 1× Taq buffer using thermocycler conditions described in supplementary table S10. PCR products were confirmed using a 1% agarose gel electrophoresis, and after purification with ExoSAP-IT® PCR Product Cleanup (Affymetrix, Inc., Santa Clara, CA, USA), they were sequenced by Sanger sequencing.

Sequences were aligned and edited using Sequencher v.4.2 (Gene Codes Corp., Ann Arbor, MI, USA). Non-redundant nucleotide sequences were deposited in European Nucleotide Archive (ENA) database under the accession numbers available in supplementary table S2. CLC Sequence Viewer v.7.5. (CLC bio, Aarhus, Denmark) was used to predict protein sequences for *in silico* analysis. BLAST searches were conducted with resulting protein sequences against UniProt database (The UniProt Consortium 2015) to ensure their identity. Protein sequences for *Danio rerio* were retrieved from UniProt database for each protein (supplementary table S4), as well as *Drosophila melanogaster* homolog sequence for each gene family (supplementary table S4). Protein sequences were aligned by gene family using the M-Coffee method, that combines several alignment algorithms (eg. MUSCLE, MAFFT and CLUSTAL) (Wallace et al. 2006) available in the T-Coffee web server (Di Tommaso et al. 2011). For each individual gene, nucleotide sequences of *Squalius* species were also aligned using M-Coffee.

Phylogenetic analysis and molecular evolution

The most appropriate model for amino acid substitution for each data set was determined with ProtTest v.3.0 (Abascal et al. 2005; Darriba et al. 2011), using both the Akaike information criteria and Bayesian information criteria. Phylogenetic trees were reconstructed for each gene family independently using Bayesian Inference method as implemented in MrBayes v.3.2.6 (Huelsenbeck and Ronquist 2001; Ronquist et al. 2012) using *D. melanogaster* protein sequences (supplementary table S4) as outgroup. 500,000 generations of Monte Carlo Markov Chain (MCMC) were run using the parameters determined in ProtTest for protein sequences as priors. Trees were sampled every 500 generations during the analysis. The first 50,000 generations were excluded as burn-in after examining the variation in log-likelihood scores over time. Phylogenetic trees were constructed using protein sequences instead of nucleotide sequences to avoid bias from the third codon rapid evolution. Protein sequences used correspond to the direct translation of nucleotide sequences under the standard genetic code. All trees were edited in FigTree v1.4.2 (A. Rambaut, University of Edinburgh, UK; http://tree.bio.ed.ac.uk/software/figtree/).

To find potential post-duplication diversification events among the proteins identified, we performed a sequence-based prediction of protein-protein interactions (PPI) using the online tool STRING v10.5 (von Mering et al. 2003; Szklarczyk et al. 2015 ; https://string-db.org/cgi/input.pl accessed in August 2018), to identify potential new functions these proteins may have acquired outside the circadian system, based on their patterns of interaction. Interactors (i.e. proteins found to be interacting with those from our study) were predicted comparing *Squalius* predicted consensus protein sequences against Actinopterygii

PPI databases (including *D. rerio, Oryzias latipes,* and *Gasterosteus aculeatus*) using a default threshold for score of 0.7.

Analysis of signatures of selection based on dN/dS

Signatures of selection were examined based on the dN/dS ratio (also known as the parameter ω) using four models implemented in HyPhy (Kosakovsky Pond et al. 2005) through the Datamonkey adaptive evolution webserver (Kosakovsky Pond & Frost 2005; Weaver et al. 2018 ; http://www.datamonkey.org/; accessed in August 2018), including (1) BUSTED (Branch-site Unrestricted Statistical Test for Episodic Diversification) (Murrell et al. 2015) that provides a gene-wide test for positive selection, i.e. it estimates one ω per gene; (2) MEME (Mixed Effects Model of Evolution) (Murrell et al. 2012) a mixed-effects maximum likelihood approach to test the hypothesis that individual sites have been subject to positive selection, i.e. estimating variable ω among sites; (3) aBSREL (adaptive Branch-Site Random Effects Likelihood) (Kosakovsky Pond et al. 2011; Smith et al. 2015) for genes whose signals of positive selection were detected with BUSTED or MEME, to test for positive selection on particular branches of the gene tree, i.e. estimating different ω for different branches; and (4) FEL (Fixed Effects Likelihood) (Kosakovsky Pond and Frost 2005b) that uses a maximum likelihood approach to infer nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny and tests the hypothesis that individual sites have been subject to negative selection.

Functional analysis and structural organisation of the predicted proteins

Homology methods available on several resources at the ExPASy Server (Gasteiger et al. 2005) were used to infer several properties of the predicted proteins from RNA sequences. Specifically, physicochemical parameters of the proteins were predicted using ProtParam (Gasteiger et al. 2005). Given the biological significance of factors related to temperature and pH, we estimated the aliphatic index (AI) and the isoelectric point (pI). We tested for differences in physicochemical parameters using several statistical tests implemented in R v.3.2.3 (R Core Team 2015). First, we checked for normality using the Shapiro-Wilk Test. Due to lack of normality, we used a Kruskal-Wallis Rank Sum Test to identify overall statistical differences in parameters across the populations. When evidence for differences were found, we conducted pairwise Wilcoxon Rank Sum Tests to compare the different groups.

To assess further functional features, a sequence-based prediction of family assignment and sequence domains based on collections of Hidden-Markov Models to support the predictions (Eddy 1998; Eddy 2011) were accomplished using HMMER web server (Finn et al. 2011; Prakash et al. 2017; Potter et al. 2018) against Pfam database (Finn et al. 2016).

Representative images of structural organization of domains and locations of sites under selection were created and edited with IBS, Illustrator for Biological Sequences (W. Liu et al. 2015).

ACKNOWLEDGEMENTS

The authors thank Carla Sousa-Santos for collaborating with fieldwork in collection of specimens of *S. aradensis*. We also acknowledge ICNF (Instituto da Conservação da Natureza e Florestas) for issuing the fishing licenses. This work was supported by FCT Strategic projects UID/BIA/00329/2013 (2015-2018) granted to cE3c by the Portuguese Foundation for Science and Technology (FCT).

AUTHOR CONTRIBUTIONS

Conceptualization and design by JMM and MMC. Investigation and data analysis by JMM. Writing, reviewing and revising the article by JMM, VCS, TFJ and MMC. Final version approved by all authors.

REFERENCES

Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Alter SE, Brown B, Stiassny MLJ. 2015. Molecular phylogenetics reveals convergent evolution in lower Congo River spiny eels. BMC Evolutionary Biology 15:224.

Barlow DJ, Thornton JM. 1988. Helix geometry in proteins. Journal of Molecular Biology 201:601–619.

Ben-Moshe Z, Alon S, Mracek P, Faigenbloom L, Tovin A, Vatine GD, Eisenberg E, Foulkes NS, Gothilf Y. 2014. The light-induced transcriptome of the zebrafish pineal gland reveals complex regulation of the circadian clockwork by light. Nucleic Acids Research 42:3750–3767.

Brito RM, Briolay J, Galtier N, Bouvet Y, Coelho MM. 1997. Phylogenetic relationships within genus Leuciscus (Pisces, Cyprinidae) in Portuguese fresh waters, based on mitochondrial DNA cytochrome b sequences. Mol. Phylogenet. Evol. 8:435–442.

Cathala G, Savouret J-F, Mendez B, West BL, Karin M, Martial JA, Baxter JD. 1983. A Method for Isolation of Intact, Translationally Active Ribonucleic Acid. DNA 2:329–335.

Chappuis S, Ripperger JA, Schnell A, Rando G, Jud C, Wahli W, Albrecht U. 2013. Role of the circadian clock gene Per2 in adaptation to cold temperature. Molecular Metabolism 2:184–193.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. Biochemistry 13:222–245.

Coelho MM, Bogutskaya NG, Rodrigues JA, Collares-Pereira MJ. 1998. Leuciscus torgalensis, and L. aradensis, two new cyprinids for Portuguese fresh waters. Journal of Fish Biology 52:937–950.

Corrà S, Salvadori R, Bee L, Barbieri V, Mognato M. 2017. Analysis of DNA-damage response to ionizing radiation in serum-shock synchronized human fibroblasts. Cell Biology and Toxicology 33:373–388.

Daiyasu H, Ishikawa T, Kuma K, Iwai S, Todo T, Toh H. 2004. Identification of cryptochrome DASH from vertebrates. Genes Cells 9:479–495.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Dávalos LM, Cirranello AL, Geisler JH, Simmons NB. 2012. Understanding phylogenetic incongruence: lessons from phyllostomid bats. Biological Reviews 87:991–1024.

Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J-M, Taly J-FO, Notredame C. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Research 39:W13–W17.

Dunlap JC. 1999. Molecular Bases for Circadian Clocks. Cell 96:271–290.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Computational Biology 7:e1002195.

Filipe AF, Araújo MB, Doadrio I, Angermeier PL, Collares-Pereira MJ. 2009. Biogeography of Iberian freshwater fishes revisited: the roles of historical versus contemporary constraints. Journal of Biogeography 36:2096–2110.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39:W29–W37.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44:D279–D285.

Foulkes NS, Whitmore D, Vallone D, Bertolucci C. 2016. Studying the Evolution of the Vertebrate Circadian Clock. In: Genetics, Genomics and Fish Phenomics. Vol. 95. p. 1–30.

Fraser NJ, Liu J-W, Mabbitt PD, Correy GJ, Coppin CW, Lethier M, Perugini MA, Murphy JM, Oakeshott JG, Weik M, et al. 2016. Evolution of Protein Quaternary Structure in Response to Selective Pressure for Increased Thermostability. Journal of Molecular Biology 428:2359–2371.

Garner G, Malcolm IA, Sadler JP, Hannah DM. 2017. The role of riparian vegetation density, channel orientation and water velocity in determining river temperature dynamics. Journal of Hydrology 553:471–485.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. The Proteomics Protocols Handbook:571–607.

Graham AM, McCracken KG. 2019. Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes EGLN1 and EPAS1 in high-altitude ducks. Heredity [Internet]. Available from: http://www.nature.com/articles/s41437-018-0173-z

Harms MJ, Thornton JW. 2013. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nature Reviews Genetics 14:559–571.

He X, Zhang J. 2005. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. Genetics 169:1157–1164.

Henriques R, Sousa V, Coelho MM. 2010. Migration patterns counteract seasonal isolation of Squalius torgalensis, a critically endangered freshwater fish inhabiting a typical Circum-Mediterranean small drainage. Conserv Genet 11:1859–1870.

Huang D, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biology 8:R183.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Hung I-C, Hsiao Y-C, Sun HS, Chen T-M, Lee S-J. 2016. MicroRNAs regulate gene plasticity during cold shock in zebrafish larvae. BMC Genomics 17.

Ikai A. 1980. Thermostability and aliphatic index of globular proteins. J. Biochem. 88:1895–1898.

Ishikawa T, Hirayama J, Kobayashi Y, Todo T. 2002. Zebrafish CRY represses transcription mediated by CLOCK-BMAL heterodimer without inhibiting its binding to DNA. Genes to Cells 7:1073–1086.

Jerônimo R, Moraes MN, de Assis LVM, Ramos BC, Rocha T, Castrucci AM de L. 2017. Thermal stress in Danio rerio : a link between temperature, light, thermo-TRP channels, and clock genes. Journal of Thermal Biology 68:128–138.

Jesus TF, Grosso AR, Almeida-Val VMF, Coelho MM. 2015. Data from: "Characterization of two Iberian freshwater fish transcriptomes, Squalius carolitertii and Squalius torgalensis, living in distinct environmental conditions" in Genomic Resources Notes Accepted 1 April 2015 to 31 May 2015. Molecular Ecology Resources 16:377.

Jesus TF, Moreno JM, Repolho T, Athanasiadis A, Rosa R, Almeida-Val VMF, Coelho MM. 2017. Protein analysis and gene expression indicate differential vulnerability of Iberian fish species under a climate change scenario. PLOS ONE 12:e0181325.

Jesus TF, Rosa IC, Repolho T, Lopes AR, Pimentel MS, Almeida-Val VMF, Coelho MM, Rosa R. 2018. Different ecophysiological responses of freshwater fish to warming and acidification. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 216:34–41.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8:275–282.

Kavembe GD, Franchini P, Irisarri I, Machado-Schiaffino G, Meyer A. 2015. Genomics of Adaptation to Multiple Concurrent Stresses: Insights from Comparative Transcriptomics of a Cichlid Fish from One of Earth's Most Extreme Environments, the Hypersaline Soda Lake Magadi in Kenya, East Africa. Journal of Molecular Evolution 81:90–109.

Khaldi N, Shields DC. 2011. Shift in the isoelectric-point of milk proteins as a consequence of adaptive divergence between the milks of mammalian species. Biology Direct 6:40.

Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczuk M, Biecek P, Polak N, Smolarczyk K, Dudek MR, Cebrat S. 2007. The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics 8:163.

Kornmann B, Schaad O, Bujard H, Takahashi JS, Schibler U. 2007. System-Driven and Oscillator-Dependent Circadian Transcription in Mice with a Conditionally Active Liver Clock. PLOS Biology 5:e34.

Kosakovsky Pond SL, Frost SDW. 2005a. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533.

Kosakovsky Pond SL, Frost SDW. 2005b. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Molecular Biology and Evolution 22:1208–1222.

Kosakovsky Pond SL, Frost SDW, Muse VS. 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics 21:676–679.

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. Molecular Biology and Evolution 28:3033–3043.

Kumar S, Tsai C-J, Nussinov R. 2000. Factors enhancing protein thermostability. Protein Engineering, Design and Selection 13:179–191.

Lahiri K, Vallone D, Gondi SB, Santoriello C, Dickmeis T, Foulkes NS. 2005. Temperature Regulates Transcription in the Zebrafish Circadian Clock.Stemple D, editor. PLoS Biology 3:e351.

Lazado CC, Kumaratunga HPS, Nagasawa K, Babiak I, Giannetto A, Fernandes JMO. 2014. Daily Rhythmicity of Clock Gene Transcripts in Atlantic Cod Fast Skeletal Muscle. PLoS ONE 9:e99172.

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. Molecular Biology and Evolution 25:1307–1320.

Li L, Ji D, Teng L, Zhang S, Li H. 2015. Identification and expression of lypc, a novel dark-inducible member of Ly6 superfamily in zebrafish Danio rerio. Gene 574:69–75.

Li Y, Li G, Wang H, Du J, Yan J. 2013. Analysis of a Gene Regulatory Cascade Mediating Circadian Rhythm in Zebrafish. PLOS Computational Biology 9:e1002940.

Lim C, Allada R. 2013. Emerging roles for post-transcriptional regulation in circadian clocks. Nature Neuroscience 16:1544–1550.

Lin C, Todo T. 2005. The cryptochromes. Genome Biol. 6:220.

Liu C, Hu J, Qu C, Wang L, Huang G, Niu P, Zhong Z, Hong F, Wang G, Postlethwait JH, et al. 2015. Molecular evolution and functional divergence of zebrafish (Danio rerio) cryptochrome genes. Scientific Reports 5:8113.

Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, Lahrmann U, Zhao Q, Zheng Y, Zhao Y, et al. 2015. IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics 31:3359–3361.

Liu Y, Merrow M, Loros JJ, Dunlap JC. 1998. How temperature changes reset a circadian oscillator. Science 281:825–829.

Machado MP, Matos I, Grosso AR, Schartl M, Coelho MM. 2016. Non-canonical expression patterns and evolutionary rates of sex-biased genes in a seasonal fish. Molecular Reproduction and Development 83:1102–1115.

Magalhães MF, Beja P, Canas C, Collares-Pereira MJ. 2002. Functional heterogeneity of dry-season fish refugia across a Mediterranean catchment: the role of habitat and predation. Freshwater Biology 47:1919–1934.

Marshall OJ. 2004. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics 20:2471–2472.

Martelo J, Grossman GD, Porto M, Filomena Magalhães M. 2014. Habitat patchiness affects distribution and microhabitat use of endangered Mira chub Squalius torgalensis (Actinopterygii, Cypriniformes). Hydrobiologia 732:93–109.

Matos IMN, Coelho MM, Schartl M. 2016. Gene copy silencing and DNA methylation in natural and artificially produced allopolyploid fish. The Journal of Experimental Biology 219:3072–3081.

McClung CR. 2011. Circadian Rhythms: Lost in Post-Translation. Current Biology 21:R400–R402.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. 2003. STRING: A database of predicted functional associations between proteins. Nucleic Acids Research 31:258–261.

Mesquita N, Coelho MM. 2002. The ichthyofauna of the small Mediterranean-type drainages of Portugal: its importance for conservation. In: Conservation of Freshwater Fishes: Options for the Future. p. 65–71.

Mesquita N, Cunha C, Carvalho GR, Coelho MM. 2007. Comparative phylogeography of endemic cyprinids in the south-west Iberian Peninsula: evidence for a new ichthyogeographic area. Journal of Fish Biology 71:45–75.

Mesquita N, Hänfling B, Carvalho GR, Coelho MM. 2005. Phylogeography of the cyprinid Squalius aradensis and implications for conservation of the endemic freshwater fauna of southern Portugal. Molecular Ecology 14:1939–1954.

Möglich A, Ayers RA, Moffat K. 2009. Structure and Signaling Mechanism of Per-ARNT-Sim Domains. Structure 17:1282–1294.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. Molecular Biology and Evolution 32:1365–1371.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genetics 8:e1002764.

Muschick M, Indermaur A, Salzburger W. 2012. Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. Current Biology 22:2362–2368.

Nakahata Y, Yoshida M, Takano A, Soma H, Yamamoto T, Yasuda A, Nakatsu T, Takumi T. 2008. A direct repeat of E-box-like elements is required for cell-autonomous circadian rhythm of clock genes. BMC Molecular Biology 9:1.

Natarajan C, Hoffmann FG, Weber RE, Fago A, Witt CC, Storz JF. 2016. Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. Science 354:336–339.

Nath A, Chaube R, Subbiah K. 2013. An insight into the molecular basis for convergent evolution in fish antifreeze Proteins. Computers in Biology and Medicine 43:817–821.

Pando MP, Pinchak AB, Cermakian N, Sassone-Corsi P. 2001. A cell-based system that recapitulates the dynamic light-dependent regulation of the vertebrate clock. Proc Natl Acad Sci U S A 98:10178–10183.

Pando MP, Sassone-Corsi P. 2002. Unraveling the mechanisms of the vertebrate circadian clock: zebrafish may light the way. BioEssays 24:419–426.

Paranjpe DA, Sharma VK. 2005. Evolution of temporal order in living organisms. Journal of Circadian Rhythms 3:7.

Passow CN, Arias-Rodriguez L, Tobler M. 2017. Convergent evolution of reduced energy demands in extremophile fish. PLoS One 12.

Pires AM, Cowx IG, Coelho MM. 1999. Seasonal changes in fish community structure of intermittent streams in the middle reaches of the Guadiana basin, Portugal. Journal of Fish Biology 54:235–249.

Ponting CP, Aravind L. 1997. PAS: a multifunctional domain family comes to light. Current biology : CB 7:R674–7.

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204.

Prakash A, Jeffryes M, Bateman A, Finn RD. 2017. The HMMER Web Server for Protein Sequence Similarity Search. Current Protocols in Bioinformatics 60:3.15.1-3.15.23.

Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nature Genetics 38:107–111.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Schmitt T. 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. Frontiers in Zoology 4:11.

Selby CP, Sancar A. 2006. A cryptochrome/photolyase class of enzymes with single-stranded DNA-specific photolyase activity. Proc Natl Acad Sci U S A 103:17696–17700.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Molecular Biology and Evolution 32:1342–1353.

Söti C, Pál C, Papp B, Csermely P. 2005. Molecular chaperones as regulatory elements of cellular networks. Current Opinion in Cell Biology 17:210–215.

Sousa-Santos C, Jesus TF, Fernandes C, Robalo JI, Coelho MM. 2019. Fish diversification at the pace of geomorphological changes: evolutionary history of western Iberian Leuciscinae (Teleostei: Leuciscidae) inferred from multilocus sequence data. Molecular Phylogenetics and Evolution 133:263–285.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research 43:D447–D452.

Tamai TK, Young LC, Whitmore D. 2007. Light signaling to the zebrafish circadian clock by Cryptochrome 1a. Proceedings of the National Academy of Sciences of the United States of America 104:14712–14717.

Tamaru T, Hattori M, Honda K, Benjamin I, Ozawa T, Takamatsu K. 2011. Synchronization of Circadian Per2 Rhythms and HSF1-BMAL1:CLOCK Interaction in Mouse Fibroblasts after Short-Term Heat Shock Pulse. PLOS ONE 6:e24521.

The UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Research 43:D204–D212.

Toloza-Villalobos J, Arroyo JI, Opazo JC. 2015. The Circadian Clock of Teleost Fish: A Comparative Analysis Reveals Distinct Fates for Duplicated Genes. Journal of Molecular Evolution 80:57–64.

Trimmel H, Weihs P, Leidinger D, Formayer H, Kalny G, Melcher A. 2018. Can riparian vegetation shade mitigate the expected rise in stream temperatures due to climate change during heat waves in a human-impacted pre-alpine river? Hydrology and Earth System Sciences 22:437–461.

Tsuchiya Y, Akashi M, Nishida E. 2003. Temperature compensation and temperature resetting of circadian rhythms in mammalian cultured fibroblasts. Genes to Cells 8:713–720.

Vallone D, Gondi SB, Whitmore D, Foulkes NS. 2004. E-box function in a period gene repressed by light. Proc Natl Acad Sci U S A 101:4106–4111.

Vatine G, Vallone D, Appelbaum L, Mracek P, Ben-Moshe Z, Lahiri K, Gothilf Y, Foulkes NS. 2009. Light Directs Zebrafish period2 Expression via Conserved D and E Boxes. PLoS Biology 7:e1000223.

Vaze KM, Sharma VK. 2013. On the adaptive significance of circadian clocks for their owners. Chronobiology international 30:413–33.

Waap S, Amaral AR, Gomes B, Manuela Coelho M. 2011. Multi-locus species tree of the chub genus Squalius (Leuciscinae: Cyprinidae) from western Iberia: new insights into its evolutionary history. Genetica 139:1009–1018.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 34:1692–1699.

Wang H. 2008a. Comparative analysis of period genes in teleost fish genomes. Journal of Molecular Evolution 67:29–40.

Wang H. 2008b. Comparative analysis of teleost fish genomes reveals preservation of different ancient clock duplicates in different fishes. Marine Genomics 1:69–78.

Wang H. 2009. Comparative genomic analysis of teleost fish bmal genes. Genetica 136:149–161.

Wang M, Zhong Z, Zhong Y, Zhang W, Wang H. 2015. The Zebrafish Period2 Protein Positively Regulates the Circadian Clock through Mediation of Retinoic Acid Receptor (RAR)-related Orphan Receptor α (Rorα). Journal of Biological Chemistry 290:4367–4382.

Weaver S, Shank SD, Spielman SJ, Li M, Muse VS, Kosakovsky Pond SL. 2018. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Molecular Biology and Evolution 35:773–777.

Weger BD, Sahinbas M, Otto GW, Mracek P, Armant O, Dolle D, Lahiri K, Vallone D, Ettwiller L, Geisler R, et al. 2011. The Light Responsive Transcriptome of the Zebrafish: Function and Regulation. PLoS ONE 6:e17080.

Whitmore D, Foulkes NS, Strähle U, Sassone-Corsi P. 1998. Zebrafish Clock rhythmic expression reveals independent peripheral circadian oscillators. Nature neuroscience 1:701–707.

Xu J, Li J-T, Jiang Y, Peng W, Yao Z, Chen B, Jiang L, Feng J, Ji P, Liu G, et al. 2017. Genomic Basis of Adaptive Evolution: The Survival of Amur Ide (*Leuciscus waleckii*) in an Extremely Alkaline Environment. Molecular Biology and Evolution 34:145–159.

Xu J, Li Q, Xu L, Wang S, Jiang Y, Zhao Z, Zhang Y, Li J, Dong C, Xu P, et al. 2013. Gene expression changes leading extreme alkaline tolerance in Amur ide (Leuciscus waleckii) inhabiting soda lake. BMC Genomics 14:682.

Yokoyama S, Tada T, Zhang H, Britt L. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proceedings of the National Academy of Sciences 105:13480–13485.

Zhang J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292–298.

Zhang W, Zhang J, Kornuc M, Kwan K, Frank R, Nimer SD. 1995. Molecular cloning and characterization of NF-IL3A, a transcriptional activator of the human interleukin-3 promoter. Mol Cell Biol 15:6055–6063.

Zhu X, Guan Y, Signore AV, Natarajan C, DuBay SG, Cheng Y, Han N, Song G, Qu Y, Moriyama H, et al. 2018. Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the Qinghai-Tibet Plateau. Proceedings of the National Academy of Sciences 115:1865–1870.

Table 1. Positively selected sites and probable impacts on protein physicochemical parameter.

Protein	Number of sites under positive selection	Sites under positive selection	p-value	Type of non-synonymous mutation	Species*	Impacted physicochemical parameter**
CRY1AA	1	K275Q	0.03	Non-conservative	*S. torgalensis*	Decrease in isoelectric point (pI)
CRY2	1	I274V	0.00	Conservative	*S. aradensis* / *S. torgalensis*	None
CRY3	1	Q96H	0.02	Non-conservative	*S. aradensis* / *S. torgalensis*	Increase in pI
PER1A	1	T254A	0.08	Non-conservative	*S. carolitertii*	Increase in aliphatic index
PER1B	1	T1256V	0.08	Non-conservative	*S. torgalensis*	Increase in aliphatic index
PER2	4	T402D	0.01	Non-conservative	*S. aradensis* / *S. pyrenaicus* (Almargem)	Decrease in pI
		T514L	0.02	Non-conservative	*S. aradensis*	Increase in aliphatic index
		R697A	0.03	Non-conservative	*S. torgalensis*	Decrease in pI; decrease in aliphatic index
		G1208T	0.09	Non-conservative	*S. torgalensis*	None
PER3	1	K429Y	0.02	Non-conservative	*S. carolitertii* / *S. pyrenaicus* (Tagus)	Decrease in pI
BMAL2	1	R20P	0.09	Non-conservative	*S. torgalensis*	Decrease in pI

*The species/populations with the mutated form was determined according to the comparison with the amino acid present in the reference (*D. rerio*) sequence.

**We include in this table only the parameters we predicted and analysed in this study; these modifications pose possible functional impacts on the protein.

Figure captions

Fig. 1. A phylogenetic tree constructed by the Bayesian Inference method for **(a)** CRY proteins with fly CRY as outgroup using the LG substitution model (Le and Gascuel 2008) with a discrete Gamma distribution (+G) with 5 rate categories; **(b)** PER proteins with fly PER as outgroup using the JTT substitution model (Jones et al. 1992) with empirical amino acid frequencies from the data (+F); **(c)** CLOCK proteins with fly CLOCK protein as outgroup using JTT substitution model (Jones et al. 1992) using a discrete Gamma distribution (+G) with 2 rate categories and empirical amino acid frequencies from the data (+F); **(d)** BMAL proteins with fly CYCLE protein as outgroup using JTT substitution model (Jones et al. 1992) using a discrete Gamma distribution (+G) with 3 rate categories. Values on branch nodes represent Bayesian posterior probabilities. Sc, *Squalius carolitertii*; SpT, *Squalius pyrenaicus* (Tagus population); St, *Squalius torgalensis*; Sa*, Squalius aradensis*; SpA, *Squalius pyrenaicus* (Almargem population); Dr, *Danio rerio*; Dm, *Drosophila melanogaster*.

Fig. 2. Domain organization of proteins with positively selected sites located inside functional domains; purple dots represent the positively selected positions and grey dots represent negatively selected positions.

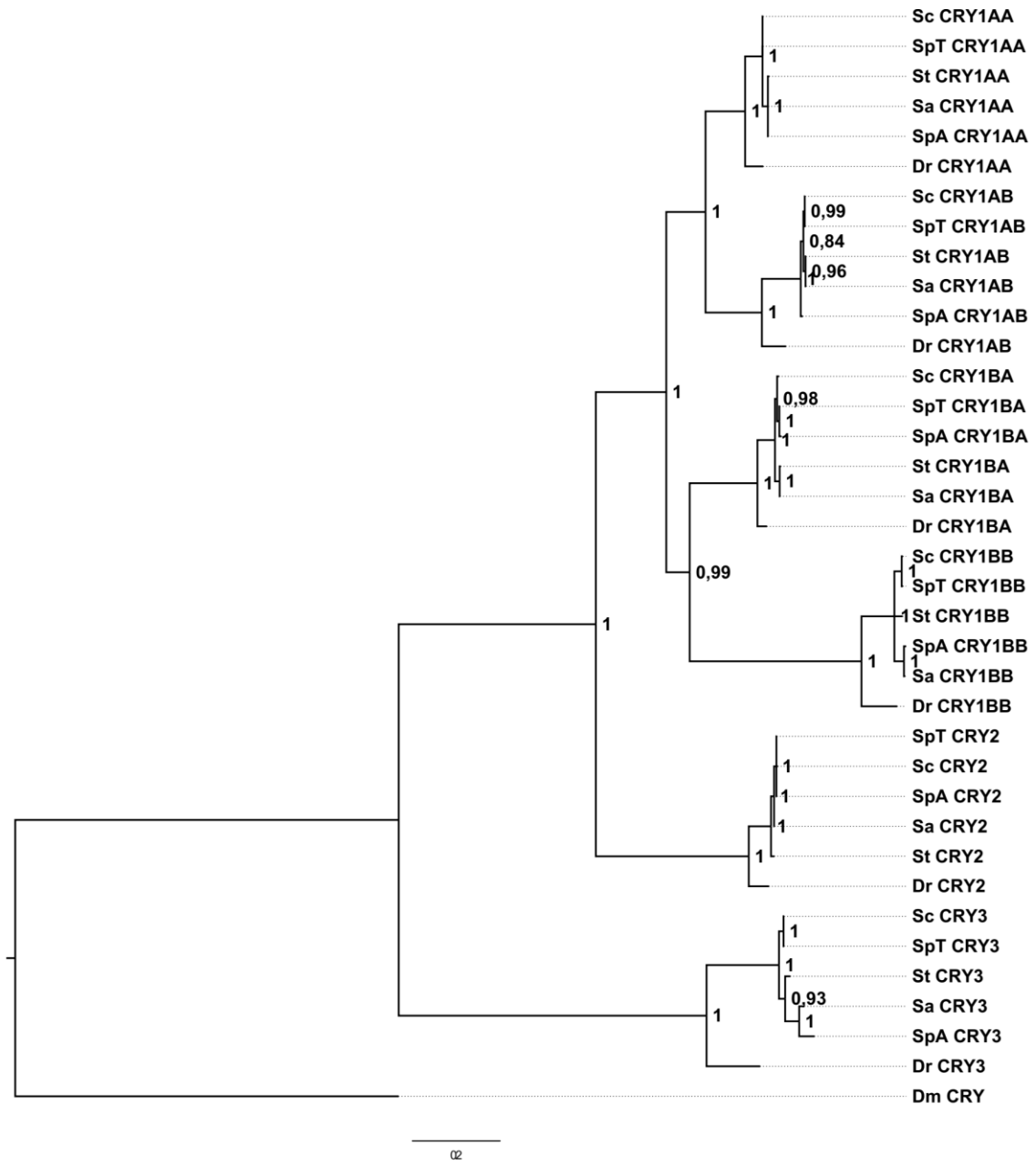
Sc CRY1AA
SpT CRY1AA
St CRY1AA
Sa CRY1AA
SpA CRY1AA
Dr CRY1AA
Sc CRY1AB
SpT CRY1AB
St CRY1AB
Sa CRY1AB
SpA CRY1AB
Dr CRY1AB
Sc CRY1BA
SpT CRY1BA
SpA CRY1BA
St CRY1BA
Sa CRY1BA
Dr CRY1BA
Sc CRY1BB
SpT CRY1BB
St CRY1BB
SpA CRY1BB
Sa CRY1BB
Dr CRY1BB
SpT CRY2
Sc CRY2
SpA CRY2
Sa CRY2
St CRY2
Dr CRY2
Sc CRY3
SpT CRY3
St CRY3
Sa CRY3
SpA CRY3
Dr CRY3
Dm CRY
0,99
0,84
0,96
0,98
0,99
0,93
1

Fig. 1 (a)

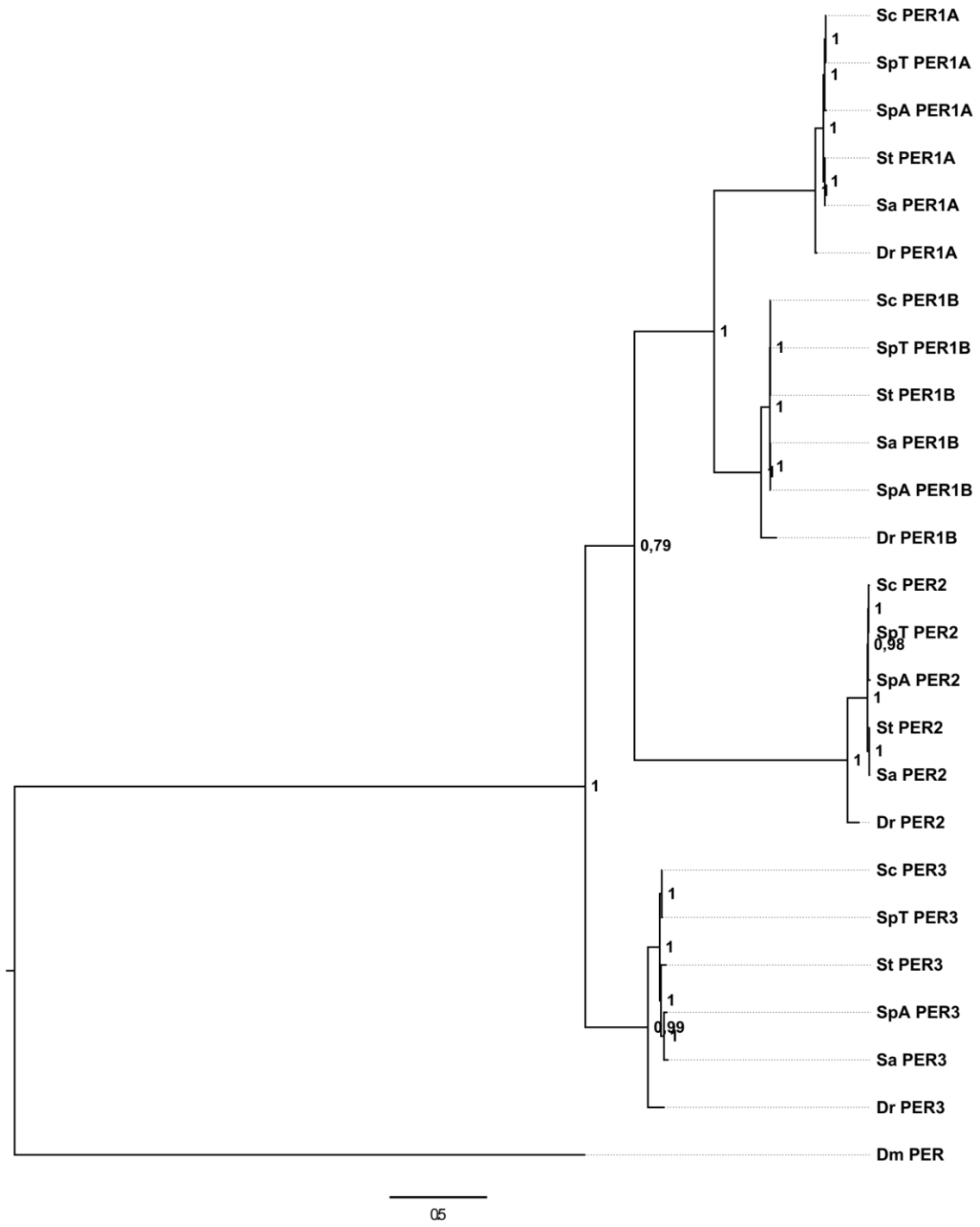
Sc PER1A
SpT PER1A
SpA PER1A
St PER1A
Sa PER1A
Dr PER1A
Sc PER1B
SpT PER1B
St PER1B
Sa PER1B
SpA PER1B
Dr PER1B
Sc PER2
SpT PER2
SpA PER2
St PER2
Sa PER2
Dr PER2
Sc PER3
SpT PER3
St PER3
SpA PER3
Sa PER3
Dr PER3
Dm PER
0,79
0,98
0,99
1
05

Fig. 1 (b)

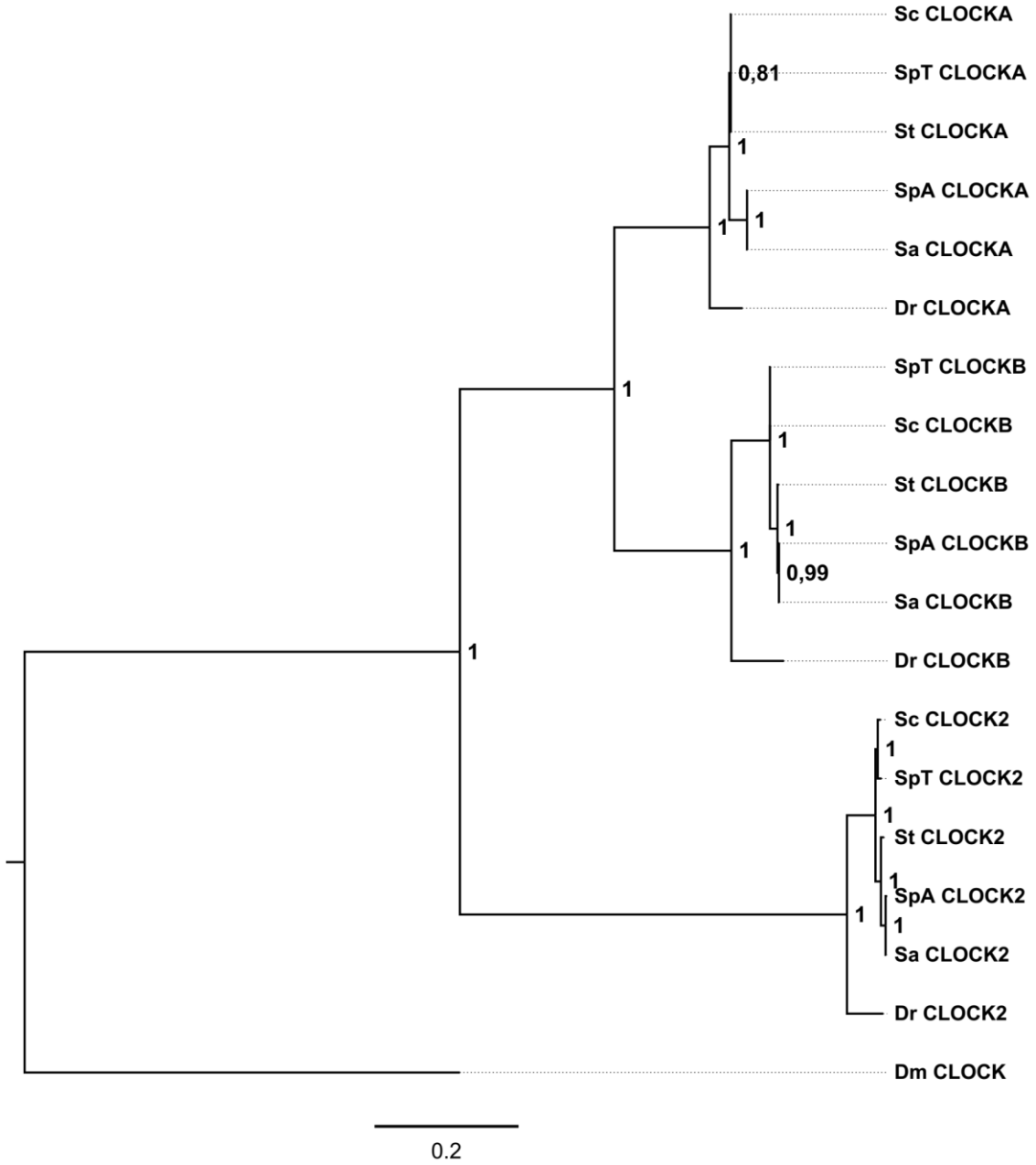
Sc CLOCKA
SpT CLOCKA
St CLOCKA
SpA CLOCKA
Sa CLOCKA
Dr CLOCKA
SpT CLOCKB
Sc CLOCKB
St CLOCKB
SpA CLOCKB
Sa CLOCKB
Dr CLOCKB
Sc CLOCK2
SpT CLOCK2
St CLOCK2
SpA CLOCK2
Sa CLOCK2
Dr CLOCK2
Dm CLOCK
0,81
1
1
1
1
1
1
1
1
0,99
1
1
1
1
1
0.2

Fig 1 (c)

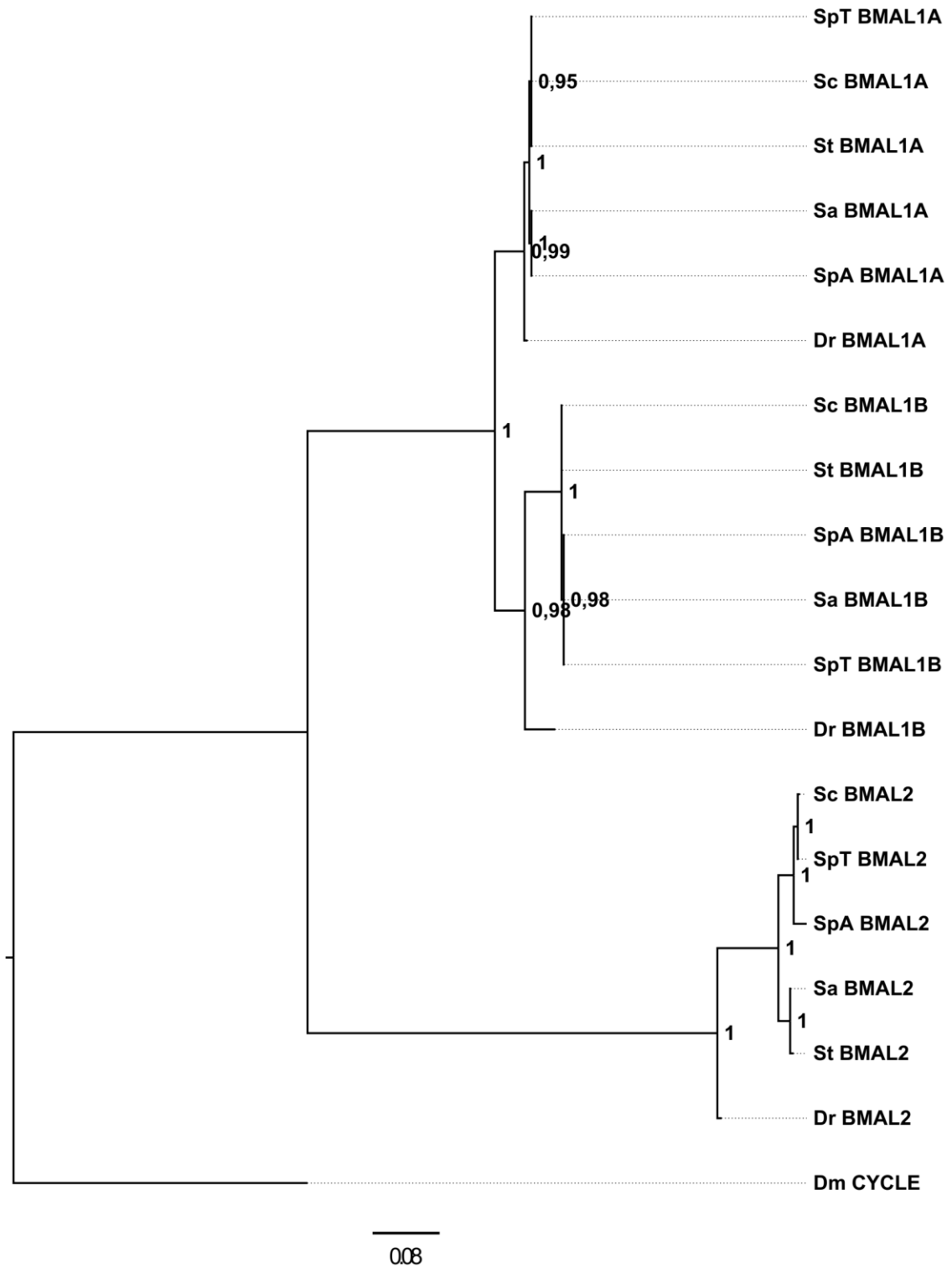
SpT BMAL1A
Sc BMAL1A
St BMAL1A
Sa BMAL1A
SpA BMAL1A
Dr BMAL1A
Sc BMAL1B
St BMAL1B
SpA BMAL1B
Sa BMAL1B
SpT BMAL1B
Dr BMAL1B
Sc BMAL2
SpT BMAL2
SpA BMAL2
Sa BMAL2
St BMAL2
Dr BMAL2
Dm CYCLE
0,95
1
1
0,99
1
1
0,98
0,98
1
1
1
1
1
0.08

Fig. 1 (d)

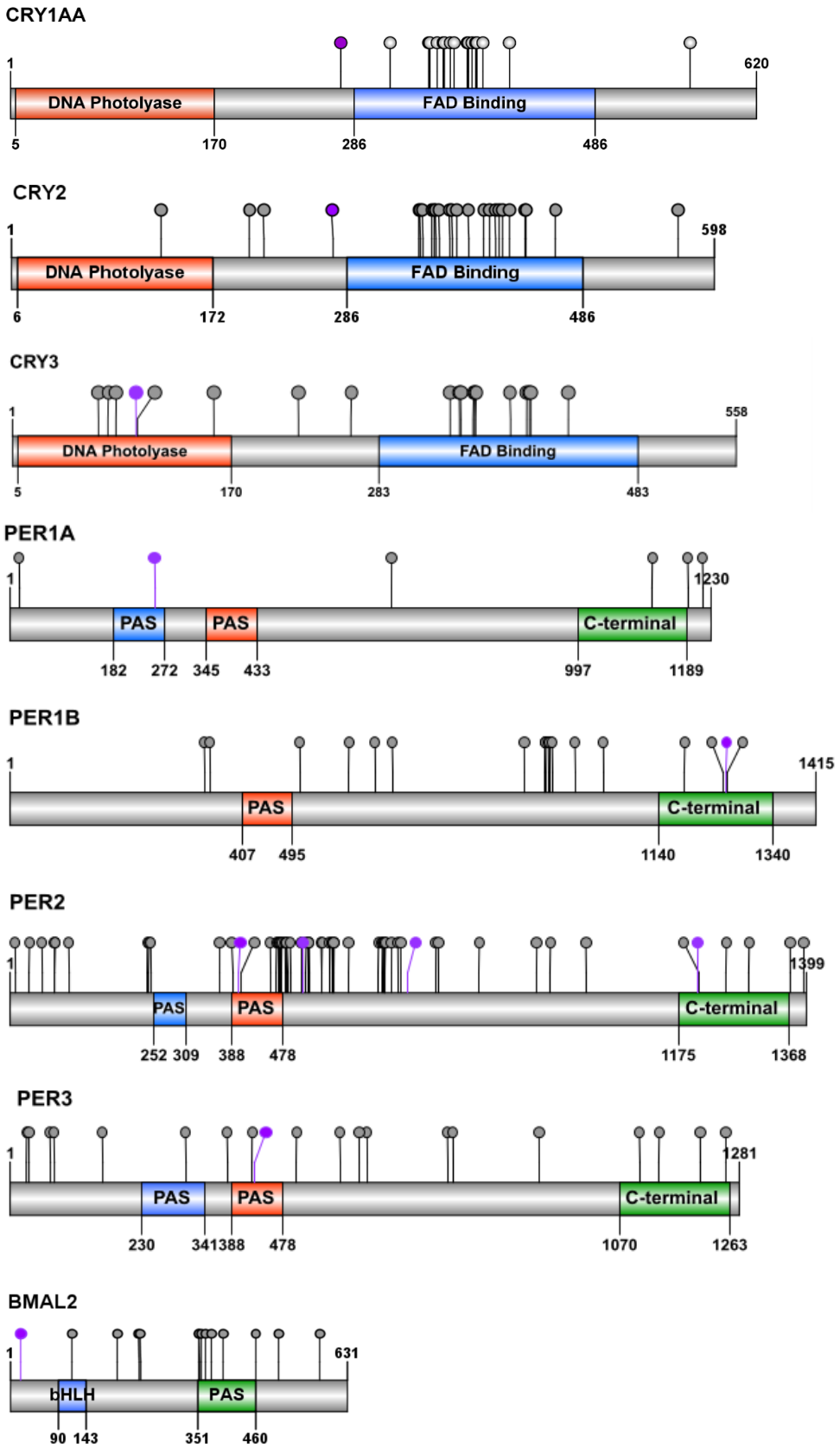
CRY1AA
1
620
DNA Photolyase
FAD Binding
5
170
286
486
CRY2
1
598
DNA Photolyase
FAD Binding
6
172
286
486
CRY3
1
558
DNA Photolyase
FAD Binding
5
170
283
483
PER1A
1
1230
PAS
PAS
C-terminal
182
272
345
433
997
1189
PER1B
1
1415
PAS
C-terminal
407
495
1140
1340
PER2
1
1399
PAS
PAS
C-terminal
252
309
388
478
1175
1368
PER3
1
1281
PAS
PAS
C-terminal
230
341
388
478
1070
1263
BMAL2
1
631
bHLH
PAS
90
143
351
460

Fig. 2